Genetic Resources

Genetic Resources (2024)
Vol. 5, Issue 9
DOI: 10.46265/genresj.2024.9
www.genresj.org
ISSN: 2708-3764

Focus and Scope of *Genetic Resources*

Genetic Resources is an open access journal disseminating global knowledge and tools used by the community of practitioners of plant and animal genetic resources involved in monitoring, collecting, maintaining, conserving, characterizing and using genetic resources for food, agriculture and forestry. ***Genetic Resources*** publishes original research, methods, strategies, guidelines, case studies and reviews as well as opinion and other papers on a variety of topics of interest on the present and future use of genetic resources. These may include the acquisition, documentation, conservation, management, assessment, characterization and evaluation of genetic resources and their link to broader biodiversity, socioeconomic practices, policy guidelines or similar, serving stakeholders within and across sectors. Occasionally, ***Genetic Resources*** publishes special issues with a focus on selected topics of interest for the genetic resources community. The journal has a focus on the European region and also welcomes contributions of wider interest from all world regions.

Cover photos (from left to right):
Goat, Greece, credit: GregMontani, Pixabay; *Brassica Oleracea*, Normandie, France, credit: ECPGR/Lorenzo Maggioni; Blueberries, credit: ChiemSeherin, Pixabay.

This journal is supported by the European Cooperative Programme for Plant Genetic Resources (ECPGR) and the European Regional Focal Point for Animal Genetic Resources (ERFP).

This journal has been conceived as part of the GenRes Bridge project. This project has received seed funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817580.

Editorial Office:

ECPGR Secretariat
c/o Alliance of Bioversity International and CIAT
Via di San Domenico 1
000153 Rome, Italy

Submissions to www.genresj.org

Editorial Board:

Genetic Resources

Genetic Resources (2023), 4 (8)
DOI: 10.46265/genresj.2023.8
www.genresj.org
ISSN: 2708-3764

TABLE OF CONTENTS

Original Articles

Short Communications

Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 1–12
DOI: 10.46265/genresj.NLVC6810
https://www.genresj.org
ISSN: 2708-3764

Leaf trichome diversity, acylsugar concentration, and their relationships to leaf area in *Solanum galapagense*

Ilan Henzler[a] **and Hamid Khazaei** *,[a,b]

[a] *World Vegetable Center, Shanhua, Tainan, 74151, Taiwan*
[b] *Current address: Natural Resources Institute Finland (Luke), Helsinki, Finland*

Abstract: Glandular trichomes are physical and chemical barriers used by some tomato wild relatives to confer resistance against insect pests and diseases transmitted by them. *Solanum galapagense* has been identified as one of the potential sources of insect pest resistance. The present study aimed to examine the trichome diversity and acylsugar concentration of 26 accessions of *S. galapagense* along with one cultivated tomato (*S. lycopersicum*) and one cherry tomato (*S. l. cerasiforme*) cultivar. The results revealed large phenotypic variation among *S. galapagense* accessions for all studied traits. The *S. galapagense* accessions had significantly higher trichome types IV density on the adaxial and abaxial surfaces of the leaf and greater acylsugar concentration but a smaller leaflet area than the cultivated tomato. The selected cherry tomato line represents greater trichome type IV density and acylsugar concentration than other groups. The acylsugar concentration was positively associated with trichome type IV but negatively associated with trichome type V on both leaf surfaces. DNA markers revealed the presence of two previously identified whitefly-resistance alleles in *S. galapagense* accessions. This study will support breeding programmes aiming to improve insect pest resistance in tomato cultivars using crop wild relatives.

Keywords: acylsugar, crop wild relatives, leaflet area, *S. Galapagense*, tomato, trichomes

Citation: Henzler, I., Khazaei, H. (2024). Leaf trichome diversity, acylsugar concentration, and their relationships to leaf area in *Solanum galapagense*. *Genetic Resources* 5 (9), 1–12. doi: 10.46265/genresj.NLVC6810.

Introduction

Cultivated tomato (*Solanum lycopersicum*) is the most valuable vegetable crop by fruit weight globally, generating revenues of US$70 billion from 187 million tonnes of fresh fruit in 2020 (FAO, 2022). Improving fruit quality and yield through domestication in this crop, has led to the loss of important plant defence characteristics (Paudel *et al*, 2019), with tomato cultivation now heavily relying on pesticides to control biotic stresses (Dari *et al*, 2016). The chemical treatments are not only costly, but they also harm the environment (Damalas and Eleftherohorinos, 2011). Developing resistant tomato cultivars could reduce the reliance on pesticides and their associated burden.

Tomato wild relatives are important sources of genetic diversity and are commonly used as reliable sources of resistance genes against biotic and abiotic stress (Ebert and Schafleitner, 2015; Khazaei and Madduri, 2022). Sources of resistance to insect pests have been identified in some tomato wild species, including *S. galapagense*, *S. habrochaites*, *S. pennellii*, *S. cheesmaniae*, and *S. pimpinellifolium* (Kennedy, 2003; Schilmiller *et al*, 2012; Leckie *et al*, 2016; Rakha *et al*, 2017b; Vosman *et al*, 2018). Among them, *S. galapagense* has been identified as one of the most promising sources of insect pest resistance (Firdaus *et al*, 2012; Lucatti *et al*, 2013). It has been the focus of most tomato breeding programmes aiming to improve biotic and abiotic stress resistance due to its close relationship to cultivated tomatoes (Vendemiatti *et al*, 2022). The *S. galapagense* species originates from the Galápagos Islands, an archipelago 1,000km west of Ecuador, where

*Corresponding author: Hamid Khazaei (hamid.khazaei@gmail.com)

Received: 04.10.2023 | Accepted: 21.12.2023 | Published online: 16.01.2024

it formed a diverse range of phenotypes due to the islands' unique ecosystem (Darwin *et al*, 2003). While genetic studies revealed a narrow genetic diversity within the *S. galapagense* germplasm (Darwin, 2009; Pailles *et al*, 2017), it presents distinct morphological characteristics. These include yellow-green foliage, orange fruit at maturity, small seed size and highly divided leaves (Darwin *et al*, 2003; Fenstemaker *et al*, 2022).

Plants have developed a variety of defence mechanisms to counter biotic and abiotic stress conditions (Levin, 1973; Oksanen, 2018). One of these is the presence of fine outgrowths, called trichomes, on the surface of flowers, fruits, stems and leaves as physical and chemical lines of defence. Numerous studies have been conducted on the nature of these epidermal outgrowths, including function, quantification and effectiveness (Glas *et al*, 2012; Vendemiatti *et al*, 2022). Seven types of trichomes are characterized on plants, four are termed glandular due to the cells at their tip which can store and secrete metabolites (Luckwill, 1943). The presence of glandular trichome types IV and VI has been associated with higher insect pest resistance (Lucatti *et al*, 2013; Firdaus *et al*, 2013; Rakha *et al*, 2017b; Zhang *et al*, 2020). These types of glandular trichomes deter insects through the release of secondary metabolites such as acylsugars, which cause behaviour changes and reduced survival in the arthropods that land on them (Antonious *et al*, 2005; Bleeker *et al*, 2011, 2012; Dias *et al*, 2016). In addition to acylsugars, other trichome metabolites such as terpenoids, methylketones and flavonoids play a key role in plant defence mechanisms (Huchelmann *et al*, 2017). Two major genomic regions conferring whitefly resistance (*Wf-1* and *Wf-2*), largely based on glandular trichomes type IV, have been identified in *S. galapagense* (accession id PRI95004; (Firdaus *et al*, 2013; Vosman *et al*, 2019)). Most likely, they regulate the formation of glandular trichome type IV on the leaf epidermis and subsequently control the accumulation of acylsugar on trichome type IV.

Trichome diversity and density, and their relationship with insect pest resistance have been investigated in tomato wild relatives, including *S. galapagense* (Firdaus *et al*, 2012; Lucatti *et al*, 2013; Rakha *et al*, 2017b). An aspect deserving further attention is to harness the genetic diversity of morphological and biochemical characteristics of large germplasm of *S. galapagense* accessions and their relationships with leaf area. So, this study aims to uncover differences in trichome diversity, leaf characteristics and acylsugar concentration in this species. This is supported by the analysis of DNA markers associated with insect pest resistance phenotypes (Firdaus *et al*, 2013).

Materials and methods

Plant material

This study was conducted on 26 accessions of *S. galapagense*, one accession of cherry tomato (*S. lycopersicum* var. *cerasiforme*, abbreviated as *S. l. cerasiforme*), and one cultivated tomato (*S. lycopersicum*). Detailed information on accession number, origin and habitat at collection sites are presented in Table 1. All *S. galapagense* accessions originated from the Galápagos Islands, Ecuador (Figure 1). More than a third were collected on Isla Isabela, the largest island. This study is the first to screen accessions VI037867, VI037869, VI045262, VI057457, VI063173, VI063178, VI063179, VI063181, VI063182 and VI063183 for insect-pest resistance-related traits. The cultivated tomato is a breeding line from the World Vegetable Center (WorldVeg) carrying multiple tomato yellow leaf curl virus resistance genes (*Ty-1/3* and *Ty-2*). The SM131 cherry tomato is a selection from accession VI063893 due to its high density of trichome type IV (unpublished data). All accessions were acquired from the WorldVeg genebank.

Seed treatment

Tomato seeds acquired from the WorldVeg genebank were treated with hydrogen chloride for 15 minutes and washed under running water. They were then treated with trisodium phosphate for one hour, washed under running water and dried in an incubator room at 60% for two days.

Growing conditions

Experiments were conducted in a glasshouse at the WorldVeg in Shanhua, Taiwan. Seeds were sown in a nursery on 25 February 2022, and after two weeks, were transplanted into 8-inch pots filled with cultivable soil collected from tomato fields. The pots were arranged in a randomized complete block design with four replicates. Plants were watered once a day in the morning and fertilized with a blend of 15–15–15 (N–P–K) at week four after transplanting. Relative humidity was about 80±15%. The temperature was set to 28±3°C during the day and to 22±2°C during the night.

Measurements

Leaflet area

The leaflet area was measured ten weeks after sowing using the third leaf from the apex. It was measured using a LI-3100 leaf area meter (LI-COR Inc., Lincoln, NE, USA). The same leaflet was also used for subsequent trichome and acylsugar measurements.

Trichome analysis

Analysis of leaf trichomes was conducted eight and ten weeks after sowing using a stereo microscope (Leica® M-Series Stereo microscopes, Ernst Leitz Wetzlar, GmbH, Germany). Leaf samples were collected at the third node from the apex using sterile forceps. The density of glandular trichome types I, IV, V and VI were measured from four randomly chosen leaflets within 1mm^2.

Table 1. Species, accession number, origin and habitat at collection sites of *Solanum* species used in this study. More information about accessions can be found at https://genebank.worldveg.org/ (accessions with 'VI' code) and http://www.ars-grin.gov/ (accessions with 'PI' code). Passport data was obtained for accessions with 'LA' code from the TGRC, http://tgrc.ucdavis.edu/. *, this accession was first classified as *S. cheesmaniae* and later reclassified as *S. galapagense*.

Species	Accession No.	Origin	Elevation (m)	Other name(s)	Habitat/phenotype
S. galapagense	VI007099	Bartolome, Galápagos Islands	15	LA0317, PI231257	Lava flow, amongst basalt rock, very arid, coastal arid zone
	VI037239	Isabela, Galápagos Islands	40	LA0436	Sandy, near lava outcrop
	VI037241	Pinta, Galápagos Islands	150	LA0526, SAL254	West side Abingdon island
	VI037339	Isabela, Galápagos Islands	5	LA1401, PI365897	Among rocks in large beach with magma and lava flows at each end, collected few metres above tide line
	VI037340	Isabela, Galápagos Islands	200	LA1408, PI379039	Ridge of Cape Berkeley volcano
	VI037867	Floreana, Galápagos Islands	-	LA1136, TL01054	Garnder near Floreana islet
	VI037868	Rabida, Galápagos Islands	10	LA1137, TL01055	On cinder ash
	VI037869	Santiago, Galápagos Islands	650	LA1141, TL01056	Interior walls of crater, purple fruit colour (Fenstemaker *et al*, 2022)
	VI045262	Santiago, Galápagos Islands	-	Selection from LA1141, TL01572	-
	VI057400	Fernandina, Galápagos Islands	-	LA483, 6201A, SAL241	-
	VI057408	Isabela, Galápagos Islands	-	Selection from LA1401	-
	VI057457	Galápagos Islands	-	LA3909	-
	VI063173	Bartolome, Galápagos Islands	15	LA0317	Lava flow, amongst basalt rock, very arid
	VI063174	Isabela, Galápagos Islands	30	LA0438, SAL192	Rocky basalt outcropping in first hills to W. (7km) from Villamil, 1km from coast
	VI063175	Isabela, Galápagos Islands	20	LA0480A, SAL238	Along coast in bay facing Cowley Islet - not far from shore, in broken terrain without shade
	VI063176	Santa Cruz, Galápagos Islands	-	LA0528, SAL256	Academy Bay
	VI063177	Fernandina, Galápagos Islands	580	LA0530, SAL258	Inside crater at edge
	VI063178	Galápagos Islands	5	-	-
	VI063179	Santiago, Galápagos Islands	6	LA0747	In lava formation near shore
	VI063180	Santiago, Galápagos Islands	3	LA0748	In lava formation
	VI063181	Isabela, Galápagos Islands	4	LA0929	Growing in lava, roots in sand, above sea level

Continued on next page

Table 1 continued

Species	Accession No.	Origin	Elevation (m)	Other name(s)	Habitat/phenotype
	VI063182	Isabela, Galápagos Islands	4	LA0930	Growing in lava, roots in sand, same as LA 0929
	VI063183	Bartolome, Galápagos Islands	-	LA1044	-
	VI063184	Isabela, Galápagos Islands	400	LA1452	On the trail from Punta Ecuador to crater rim - mid-elevation in longer of two lava flows
	VI063185	Isabela, Galápagos Islands	100	LA1627	Volcanic cone above Darwin's salt lake - likely Tagus Cove
	VI063187*	Santiago, Galápagos Islands	10	LA1411, PI379040	On soft bright red rock formation, margin of beach
S. l. cerasiforme	VI063893	Fernandina, Galápagos Islands	-	SM131	-
S. lycopersicum	CLN3682C	Breeding line	-	AVTO1424	Pedigree: CLN3682F1-10-3-4-27-3-16

The number of trichomes was counted from four different microscopic fields at 5X magnification and converted to the number per mm^2 using a standard scale. The identification of trichome types on the leaf surface followed a schematic drawn by Luckwill (1943). After measuring trichomes on the adaxial surface, the leaflets were flipped to measure the trichomes on the abaxial surface.

Acylsugar concentration

Analysis of acylsugar content was conducted at eight and ten weeks after sowing. Polyethylene vials were used to collect four 3±1cm lateral leaflets from each plant at the third node from the apex. Samples were dried in an incubator at 29°C for three days before washing them with 3ml methanol. Of this suspension, 100μl was added to 100μl 6M Ammonium Hydroxide in 96 well ELISA plates with two biological replicates, following a protocol developed by Martha Mutschler (Savory, 2004). The samples were incubated overnight and left to dry under the hood for three days before adding 200μl PGO reagent to each well and placing it on an orbital shaker. After three hours, absorbance values at 490nm were measured using BioTek's uQuant (Agilent Technologies Inc., Santa Clara, CA, USA) and converted into acylsugar concentration using a sucrose standard curve.

DNA extraction and DNA marker assay

Genomic DNA was extracted from 10-week-old plants using the CTAB method (Doyle and Doyle, 1990). The *Wf-1* and *Wf-2* detailed marker sequences presented in Firdaus *et al* (2013) were used for genotyping the studied germplasm for the presence/absence of corresponding bands. The term '*Wf*' stands for whitefly and represents markers previously identified for whitefly resistance. These markers are located in tomato chromosomes 2 and 9, respectively. Purified DNA samples were digested with restriction enzymes DDeI and HpyCH4IV for *Wf-1* and *Wf-2* markers, respectively. Digested samples were amplified along with marker-specific primers using PCR reactions as described by Mahfouze and Mahfouze (2019). The PCR-amplified samples were run on a 5% acrylamide gel for 30 minutes at 100V and stained using an ETBR-out stain. The gel was scanned in a Bio-1000F scanner, and the amplified bands visualized using Microtek MiBio Fluo software (both from MicroTek International, Inc., Hsinchu, Taiwan).

Statistical analysis

The R statistical package (R Core Team, 2021) was used for data analysis. Correlation analysis was performed to determine the relationships between morphological measurements. The dataset was subjected to a one-way analysis of variance (ANOVA), and the SEM (standard error of means) was calculated. Principal component analysis (PCA) was employed to illustrate relationships between accessions and leaf morphological measurements. The online mapping tool at maps.co was used to plot the coordinates of accessions in Figure 1 (https://maps.co/). The geographic coordinates of *S. galapagense* accessions were obtained from the Tomato Genetics Resource Center (C.M. Rick TGRC, https://tgrc.ucdavis.edu/) and the WorldVeg (https://genebank.worldveg.org/#/) genebank databases.

Results

Trichome densities varied significantly a m ong studied germplasm (Table 2 and Supplemental Table 1). Trichome type IV density ranged from 6.3 to 13.5 for the abaxial and 0.7 to 10.9 for the adaxial surface of *S. galapagense* accessions, while the cultivated tomato (CLN3682C) had none on either surface. Accessions VI057408, VI063174, VI063177, VI063185 and VI057400 had the greatest number of trichome type IV on both surfaces. The cherry tomato (VI063893) had 22% greater (on both surfaces) trichome type IV compared to the average values of *S. galapagense* accessions. Within *S. galapagense* accessions, trichome type VI varied from 0.4 to 2.7 on the abaxial and 0.3 to 2.8 on the adaxial side. For the abaxial surface, this was 30% and 14% lower than cultivated and cherry tomato cultivars, respectively. For the adaxial surface, it was 94% lower and 30% higher than cultivated and cherry tomatoes, respectively (Table 2). Most studied accessions had fewer trichomes on the adaxial than on the abaxial side, with 13% less for type IV and 40% less for type VI. Comparing 8- and 10-week trichome phenotyping at the abaxial surface, *S. galapagense* trichome densities remained stable with a 3% increase for type IV, and a 12% decrease in type VI.

Acylsugar concentration varied significantly ($P < 0.001$) among *S. galapagense* accessions, ranging from 5.43 to 58.03 μmol/g. The cultivated tomato cultivar (CLN3682C) showed a very low acylsugar concentration of 0.94μmol/g and the cherry tomato (VI063893) showed a moderate level of 21.02μmol/g (Table 2). On average, 10-week-old *S. galapagense* accessions had 45% greater acylsugar concentrations than 8-week-old plants. Accessions VI063181, VI037869 and VI045262 had the highest concentration of acylsugar, all above 50μmol/g.

The leaflet area varied significantly ($P < 0.001$) among *S. galapagense* accessions, ranging from 2.43 to 9.15cm^2. Leaflet area was significantly greater for cultivated tomato, with 14.36cm^2 than for all *S. galapagense* accessions. On average, cherry tomato leaflets were 10% larger than *S. galapagense* accessions (Table 2).

Correlations between trichome types and acylsugar concentration are presented in Table 3. Acylsugar concentration was positively associated with trichome type IV and negatively associated with trichome type V. The negative correlation between acylsugar and trichome type VI was only significant at the 10-week-old plant stage. In addition, leaflet area was negatively correlated with trichome IV density of abaxial surface

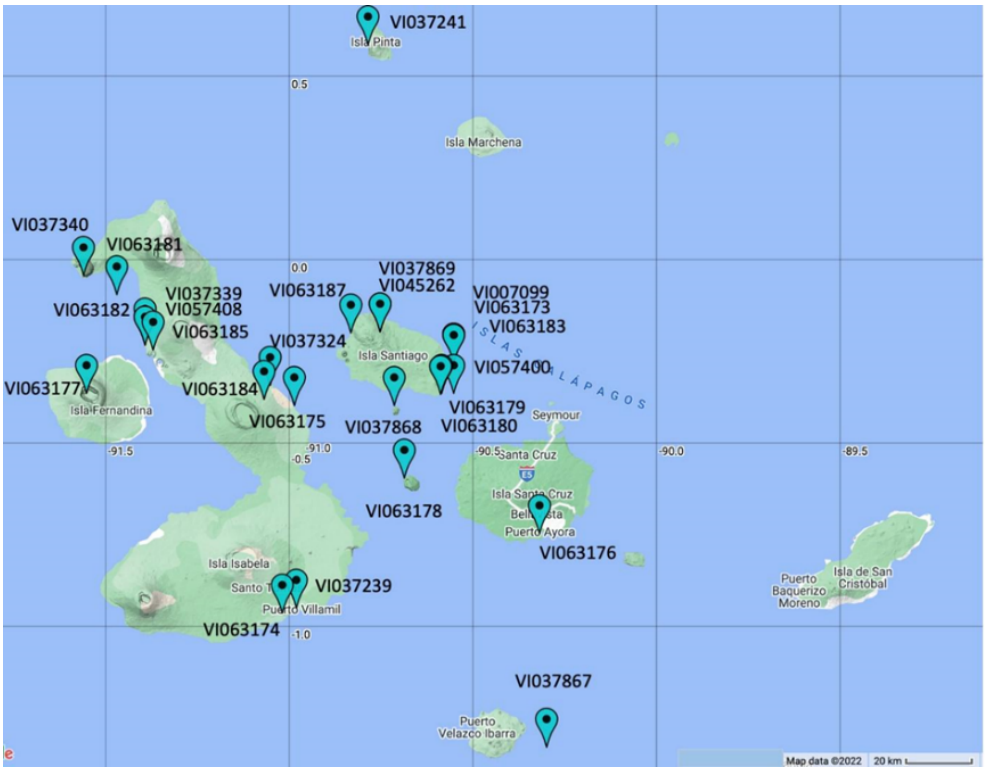

Figure 1. Geographical distribution of *S. galapagense* accessions in the Galápagos Islands.

($P < 0.001$, Figure 2A) and acylsugar concentration ($P = 0.078$, Figure 2B).

The PCA plot (Figure 3) showed three clusters based on acylsugar, trichome type and leaf area measurements. The first group consisted of *S. galapagense* accessions and the cherry tomato genotype (VI063893), which was correlated with high acylsugar concentration as well as high levels of trichome type IV on the adaxial surface. The cultivated tomato (CLN3682C) was separated from *S. galapagense* group mainly due to its greater leaflet area and high levels of trichome type V. One of the *S. galapagense* accessions, VI007099, was grouped closer to the cultivated tomato mainly due to its larger leaflet area (its leaf morphology was similar to cultivated tomatoes) compared to other accessions from the Galápagos Islands.

The DNA marker results are presented in Supplemental Table 3. All accessions of *S. galapagense* showed both resistance bands at 46 and 40bp for the *Wf-2* marker (indicated by '++') while both *S. lycopersicum* cultivars (CLN3682C and VI063893) showed the susceptible band at 87bp (indicated by '-'). The *Wf-1* marker showed a resistance band at 139bp (indicated by '+'). Both *S. lycopersicum* cultivars showed the susceptible band at 157bp (indicated by '-').

Discussion

In this study, we screened a large germplasm collection of *S. galapagense* accessions revealing a considerable phenotypic variation in glandular trichome density and acylsugar concentration. Our results indicated that more than 50% of *S. galapagense* accessions had acylsugar concentration more than 30μmol/g – on average, 34 times greater than the studied *S. lycopersicum* (CLN3682C) genotype. However, not all *S. galapagense* had high concentrations and trichomes types IV and VI density. There was a 12-fold difference between the greatest and lowest *S. galapagense* accessions for acylsugar concentration. However, only a 2-fold variation was observed for trichome type IV density in this species. Our results also showed that tomato plants with smaller leaves had higher densities of trichome type IV.

The selection of tomato plants with higher densities of glandular trichomes IV and VI has shown to be an effective criterion for obtaining superior resistance to insect pests. This has been widely proven in resistance to spider mites (*Tetranychus urticae,* Andrade *et al* (2017); Rakha *et al* (2017a); De Souza-Marinke *et al* (2022)), whitefly (*Bemisia tabaci,* Lucatti *et al* (2013); Rakha *et al* (2017b)), thrips (Escobar-Bravo *et al*, 2017), and cabbage looper caterpillar (Mymko and Avila-Sakar, 2019). These studies generally attribute insect-pest resistance to the excretion of acylsugars by the tip of glandular trichome type IV, thereby acting as a biopesticide.

Several studies have reported genetic diversity for trichome density and acylsugar concentration in tomato (Lucatti *et al*, 2013; Baier *et al*, 2015; Rakha *et al*, 2017b). These have allowed tomato breeding programmes globally to exploit the available diversity and improve insect-pest resistance cultivars. We screened a relatively larger number of *S. galapagense* accessions (26) compared to previous studies, which

Table 2. Mean ± SD (standard deviation) for trichome IV and VI measurements (abaxial and adaxial surfaces) at ten weeks after sowing and acylsugar concentration and leaflet area on 26 *Solanum galapagense* accessions along with one cherry tomato and one cultivated tomato genotype. SEM, standard error of means.

		Trichome types per mm^2 – Abaxial		Trichome types per mm^2 – Adaxial		Acylsugar (μmol/g)	Leaflet area (cm^2)
Species	**Accession No.**	**IV**	**VI**	**IV**	**VI**		
S. galapagense	VI063181	8.0±5.7	1.6±1.3	7.1±1.6	2.6±0.9	58.02±8.32	8.01±1.84
	VI037869	10.5±1.6	0.6±0.4	9.1±1.8	0.2±0.5	57.68±15.42	6.12±2.67
	VI045262	11.3±1.2	1.0±0.4	8.7±1.3	0.4±0.1	54.02±11.27	5.95±1.56
	VI057408	13.5±1.3	1.0±0.5	9.1±0.7	2.8±0.7	46.64±12.57	3.83±0.65
	VI063184	10.1±1.1	1.1±0.7	7.5±1.8	1.4±0.3	45.56±4.43	5.92±0.95
	VI063187	9.2±1.3	1.0±0.8	8.6±1.4	0.5±0.2	44.25±9.63	3.81±1.45
	VI063185	12.0±0.5	1.5±0.3	9.0±3.0	1.1±0.9	43.49±5.71	5.15±0.45
	VI057400	11.8±0.9	0.9±0.6	10.8±0.8	0.3±0.3	43.07±4.51	4.40±0.18
	VI063177	12.1±1.0	1.1±0.5	9.7±2.5	0.4±0.1	41.75±11.28	5.15±1.46
	VI037241	10.1±1.4	0.3±0.1	6.0±1.4	0.5±0.2	34.00±0.01	2.88±0.17
	VI037339	7.7±0.6	0.7±1.5	9.3±2.5	0.3±0.0	32.40±0.01	6.89±0.0
	VI063182	11.1±1.9	0.8±0.3	8.5±1.0	0.6±0.3	32.39±6.25	6.15±1.78
	VI057457	10.0±2.7	2.5±1.5	9.0±1.2	0.3±0.0	31.87±6.97	2.65±1.01
	VI063174	12.4±0.9	1.0±0.5	10.8±1.5	1.2±0.6	30.58±8.12	6.81±3.33
	VI063179	6.2±5.5	2.0±1.8	4.3±0	1.3±0.0	25.20±1.86	6.59±0.83
	VI037340	11.3±1.3	1.1±0.9	8.5±1.5	0.8±0.6	24.65±2.52	7.07±1.03
	VI037239	9.7±0.7	1.2±0.7	9.5±1.1	1.1±0.2	24.60±0.01	6.60±1.2
	VI063175	12.0±1.4	2.6±1.6	7.5±1.1	1.6±1.4	22.53±5.67	4.60±1.62
	VI037868	10.6±2.3	1.8±0.4	9.5±1.6	1.4±0.5	22.08±5.05	5.40±1.43
	VI063183	10.3±1.5	2.4±1.4	7.2±1.7	0.8±0.1	22.07±1.52	2.43±1.66
	VI037867	9.0±0.4	1.5±0	7.7±1.8	0.7±0.5	18.46±1.81	5.17±0.38
	VI063176	9.3±5.3	0.9±0.9	7.7±5.4	0.5±0.5	16.06±4.36	4.37±1.86
	VI063173	11.6±0.5	2.4±1.6	9.4±2.8	1.6±1.2	15.29±4.70	5.46±1.92
	VI063180	10.3±1.8	1.4±0.5	8.1±2.5	0.3±0.3	15.23±2.15	7.07±1.77
	VI063178	10.0±1.3	1.6±1.5	8.7±2.7	1.1±0.3	14.8±3.61	5.32±0.98
	VI007099	6.6±2.9	1.6±0.5	0.6±0.8	1.8±0.5	5.43±5.88	9.15±5.5
	Range (*S. galapagense*)	6.2–13.5	0.3–2.6	0.6–10.9	0.2–2.8	5.43–58.02	2.43–9.15
	Mean (*S. galapagense*)	**10.3±1.7**	**1.4±0.6**	**8.2±2.0**	**1.0±0.7**	**31.6±14.9**	**5.50±1.59**
S. l. cerasiforme	VI063893	**12.6±1.5**	**1.6±0.4**	**10.0±2.2**	**0.7±0.3**	**21.02±15.93**	**6.10±0.94**
S. lycopersicum	CLN3682C	**0.0**	**2.0±0.8**	**0.0**	**4.6±2.3**	**0.94±0.23**	**14.36±3.64**
SEM		**1.7**	**0.2**	**1.6**	**0.2**	**16.95**	**1.3**

Table 3. Correlations between trichome types and acylsugar (AS) concentration at 8-week (N = 26) and 10-week (N = 28) intervals on abaxial surface. *, $P < 0.05$; **, $P < 0.01$. Data for 8-week measurements is presented in Supplemental Table 2.

Trichome type	AS (8-week-old)	AS (10-week-old)
Type I	-0.140	-0.238
Type IV	0.473*	0.410*
Type V	-0.540**	-0.459*
Type VI	-0.347	-0.471*

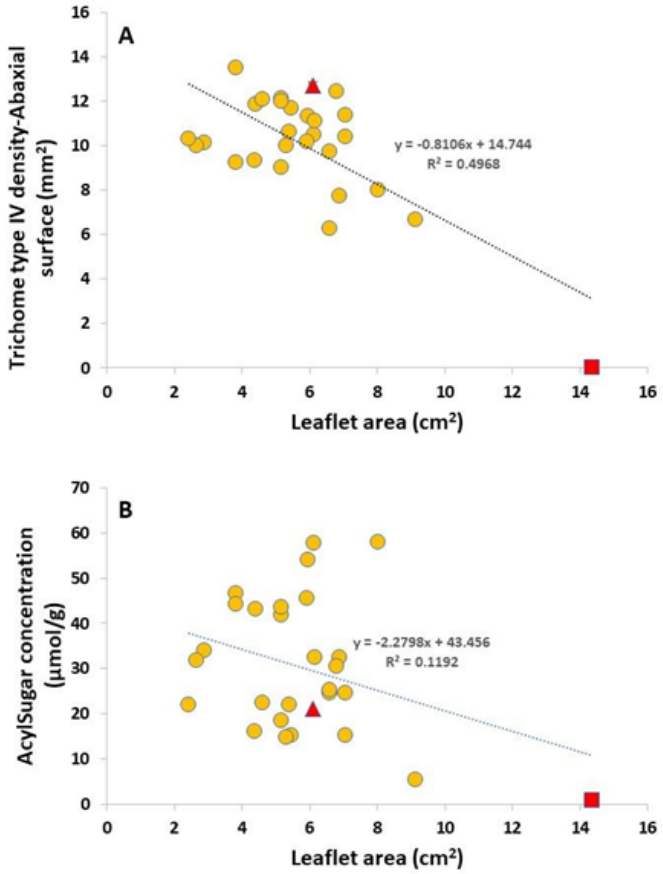

Figure 2. Relationships between trichome type IV (**A**) and acylsugar concentration (**B**) with leaflet area at 10-week-old seedlings (N = 28). The red triangle and red square represent cherry and cultivated tomato cultivars, respectively.

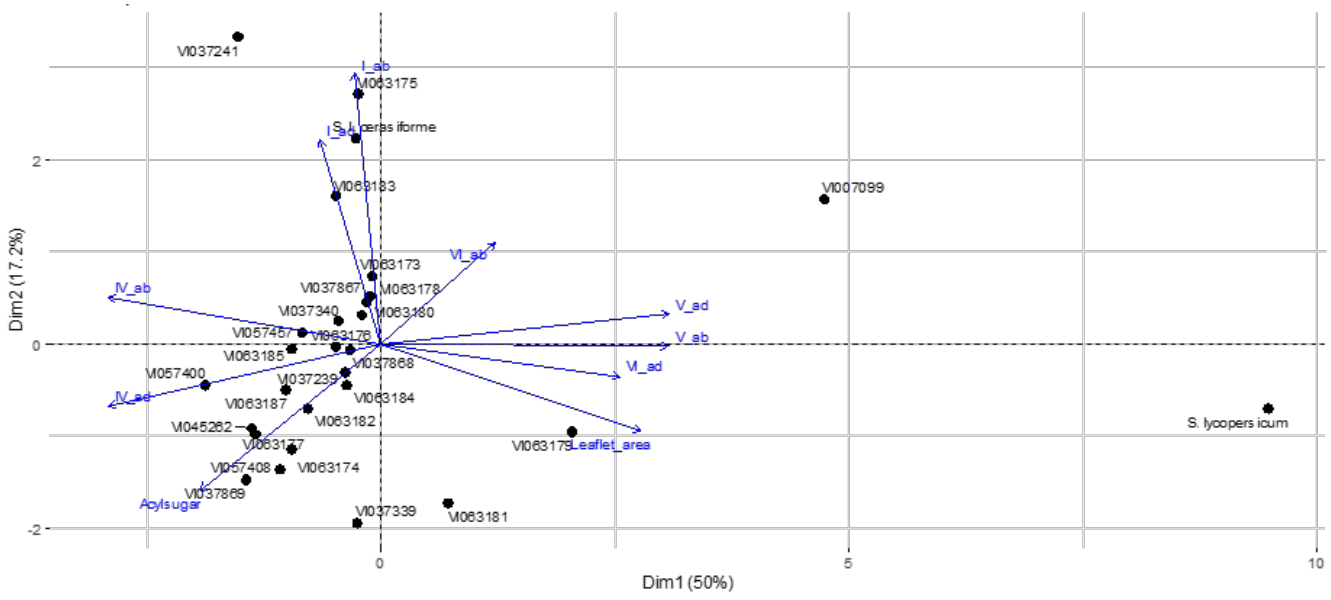

Figure 3. The biplot illustrates the principal components (PC) analysis for the 26 accessions of *Solanum galapagense* and two *S. lycopersicum* lines with measurements as vectors. Vectors that are close together are correlated in terms of the leaf measurements. ad, adaxial surface; ab, abaxial surface. I, IV, V, and VI refer to trichome types.

only characterized 10 (Rakha *et al*, 2017b) or 11 accessions (Lucatti *et al*, 2013). Our study revealed a wider variation for trichome type IV, however, smaller values for acylsugar concentration compared to a similar study by Rakha *et al* (2017b). The cherry tomato genotype (VI063893) was previously characterized by high trichome type IV density (unpublished data). Our results confirm that its trichome type IV density was higher than 96% of studied *S. galapagense* accessions (Table 2). This cherry tomato genotype may be used as a source of insect resistance in cherry and cultivated tomato germplasm.

A moderate positive correlation ($P < 0.05$) was observed between acylsugar concentration and trichome type IV at two different sampling times. Previous studies also reported a similar trend (Lucatti *et al*, 2013; Rakha *et al*, 2017b). A possible explanation for this could be the poor phenotyping of trichomes under the microscope, as counting the number of trichomes is an inherently delicate task. This difficulty highlights the need for high throughput methods to measure trichomes. Another explanation for the lack of strong correlation between trichome IV density and acylsugar could be that acylsugar production is not solely linked to trichome density but also their metabolic activity, whereby the same trichome types in different accessions produce varying levels of acylsugar (Zhang *et al*, 2008; Bergau *et al*, 2015). Following this reasoning, isolated trichomes could be tested for metabolic activity through GC-MS as described for *L. hirsutum* by Fridman *et al* (2005).

A negative correlation was observed between trichome type IV and leaflet area, a trend that has been reported in other plant species, including *S. berthaultii* Hawkes (Pelletier, 1990) and silver birch (*Betula pendula* Roth) (Lihavainen *et al*, 2017). Mymko and Avila-Sakar (2019) reported that unexpanded leaves had greater trichome density and resistance than expanded (larger) leaves at different growth stages of tomatoes. In our study, leaflet area was one of the main drivers in allocating tomato species into three different groups (Figure 3). Accession VI007099 had the greatest leaflet area and lowest trichome density and acylsugar concentration among *S. galapagense* and presented leaf morphology between wild and cultivated tomatoes. On the other hand, accession VI063181 had the second-largest leaflet area among *S. galapagense* accessions but also the greatest acylsugar concentration among all accessions. This controversy was also evident by the weak correlation between leaflet area and acylsugar concentration (Figure 2). These results suggest that acylsugar concentration may not be derived by leaf size and trichome type IV in *S. galapagense* germplasm.

No clear pattern between leaf trichome measurements and accessions' geographical origin in the Galápagos Islands was observed. The only accession from Pinta Island, VI037241, was separated from other *S. galapagense*. Pinta Island is the northernmost of the main islands and has a semi-arid ecosystem (Hunter and Gibbs, 2014). The main driver for the deviation of other *S. galapagense* from the core cluster (Figure 3) was leaflet area (VI063181, VI063179, VI037339 and VI007099) and higher trichome type VI density (VI063175 and VI063183). The three accessions with the highest trichome type IV density were all found among volcanic rock with VI057408 and VI06374 found within 1km of the sea on Isla Isabela and VI063177 inside a volcanic crater on Isla Fernandina. An explanation for this correlation could be the presence of silicon in volcanic soil as it has been shown to boost plants' resistance to pests by accumulating magnesium at the base of trichomes (Ma, 2004).

Some *S. galapagense* accessions originating from the Galápagos Islands have been exposed to dry and saline growing conditions (Pailles *et al*, 2020) and high insect pressure (Peck, 2008), thus may represent a generous source of alleles that can be explored to improve biotic and abiotic stress. As this species can easily hybridize with cultivated tomatoes (Rick, 1961), they have been used as donors for stress tolerance genes, which could be transferred into commercial varieties by introgression breeding (Zamir, 2001). For example, VI037339 (LA1401) and VI007099 accessions have already been utilized as donors of high trichome IV density into modern cultivated tomato cultivars through interspecific crosses (Andrade *et al*, 2017; Rakha *et al*, 2017b; DaSilva *et al*, 2019; Vendemiatti *et al*, 2022). However, in this study, these accessions were not among those with the highest trichome type IV density and acylsugar concentration. This could be due to the genotype-by-environment interactions as the experiments were carried out under different growing conditions.

From the analysis of DNA markers, we could see that most *S. galapagense* accessions were homozygous for *Wf-1* and *Wf-2* but neither S. lycopersicum cultivars. This suggests that the morpho-chemical measurements in this study were linked to the genetic background of *S. galapagense* accessions. However, the cherry tomato genotype VI063893 (SM131), which showed high levels of trichome type IV and acylsugar, did not amplify either band of the DNA markers. This was not surprising as those DNA markers were developed from an interspecific population derived from *S. galapagense* (Firdaus *et al*, 2013). A potential reason why the cultivated tomato accession (CLN3682C) did not show the resistance markers is that the related genes may have been lost during domestication. We conclude that *Wf-1* and *Wf-2* may be more suitable to be used in genetic materials derived only from *S. galapagense*. The other possibility is that the source(s) of resistance in VI063893 is non-allelic to *Wf-1* and *Wf-2*.

In conclusion, our study focused on screening a large *S. galapagense* germplasm, supporting breeding programmes aiming to improve insect-pest resistance in tomatoes using crop wild relatives. The ultimate goal is to develop tomato cultivars with insect-pest resistance-related traits that help farmers reduce pesticide use

and produce a high-quality and chemical-free tomato crop. The glandular trichome density and chemistry are highly affected by the genotype by environment interactions (Wang *et al*, 2021). This needs to be considered when selecting these traits under field conditions.

Supplemental data

Supplemental Table 1. Mean ± SD (standard deviation) for trichomes I and V measurements (abaxial and adaxial surfaces)
Supplemental Table 2. Trichome density and acylsugar concentration at 8-week-old seedlings of *S. galapagense*
Supplemental Table 3. DNA marker assay for *B. tabaci* whitefly resistance alleles Wf-1 and Wf-2. N, missing data

Data availability

The data that support this study will be shared upon reasonable request to the corresponding author.

Conflicts of interest

The authors declare no conflicts of interest

Author contributions

IH and HK did the conceptualization, data curation and formal analysis, led the methodology, and created the original draft of the manuscript.

Acknowledgement

We would like to thank Yun-che Hsu (Grace) and Jean Lin for their kind assistance during the experiments. We also thank Dr Roland Schafleitner (Flagship Leader, Vegetable Diversity & Improvement) and Dr Maarten van Zonneveld (Genebank manager) for their valuable suggestions during the experiments. In addition, the first author would like to thank the National Cheng Kung University (NCKU), Taiwan for its support.

Funding

Financial support was provided by long-term strategic donors to the World Vegetable Center: Taiwan, UK aid from the UK government, the United States Agency for International Development (USAID), the Australian Centre for International Agricultural Research (ACIAR), Germany, Thailand, Philippines, Korea, and Japan.

References

Andrade, M. C., Silva, A. A. D., Neiva, I. P., Oliveira, I. R. C., De Castro, E. M., Francis, D. M., and Maluf, W. R. (2017). Inheritance of type IV glandular trichome density and its association with whitefly resistance from *Solanum galapagense* accession LA1401. *Euphytica* 213, 52–52. doi: https://doi.org/10.1007/s10681-016-1792-1

Antonious, G., Kochhar, F., Simmons, T. S., and M, A. (2005). Natural products: seasonal variation in trichome counts and contents in *Lycopersicum hirsutum f. glabratum*. *Journal of Environmental Science and Health* 40, 619–631. doi: https://doi.org/10.1081/PFC-200061567

Baier, J. E., Resende, J. T. V., Faria, M. V., Schwarz, K., and Meert, L. (2015). Indirect selection of industrial tomato genotypes that are resistant to spider mites (*Tetranychus urticae*). *Genetics and Molecular Research* 14, 244–252. doi: http://dx.doi.org/10.4238/2015.January.16.8

Bergau, N., Bennewitz, S., Syrowatka, F., Hause, G., and Tissier, A. (2015). The development of type VI glandular trichomes in the cultivated tomato *Solanum lycopersicum* and a related wild species *S. habrochaites*. *BMC Plant Biology* 15, 289–289. doi: https://doi.org/10.1186/s12870-015-0678-z

Bleeker, P. M., Diergaarde, P. J., Ament, K., Schütz, S., Johne, B., Dijkink, J., Hiemstra, H., De Gelder, R., De Both, M. T. J., Sabelis, M. W., Haring, M. A., and Schuurink, R. C. (2011). Tomato-produced 7-epizingiberene and R-curcumene act as repellents to whiteflies. *Phytochemistry* 72, 68–73. doi: https://doi.org/10.1016/j.phytochem.2010.10.014

Bleeker, P. M., Mirabella, R., Diergaarde, P. J., Vandoorn, A., Tissier, A., Kant, M. R., Prins, M., De Vos, M., Haring, M. A., and Schuurink, R. C. (2012). Improved herbivore resistance in cultivated tomato with the sesquiterpene biosynthetic pathway from a wild relative. *Proceedings of the National Academy of Sciences of the United States of America* 109, 20124–20129. doi: https://doi.org/10.1073/pnas.1208756109

Damalas, C. A. and Eleftherohorinos, I. G. (2011). Pesticide exposure, safety issues, and risk assessment indicators. *International Journal of Environmental Research and Public Health* 8, 1402–1419. doi: https://doi.org/10.3390/ijerph8051402

Dari, L., Addo, A., and Dzisi, K. A. (2016). Pesticide use in the production of tomato (*Solanum lycopersicum* L.) in some areas of Northern Ghana. *African Journal of Agricultural Research* 11, 352–355. doi: https://doi.org/10.5897/AJAR2015.10325

Darwin, S. C. (2009). The systematics and genetics of tomatoes on the Galápagos Islands (*Solanum*, Solanaceae). Ph.D. thesis, University College London, UK.

Darwin, S. C., Knapp, S., and Peralta, I. E. (2003). Taxonomy of tomatoes in the Galápagos Islands: Native and introduced species of *Solanum* section *Lycopersicon* (Solanaceae). *Systematics and Biodiversity* 1(1), 29–53. doi: https://doi.org/10.1017/S1477200003001026

DaSilva, A. A., Carvalho, R. D. C., Andrade, M. C., Zeist, A. R., Resende, J. T. V., and Maluf, W. R. (2019). Glandular trichomes that mediate resistance to green peach aphid in tomato genotypes from the cross between *S. galapagense* and *S. lycopersicum*.

Acta Scientiarum Agronomy 41. doi: https://doi.org/10.4025/actasciagron.v41i1.42704

De Souza-Marinke, L., De Resende, J. T. V., Hata, F. T., Dias, D. M., De Oliveira, L. V. B., Ventura, M. U., Zanin, D. S., and Filho, R. B. D. L. (2022). Selection of tomato genotypes with high resistance to *Tetranychus evansi* mediated by glandular trichomes. *Phytoparasitica* 50, 629–643. doi: https://doi.org/10.1007/s12600-022-00984-6

Dias, D. M., Resende, J. T. V., Marodin, J. C., Matos, R., Lustosa, I. F., and Resende, N. C. V. (2016). Acyl sugars and whitefly (*Bemisia tabaci*) resistance in segregating populations of tomato genotypes. *Genetics and Molecular Research* 15(2). doi: http://dx.doi.org/10.4238/gmr.15027788

Doyle, J. J. and Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.

Ebert, A. W. and Schafleitner, R. (2015). Utilization of wild relatives in the breeding of tomato and other major vegetables. *Crop Wild Relatives and Climate Change* 141-172. doi: https://doi.org/10.1002/9781118854396.ch9

Escobar-Bravo, R., Klinkhamer, P. G. L., and Leiss, K. A. (2017). Induction of jasmonic acid-associated defenses by thrips alters host suitability for conspecifics and correlates with increased trichome densities in tomato. *Plant and Cell Physiology* 58(3), 622–634. doi: https://doi.org/10.1093/pcp/pcx014

FAO (2022). FAOSTAT - Food and Agriculture Organization of the United Nations. url: http://faostat.fao.org. accessed date: 2023-03-23

Fenstemaker, S., Sim, L., Cooperstone, J., Francis, and D (2022). *Solanum galapagense*-derived purple tomato fruit color is conferred by novel alleles of the *anthocyanin fruit* and *atroviolacium* loci. *Plant Direct* 6(4). doi: https://doi.org/10.1002/pld3.394

Firdaus, S., Van Heusden, A. W., Hidayati, N., Supena, E. D. J., Mumm, R., De Vos, R. C. H., Visser, R. G. F., and Vosman, B. (2013). Identification and QTL mapping of whitefly resistance components in *Solanum galapagense*. *Theoretical and Applied Genetics* 126, 1487–1501. doi: https://doi.org/10.1007/s00122-013-2067-z

Firdaus, S., Van Heusden, A. W., Hidayati, N., Supena, E. D. J., Visser, R. G. F., and Vosman, B. (2012). Resistance to *Bemisia tabaci* in tomato wild relatives. *Euphytica* 187, 31–45. doi: https://doi.org/10.1007/s10681-012-0704-2

Fridman, E., Wang, J., Iijima, Y., Froehlich, J. E., Gang, D. R., Ohlrogge, J., and Pichersky, E. (2005). Metabolic, genomic, and biochemical analyses of glandular trichomes from the wild tomato species *Lycopersicon hirsutum* identify a key enzyme in the biosynthesis of Methylketones. *The Plant Cell* 17, 1252–1267. doi: https://doi.org/10.1105/tpc.104.029736

Glas, J., Schimmel, B., Alba, J., Escobar-Bravo, R., Schuurink, R., and Kant, M. (2012). Plant glandular trichomes as targets for breeding or engineering of resistance to herbivores. *International Journal of Molecular Sciences* 13, 17077–17103. doi: https://doi.org/10.3390/ijms131217077

Huchelmann, A., Boutry, M., and Hachez, C. (2017). Plant glandular trichomes: Natural cell factories of high biotechnological interest. *Plant Physiology* 175, 6–22. doi: https://doi.org/10.1104/pp.17.00727

Hunter, E. A. and Gibbs, J. P. (2014). Densities of Ecological Replacement Herbivores Required to Restore Plant Communities: A Case Study of Giant Tortoises on Pinta Island. *Galápagos. Restoration Ecology* 22, 248–256. doi: https://doi.org/10.1111/rec.12055

Kennedy, G. G. (2003). Tomato, pests, parasitoids, and predators: tritrophic interactions involving the genus *Lycopersicon*. *Annual Review of Entomology* 48, 51–72. doi: https://doi.org/10.1146/annurev.ento.48.091801.112733

Khazaei, H. and Madduri, A. (2022). The role of tomato wild relatives in breeding disease-free varieties. *Genetic resources* 3(6), 64–73. doi: https://doi.org/10.46265/genresj.PSES6766

Leckie, B. M., D'ambrosio, D. A., Chappell, T. M., Halitschke, R., De Jong, D. M., Kessler, A., Kennedy, G. G., and Mutschler, M. A. (2016). Differential and synergistic functionality of acylsugars in suppressing oviposition by insect herbivores. *PLoS ONE* 11(4). doi: https://doi.org/10.1371/journal.pone.0153345

Levin, D. A. (1973). The role of trichomes in plant defense. *The Quarterly Review of Biology* 48, 3–15. url: https://www.journals.uchicago.edu/doi/epdf/10.1086/407484.

Lihavainen, J., Ahonen, V., Keski-Saari, S., Sõber, A., Oksanen, E., and Keinänen, M. (2017). Low vapor pressure deficit reduces glandular trichome density and modifies the chemical composition of cuticular waxes in silver birch leaves. *Tree Physiology* 37, 1166–1181. doi: https://doi.org/10.1093/treephys/tpx045

Lucatti, A. F., Van Heusden, A. W., De Vos, R. C., Visser, R. G. F., and Vosman, B. (2013). Differences in insect resistance between tomato species endemic to the Galapagos Islands. *BMC Evolutionary Biology* 13, 175–175. doi: https://doi.org/10.1186/1471-2148-13-175

Luckwill, L. C. (1943). The genus *Lycopersicon*: An historical, biological and taxanomic survey of the wild and cultivated tomatoes (U.K: Aberdeen University Press).

Ma, J. F. (2004). Role of silicon in enhancing the resistance of plants to biotic and abiotic stresses. *Soil Science and Plant Nutrition* 50(1), 11–18. doi: https://doi.org/10.1080/00380768.2004.10408447

Mahfouze, S. A. and Mahfouze, H. A. (2019). A Comparison between CAPS and SCAR markers in the detection of resistance genes in some tomato genotypes against *Tomato Yellow Leaf Curl Virus* and whitefly. *Jordan Journal of Biological Sciences* 12, 123–133. url: https://jjbs.hu.edu.jo/files/vol12/n2/Paper%20number%201.pdf.

Mymko, D. and Avila-Sakar, G. (2019). The influence of leaf ontogenetic stage and plant reproductive phenology on trichome density and constitutive resistance in six tomato varieties. *Arthropod-Plant Interactions* 13, 797–803. doi: https://doi.org/10.1007/s11829-019-09690-3

Oksanen, E. (2018). Trichomes form an important first line of defence against adverse environment-New evidence for ozone stress mitigation. *Plant, Cell & Environment* 41, 1497–1499. doi: https://doi.org/10.1111/pce.13187

Pailles, Y., Awlia, M., Julkowska, M., Passone, L., Zemmouri, K., Negrão, S., Schmöckel, S. M., and Tester, M. (2020). Diverse traits contribute to salinity tolerance of wild tomato seedlings from the Galapagos Islands. *Plant Physiology* 182, 534–546. doi: https://doi.org/10.1104/pp.19.00700

Pailles, Y., Ho, S., Pires, I. S., Tester, M., Negrão, S., and Schmöckel, S. M. (2017). Genetic diversity and population structure of two tomato species from the Galapagos Islands. *Frontiers in Plant Science* 8, 138–138. doi: https://doi.org/10.3389/fpls.2017.00138

Paudel, S., Lin, P. A., Foolad, M. R., Ali, J. G., Rajotte, E. G., and Felton, G. W. (2019). Induced plant defenses against herbivory in cultivated and wild tomato. *Journal of Chemical Ecology* 45, 693–707. doi: https://doi.org/10.1007/s10886-019-01090-4

Peck, S. B. (2008). Galápagos Islands Insects: Colonization, Structure, and Evolution. In *Encyclopedia of Entomology,* ed. Capinera, J. L., (Dordrecht: Springer).

Pelletier, Y. (1990). The effect of water stress and leaflet size on the density of trichomes and the resistance to Colorado potato beetle larvae (*Leptinotarsa decemlineata* [say]) in *Solanum berthaultii* hawkes. *The Canadian Entomologist* 122(6), 1141–1147. doi: https://doi.org/10.4039/Ent1221141-11

R Core Team (2021). R: a language and environment for statistical computing. R Foundation for Statistical Computing. url: https://www.R-project.org.

Rakha, M., Bouba, N., Ramasamy, S., Regnard, J. L., Hanson, and P (2017a). Evaluation of wild tomato accessions (*Solanum* spp.) for resistance to two-spotted spider mite (*Tetranychus urticae* Koch) based on trichome type and acylsugar content. *Genetic Resources and Crop Evolution* 64, 1011–1022. doi: https://doi.org/10.1007/s10722-016-0421-0

Rakha, M., Hanson, P., and Ramasamy, S. (2017b). Identification of resistance to *Bemisia tabaci* Genn. in closely related wild relatives of cultivated tomato based on trichome type analysis and choice and no-choice assays. *Genetic Resources and Crop Evolution* 64, 247–260. doi: https://doi.org/10.1007/s10722-015-0347-y

Rick, C. M. (1961). Biosystematic studies on Galápagos tomatoes.

Savory, E. A. (2004). Modification of the PGO Assay for Use in Acylsugar Quantification (Ithaca, NY, USA) 36-36.

Schilmiller, A. L., Charbonneau, A. L., and Last, R. L. (2012). Identification of a BAHD acetyltransferase that produces protective acyl sugars in tomato trichomes. *Proceedings of the National Academy of Sciences of the United States of America* 109, 16377–16382. doi: https://doi.org/10.1073/pnas.1207906109

Vendemiatti, E., Therezan, R., Vicente, M. H., Pinto, M. D. S., Bergau, N., Yang, L., Bernardi, W. F., De Alencar, S. M., Zsögön, A., Tissier, A., Benedito, V. A., and Peres, L. E. P. (2022). The genetic complexity of type-IV trichome development reveals the steps towards an insect-resistant tomato. *Plants* 11(10), 1309–1309. doi: https://doi.org/10.3390/plants11101309

Vosman, B., Kashaninia, A., Van't Westende, W., Meijer-Dekens, F., Van Eekelen, H., Visser, R. G. F., De Vos, R. C. H., and Voorrips, R. E. (2019). QTL mapping of insect resistance components of *Solanum galapagense*. *Theoretical and Applied Genetics* 132, 531–541. doi: https://doi.org/10.1007/s00122-018-3239-7

Vosman, B., Van't Westende, W. P. C., Henken, B., Van Eekelen, H. D. L. M., De Vos, R. C. H., and Voorrips, R. E. (2018). Broad spectrum insect resistance and metabolites in close relatives of the cultivated tomato. *Euphytica* 214, 46–46. doi: https://doi.org/10.1007/s10681-018-2124-4

Wang, X., Shen, C., Meng, P., Tan, G., and Lv, L. (2021). Analysis and review of trichomes in plants. *BMC Plant Biology* 21, 70–70. doi: https://doi.org/10.1186/s12870-021-02840-x

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nature Reviews Genetics* 2, 983–989. doi: https://doi.org/10.1038/35103590

Zhang, X., Thacker, R. R., and Snyder, J. C. (2008). Occurrence of 2,3-dihydrofarnesoic acid, a spidermite repellent, in trichome secretions of L*ycopersicon esculentum* × *L. hirsutum* hybrids. *Euphytica* 162, 1–9. doi: https://doi.org/10.1007/s10681-007-9489-0

Zhang, Y., Song, H., Wang, X., Zhou, X., Zhang, K., Chen, X., Liu, J., Han, J., Wang, and A (2020). The roles of different types of trichomes in tomato resistance to cold, drought, whiteflies, and *Botrytis*. *Agronomy* 10, 411–411. doi: https://doi.org/10.3390/agronomy10030411

Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 13–28
DOI: 10.46265/genresj.REJR6896
https://www.genresj.org
ISSN: 2708-3764

European genetic resources conservation in a rapidly changing world: three existential challenges for the crop, forest and animal domains in the 21st century

François Lefèvre [*,a], **Danijela Bojkovski** [b], **Magda Bou Dagher Kharrat** [c,d], **Michele Bozzano** [d], **Eléonore Charvolin-Lemaire** [e], **Sipke J Hiemstra** [f], **Hojka Kraigher** [g], **Denis Laloë** [e], **Gwendal Restoux** [e], **Suzanne Sharrock** [h], **Enrico Sturaro** [i], **Theo van Hintum** [f], **Marjana Westergren** [g], **Nigel Maxted** [j] **and GenRes Bridge Expert Panel** [k]

[a] *INRAE, Ecologie des Forêts Méditerranéennes, URFM, Domaine Saint Paul Agroparc, 84914, Avignon, France*
[b] *Biotechnical Faculty, Department of Animal Science, Jamnikarjeva 101, 1000, Ljubljana, Slovenia*
[c] *Laboratory of biodiversity and functional genomics, Faculty of science, Saint Joseph University, Beirut, Lebanon*
[d] *European Forest Institute, Sant Pau Art Nouveau Site, Carrer Sant Antoni M. Claret, 167, 08025, Barcelona, Spain*
[e] *GABI, AgroParisTech, INRAE, Université Paris-Saclay, 78350, Jouy-en-Josas, France*
[f] *Centre for Genetic Resources, the Netherlands, Wageningen University & Research, Radix Building 107, Droevendaalsesteeg 1, 6708 PB, Wageningen, the Netherlands*
[g] *Department of Forest Physiology and Genetics, Slovenian Forestry Institute, Večna pot 2, 1000, Ljubljana, Slovenia*
[h] *Botanic Gardens Conservation International, Descanso House, 199 Kew Road, Richmond, TW9 3BW, UK*
[i] *Department of Agronomy, Food, Natural Resources, Animals and the Environment DAFNAE, Università degli Studi di Padova, Viale dell'Università 16, 35020, Legnaro (PD), Italy*
[j] *School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK*
[k] *Full list available at the end of the article*

Abstract: Even though genetic resources represent a fundamental reservoir of options to achieve sustainable development goals in a changing world, they are overlooked in the policy agenda and severely threatened. The conservation of genetic resources relies on complementary *in situ* and *ex situ* approaches appropriately designed for each type of organism. Environmental and socioeconomic changes raise new challenges and opportunities for sustainable use and conservation of genetic resources.

Aiming at a more integrated and adaptive approach, European scientists and genetic resources managers with long experience in the agricultural crop, animal and forestry domains joined their expertise to address three critical challenges: (1) how to adapt genetic resources conservation strategies to climate change, (2) how to promote *in situ* conservation strategies and (3) how can genetic resources conservation contribute to and benefit from agroecological systems. We present here 31 evidence-based statements and 88 key recommendations elaborated around these questions for policymakers, conservation actors and the scientific community.

We anticipate that stakeholders in other genetic resources domains and biodiversity conservation actors across the globe will have interest in these crosscutting and multi-actor recommendations, which support several biodiversity conservation policies and practices.

Keywords: Agroecology, climate change, *in situ* conservation, multi-actor engagement, policy

Citation: Lefèvre, F., Bojkovski, D., Bou Dagher Kharrat, M., Bozzano, M., Charvolin-Lemaire, E., Hiemstra, S. J., Kraigher, H., Laloë, D., Restoux, G., Sharrock, S., Sturaro, E., van Hintum, T., Westergren, M., Maxted, N., GenRes Bridge Expert Panel (2024). European genetic resources conservation in a rapidly changing world: three existential challenges for the crop, forest and animal domains in the 21st century. *Genetic Resources* 5 (9), 13–28. doi: 10.46265/genresj.REJR6896.

Received: 26.05.2023 | Accepted: 20.02.2024 | Published online: 11.03.2024

Introduction

Genetic resources are at the crossroads of multiple policy agendas, in particular biodiversity conservation and sustainable development goals (FAO (2019), e.g. p. 3 about diversity loss in production systems; IPBES (2019), e.g. SPM-A6 and p. 247 about erosion of genetic resources diversity; IPCC (2019), e.g. SPMB6.2; CBD (2020), e.g. Aichi Targets 13, 14 and 16; CBD (2022), e.g. Targets 4 and 13). In the Common International Classification of Ecosystem Services (Haines-Young and Potschin, 2018), the term 'genetic resources' is not used but genetic resources explicitly appear both as provisioning services under the term 'genetic material', as regulation and maintenance services under the term 'gene pool', and could also be considered as cultural services in the class of "characteristics or features of living systems that have an option or bequest value". This classification reveals the multiple values of genetic resources: direct use value of well-characterized genetic material, option value of the genetic diversity, bequest value of biodiversity components. However, the threat to and the erosion of genetic resources diversity, both in wild populations and in production systems, is now widely documented (FAO, 2019; IPBES, 2019; CBD, 2020) and the related Aichi target of safeguarding genetic diversity has not yet been achieved (CBD, 2020). The Kunming-Montreal Global Biodiversity Framework calls for "Target 4: Ensure urgent management actions [...] for the recovery and conservation of species [...] to significantly reduce extinction risk, as well as to maintain and restore the genetic diversity within and between populations of native, wild and domesticated species to maintain their adaptive potential including through *in situ* and *ex situ* conservation" (CBD, 2022). Despite their critical importance for sustainable development, on the one hand, and the ongoing erosion of their diversity, on the other hand, genetic resources are largely overlooked by policymakers. Three reasons may explain this paradox. First, the role of within-species genetic diversity remains poorly understood and appreciated in biodiversity conservation (Hoban *et al*, 2020). Second, few recognize the existential importance of within-species genetic diversity in sustaining continued crop, forest and animal production. Third, the term 'resources' does not explicitly refer to the notions of diversity, which in a changing world is valued over quantity, and rather focuses on the use aspects.

The important role of diversity between and within crop[1], animal and forest genetic resources for maintaining production has been recognized for centuries, but the actual term 'genetic resources' was only coined in 1967 at the International Conference on Crop Plant Exploration and Conservation (Frankel and Bennett, 1970). It received a common definition a nd global consideration in the Convention on Biological Diversity (CBD (1992), Article 2): "genetic resources means genetic material of actual or potential value". Thus, genetic resources refer to genetic diversity of actual or potential use value between and within species, with a continuum from domestic gene pools (varieties, breeds, isolates) to wild populations. The evolutionary processes during domestication are mainly driven by targeted human interventions such as selection, migration and hybridization. In the case of partially anthropized systems where populations are exploited and managed through natural regeneration systems (e.g. many forests, fisheries and grazed areas), management practices indirectly shape genetic resources by interfering with natural evolutionary and ecological processes. The domestic and wild gene pools are often connected in the landscape where they develop three types of interactions: (1) competition for land (Grau *et al*, 2013), (2) ecological interactions (Pozo *et al*, 2021) and (3) possible gene flow between domestic gene pools and their wild relatives (Ellstrand and Rieseberg, 2016). Thus, genetic resources conservation has to be considered in the context of social-ecological systems, where humans directly or indirectly sustain genetic resources and humankind benefits substantially from their genetic diversity maintenance and utilization.

The communities working on genetic resources have historically tended to be defined by the scope of their taxonomic coverage, each specializing in crop, forestry, domesticated animal, fish, microbe or pollinator genetic diversity conservation and use, the linking of conservation with use of the conserved resource setting them apart from the broader biodiversity conservation community. The crop, forestry and domesticated animal domains have worked largely independently to develop conservation and use actions specifically designed within their respective contexts, without sharing experience and benefiting from mutually advantageous collaboration. To fill this gap, the European Union's Horizon 2020 'GenRes Bridge' project brought together for the first time the European crop, forestry and domesticated animal genetic resources networks (http://www.genresbridge.eu/).

Three individual networks have been coordinating and facilitating genetic resources conservation and use in Europe for more than 25 years within their respective domains: the European Cooperative Programme for Plant Genetic Resources (ECPGR, https://www.ecpgr.org/), the European Regional Focal Point for Animal Genetic Resources (ERFP, https://www.animalgeneticresources.net/), and the European Forest Genetic Resources Programme (EUFORGEN, https://www.euforgen.org/). The three networks joined forces in the GenRes Bridge project to elaborate a *Genetic Resources Strategy for Europe* speaking with a stronger policy 'voice' and

*Corresponding author: François Lefèvre (francois.lefevre.2@inrae.fr)

[1] In this article, the term 'crop genetic resources' encompasses plants used for agricultural production, including crop wild relatives and wild food plants, and is used instead of the more common 'plant genetic resources' to avoid confusion with the forest domain, which deals with the genetic resources of forest trees and other woody plants.

facilitating more effective implementation. This strategy consists of a comprehensive overarching framework of appropriate coordinated actions to conserve and sustainably use genetic resources (GenRes Bridge Project Consortium, ECPGR, ERFP and EUFORGEN, 2021), and three derived domain-specific documents accounting for respective contexts (ECPGR, 2021; ERFP, 2021; EUFORGEN, 2021).

Although the biological and socioeconomic contexts of conservation and sustainable use of genetic resources differ for agricultural crop, animal farming and forestry domains, from the biological point of view, the coexistence of human-directed and natural evolutionary processes are common to all domains of genetic resources. Furthermore, from the socioeconomic point of view, sustainable development depends on continued access to a combined set of genetic resources from each domain and combined production systems (e.g. agroforestry). Finally, genetic resources conservation and sustainable use in all domains are currently facing common challenges in the context of environmental, socioeconomic and legal changes. Therefore, joining expertise from different domains, with various social-ecological contexts, will help effectively address these challenges for sustainable use and conservation of genetic resources. This paper illustrates crosscutting and integrated solutions to three existential challenges for genetic resources in the 21^{st} century:

1. How to adapt genetic resources conservation strategies to climate change
2. How to promote *in situ* conservation strategies (with common objectives despite diverse modalities across domains)
3. How can genetic resources conservation contribute to and benefit from agroecological systems

We here provide general arguments and recommendations reusable by different genetic resources and conservation communities.

Methodology

The three challenges were addressed during three workshops engaging a global panel of 43 invited experts on genetic resources, i.e. scientists in conservation science and practitioners, from 16 countries and one international organization, with balanced representation of the three domains. To develop policy-relevant conservation science, we first identified **evidence-based statements** common to all genetic resources domains, beyond biological and socioeconomic specificities. These statements were based on the reports of international agencies and platforms (FAO, 2019; IPBES, 2019; IPCC, 2019; CBD, 2020, 2022) and workshop participants' expertise. Then, each evidence-based statement was deconstructed and reviewed, and **key arguments and recommendations** were derived for each of the three prime target audiences: policymakers, conservation actors and the scientific community. Final statements and recommendations were elaborated through online collaboration.

These statements and recommendations have broad general interest not only for other genetic resources domains, e.g. fisheries or industrial microbiology, but also for other biodiversity conservation programmes accounting for genetic diversity at global, regional and national levels. Here, we present a list of 31 evidence-based statements, and 88 arguments and key recommendations related to the three challenges. We then briefly analyze the targeted audiences and describe how these particular statements and recommendations were considered in the *Genetic Resources Strategy for Europe*. Finally, we propose some perspectives building on the inter-domain collaborative experience.

Results

How to adapt genetic resources conservation strategies in the context of climate change

Ten statements (CC1 to CC10) and 26 recommendations on this challenge are given in Table 1.

The first three statements, CC1 to CC3, raise the point that, in the context of climate change, the diversity of genetic resources is both at risk while also representing a reservoir of options to sustain agriculture and forestry in the face of multiple uncertainties (Koskela *et al*, 2007; FAO, 2015). Therefore, to better use genetic resources, we need to explore and characterize their diversity and potential benefits using both *in situ* material and *ex situ* collections. Scientists and actors on genetic resources in all domains agree on the severe level of threats of erosion and extinction currently impacting genetic resources diversity. Efforts to improve the conservation, characterization and use of genetic resources need to be actively promoted, even if there is still a lack of quantitative assessment of these threats (IPBES, 2019).

A second set of statements, CC4 to CC6, stresses the need for raising awareness on genetic resources diversity, conservation and use issues, and for better sharing science-based knowledge with multiple actors and policymakers involved. This lack of knowledge sharing was identified as a limiting factor in genetic resources conservation and use. The related recommendations aim to support the 'chain of knowledge' from science to policy decisions, on the one hand, and to facilitate exchanges of information or material among local expert communities in genetic resources, on the other hand, both needed to adapt genetic resources conservation and use strategies in the context of climate change.

Table 1. Statements and recommendations on how to adapt genetic resources conservation strategies in the context of climate change. Prime target audience: P, policymakers; S, scientific community; C, conservation actors (other than P and S).

Statements		Arguments and recommendations	
CC1	Climate change poses a significant threat to genetic diversity.	CC1.1	Immediate conservation action is needed now to prevent loss of genetic diversity and political commitment linked to policy action is required to support this initiative. [P, C]
		CC1.2	The diversity of climate change-related threats needs clarification, and their mitigation demands a diversity of responses. [S]
		CC1.3	Threats from climate change need to guide future genetic resources conservation and use strategy developments and prioritize mitigating action implementation. [C]
CC2	Genetic diversity provides resilience in the face of unexpected change.	CC2.1	Social and economic studies are required to evaluate how genetic resources diversity mitigates threats to food security and other contributions of agriculture and forests to people. [S]
		CC2.2	Studies are required to provide concrete examples of the benefits provided by genetic diversity in the agroecosystems and the values of such ecological, social and economic benefits. [C, S]
CC3	In order to deploy sources of resilience, diversity has to be identified and characterized.	CC3.1	Characterization of genetic resources and sharing of this information in a standardized manner are essential. [C, S]
		CC3.2	Improved availability of more standardized scientific information on genetic and phenotypic diversity is required. [S]
		CC3.3	Predictive characterization may also be used to speed up identification of desired traits. [S]
CC4	Genetic resource-related policies should be based upon relevant scientific findings.	CC4.1	Science provides evidence-based insights that are essential in defining effective policies. [P, S]
		CC4.2	Increased collaboration between scientists and policymakers could improve the uptake of scientific messages in policy decisions. [P, S]
CC5	Awareness of the importance of genetic diversity for the survival of humankind should be raised.	CC5.1	Public and political support for genetic resources conservation is essential to secure appropriate funding. [P]
		CC5.2	The general public, but also policymakers, are rarely aware of the important role of the diversity provided by genetic resources in adaptation to the changing climate and changing demands from society. [P, C, S]
CC6	Cooperation between formal genetic resources conservation, breeding programmes and community-based conservation initiatives should be improved.	CC6.1	Community-based activities can play an important role in the identification of resilient genetic resources suitable for the changing environment. [C]
		CC6.2	Link between the formal genetic resources management systems with local initiatives is often weak, and access to each other's genetic resources is often limited. [C]

Continued on next page

Table 1 continued

Statements		Arguments and recommendations	
CC7	Early signs of potential future needs and threats to genetic diversity and genetic resources use have to be detected.	CC7.1	Foresight studies (Horizon scanning exercises) can produce scenarios to guide long-term strategies for genetic resources conservation and use. [C, S]
		CC7.2	Studies should address possible relevant socioeconomic changes and technological advances. [S]
		CC7.3	Scenarios of change should also consider biotic and abiotic hazards, as well as external contingencies. [S]
CC8	Periodic monitoring of the actual impacts of climate change on genetic diversity and associated organisms is required.	CC8.1	Given the climate crisis and associated uncertainties, regular monitoring allows tracking of changes and development of scenarios. [C]
		CC8.2	Based on the knowledge gained from monitoring, prioritization of actions can and should be made. [C]
		CC8.3	Monitoring should not only include genetic resources but also associated organisms, like pollinators. [C]
CC9	Communication between all practitioners involved in genetic resources management and use (genebank managers, *in situ* network managers, breeders, farmers and foresters, protected area managers, etc.), policymakers and scientists, needs to be improved.	CC9.1	Coordination of genetic resource-related actions will improve with better communication. [P, C, S]
		CC9.2	This will lead also to establishing and reinforcing collaboration between multiple actors involved in genetic resources conservation and sustainable use. [P, C, S]
		CC9.3	Multiple actors will contribute to awareness raising on the basis of values and trade-offs between all relevant values. [P, C, S]
CC10	The traits related to adaptation of genetic resources to climate change need to be given more attention in research.	CC10.1	Tolerance to climate-related hazards (heat, drought, etc.), and resistance to existing and emerging pests and diseases will become essential in adaptation to climate change; more knowledge about these traits will become essential to allow adaptation. [S]
		CC10.2	Genetic resources will benefit from basic research on these traits and their use as study objects should be promoted. [S]
		CC10.3	The way how these traits can support adaptation of agroecosystems, or help to diversify these systems, should be assessed. [S]

Table 2. Statements and recommendations on how to promote *in situ* conservation strategies. Prime target audience: P, policymakers; S, scientific community; C, conservation actors (other than P and S).

Statements		Arguments and recommendations	
IS1	Dynamic *in situ* conservation strategies integrate adaptation to global change into the conservation process.	IS1.1	Genetic resources are kept in the productive environment allowing exposure to change and stress situations. [C]
		IS1.2	*In situ* conserved genetic resources stay useful in a changed environment. [C]
IS2	*In situ* conservation continuously contributes to multiple ecosystem services and benefits to people.	IS2.1	*In situ* conservation with management contributes to rural development. [P, C]
		IS2.2	*In situ* conservation provides a broad range of diversity to users. [P, C]
		IS2.3	*In situ* conservation also contributes to regulation and maintenance as well as cultural ecosystem services. [P, C]
		IS2.4	*In situ* conservation allows for better dynamic reactions to different drivers of change, including market needs and new market niche exploration. [P, C]
IS3	Effective and efficient *in situ* conservation and sustainable use of genetic diversity rely on the participation of multiple actors and coordinated efforts.	IS3.1	Key actors and potential new actors should be identified/recognized and involved in genetic resources strategies. [P, C, S]
		IS3.2	*In situ* conservation programmes should be designed based on a participatory approach involving all actors. [P, C, S]
		IS3.3	All actors need to be financially supported and incentives should rely on available scientific proofs. [P, C, S]
IS4	Coordination of efforts by the various actors involved in dynamic *in situ* conservation is needed to ensure that long-term objectives are reached.	IS4.1	Actions are needed to strengthen the links between all actors (practitioners, scientists, etc.) in *in situ* management of genetic resources. [C, S]
		IS4.2	Strategical recommendations, guidelines and directives should be tested by practitioners in collaboration with scientists and extension services before general implementation. [C,S]
IS5	Coordinated and standardized national inventories of *in situ* genetic resources have to be prepared and made accessible.	IS5.1	Inventories of *in situ* genetic resources improve our knowledge about what and where they are still maintained or cultivated, thus providing a resource from where important traits can be identified for plant and animal improvement by breeders or direct utilization by farmers. [C]

Continued on next page

Table 2 continued

Statements		Arguments and recommendations	
		IS5.2	Inventories of *in situ* genetic resources are needed for planning more systematic crop-collecting missions addressing possible gaps and for designing on-farm conservation and management projects. [C]
		IS5.3	Data structure of the national inventories of *in situ* genetic resources should feed the appropriate European information systems. [C, S]
IS6	Active genetic management including selective breeding for performance traits and evolution-oriented forest management can contribute to *in situ* conservation 'in use' of genetic resources.	IS6.1	Knowledge of the qualitative and quantitative impacts of agricultural and forestry practices on evolutionary processes needs to be improved and shared with practitioners. [C, S]
		IS6.2	Information is needed on the potential role of active genetic management on performance and adaptive traits to improve self-sustainability of *in situ* genetic resources within the constraints of their typical characteristics. [C, S]
IS7	New operational tools for the *in situ* conservation of genetic resources have to be developed and practically applied in all domains to increase our understanding and capacity to develop more efficient strategies for genetic resources conservation and use.	IS7.1	There is a need for operational tools for *in situ* characterization, evaluation, management and monitoring of genetic resources. [S]
		IS7.2	The development of tools is a dynamic process, both for the update and the uptake, in which the three domains could share experiences and innovations. [S]
IS8	A commitment and a concept for long-term genetic monitoring are needed to guide *in situ* conservation and sustainable use of genetic resources.	IS8.1	Genetic monitoring is an efficient tool to characterize and detect changes in genetic diversity over time. [C, S]
		IS8.2	The standardization and/or comparability of genetic monitoring information over time must be ensured to allow proper assessment of the changes (independently of the new tools). [S]
		IS8.3	Sufficient resources should be committed to implementing long-term genetic monitoring. [P]
IS9	The complementarity between *in situ* and *ex situ* techniques can contribute to increasing the systematic coverage of genetic diversity under conservation as well as the efficiency of genetic resources conservation.	IS9.1	There is a need to investigate explicitly the multiple advantages, and risks, of combining *in situ* and *ex situ* strategies: provide insurance and backup, facilitate access to material, provide additional material for reinforcement *in situ*, etc. [S]
		IS9.2	Suitable methods and tools to integrate dynamic and static conservation approaches should be developed for all domains. [S]
		IS9.3	Such integrated approaches will allow opportunities for a wider range of stakeholders, including local communities, to participate in different networks at various levels. [C, S]

Continued on next page

Table 2 continued

Statements		Arguments and recommendations	
IS10	Opportunities for the protection with utilization and valorization of the diversity of genetic resources in various ecosystems should be promoted.	IS10.1	Characterization of the genetic diversity of genetic resources available in cultivated areas, protected areas, rural and urban spaces, and private and public gardens, is needed. [C, S]
		IS10.2	Using these spaces for conservation of both wild and cultivated diversity should be promoted and supported. [P]
IS11	Long-term conservation policies, strategies and programmes are needed to ensure dynamic *in situ* conservation of genetic resources diversity.	IS11.1	Long-term perspective of genetic resources conservation strategies must clearly appear in the related EU policies, strategies and programmes, to support adaptive dynamics in the *in situ* conservation devices. [P]
		IS11.2	Long-term policy support for *in situ* management and monitoring is needed. [P]
IS12	Cooperation within and across domains at the European scale to develop dynamic *in situ* conservation strategies of genetic resources is needed.	IS12.1	Dynamic *in situ* conservation strategies of genetic resources can use very diverse methods and tools; sharing experiences and research efforts across domains, geographic areas or species, is needed. [C, S]
		IS12.2	Dynamic *in situ* strategies provide opportunities to combine multiple genetic resources targets in the same conservation action, including trans-domain actions. [C, S]
		IS12.3	Following an adaptive management framework, permanent upgrading should be incorporated into the strategies, including complementarity of *in situ* and *ex situ* conservation. [C, S]

Finally, statements CC7 to CC10 illustrate the fact that climate change is forcing us to revise our vision, tools and methods for sustainable genetic resources management particularly as genetic resources are evolving in an unpredictable and dynamic context. More than on current diversity per se, the focus should be put on its trajectory and the drivers of this trajectory. Putting genetic resources management in such a dynamic perspective also requires monitoring strategies. To support innovative approaches, the recommendations provide a list of actions to develop governance and decision support tools like indicators or scenarios and new target traits of interest for research.

How to promote *in situ* conservation strategies

Twelve statements (IS1 to IS12) and 31 arguments and recommendations on this challenge are given in Table 2.

The first two statements emphasize key characteristics of *in situ* conservation: framed in a dynamic perspective as mentioned above, genetic diversity continuously evolves in diverse and changing biotic and abiotic environments, and it combines conservation with sustained local use benefits. This integrative conservation approach illustrates the socioecological dimension of genetic resources: ecological adaptation interlaces with contributions to people. Based on these statements, the European experts from the different domains derived some general 'arguments' to explain and promote the integrative *in situ* approach in various contexts, rather than specific recommendations for each domain.

Combining conservation and use, *in situ* conservation strategies systematically rely on multiple and diverse actors, either directly managing genetic resources or indirectly controlling the environment in which they are left evolving. To ensure effective and efficient *in situ* conservation, multiple actors must be coordinated not only locally, but also at national level to establish networks of local initiatives. These points are raised in statements IS3 to IS5. The related recommendations aim to create and support multi-actor engagement, coordinate their actions with appropriate science-based guiding tools and monitor jointly the development of actions and the diversity of genetic resources that result from these actions.

Statements IS6 to IS8 identify three specific aspects of *in situ* conservation on which knowledge, i.e. both scientific knowledge and practitioners' expertise, must urgently be expanded. The first aspect is a quantitative assessment of the potential role that ecosystem management (agriculture or forestry practices) can play as evolutionary drivers of genetic resources diversity. The second aspect is the need to develop operational tools specifically dedicated to *in situ* genetic resources and actors involved. The third aspect is the need for standardized, long-term monitoring programmes applicable to all three domains, in the framework of international initiatives towards global genetic diversity monitoring (Hoban *et al*, 2022). The related recommendations represent priority actions in these fields.

Statements IS9 and IS10 reveal other actions which *in situ* conservation could easily complement with great benefits: association with other genetic resources conservation approaches (i.e. *ex situ* conservation) or other compatible land uses (e.g. protected, cultivated or even urban areas). The related recommendations aim to support this integrative approach of *in situ* genetic resources conservation in a broader framework.

Finally, the last two statements and related recommendations raise the fact that benefiting from all integrative dimensions of *in situ* conservation requires long-term programmes and support, as well as large-scale cooperation.

How can genetic resources conservation contribute to and benefit from agroecological systems

Nine statements (AE1 to AE9) and 31 recommendations on this challenge are given in Table 3. The challenge here is to search for mutual opportunities between the emerging interest to apply agroecological principles in the development of agricultural and forestry systems and genetic resources conservation.

During the discussions, European experts highlighted the potential benefit of genetically diverse resources in the agroecology framework (Chable *et al*, 2020). To reach this benefit, the recommendations related to the first statement AE1 focus on the identification and contextualization of genetic resources in agroecological systems, and on sharing this information. The second statement AE2 highlights the key role of genetic resources managers in agroecology. Five recommendations explain how to support these actors.

The next five statements, AE3 to AE7, underline the specific relevance of the local scale (landscape, territory) to develop synergies between genetic resources conservation, agroecology, and resilience/sustainability of agriculture and forestry systems. Indeed, local actors involved in agroecology have the capacity to implement integrative *in situ* conservation within-sites as proposed in the previous sections, while the diversity of social-ecological contexts among localities contributes to maintaining between-sites diversity. This idealistic view relies on the engagement of multiple actors in multiple sites, which cannot be achieved without searching for win-win solutions: altogether, 16 recommendations related to the above statements were proposed.

Finally, statement AE8 stresses the need for a holistic and social-ecological approach to consider all levels of diversity together, and statement AE9 raises the particular importance of data management for this challenge. Indeed, various types of data should be handled jointly: different kinds of genetic resources, multiple uses and related genetic resources characteristics, biological to socioeconomic data or regulation information, georeferencing, etc.

Table 3. Statements and recommendations on how genetic resources conservation can contribute to and benefit from the agroecology transition. Prime target audience: P, policymakers; S, scientific community; C, conservation actors (other than P and S).

Statements		Arguments and recommendations	
AE1	Diverse genetic resources are key elements in the agroecology framework.	AE1.1	All component species and their role in each agroforestry system need to be identified. [C]
		AE1.2	Long-term study cases should be established in different geographical and socioeconomical contexts to analyze and demonstrate the impact of genetic resources on agroecology systems across different time scales. [C, S]
		AE1.3	The use of a broad diversity of genetic resources, and the exchange of genetic resources and related information should be promoted. [P]
		AE1.4	Increased knowledge of the (epi)genetic variability of genetic resources will favour their integration into agroecological systems. [S]
AE2	Genetic resources managers have a key role to play in the agroecological transition.	AE2.1	Policy support must be implemented with a long-term view and connected with public support. [P]
		AE2.2	Ecological performance (multi-criteria performance evaluation) must be integrated into the value chain labelling process. [P]
		AE2.3	Further research is needed on how to manage genetic resources for a transition from an intensive (standard) production system to a more ecologically oriented mode of production. [S]
		AE2.4	The long-term benefits of ecological performance/sustainability when all three domains are considered should be further investigated and knowledge communicated to the end-users. [C, S]
		AE2.5	Scientists and genetic resources managers together should propose decision-making tools, identify actors and consider geographical information supporting sustainable use of genetic resources in the agroecological transition. [C, S]
AE3	Research, policy, managers and users' communities on genetic resources must be connected.	AE3.1	Demonstrations of how useful genetic resources are for farmers, forest managers, and their respective systems, are needed. [C, S]
		AE3.2	The views of the users must be taken into account in the design and the analysis of study cases and in the implementation of the strategy. [C]
		AE3.3	Research has to produce a synthesis of results and knowledge for stakeholders. [S]
		AE3.4	Common terminology must be shared across the different communities involved. [C, S]

Continued on next page

Table 3 continued

Statements		Arguments and recommendations	
AE4	The agroecology framework provides an opportunity to look at landscape/territory scale which is also relevant for genetic resources management	AE4.1	Management and research activities must consider landscape/territory scale as the way to associate genetic resources with ecosystem services and identify multiple values of genetic resources. [C, S]
		AE4.2	Maintenance of diversity rather than specific unicity of genetic resources must be supported to avoid negative side effects of decreased diversity. [P]
		AE4.3	Conservation of diversity at local scale can be costly, so there is a need to involve multiple actors to share costs and to work on social organizations. [P]
		AE4.4	Actions that connect across territories should be supported: locally appropriate genetic resources may be non-local, may need to be imported (incl. from *ex situ* genebanks); reciprocally each territory may handle genetic resources of poor local value but high value for elsewhere. [P]
AE5	Human dimension and local knowledge are important for sustainable use of genetic resources and cultural heritage	AE5.1	Consideration and characterization of local knowledge and traditional use have to be accounted for in the characterization of genetic resources. [C, S]
		AE5.2	Participatory approaches must be supported and developed. [P, C, S]
AE6	Integrated genetic resources management contributes to increasing biodiversity as a factor of resilience of production systems in the agroecology framework	AE6.1	Case studies can be used to improve knowledge of the respective roles of the different levels of diversity in agroecological systems: from the within-crop/breed/population/species diversity to the between-crop/breed/population/species diversity. [C, S]
		AE6.2	Traceability of genetic resources uses is needed to analyze crisis situations and document the role of diversity in buffering changes and unexpected disturbances. [C, S]
		AE6.3	Research should further investigate the conditions where diversity can be beneficial or detrimental to productivity. [S]
		AE6.4	The agroecology framework should be implemented with the aim to optimize both the production and the management of diversity, either for conservation or for preserving variability for future selection. [C, S]
		AE6.5	Scenarios of complementarity between agroecology and genetic resources conservation must be evidenced. [C, S]
AE7	The agroecology framework takes advantage of local context specificities.	AE7.1	There is no single solution to be applied everywhere: it is important to find a way to share experience/methods/tools from local to global level. [P, C, S]
AE8	A holistic approach is needed to consider all levels of diversity and time scale from an agroecological perspective.	AE8.1	Implementation of management should be based on ecological considerations with an emphasis on the links between the three domains (forest, crops, animals), natural diversity (wildlife, micro-organisms, soils, etc) and human dimension. [C]

Continued on next page

Table 3 continued

Statements		Arguments and recommendations	
		AE8.2	Implementation of an indicators system based on multicriteria including the assessment of diversity, including genetic resources, at different levels (e.g. FAO grid) is recommended. [C, S]
AE9	Reliable and abundant data are needed to support a better valorization of the genetic resources into an agroecological framework.	AE9.1	The abundant data present in individual databases should be made broadly accessible through global portals respecting FAIR principles. [C, S]
		AE9.2	Data georeferencing can be implemented to favour making links among databases and information systems. [C, S]
		AE9.3	Genetic resources managers and researchers together have to identify all relevant data related to genetic resources that can be useful for sustainable deployment and conservation of genetic resources for the agroecology transition (from biological to socioeconomic data or regulation information). [C, S]
		AE9.4	Data about the societal impact of genetic resources need to be measured and metrics have to be defined. [C, S]

All these data need to be standardized and broadly accessible by applying the FAIR principles (Findable, Accessible, Interoperable, Reusable https://www.go-fair.org/fair-principles/).

Synthetic overview: targeted audiences and action plan

Each of the 88 (altogether) arguments and recommendations has one or multiple target audiences drawn from policymakers, conservation actors and scientists. In the consensus reached by the international experts of the different genetic resources domains, policymakers are called to be highly concerned by one-third of the arguments and recommendations, with a slightly higher proportion for climate change and *in situ* issues, while more than half of them concern practitioners and scientists (Table 4). The proportion of arguments and recommendations addressed to conservation actors is higher for the *in situ* and the agroecology issues because addressing both issues requires engaging a broad range of local actors, not only genetic resources specialists. Overall, more than two-thirds of the arguments and recommendations addressed to scientists are jointly addressed to other audiences, reflecting the need to reinforce participatory approaches, co-development and policy support activities in research.

All of these arguments and recommendations are picked up by the *Genetic Resources Strategy for Europe*, which defines a comprehensive action plan at national and European levels with three main objectives: (1) strengthening and widening actions for genetic resources conservation and sustainable use, (2) enabling transformative change and (3) reinforcing international cooperation. Each objective of the action plan is subdivided into several sections. Figure 1 shows that the action plan of the Strategy addresses most of the arguments and recommendations at multiple levels, and it also shows that all aspects of the action plan are needed to respond to the three new challenges reviewed for genetic resources.

Perspectives

Despite contextual differences between crop, forest and animal genetic resources, the international panel of experts drawn from these domains recognized the emergence of a new era for genetic resources conservation and sustainable use in a context of potentially existential environmental and socioeconomic changes. Such changes that trigger multiple uncertainties require adaptive responses. In this era, the broad diversity of genetic resources can provide solutions to multiple issues, but only if the threats to diversity are effectively mitigated. Multiple conservation actions and sustainable uses must be considered in an integrated way.

For each of the three challenges under discussion, sharing expert views across domains resulted in a comprehensive list of general recommendations that are equally strategic for each genetic resources community in Europe. Beyond personal scientific expertise within the panel, the general arguments and recommendations made here also feed on the scientific evidence provided by the international conventions (the Convention on Biological Diversity CBD), platforms (the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, IPBES) and organizations (the UN Food and Agriculture Organization, FAO) addressing genetic resources and biodiversity conservation issues. Other genetic resources communities around the world could benefit from this work and reuse these evidence-based arguments and recommendations in two ways. Firstly, other communities can use them as a benchmark for the review of their own internal work plan. Secondly, these tables provide an opportunity to identify possible collaborative actions involving multiple genetic resources and conservation communities together rather than each one independently. Furthermore, all genetic resources activities have value through methodological application in the broader genetic and biodiversity conservation context. For instance, the crop wild relative population management guidelines (Iriondo *et al*, 2021) and conservation planning toolkits (Maxted *et al*, 2015; Brehm *et al*, 2017) are equally applicable for genetic or taxon-based biodiversity wild plant or animal conservation planning and *in situ* implementation. Therefore, these arguments and recommendations can feed the implementation and update of the FAO Global Plan of Action on genetic resources in the different domains, the International Treaty on Plant Genetic Resources for Food and Agriculture, and the CBD.

To address the three challenges, the experts agreed on the need for innovative solutions requiring cross-cutting collaborations, multi-actor engagement and policy support for effective implementation. Actions addressing genetic resources should engage conservation managers and users of genetic resources together with scientists from life sciences and social sciences addressing new research questions related to genetic resources, and decision-makers at multiple policy levels possibly benefiting from or influencing genetic resources conservation and sustainable use. Key are local actors concerned with *in situ* genetic resources management in production systems and conservation programmes.

With these recommendations, the panel of experts urges scientists to collaborate proactively with policymakers and a broad range of actors at local, national and international levels. Engaging new actors will depend on the capacity of the genetic resources communities to raise awareness of genetic resources values, share academic and non-academic knowledge and expertise on threats and solutions, co-develop efficient tools and advocate for supportive regulations. The *Genetic Resources Strategy for Europe* and the related action plan can help address the three challenges mentioned here and many more. In turn, disseminating the recommendations to their respective audiences will help support the uptake of the strategy. The panel of experts hopes that the statements and recommendations jointly

Table 4. Number of arguments and recommendations addressed to different target audiences for each challenge. CC, climate change challenge; IS, *in situ* challenge; AE, agroecology context. *, for each challenge, the sum of percentages is more than 100% because each recommendation may have multiple target audiences. **, out of the 59 recommendations to scientists, 41 (69%) are jointly addressed to other audiences.

	Target audience			
Challenge	**Policymakers**	**Conservation actors**	**Scientific community**	**Total**
CC	8 (31%)*	14 (54%)	18 (69%)	26
IS	11 (35%)	22 (71%)	19 (61%)	31
AE	8 (26%)	21 (68%)	22 (71%)	31
Total	27	57	59**	88

formulated by genetic resources experts in the three domains will be sufficiently integrated into future agricultural, food security, ecological, social and political policy across Europe and broader global fora to influence these sectors' policies. The implementation of the recommendations and follow-up with policymakers will require a tailored approach taking into account the specificities of each domain: this is achieved in the sectorial *Plant, Animal and Forest Genetic Resources Strategies for Europe* (respectively, ECPGR (2021); ERFP (2021); EUFORGEN (2021)).

This was the first time the three genetic resources communities (agricultural crop, animal and forestry domains) have come together at a continental level to investigate the similarities and dissimilarities between the three domains and to investigate if closer linkages could produce beneficial synergies. There is wide agreement within the panel of experts that the process itself has proven beneficial: it has shown that similarities outweigh differences, and that speaking with one unified voice is more effective in the policy context. The challenges posed by climate change, the benefit in this context of *in situ* conservation combined with sustainable use, and the need for locally adapted diversity have become so predominant, that the genetic resources communities should strive for continuous collaboration with mutual benefits.

Authors' contributions

FL, DB, MBDK, MB, ECL, SJH, HK, DL, GR, SS, ES, TVH, MW and NM organized and chaired the three GenRes Bridge Expert Panel workshops and wrote the text of the manuscript; the Expert Panel elaborated the

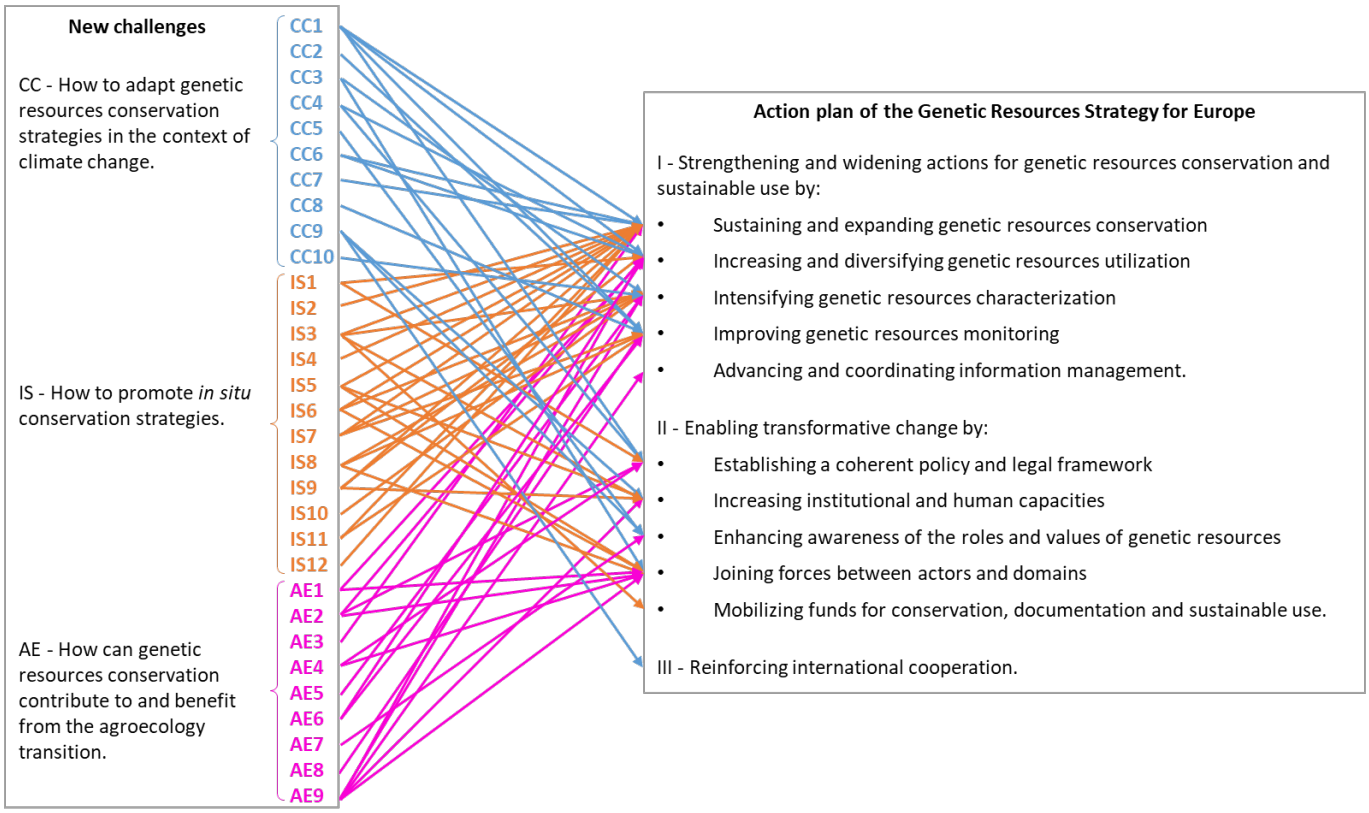

Figure 1. Main sections of the action plan of the *Genetic Resources Strategy for Europe* where the recommendations for the three new challenges are considered, here grouped by statement.

statements and recommendations (Tables 1-3) during the workshops and reviewed the manuscript.

GenRes Bridge Expert Panel contributors

Ricardo Alia (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-Centro Superior de Investigaciones Científicas, Spain); Hysen Bytyqi (University of Prishtina, Kosovo); Montserrat Castellanos Moncho (Ministry of Agriculture, Fisheries and Foood, Spain); Joži J. Cvelbar (Ministry for Agriculture, Forestry and Food, Slovenia); Suzana Đorđević-Milošević (Singidunum University, Serbia); Edoardo Esposito (European Forest Institute, Spain); Anna-Maria Farsakoglou (European Forest Institute, Spain); Jesús Fernández Martín (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-Centro Superior de Investigaciones Científicas, Spain); Gustavo Gandini (University of Milan, Italy); Ewa Hermanowicz (Forest Stewardship Council, Germany); Mervi Honkatukia (Farm Animals, Nordic Genetic Resource Center NordGen, Norway); Ivan Kreft (Nutrition Institute, Slovenia); Nataša Lovrić (European Forest Institute, Finland); Joana Magos Brehm (University of Birmingham, UK); Daniel Martín-Collado (Centro de Investigación y Tecnología Agroalimentaria de Aragón, Spain); Claudio Niggli (ProSpecieRara, Switzerland); Eduardo Notivol (Centro de Investigación y Tecnología Agroalimentaria de Aragón, Spain); Lorenzo Raggi (Dipartimento di Scienze Agrarie Alimentari e Ambientali, Università degli Studi di Perugia, Italy); Mari Rusanen (Natural Resources Institute, Finland); Stefan Schröder (Federal Office for Agriculture and Food, Germany); Paul Smith (Botanic Gardens Conservation International, UK); Katja Kavčič Sonnenschein (Slovenian Forestry Institute, Slovenia); Michèle Tixier-Boichard (National Research Institute for Agriculture, Food and Environment, France); Branislav Trudic (Forestry Division, Food and Agriculture Organization of United Nations, Italy); Luis Pablo Ureña (Institute of Agricultural Research and Training, Spain); Jelka Šuštar Vozlič (Agricultural Institute of Slovenia, Slovenia); Sharon Walshe (Department of Agriculture, Food and Marine, Ireland); Henri Woelders (Wageningen University and Research, The Netherlands); Frank Wolter (Nature and Forest Agency, Luxembourg).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817580, GenRes Bridge project.

References

Brehm, J. M., Kell, J., Thormann, I., Gaisberger, H., Dulloo, E., and Maxted, N. (2017). Interactive Toolkit for Crop Wild Relative Conservation Planning version 1.0 (University of Birmingham, Birmingham, UK and Bioversity International, Rome, Italy). url: http://www.cropwildrelatives.org/conservation-toolkit/.

CBD (1992). Text of the Convention on Biological Diversity. United Nations. url: https://www.cbd.int/convention/text/.

CBD (2020). Global Biodiversity Outlook 5 (Montreal, Canada: Convention on Biological Diversity), 208p. url: https://www.cbd.int/gbo5.

CBD (2022). Kunming-Montreal Global Biodiversity Framework. CBD/COP/DEC/15/4 (Montreal, Canada: Convention on Biological Diversity). url: https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-04-en.pdf.

Chable, V., Nuijten, E., Costanzo, A., Goldringer, I., Bocci, R., Oehen, B., Rey, F., Fasoula, D., Feher, J., Keskitalo, M., Koller, B., Omirou, M., Mendes-Moreira, P., Van Frank, G., Jika, A. K. N., Thomas, M., and Rossi, A. (2020). Embedding Cultivated Diversity in Society for Agro-Ecological Transition. *Sustainability* 12, 784–784. doi: https://doi.org/10.3390/su12030784

ECPGR (2021). Plant Genetic Resources Strategy for Europe. (Rome, Italy: European Cooperative Programme for Plant Genetic Resources), 71p. url: https://www.ecpgr.cgiar.org/fileadmin/bioversity/publications/pdfs/PGR_STRATEGY_LP_22_Nov_revised.pdf.

Ellstrand, N. C. and Rieseberg, L. H. (2016). When gene flow really matters: gene flow in applied evolutionary biology. *Evolutionary Applications* 9, 833–836. doi: https://doi.org/10.1111/eva.12402

ERFP (2021). Animal Genetic Resources Strategy For Europe 34p. url: https://www.animalgeneticresources.net/wp-content/uploads/2022/03/Final_AnGR-Strategy-_022022.pdf.

EUFORGEN (2021). Forest Genetic Resources Strategy For Europe (Barcelona, Spain: European Forest Institute), 64p. url: http://www.euforgen.org/FgRStrategy4Europe.

FAO (2015). Coping with climate change - the roles of genetic resources for food and agriculture (Rome, Italy: Food and Agriculture Organization), 110p. url: https://www.fao.org/3/i3866e/i3866e.pdf.

FAO (2019). The State of the World's Biodiversity for Food and Agriculture, ed. Bélanger, J. and Pilling, D. (Rome, Italy: FAO Commission on Genetic Resources for Food and Agriculture). url: https://www.fao.org/3/CA3129EN/CA3129EN.pdf.

Frankel, O. H. and Bennett, E. (1970). Genetic Resources in Plants – Their Exploration and Conservation, International Biological Programme, Handbook II (Oxford: Blackwell), 554p.

GenRes Bridge Project Consortium, ECPGR, ERFP and EUFORGEN (2021). Genetic Resources Strategy for Europe. url: http://www.genresbridge.eu/GRS4E.

Grau, R., Kuemmerle, T., and Macchi, L. (2013). Beyond 'land sparing versus land sharing': environmental heterogeneity, globalization and the balance between agricultural production and nature conservation. *Current Opinion in Environmental Sustainability* 5, 477–483. doi: https://doi.org/10.1016/j.cosust.2013.06.001

Haines-Young, R. and Potschin, M. B. (2018). Common International Classification of Ecosystem Services (CICES) V5.1 and Guidance on the Application of the Revised Structure. url: https://cices.eu/content/uploads/sites/8/2018/01/Guidance-V51-01012018.pdf.

Hoban, S., Archer, F. I., Bertola, L. D., Bragg, J. G., Breed, M. F., Bruford, M. W., Coleman, M. A., Ekblom, R., Funk, W. C., Grueber, C. E., Hand, B. K., Jaffé, R., Jensen, E., Johnson, J. S., Kershaw, F., Liggins, L., Macdonald, A. J., Mergeay, J., Miller, J. M., Muller-Karger, F., O'Brien, D., Paz-Vinas, I., Potter, K. M., Razgour, O., Vernesi, C., and Hunter, M. E. (2022). Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. *Biological Reviews* 97, 1511–1538. doi: https://doi.org/10.1111/brv.12852

Hoban, S., Bruford, M., Jackson, J. D., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., Liggins, L., Mergeay, J., Moharrek, F., O'Brien, D., Ogden, R., Palma-Silva, C., Pierson, J., Ramakrishnan, U., Simo-Droissart, M., Tani, N., Waits, L., and Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation* 248(108654). doi: https://doi.org/10.1016/j.biocon.2020.108654

IPBES (2019). Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, ed. Brondízio, E. S., Settele, J., Díaz, S., and Ngo, H. T. url: https://ipbes.net/global-assessment.

IPCC (2019). Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems, ed. Shukla, P. R., Skea, J., Buendia, E. C., Masson-Delmotte, V., Pörtner, H. O., Roberts, D. C., Zhai, P., Slade, R., Connors, S., van Diemen, R., Ferrat, M., Haughey, E., Luz, S., Neogi, S., Pathak, M., Petzold, J., Pereira, J. P., Vyas, P., Huntley, E., Kissick, K., Belkacemi, M., and Malley, J. 864p. url: https://www.ipcc.ch/site/assets/uploads/2019/11/SRCCL-Full-Report-Compiled-191128.pdf.

Iriondo, J. M., Brehm, J. M., Dulloo, M. E., and Maxted, N. (2021). Crop Wild Relative Population Management Guidelines. Farmer's Pride: Networking, partnerships and tools to enhance in situ conservation of European plant genetic resources . url: https://more.bham.ac.uk/farmerspride/wp-content/uploads/sites/19/2021/07/Crop_Wild_Relative_Population_Management_Guidelines.pdf.

Koskela, J., Buck, A., and Du, T. (2007). Climate change and forest genetic diversity: Implications for sustainable forest management in Europe (Rome, Italy: Bioversity International), 111p. url: https://www.euforgen.org/fileadmin/bioversity/publications/pdfs/1216.pdf.

Maxted, N., Kell, S., and Brehm, J. M. (2015). National Level Conservation and Use of Landraces Draft Technical Guidelines volume 14. (Rome, Italy: Food and Agriculture Organization of the United Nations). url: http://www.fao.org/3/a-mm564e.pdf.

Pozo, R. A., Cusack, J. J., Acebes, P., Malo, J. E., Traba, J., Iranzo, E. C., Morris-Trainor, Z., Minderman, J., Bunnefeld, N., Radic-Schilling, S., Moraga, C. A., Arriagada, R., and Corti, P. (2021). Reconciling livestock production and wild herbivore conservation: challenges and opportunities. *Trends in Ecology and Evolution* 36, 750–761. doi: https://doi.org/10.1016/j.tree.2021.05.002

Genetic Resources

SHORT COMMUNICATION

Genetic Resources (2024), 5 (9), 29–35
DOI: 10.46265/genresj.WAOT8693
https://www.genresj.org
ISSN: 2708-3764

The first draft genome sequence of Russian olive (*Elaeagnus angustifolia* L.) in Iran

Leila Zirak, Reza Khakvar * and Nadia Azizpour

Department of Plant Protection, Faculty of Agriculture, University of Tabriz, 5166616471, Tabriz, Iran

Abstract: Russian olive (*Elaeagnus angustifolia* L.) is a native tree species of Iran and the Caucasus region growing in both wild habitats and cultivated settings. The area under cultivation of this tree has been increasing in recent years due to its ability to withstand drought and soil salinity. Revealing the complete genome of this tree holds great importance. To achieve this, a local cultivar of Russian olive was sampled from the northwest region of Iran for whole genome sequencing using the Illumina platform resulting in approximately 6GB of raw data. A quality check of the raw data indicated that approximately 45,011,388 read pairs were obtained from sequences totalling around 6.7×10^9bp with CG content of 31%. To assemble the genome of the Russian olive tree, the raw data was aligned to a reference sequence of the jujube (*Ziziphus jujuba*) genome, which is the taxonomically closest plant to the Russian olive. Assembly of alignments yielded a genome size of 553,696,299bp consisting of 339,701 contigs. The N50 value was 5,300 with an L50 value of 24,921 and GC content of the Russian olive genome was 31.5%. This research represents the first report on the genome of the Iranian cultivar of the Russian olive tree.

Keywords: Russian olive (*Elaeagnus angustifolia* L.), Genome, WGS, Iran

Citation: Zirak, L., Khakvar, R., Azizpour, N. (2024). The first draft genome sequence of Russian olive (*Elaeagnus angustifolia* L.) in Iran. *Genetic Resources* 5 (9), 29–35. doi: 10.46265/genresj.WAOT8693.

Introduction

The Russian olive (*Elaeagnus angustifolia* L.) is a deciduous tree growing abundantly in several areas in Iran (Mozaffarian, 2009). It belongs to the Elaeagnaceae family and is native to western and central Asia including Iran, southern Russia, Turkey, Kazakhstan (Lamers and Khamzina, 2010) and China (Huang *et al*, 2010). Recently, it has also been cultivated in North America (Mineau *et al*, 2012). The Russian olive is a drought-resistant species and plays an important ecological role in Iran's dry climate. It is widely cultivated in Iran but also grows wildly. Since 97% of lands in Iran are arid or semi-arid, many artificial afforestation and urban green space projects have been devised, especially in the drainage basin of Lake Urmia, which is at risk of drying out completely, with Russian olive being a key species in these efforts (Tabatabaei, 2010). About 10,000ha of arid or semi-arid lands are cultivated with Russian olive tree in East Azerbaijan province in the northwest, the most important Russian olive cultivation area in Iran (Mozaffarian, 2009).

The whole genome of a local cultivar of Russian olive (Tabriz cultivar) was sequenced using the next-generation sequencing (NGS) method. Prior to this research, there was no available information on the genome of this species. Due to the importance of the Russian olive in the construction of artificial forests and urban green spaces in the northwest of Iran, the information obtained from sequencing the genome could help characterize the Iranian cultivar. The full genome sequence obtained in this research is the first report for the Russian olive tree genome worldwide.

Materials and methods

Plant sampling and DNA extraction

For sampling, a tree was randomly selected as a representative sample of this cultivar (*E. angustifolia* cultivar Tabriz) (Figure 1) from the Eynali artificial afforestation area located in Tabriz city in the northwest

*Corresponding author: Reza Khakvar
(khakvar@tabrizu.ac.ir; khakvar@gmail.com)

Received: 20.10.2023 | Accepted: 13.03.2024 | Published online: xx.04.2024

of Iran. About 5g of leaf tissues were separated, crushed using liquid nitrogen and then used for DNA extraction (Murray and Thompson, 1980). Extracted DNA was dissolved in 100ml distilled sterile water and stored at -20°C.

Next-generation sequencing analyses

About 200μl of the DNA solution, with a total amount of 10μg DNA, extracted from Russian olive leaf samples was purified and used for library preparation. The concentration and purity of DNA was determined using a Qubit fluorometer. The concentration of purified DNA was 43.20ng/μl and evaluated appropriately for the whole genome sequencing process. The Illumina 1.9 Novaseq 6000 platform was used to generate paired-end libraries by Novogene (Beijing, China). Finally, about 6GB of raw data was obtained. In total, 45,011,388 read pairs in about 6.7×10^9bp sequences with CG content of 31% were obtained from Russian olive genome sequencing. Each raw read length was 150bp and the insert size was 350bp.

Genome qualification, reference genome preparation and sequence alignment

The quality of Illumina raw data was checked by FastQC software version 0.73 (Brown *et al*, 2017). For reference genome preparation, the common jujube (*Ziziphus jujuba* (2n=24)) genome, with 405,637Mbp size and GC content of 33.084% comprising of 12 full-length chromosomes submitted in the NCBI genome database (accession numbers NC_063287, NC_063288, NC_063289, NC_063290, NC_063291, NC_063292, NC_063293, NC_063294, NC_063295, NC_063296, NC_063297 and NC_063298), a 365,812bp mitochondrion sequence (CM036902) and a 161,185bp chloroplast sequence (CM036903), was used Yang *et al* (2023). The NGS raw data was aligned to the reference genome using Bowtie2 software version 2.5.0 (Langmead and Salzberg, 2012).

Genome assembly

For the genome preparation, all aligned reads which resulted from alignment analysis, were used for the assembly by metaSPAdes software version 3.15.4 (Nurk *et al*, 2017). For nucleotide sequence clustering and to improve the performance of sequence analyses, all contigs were clustered using CD-HIT-EST software version 4.8.1 (Fu *et al*, 2012).

Genome annotation

To characterize proteins related to the Russian olive genome, contigs were annotated using InterProScan functional annotation software version 5.59-91.0 (Jones *et al*, 2014). The annotation results using InterProScan are summarized in Supplemental Table 1. In addition, for genes and proteins sequence prediction, all contigs were subjected to another annotation method using GhostKOALA tool of the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (https://www.kegg.jp/ghostkoala/).

Results

Russian olive genome information

Since no Russian olive genome has been submitted to the NCBI genome database so far, the NGS raw data were aligned to the reference genome prepared with the common jujube (*Z. jujuba*) genome. The jujube tree is a species taxonomically close to the Russian olive and its genome is available in the NCBI genome database. The mapping rate was 96.4%. Assembly of aligned reads resulted in a genome in contig level with 553,696,299bp size consisting of 339,701 contigs with N50 = 5,300bp, L50 = 24,921bp and GC content of 31.5%. The genome coverage was 442.0x. Finally, the *E. angustifolia* cultivar Tabriz genome was deposited in the NCBI genome database under the whole genome accession number JAIFOS000000000, BioProject accession number PRJNA744085, BioSample accession number SAMN20079343 and Assembly accession number GCA 019593565.

Genome annotation results using InterProScan

For genome annotation, two methods were used. Initially, functional analysis of proteins and nucleotides was conducted using InterProScan software, which classifies t hem i nto f amilies a nd p redicts d omains and important sites. To classify proteins, InterProScan uses predictive models, known as signatures, provided by several different databases. According to InterProScan results, a total of 496,838 proteins were predicted in the Russian olive genome. Among all proteins, 106,757 proteins shared consensus disorder prediction. The intrinsic disorder (ID) is recognized as an important feature of protein sequences. The consensus-based prediction of disorder in protein was done using the MobiDB-lite method which has been integrated with the InterPro database (Necci *et al*, 2017). About 1,454 proteins remained uncharacterized (Supplemental Table 1).

Genome annotation results using GhostKOALA

The assembled genome was subjected to annotation using the GhostKOALA server to characterize individual gene functions. The protein sequences that were used for GhostKOALA analysis were provided using the MetaGenMark online web tool (Zhu *et al*, 2010). The KEGG GENES database (Kanehisa *et al*, 2016) searches indicated that 54,162 proteins (about 17% of whole proteins) acquired original KO numbers, 178,897 proteins acquired second-best KO numbers and 85,713 proteins could not be matched with any characterized proteins and were therefore considered as uncharacterized proteins. The GhostKOALA annotation results are summarized in Table 1.

Figure 1. The Russian olive cultivar Tabriz subject to NGS analyses in this study.

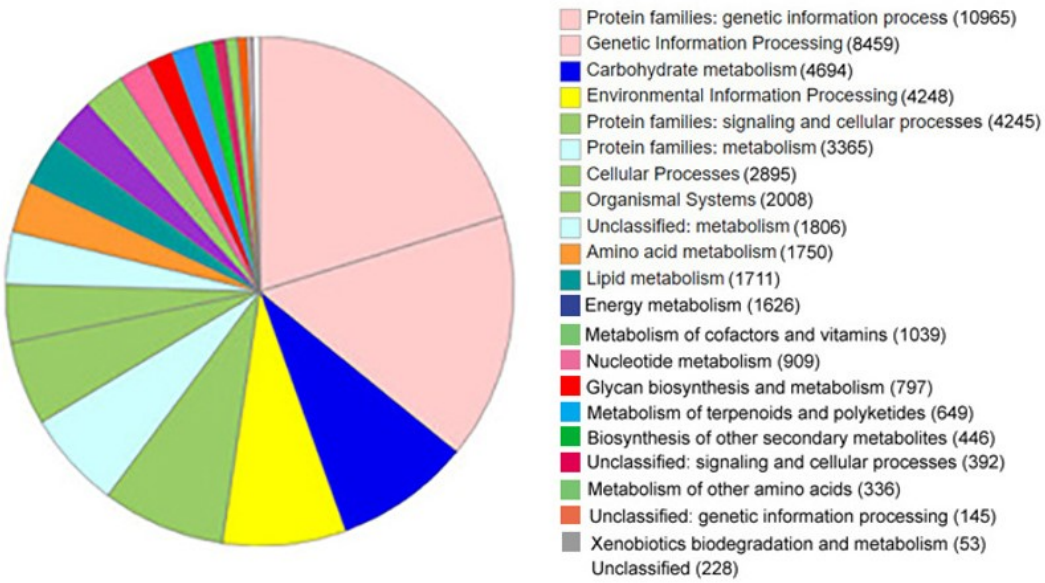

Figure 2. Functional category and pathways for predicted proteins of Russian olive tree prepared with GhostKOALA.

Table 1. Functional classification predictions of proteins annotated in the Russian olive genome based on KEGG BRITE classification.

Pathway	Protein description	No. of predicted proteins
Genes and proteins	Ribosomal Proteins	153
	RNA polymerases	34
	DNA polymerases	25
	Aminoacyl-tRNA synthetases	28
	Enzymes of 2-oxocarboxylic acid metabolism	33
	Dioxygenases	2
	Photosynthetic and chemosynthetic capacities	4
Orthologs, modules and networks	KEGG Orthology (KO)	8,957
Protein families: metabolism	Enzymes	3,712
	Protein kinases	315
	Protein phosphatases and associated proteins	204
	Peptidases and inhibitors	379
	Glycosyltransferases	148
	Lipopolysaccharide biosynthesis proteins	25
	Peptidoglycan biosynthesis and degradation proteins	29
	Lipid biosynthesis proteins	67
	Polyketide biosynthesis proteins	7
	Prenyltransferases	28
	Amino acid-related enzymes	64
	Cytochrome P450	67
	Photosynthesis proteins	56
Protein families: genetic information processing	Transcription factors	410
	Transcription machinery	229
	Messenger RNA biogenesis	308
	Spliceosome	256
	Ribosome	154
	Ribosome biogenesis	259
	Transfer RNA biogenesis	187
	Translation factors	84
	Chaperones and folding catalysts	165
	Membrane trafficking	787
	Ubiquitin system	441
	Proteasome	52
	DNA replication proteins	132
	Chromosome and associated proteins	678
	DNA repair and recombination proteins	309
	Mitochondrial biogenesis	269

Continued on next page

Table 1 continued

Pathway	Protein description	No. of predicted proteins
Protein families: signaling and cellular processes	Transporters	595
	Secretion system	79
	Two-component system	40
	Cilium and associated proteins	181
	Cytoskeleton proteins	222
	Exosome	333
	G protein-coupled receptors	130
	Cytokine receptors	9
	Pattern recognition receptors	6
	Nuclear receptors	14
	Ion channels	125
	GTP-binding proteins	68
	Cytokines and growth factors	21
	Cell adhesion molecules	50
	CD molecules	55
	Proteoglycans	16
	Glycosaminoglycan binding proteins	46
	Glycosylphosphatidylinositol (GPI)-anchored proteins	21
	Lectins	23
	Domain-containing proteins not elsewhere classified	241
	Other proteins	65

An overview of putative functions of annotated proteins is given in Figure 2.

Discussion

Ecological importance of the Russian olive tree in Iran

The Russian olive is a long-lived tree that can live up to 100 years and tolerate a wide range of hard environmental conditions such as severe drought, flood and high salinity or alkalinity of the soils (Asadiar *et al*, 2013). This tree produces edible fruits with high medicinal properties. Russian olive fruits have antioxidant activities and anti-inflammatory properties. Fruit kernel powder is used in the treatment of acute and chronic inflammations, such as arthritis (Tabatabaei, 2010; Wang *et al*, 2013). The climate of Iran is mostly arid or semi-arid and is strongly affected by depleting water resources, as a result of rising demand, salinization, ground water overexploitation and increasing drought frequency. Therefore, plants that could withstand harsh environmental conditions and have low water consumption have been considered for cultivation in several regions. The Russian olive is growing as a wild plant in all areas with a dry climate in Iran; however, it also serves as the main species in many artificial forestation projects. The climatic and ecological benefits provided by the Russian olive in Iran underline the importance of exploring the genomic characteristics of its Iranian cultivar.

Genomic characteristics of the Russian olive cultivar Tabriz

Before this study, no information about the Russian olive genome was available. Therefore, the common jujube genome was used for the Russian olive genome preparation, since it is the closest taxonomical relative to the Russian olive and its genome is available in NCBI. The common jujube belongs to the Rhamnaceae family, which along with the Elaeagnaceae family belongs to the Rosales order. The genome of *Z. jujuba* comprises 12 chromosomes with an average size of 405.637Mb and GC content of about 33%. Alignments of NGS reads obtained from Russian olive to the jujube whole genome sequence resulted in a 553,696,299bp genome composed of 339,701 contigs. The GC content of the new genome was 31.5% which was nearly the same as the jujube genome GC content.

Annotation analysis was accomplished by several programmes. At last, two methods based on the online GhostKOALA web server and InterProScan software were found suitable for Russian olive genome annotation. The genome functional annotation using online KEGG mapper reconstruction resulted in 3,186 proteins for metabolism pathways in the genome including 647 involved in carbohydrate metabolism pathways, 345 in energy metabolism pathways, 381 proteins for lipid metabolism pathways, 185 proteins for nucleotide metabolism pathways, 686 proteins for amino acid metabolism pathways, 258 proteins for glycan biosynthesis and metabolism pathways, 296 proteins for metabolism of cofactors and vitamins pathways, 122 proteins for metabolism of terpenoids and polyketides pathways, 144 proteins for biosyntheses of other secondary metabolites pathways and 122 proteins for xenobiotics biodegradation and metabolism pathways.

Also, for the transcription and translation systems, 35 RNA polymerases, 34 basal transcription factors, 105 spliceosomes, 124 ribosomes, 33 proteins for aminoacyl-tRNA biosynthesis pathways, 83 nucleocytoplasmic transport proteins, 58 mRNAs surveillance and 62 ribosomes of biogenesis in eukaryotes were obtained.

The folding, sorting and degradation systems included 31 export proteins, 38 proteasomes, 70 RNA degradation proteins and 237 other proteins. The replication and repair systems include 43 DNA replication proteins, 34 base excision repair proteins, 42 nucleotide excision repair proteins, 54 homologous recombination proteins and 11 non-homologous end-joining proteins. The membrane transport includes 114 ABC transporters and 9 proteins for phosphotransferase system PTS. Also, 1600 signal transduction proteins, 497 proteins for cell growth and death pathways, 272 cellular community proteins and 8,494 other proteins exist in the genome.

Conclusion

The Russian olive is an ecologically important tree serving as vegetation in Iran's arid climate. It is also known as an important medicinal plant in Iranian traditional medicine. However, genetic information about this species remains sparse. In this research, wedescribe the genome of an Iranian cultivar of the Russian olive by using the jujube genome as a reference, since it is the closest species with a characterized genome. As a result, a full-size genome with 553.7Mb size in contig level was obtained, which can provide the foundations for the chromosomal sequence of this species. Russian olive is one of the most important horticultural tree species in the northwest of Iran and its genome characterization serves as a key step towards broader research to characterize the genome of other important plant species of Iran.

Supplemental data

Supplemental Table 1. Uncharacterized annotated proteins of Russian olive

Author contribution

Leila Zirak executed both the laboratory experiments and the subsequent bioinformatics analyses; Reza Khakvar, who supervised the project, was responsible for the validation of all experimental data and contributed to the editing of the final manuscript draft; Nadia Azizpour provided assistance in the collection of samples

and the extraction of DNA, and in the composition of the initial manuscript draft.

Conflict of interest statement

The authors declare no conflicts of interest.

References

Asadiar, L. S., Rahmani, F., and Siami, A. (2013). Assessment of genetic diversity in the Russian olive (*Elaeagnus angustifolia*) based on ISSR genetic markers. *Revista Ciencia Agronomica* 44(2), 310–316. doi: https://doi.org/10.1590/S1806-66902013000200013

Brown, J., Pirrung, M., and Mccue, L. A. (2017). FQC dashboard: Integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* 33(19), 3137–3139. doi: https://doi.org/10.1093/bioinformatics/btx373

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23), 3150–3152. doi: https://doi.org/10.1093/bioinformatics/bts565

Huang, Z., Liu, M., Chen, B., Uriankhai, T., Xu, C., and Zhang, M. (2010). Distribution and interspecific correlation of root biomass density in an arid *Elaeagnus angustifolia* - *Achnatherum splendens* community. *Acta Ecologica Sinica* 30(1), 45–49. doi: https://doi.org/10.1016/j.chnaes.2009.12.008

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., Mcanulla, C., Mcwilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9), 1236–1240. doi: https://doi.org/10.1093/bioinformatics/btu031

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* 44, 457–462. doi: https://doi.org/10.1093/nar/gkv1070

Lamers, J. P. and Khamzina, A. (2010). Seasonal quality profile and production of foliage from trees grown on on degraded cropland in arid Uzbekistan, Central Asia. *Journal of Animal Physiology and Animal Nutrition* 94, 77–85. doi: https://doi.org/10.1111/j.1439-0396.2009.00983.x

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie. 2. *Nature Methods* 9, 357–359. doi: https://doi.org/10.1038/nmeth.1923

Mineau, M. M., Baxter, G. V., Marcarelli, A. M., and Minshall, G. W. (2012). An invasive riparian tree reduces stream ecosystem efficiency via a recalcitrant organic matter subsidy. *Ecology* 93(7), 1501–1508. doi: https://doi.org/10.1890/11-1700.1

Mozaffarian, V. (2009). Trees and shrubs of Iran (Iran Farhang Pree), 2nd edition, 1054p.

Murray, M. G. and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* 8(19), 4321–4326. doi: https://doi.org/10.1093/nar/8.19.4321

Necci, M., Piovesan, D., Dosztányi, Z., and Tosatto, S. C. E. (2017). MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* 33(9), 1402–1404. doi: https://doi.org/10.1093/bioinformatics/btx015

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27, 824–834. doi: https://doi.org/10.1101/gr.213959.116

Tabatabaei, M. (2010). Senjed = *Elaeagnus angustifolia* L. (Elaeagnaceae) (Tehran, Iran: SANA Publication), 92p.

Wang, Y., Guo, T., Li, J. Y., Zhou, S. Z., Zhao, P., and Fan, M. T. (2013). Four flavonoid glycosides from the pulps of *Elaeagnus angustifolia* and their antioxidant activities. *Advanced Materials Research* 756, 16–20. doi: https://doi.org/10.2991/iccia.2012.415

Yang, M., Han, L., Zhang, S., Dai, L., Li, B., Han, S., Zhao, J., Liu, P., Zhao, Z., and Liu, M. (2023). Insights into the evolution and spatial chromosome architecture of jujube from an updated gapless genome assembly. *Plant Community* 4(6), 100662. doi: https://doi.org/10.1016/j.xplc.2023.100662

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* 38(12), e132. doi: https://doi.org/10.1093/nar/gkq275

Genetic Resources

SHORT COMMUNICATION

Genetic Resources (2024), 5 (9), 36–44
DOI: 10.46265/genresj.WQZS1824
https://www.genresj.org
ISSN: 2708-3764

A public mid-density genotyping platform for cultivated blueberry (*Vaccinium* spp.)

Dongyan Zhao[a], Manoj Sapkota[a], Jeff Glaubitz[b], Nahla Bassil[c], Molla F Mengist[d], Massimo Iorizzo[d], Kasia Heller-Uszynska[e], Marcelo Mollinari[f], Craig T Beil[a] and Moira J Sheehan [*,a]

[a] *Breeding Insight, Cornell University, Ithaca, 14853, NY, USA*
[b] *Institute of Biotechnology, Cornell University, Ithaca, 14853, NY, USA*
[c] *National Clonal Germplasm Repository, USDAARS, OR, 97333, Corvallis, USA*
[d] *Plants for Human Health Institute, North Carolina State University, NC, 28081, Kannapolis, USA*
[e] *Diversity Arrays Technology, ACT 2617, Bruce, Australia*
[f] *North Carolina State University, Campus Box 7609, NC, Raleigh, 27695, USA*

Abstract: Small public breeding programmes have many barriers to adopting technology, particularly creating and using genetic marker panels for genomic-based decisions in selection. Here we report the creation of a DArTag panel of 3,000 loci distributed across the tetraploid genome of blueberry (*Vaccinium corymbosum*) for use in molecular breeding and genomic prediction. The creation of this marker panel brings cost-effective and rapid genotyping capabilities to public and private breeding programmes. The open access provided by this platform will allow genetic data sets generated on the marker panel to be compared and joined across projects, institutions and countries. This genotyping resource has the power to make routine genotyping a reality for any breeder of blueberry.

Keywords: Vaccinium spp., amplicon-sequencing, plant breeding, DArTag genotyping, microhaplotype

Citation: Zhao, D., Sapkota, M., Glaubitz, J., Bassil, N., Mengist, M. F., Iorizzo, M., Heller-Uszynska, K., Mollinari, M., Beil, C. T., Sheehan, M. J. (2024). A public mid-density genotyping platform for cultivated blueberry (*Vaccinium* spp.). *Genetic Resources* 5 (9), 36–44. doi: 10.46265/genresj.WQZS1824.

Introduction

Molecular techniques have been employed for nearly four decades to enhance and speed the breeding efforts for major staple food crops like tomato, maize and barley (Tanksley (1983); Helentjaris *et al* (1985); Feuerstein *et al* (1990) and reviewed in Hasan *et al* (2021)). Over time, molecular biology techniques have been paired with high-quality phenotypic data to perform genome-wide association studies (GWAS), genomic selection and prediction, further fuelling breeding for quantitative or complex traits (Eathington *et al*, 2007; Heffner *et al*, 2009; Lorenzana and Bernardo, 2009). While these achievements are significant, many crop species grown for human consumption are still unable to apply these techniques in breeding efforts. Many breeders would like to adopt molecular breeding tools and techniques, but sometimes doing so is hampered by large barriers-to-entry challenges. The range of barriers, and how surmountable they are, varies from species to species and is impacted by species-specific challenges in logistics, technical know-how, biology and the growing environment.

Blueberries (*Vaccinium* spp) are native to North America and are a relatively recent crop, having been cultivated only since 1916 (USHBC, 2021). The United States (US) is the largest global producer of blueberries (FAOSTAT, 2021). In 2022, the US produced over 282 million kilograms (622 million pounds) of cultivated blueberries and harvested 35.2 million kilograms (77.6 million pounds) of wild blueberries, which amounted to a total crop value of USD1.04

*Corresponding author: Moira J Sheehan (moirasheehan@cornell.edu)

Received: 14.02.2024 | Accepted: 02.04.2024 | Published online: 11.04.2024

billion (NASS, 2023). Blueberries are considered a 'superfruit' for human nutrition due to their high levels of essential nutrients, fibre and antioxidants. Because of their nutritional value, blueberries are produced for a wide range of markets including fresh eating (and U-pick), frozen whole berries, frozen juice, powders and dried leaves for herbal tea.

Cultivated blueberries are categorized by growing region and chilling requirements. In Northern states, most varieties are Northern highbush types (NHB; *V. corymbosum*) that only flower after about 800–1,000 hours of exposure to temperatures between 0°C–7°C (32°F–45°F) (Hancock, 2009). The Southern highbush types (SHB) are complex hybrids between *V. corymbosum* with the evergreen species *V. darrowii* native to Florida. Southern highbush varieties have reduced chilling requirements (200–300 hours) and enhanced adaptation to Southern climates and soils (Hancock, 2009). The half-high blueberry (HHB) is derived from crosses between NHB with *V. angustifolium*, a wild Northern species. Half-high blueberry is preferred for commercial environments that require varieties with enhanced hardiness. Unlike NHB, SHB and HHB which are tetraploid types, the fourth cultivated type, rabbiteye (RE; *V. virgatum*), is hexaploid. Rabbiteye blueberry, known for its high vigour and heat tolerance, is native to the Southeastern US (Edger *et al*, 2022). Although most of the breeding efforts are focused on these four cultivated types, some pre-breeding work has included parents from wild species (also known as lowbush blueberry) of the *Cyanococcus* section of the subgenus *Vaccinium*.

Blueberry breeding is a long and tedious process (Gallardo *et al*, 2018). Traditional breeding approaches can take 9 to 20 years from crossing and testing to the release of new cultivars (Gallardo *et al*, 2018). Some of the breeding challenges are that cultivated blueberries are perennials, outcrossing, highly heterozygous and autotetraploid, where random chromosome pairing during meiosis predominates (Qu and Hancock, 1995; Qu *et al*, 1998; Lyrene *et al*, 2003). Traditional biallelic SNP marker systems designed for inbred or diploid species often fall short when applied to heterozygous and polyploid species due to their inability to identify multiallelic dosages accurately. A more sophisticated genotyping system is needed to address the unique challenges posed by blueberry's autotetraploid nature, yet the investment cost and reliance upon skilled bioinformatics support for each genotyping run make this a high-risk endeavour for breeders.

The first and most tractable place to build capacity and tools for molecular breeding is to create a rapid genotyping pipeline that fits within both the breeding and selection cycles and can deliver on the breeder's objectives (Hawkins and Yu, 2018; Mejia-Guerra *et al*, 2021) . As stated here, a pipeline refers to a complete workflow starting with a genetic marker platform, vendors for services and bioinformatic tools to transform returned raw data into a usable format for breeders. There are several factors to consider when choosing a genetic marker platform: cost per data point, vendor services, turnaround times and what genetic analyses can be done with the resulting data. For blueberry, we hypothesized that a targeted-amplicon sequenced-based approach would be the most beneficial for breeders. Unlike Genotyping-by-Sequencing (GBS), targeted amplicon-based genotyping technologies such as DArTag (Diversity Array Technology - DArT), Flex-Seq (RAPiD Genomics), and Capture-Seq (LGC Genomics) have low missing data rates and query the same loci in all samples across genotyping projects, allowing new data to be easily appended to existing data (Darrier *et al*, 2019; Telfer *et al*, 2019; Wang *et al*, 2020). The amount of data returned is in the tens of thousands or less, rather than the millions of reads from GBS, simplifying downstream bioinformatics processing (Darrier *et al*, 2019; Milner *et al*, 2019). This in turn speeds up the analysis time for marker-assisted selection (MAS), introgression tracking, linkage mapping, GWAS and genomic prediction (Darrier *et al*, 2019).

Here, we report the creation of a mid-density DArTag panel of 3,000 marker loci distributed across the blueberry genome for use in molecular breeding and genomic prediction. DArTag is a hybridization/amplicon-based targeted genotyping platform developed by DArT (Blyton *et al* (2023); https://www.diversityarrays.com/services/targeted-genotying/) available to the public.

Materials and methods

Germplasm selection and whole-genome sequencing of a blueberry diversity panel

A total of 31 cultivated blueberry accessions focused on elite North American breeding lines were selected for skim sequencing. This panel consisted of 12 NHB, 10 SHB, 2 NHB x SHB hybrids, 5 RE, and 1 RE x SHB accessions (Supplemental Table 1, entries marked with asterisks). Two biological replicates of each sample in the discovery panel were processed, where the sequencing libraries (average insert DNA size of 300bp) were prepared using either Illumina Nextera WGS library prep at the Genomics Facility of Cornell Institute of Biotechnology or NEBNext Ultra DNA Library Prep Kit at Novogene. Whole-genome sequencing was done using Illumina NovaSeq 6000 at Novogene (https://en.novogene.com).

SNP discovery and selection of 3K marker loci for building DArTag genotyping panel

Raw FASTQ sequences were processed by removing residual adapter sequences and low-quality bases using Trimmomatic (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:30) (Bolger *et al*, 2014). Cleaned reads were then aligned to the haploid set (i.e., the first set out of the four homologous chromosomes) of the blueberry reference genome as described by Colle *et al* (2019) using BWA-MEM (Li, 2013). Structural variants (SNPs and indels) were called using the DNAseq

pipeline developed by Sentieon (https://www.sentieon.com). A total of 600K SNPs were discovered in the diver-sity panel. A high-confidence set of 10K SNPs (Figure 1) was then identified using the following criteria: (1) not located within 5bp from an indel, (2) QUAL > 30, (3) minimum and maximum read depths of 20 and 1,500, respectively, (4) at each heterozygous site, at least one read supporting the reference allele and two reads supporting the alternative allele, (5) no missing genotype per SNP position, (6) with a minor allele frequency greater than 0.25, (7) not located in transposable elements or within 1Kb of chromosome termini and (8) even genomic distribution and mostly located in genic regions. The 10K SNPs were submitted for QC to DArT (Diversity Arrays Technology Pty Ltd, www.diversityarrays.com), from which a 3K SNP set was selected. Additionally, a few experimentally validated SNPs were also force-included in the panel.

Custom oligo probes were then synthesized, and genotyping was done at DArT. A total of 1,445 and 1,555 marker loci (Supplemental File 1) were designed to produce amplicons from the plus and minus strands based on the reference genome, respectively (Colle *et al*, 2019). Based on the 'Draper' reference genome and gene assembly v1.0 from Colle *et al* (2019), 97% (2,924) reside in genic regions, with only 3% (76) residing in non-genic regions (Supplemental File 1). Among the 3,000 loci selected, each chromosome harbors between 219 loci on Chr07.1 to 296 loci on Chr02 with an average of 250 loci per chromosome. In addition, there is a positive correlation (R^2=0.70) between the number of genes on a chromosome and the number of targeted loci on that chromosome, indicating that chromosomes with more genes have better marker coverage (Supplemental File 1). The DArTag genotyping technology produces multi-allelic data as 54bp and 81bp amplicons (referred to as microhaplotypes in this study) encompassing the 3K target SNP sites, therefore, we refer to these target sequences as marker loci.

Selection of samples for validating the DArTag panel and genotyping results

The DArTag genotyping assay consists of four steps based on principles described in Krishnakumar *et al* (2008) and implemented as described in Zhao *et al* (2023). Briefly, the pool of 3,000 blueberry oligos, each targeting one genetic variant plus adjacent flanking sequence, is hybridized to denatured gDNA in step 1, followed by SNP/INDEL copying into DArTag molecules by DNA polymerase in step 2. After ligation into circular molecules also in step 2, and nuclease treatment to remove uncircularized molecules in step 3, DArTag products are subsequently amplified i n s tep 4 with the simultaneous addition of sample unique barcodes used downstream for demultiplexing. The products of DArTag assay, after purification a nd quantification, are sequenced on NGS platforms (e.g. NovaSeq 6000, Illumina) with a depth of around 200x, demultiplexed and the genetic variants are detected using the DArT proprietary analytical pipeline.

The blueberry 3K marker panel was tested using a set of 375 samples, including: (1) a diverse set of cultivated blueberries (n = 171), (2) a 'Draper' x 'Jewel' (DxJ) F_1 population (n = 175), (3) wild *Vaccinium* species and other interspecific hybrids (*Vaccinium* subgenus) (n = 24), and (4) a small number of cultivated cranberry varieties (n = 5) (*Oxycoccus* subgenus) (Supplemental Table 1). The raw genotyping data included FASTQ and the missing allele discovery count (MADC) file (Supplemental File 2).

The MADC file was first filtered at the microhaplotype level. A microhaplotype was retained if it was present in at least 10 samples and each sample had at least 2 reads detected. First, samples with ≥ 95% missing data were removed. Then, filtering of marker loci was based on ≥ 10 samples with each having ≥ 10 reads for each marker locus per sample. All SNPs, including both target and off-target SNPs were extracted from all remaining marker loci for downstream analyses. Principal component analysis was conducted using read count data from all samples using AddPCA function in polyRAD (Clark *et al*, 2019) and plotted using ggplot2.

Genetic map construction

The DxJ F_1 population was derived from a 'Draper' x 'Jewel' cross. 'Draper' is a NHB variety released by Michigan State University in 2004, whereas 'Jewel' is a SHB variety released by the University of Florida in 1999. The true parental plants that were used to make the DxJ cross are no longer available, so we genotyped five Draper accessions and five Jewel accessions from across several public programmes. Genotype dosage calls for each SNP in the DxJ population were determined with updog software (Gerard *et al*, 2018). A PCA was performed in polyRAD and identified 1 4 D xJ progenies that do not appear to be true F_1s (Supplemental Figure 1A). Before mapping, these 14 individuals were removed leaving 161 DxJ F_1 progeny and the most similar parents to the true parents, which were not available, 'Draper_2004.001_S10-42' and 'Jewel_2157.001_G04-01' were identified a s p roxy p arents. (Supplemental Figure 1B). Of the 8,955 SNPs detected, 4,918 were non-informative in the DxJ F_1 population and were removed from further mapping on the 'true' 161 F1s in the DxJ population. The average missing data for this population was 15% (range 6–26%) (Supplemental Figure 2). Marker loci with > 5% missing rate (n = 840), and that did not fit expected Mendelian segregation (n = 1,203) were also removed from further analysis leaving 1,994 markers available for map construction. To construct the F_1 population genetic map MAPpoly2 was used (https://github.com/mmollina/mappoly2; Mollinari and Garcia (2019); Mollinari *et al* (2020)). A recombination fraction matrix was calculated and used to cluster the markers into linkage groups. Screening SNPs based on recombination frequency via the rf filter function eliminated additional SNPs (n = 497). For each linkage

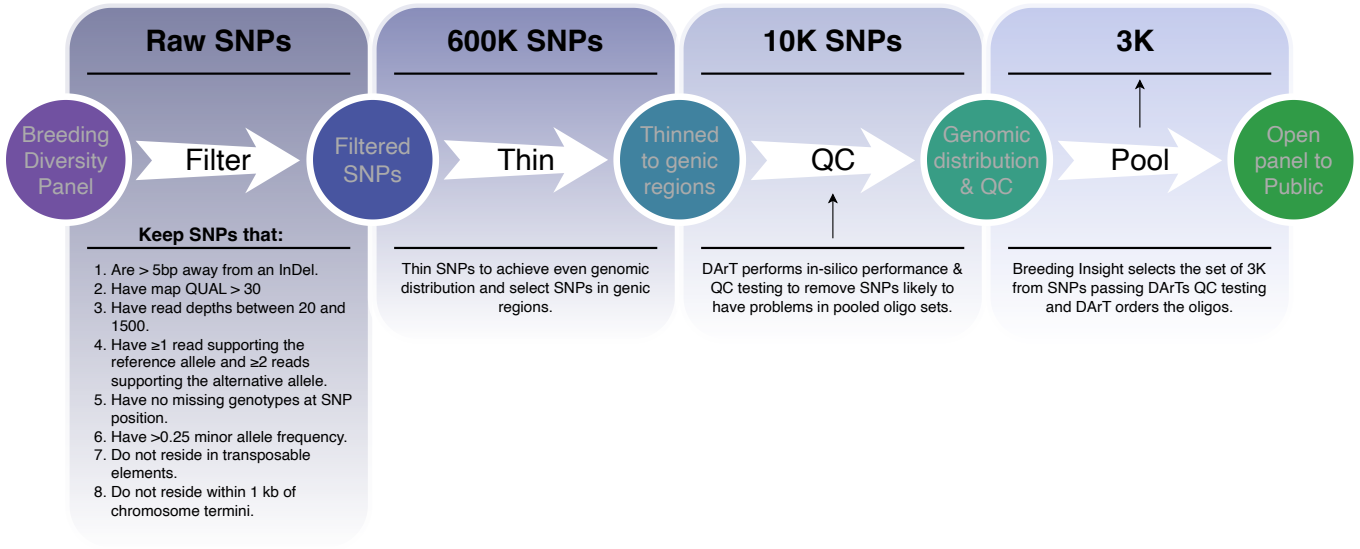

Figure 1. Filters and criteria applied to produce the 3K DArTag marker loci panel from the WGS of the blueberry diversity panel. Note that the 3K marker loci selection contains the 3K target SNPs discovered from the SNP discovery using a diversity panel of 31 blueberry lines. Abbreviations: K is thousands.

group, genomic order (physical position) of the markers was used to perform phasing and generate the genetic map. The construction of the genetic map involved initially creating individual maps for each parent, which were then integrated into a comprehensive HMM model using the merge_single_parents_maps function, resulting in a consolidated map. Additional unmapped markers were incorporated using the augment_phased_map function, which adds markers with redundant map information. The final F_1 genetic map was constructed with 1,301 unique (1,487 total) markers (Supplemental Figure 3B). Lastly, the haplotypes of the F_1 individuals were reconstructed by employing genotypic conditional probabilities through the calc_homoprob function (example shown in Supplemental Figure 3C).

Results

Validation of the 3K blueberry DArTag panel and genotyping results

To assess the quality and completeness of data, a validation set of 375 samples was genotyped using the 3K DArTag panel to (1) assess diversity among cultivated blueberries, (2) construct a genetic linkage map, and (3) evaluate its usefulness across species and subgenera.

DArT generates genotyping results in several formats, among which the MADC format (missing allele discovery count) provides all the microhaplotypes (54–81bp) discovered based on amplicons for the 3K marker loci. These microhaplotypes contain target SNPs per assay design as well as off-target SNPs, which are present in flanking amplicon sequences. To better distinguish these microhaplotypes, those matching the reference and alternative alleles at the target SNP site and containing no other variant nucleotide are denoted as Ref and Alt microhaplotypes, respectively. Additional haplotypes that contain off-target SNPs in variant nucleotides in the flanking sequences are denoted as RefMatch (when target SNP matches Ref) and AltMatch (target SNP matches Alt) with consecutive numbering for uniqueness (Figure 2). The MADC report (Supplemental File 2) was filtered at the microhaplotype level by requiring at least 5% of total samples, each having a minimum of 2 reads to retain a RefMatch or AltMatch. Out of 16,340 RefMatch and AltMatch, 8,370 were filtered out due to high missing data and 7,970 remained.

Panel effectiveness in extant accessions

The marker loci detection rate was determined at both sample and marker levels, respectively. All 375 samples contained data from $\geq$ 25% marker loci, therefore, no samples were removed. About 95% (n = 355) of total samples have data from $\geq$ 75% marker loci, indicating the high detection efficiency of the marker panel. At the marker level, data presence ranged from 5% to 100% in samples. It is worth noting that 1,722 (57%) marker loci were detected in $\geq$ 95% of samples and 299 (10%) marker loci were detected in all the samples surveyed, representing the most conserved marker loci in the blueberry genome and its related species. A total of 101 marker loci with data in $<$ 5% of total samples were excluded for downstream analyses. The average missing data for each cultivated blueberry type was as follows: 19% for NHB (range 8–24%), 18% for SHB (range 13–26%), 20% for HHB (range 18–21%), and 21% for RE (range 17–24%) (Supplemental Figure 2). Wild species from the *Vaccinium* subgenus had a missing data rate ranging from 18–56% (Supplemental Figure 2), whereas the five cranberry samples (*Oxycoccus* subgenus) exhibited the highest missing rates ranging from 54–73%. Marker loci that worked across subgenera

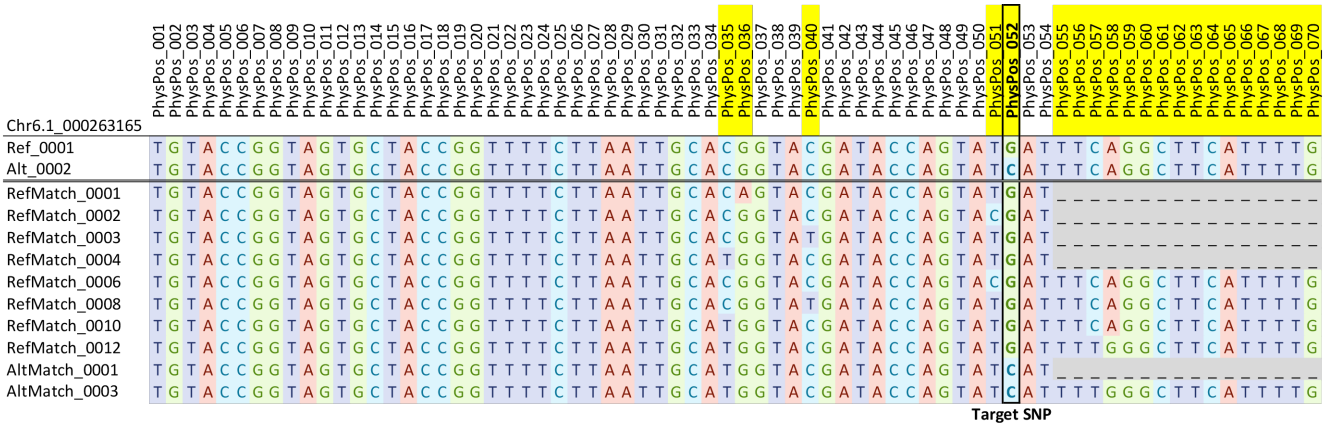

Figure 2. Example of DArTag sequencing reads from blueberry locus Chr6.1_000263165. Each sequence is a microhaplotype detected in breeding material tested on the panel. The DArTag assay was designed to detect the target SNP and distinguish the Reference allele from the Alternative allele. Additional variant positions (yellow fill) distinguish the individual microhaplotypes. PhysPos refers to the physical nucleotide position within the sequencing read from left to right. Newly discovered haplotypes are named with incrementing left-padded numbers with a prefix of 'RefMatch' or 'AltMatch' depending on which allele they match the Ref or Alt nucleotide at the Target SNP, respectively.

are likely linked to conserved regions of the blueberry and cranberry genomes.

Creation of a linkage map

A bi-parental population of 'Draper' (NHB) and 'Jewel' (SHB) (DxJ) was genotyped to test if the 3K DArTag panel can be used to generate a linkage map. The population was created by Michigan Blueberry Grower Marketing and clones of the parents, 'Draper' and 'Jewel', were distributed widely to researchers and growers nationwide. The true parents of the DxJ population were not available to be genotyped so we genotyped five different samples of both 'Draper' and 'Jewel'. Genetic evidence supported that 'Draper_2004.001 S10-42' and 'Jewel_2157.001_G04-01' were close proxies for the true parents (see Materials and Methods; Supplemental Figure 1).

The final DxJ F_1 linkage map consisted of 12 linkage groups with 1,301 markers and a total length of 1,368.6cM (average density of 0.96 markers/cM) from 161 progeny (Figure 3; Supplementary Figure 3). Linkage group length ranged from 90.50cM to 148.30cM, with an average of 114.05cM. Markers were well distributed throughout the 12 linkage groups. Supplemental File 3 contains linkage groups with marker order, positions in cM, and parental phasing. Additionally, haplotypes (indicating recombination events) for all individuals in F_1 population were reconstructed (Supplemental Figure 3C) based on the genotypic conditional probabilities.

Discussion and conclusion

The blueberry DArTag panel is now publicly available and open for any researcher or breeder to order through DArT (https://www.diversityarrays.com). The panel was designed on the legacy technology to produce 54bp reads but worked equally well with the current technology (81bp reads) with the caveat that some residual adapter sequences may be included (read-through of the entire fragment into the adapter). Raw data in FASTQ can be requested as can the Missing Allele Discovery File (MADC) that indicates the read depth of each microhaplotype in each sample. The high detection rate and repeatability make this panel suitable for genetic map construction, marker-assisted selection, whole-genome association mapping, reconstruction of recombination patterns, allele dosage estimation and parental confirmation in North American cultivated NHB, SHB, RE, and HHB, with some limited application in other *Vaccinium* species. The efficacy of the panel on breeding materials outside of North America has not been tested at this time.

The DArTag assay can be processed from blueberry gDNA or leaf tissue to genotyping data extraction in a 3–4-week turnaround time. The DArT genotyping data report comprises allele dose calls and raw data with custom report formats available upon request. One benefit that DArTag has over fixed array platforms is the ability to update and improve the marker panel as required over time. The panel is a pool of 3,000 oligos, one per locus, which is used to generate the sequencing libraries from the assayed material. Because the pool is created from individual oligo stocks, the removal of suboptimal loci or the addition of new loci can be easily done by creating a new pool. To determine which loci should be considered for removal, extensive genotyping (> 10,000 samples) is underway to identify and remove those loci that consistently underperform or fail. Independently, as new significant QTL markers and/or markers specific to other germplasm are detected, they can be targeted for inclusion in the original pool in the next version(s) of the panel. DArT offers re-pooling services once per year at low or no cost, but more frequent requests could result in labour surcharges being applied (Andrzej Kilian, personal communication). Researchers interested

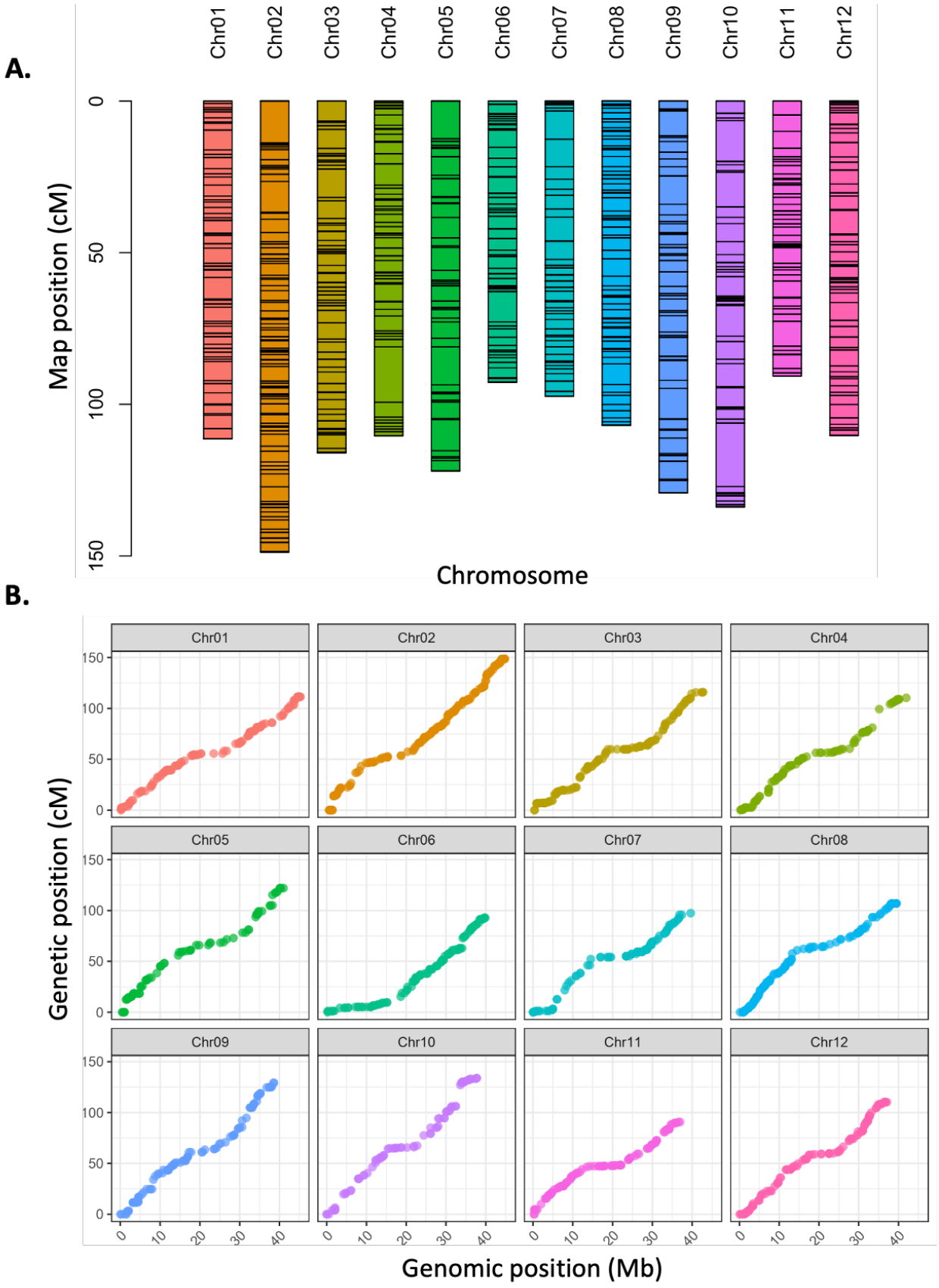

Figure 3. Genetic map of the DxJ bi-parental F1 population. A) Marker distribution across 12 linkage groups of the blueberry genome. Scale bar is shown in cM. B) Relationship plots of genetic distance (cM) to physical distance (Mb) for each of the 12 linkage groups.

in initiating projects with DArT are encouraged to contact DArT directly for consultation.

Another benefit of genotyping using the blueberry 3K DArTag panel is the ability to detect and catalogue all microhaplotypes into a fixed allele database, which will improve combining data sets across genotyping projects (manuscript in preparation). If after testing on thousands of samples, there are too few markers for GWAS for a given trait of interest, additional DArTag panels can be made to complement this panel, or larger platforms like the Flex-Seq 22K panel (Flex-Seq Panel Code: FS_1903) from RAPiD Genomics could be utilized (Nahla Bassil, personal communication). Another option is to add the required loci to the existing panel up to the technical limit of 7K, which is a more cost-effective option for the routine genotyping service with scalability.

We chose to create a panel of 3,000 marker loci due to cost and technical reasons, but smaller complementary panels can be made at lower up-front and downstream usage costs. The practical upper limit for the maximum number of probes on a DArTag panel is 7,000 loci, though the optimal maximum may differ by species and genome complexity, and read depth required to sufficiently c all g enotypes (Andrzej Kilian DArT, personal communication). The blueberry breeding community could decide to create a complementary 3K panel to result in more detailed genotypic data, however, this would nearly double the cost of genotyping per sample.

Data availability statement

The FASTQ files from the whole-genome skim sequencing for the 31 blueberry accessions used for identifying the candidate SNP variants are housed in the NCBI Short Read Archive under the BioProject ID PRJNA1020150. The targeted regions used to create the 3K DArTag markers are available on DRYAD (pre-publication URL: https://datadryad.org/stash/share/UeW2RMVU2bbxTKm0SBMuKp6VJVFshE72Um9maVAKqjA; DOI: 10.5061/dryad.j6q573nnc). The code and data for the construction of the F_1 map in MAPpoly2 are available in our GitHub repository for those interested in reproducing our analysis (https://github.com/Breeding-Insight/Blueberry_DArTag_Panel_paper#blueberry_dartag_panel_paper).

Supplemental data

Supplemental Table 1. Accessions used in the construction and testing of the blueberry 3K DArTag panel
Supplemental Figure 1. Principle Component Analy-sis (PCA) plots of the 'Draper' x 'Jewel' F_1 population
Supplemental Figure 2. Missing data rates for different grouped subsets of genetic material
Supplemental Figure 3. Blueberry Genetic map construction for the F1 population
Supplemental File 1. Genomic information of the blueberry 3K DArTag marker panel
Supplemental File 2. MADC report for the 375 samples used to validate the 3K DArTag panel
Supplemental File 3. Linkage group with their marker order, positions in cM, and parental phasing information where P1 represents 'Draper' and P2 represents 'Jewel'.

Acknowledgements

Breeding Insight is acknowledged for project design, marker development, curation and data processing. Diversity Arrays Technology created the oligo array, provided sequencing services and contributed to the manuscript. We thank Chad Finn, Nahla Bassil, Ebrahiem Babiker, Massimo Iorizzo and Mark Ehlenfeldt for providing germplasm. We also thank Alexandra Casa and Sharon Mitchell for careful reading and revision of the manuscript. Breeding Insight was funded for this work 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).

Marcelo Mollinari was funded by a USDA NIFA-awarded AFRI grant (project number: 2022-67013-36269).

Author contributions

DZ, NB, and MJS contributed to experimental design and planning. DZ, NB and MJS selected the diversity panel for WGS. NB collected and prepared all plant materials used in the study. DZ performed all the WGS analyses, SNP database creation, filtering pipelines, and quality control analyses to create the 3K panel. KHU managed the panel creation at Diversity Arrays Technology. DZ, MS and MM executed the data analyses and genetic mapping. MFG and MI assisted with 'Draper' and 'Jewel' parental identification. DZ, MS and MJS wrote the initial draft of the manuscript. CB managed experiments and communication among all authors involved. All authors contributed to reviewing the manuscript.

Conflict of interest statement

The authors have no conflicts of interest to report.

References

Blyton, M. D. J., Brice, K. L., Heller-Uszynska, K., Pascoe, J., Jaccoud, D., Leigh, K. A., and Moore, B. D. (2023). A new genetic method for diet determination from faeces that provides species level resolution in the koala. *bioRxiv* 2023.02.12.528172. doi: https://doi.org/10.1101/2023.02.12.528172

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15). doi: https://doi.org/10.1093/bioinformatics/btu170

Clark, L., Lipka, A., and Sacks, E. (2019). polyRAD: Genotype calling with uncertainty from sequenc-

ing data in polyploids and diploids. *G3 Genes—Genomes—Genetics* 9(3), 663–673. doi: https://doi.org/10.1534/g3.118.200913

Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., Wisecaver, J. H., Yocca, A. E., Alger, E. I., Tang, H., and Xiong, Z. (2019). Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience* 8, 12–12. doi: https://doi.org/10.1093/gigascience/giz012

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., Ramsay, L. D., Halpin, C., Mascher, M., Fleury, D. L., Landridge, P., Stein, N., and Waugh, R. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front Plant Sci* 10, 1–14. doi: https://doi.org/10.3389/fpls.2019.00544

Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., and Bull, J. K. (2007). Molecular markers in a commercial breeding program. *Crop Sci* 47, 154–163. doi: https://doi.org/10.2135/cropsci2007.04.0015IPBS

Edger, P. P., Iorizzo, M., Bassil, N. V., Benevenuto, J., Ferrão, L. F., Giongo, L., Hummer, K. E., Lawas, L. F., Leisner, C. P., Li, C., Munoz, P., Ashrafi, H., Atucha, A., Babiker, E. M., Canales, E., Chagne, D., Devetter, L., Ehlenfeldt, M. K., Espley, R. V., Gallardo, K., Gunther, C. S., Hardigan, M. A., Hulse-Kemp, A. M., Jacobs, M. L., Lila, M., Luby, C. H., Main, D., Mengist, M. F., Owens, G. L., Perkins-Veazie, P., Polashock, J. J., Pottorff, M., Rowland, L. J., Sims, C. A., Song, G., Spencer, J., Vorsa, N., Yocca, A. E., and Zalapa, J. E. (2022). There and back again; historical perspective and future directions for Vaccinium breeding and research studies. *Horticulture Research* 9. doi: https://doi.org/10.1093/hr/uhac083

FAOSTAT (2021). Food and Agriculture Organization of the United Nations Statistics Division (FAOSTAT). Click Item as Blueberries, Area as United States and From Year 2016 To Year 2021.

Feuerstein, U., Brown, A. H. D., and Burdon, J. J. (1990). Linkage of rust resistance genes from wild barley (Hordeum spotaneum) with isozyme markers. *Plant Breeding* 104, 318–324. doi: https://doi.org/10.1111/j.1439-0523.1990.tb00442.x

Gallardo, R. K., Zhang, Q., Klingthong, P., Dossett, M., Polashock, J. J., Rodriguez-Saona, C., Vorsa, N., Edger, P., Scherm, H., Ashrafi, H., Babiker, E. M., Finn, C. E., and Iorizzo, M. (2018). Breeding trait priorities of the blueberry industry in the United States and Canada. *HortScience* 53, 1021–1028. doi: https://doi.org/10.21273/HORTSCI12964-18

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genet* 210, 789–807. doi: https://doi.org/10.1534/genetics.118.301468

Hancock, J. (2009). Highbush blueberry breeding. *Latvian J of Agron* 12, 35–38.

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 Biotech* 19, 1–26. doi: https://doi.org/10.1186/s43141-021-00231-1

Hawkins, C. and Yu, L. X. (2018). Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. *The Crop Journal* 6, 565–575. doi: https://doi.org/10.1016/j.cj.2018.01.006

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci* 49, 1–12. doi: https://doi.org/10.2135/cropsci2008.08.0512

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: https://doi.org/10.1007/BF00020093

Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., and Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. *PNAS* 105, 9296–9301. doi: https://doi.org/10.1073/pnas.0803240105

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v2. url: https://arxiv.org/abs/1303.3997.

Lorenzana, R. and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120, 151–161. doi: https://doi.org/10.1007/s00122-009-1166-3

Lyrene, P. M., Vorsa, N., and Ballington, J. R. (2003). Polyploidy and sexual polyploidization in the genus Vaccinium. *Euphytica* 133, 27–36. doi: https://doi.org/10.1023/A:1025608408727

Mejia-Guerra, M. K., Zhao, D., Sheehan, M. J., Yu, L. X., and Kole, C. (2021). Genomic resources for breeding in alfalfa: availability, utility, and adoption. In The Alfalfa Genome, Compendium of Plant Genomes, Springer, Cham, 177-189.

Milner, S. G., Jost, M., Taketa, S., Mazon, E. R., Himmelbach, A., Oppermann, M., Weise, S., Knupffer, H., Basterrechea, M., König, P., Schüler, D., Sharma, R., Pasam, R. K., Rutten, T., Guo, G., Xu, D., Zhang, Z., Herren, G., Müller, T., Krattinger, S. G., Keller, B., Jiang, Y., González, M. Y., Zhao, Y., Habekuß, A., Fäber, S., Ordon, F., Lange, M., Börner, A., Graner, A., Reif, J. C., Scholz, U., Mascher, M., and Stein, N. (2019). Genebank genomics reveals the diversity of a global barley collection. *Nat Genet* 51, 319–326. doi: https://doi.org/10.1038/s41588-018-0266-x

Mollinari, M. and Garcia, A. A. F. (2019). Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models. *G3 Genes—Genomes—Genetics* 3, 3297–3314. doi: https://doi.org/10.1534/g3.119.400378

Mollinari, M., Olukolu, B. A., Pereira, G. S., Khan, A., Gemenet, D., Yencho, G. C., and Zeng, Z.

(2020). Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. *G3 Genes—Genomes—Genetics* 3, 281–292. doi: https://doi.org/10.1534/g3.119.400620

NASS (2023). Noncitrus Fruits and Nuts 2022 Summary (National Agricultural Statistics Service). url: https://downloads.usda.library.cornell.edu/usda-esmis/files/zs25x846c/zk51wx21m/k356bk214/ncit0523.pdf.

Qu, L., Hancock, J., and Whallon, J. (1998). Evolution in an autopolyploid group displaying predominantly bivalent pairing at meiosis: genomic similarity of diploid Vaccinium darrowi and autotetraploid V. corymbosum (Ericaceae). *Am J Bot* 85, 698–703. doi: https://doi.org/10.2307/2446540

Qu, L. and Hancock, J. F. (1995). Nature of 2n gamete formation and mode of inheritance in interspecific hybrids of diploid Vaccinium darrowi and tetraploid V. corymbosum. *Theor Appl Genet* 91, 1309–1315. doi: https://doi.org/10.1007/BF00220946

Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Mol Biol Rep* 1, 3–8. doi: https://doi.org/10.1007/BF02680255

Telfer, E., Graham, N., Macdonald, L., Li, Y., Klápště, J., Resende, M., Neves, L. G., Dungey, H., and Wilcox, P. (2019). A high-density exome capture genotype-by-sequencing panel for forestry breeding in Pinus radiata. *PLoS One* 14, 222640–222640. doi: https://doi.org/10.1371/journal.pone.0222640

USHBC (2021). History of highbush blueberries (U.S. Highbush Blueberry Council). url: https://blueberry.org/about-blueberries/history-of-blueberries/.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., Vincente, F. S., Prasanna, B. M., Li, X., and Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci Rep* 10. doi: https://doi.org/10.1038/s41598-020-73321-8

Zhao, D., Mejia-Guerra, K. M., Mollinari, M., Samac, D. A., Irish, B. M., Heller-Uszynska, K., Beil, C. T., and Sheehan, M. J. (2023). A public mid-density genotyping platform for alfalfa (Medicago sativa L.). *Genet Resourc J* 4(8), 55–63. doi: https://doi.org/10.46265/genresj.EMOR6509

Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 45–60
DOI: 10.46265/genresj.DBNO8764
https://www.genresj.org
ISSN: 2708-3764

Identification of genetically plastic forms among Belarusian ancient flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) varieties using the Linum Insertion Sequence LIS-1

Maria Parfenchyk *, Valentina Lemesh, Elena Lagunovskaya, Valentina Sakovich, Andrei Buloichik, Elena Guzenko and Lyubov Khotyleva

Institute of Genetics and Cytology, National Academy of Sciences of Belarus, 27, Akademicheskaya Str, Minsk, 220072, Republic of Belarus

Abstract: The Linum Insertion Sequence 1 (LIS-1) occurs in the genetically plastic flax genotypes in response to the lack or excess of mineral and water nutrition, but also naturally, and can be transmitted to the progeny. We have analyzed 21 ancient Belarusian varieties of flax *Linum usitatissimum* convar. *elongatum* Vav. et Ell. The LIS-1 presence or absence was checked for individual plants in at minimum two generations with primer-specific polymerase chain reaction (PCR) and agarose gel electrophoresis. The studied flax varieties formed four groups: non-responsive varieties (LIS-1 was not found, group NR); responsive, which formed and completely lost the insertion (group R0); responsive, which formed and retained LIS-1 (group R1); and responsive unstable (group R2). A statistically significant difference was found in 'plant height' ($p < 0.05$), 'technical length of the stem' ($p < 0.05$) between R0 and NR, and R2 and NR LIS-1 groups. The machine learning algorithm random forest classifier was used to predict the presence, absence or heterozygosity of LIS-1 in flax plants based on their growth and reproductive characteristics. As a result, the accuracy of the prediction was 98% on test data. In terms of sources for the selection of fibre flax varieties adaptive to environmental challenges, the most promising group consists of responsive varieties that have formed LIS-1 insertion (R0, R1 and R2 groups).

Keywords: Flax, *Linum usitatissimum* convar *elongatum*, linum insertion sequence (LIS1), local varieties, machine learning, random forest classifier

Citation: Parfenchyk, M., Lemesh, V., Lagunovskaya, E., Sakovich, V., Buloichik, A., Guzenko, E., Khotyleva, L. (2024). Identification of genetically plastic forms among Belarusian ancient flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) varieties using the Linum Insertion Sequence LIS-1. *Genetic Resources* 5 (9), 45–60. doi: 10.46265/genresj.DBNO8764.

Introduction

Flax (*Linum usitatissimum* L.) has been one of the most important industrial crops for several millennia, the fibre and oil of which are used in various industries around the world (Sa *et al*, 2021). There are four convarieties of cultivated flax: *crepitans*, *elongatum*, *mediterraneum* and *usitatissimum*. Convariety *elongatum* is characterized by a plant height exceeding 70cm and side branches occupying only the upper one-third or less of the entire stem length; if the plant height falls below 70cm, stem branches are located in the upper one-fifth of the entire stem length (Diederichsen, 2019). It is common mostly in Eastern Europe (Vavilov, 1926). In Belarus, *Linum* has been one of the major technical crops for many decades. Linseed varieties are grown for oil, flax varieties are grown for fibre, and flax varieties that tend to be intermediate between oilseed and fibre types may be used to produce high oilseed yields and good-quality fibre (Rachinskaya *et al*, 2011). Flax can grow in many environments but prefers cool weather and well-drained soils with good water-holding capacity (Ehrensing, 2008).

*Corresponding author: Maria Parfenchyk (maria.parfenchyk@gmail.com)

Received: 11.07.2023 | Accepted: 07.04.2024 | Published online: 07.05.2024

Experimental studies have shown that changes in soil nutrients and water regime during flax cultivation lead to phenotypic changes, which are accompanied by genomic rearrangements (Evans *et al*, 1966; Cullis, 1976, 1981; Goldsbrough *et al*, 1981). One of the heritable genomic changes in response to nutrient and water stress is the appearance of Linum Insertion Sequence-1 (LIS-1). A more detailed study of the occurrence of LIS-1 showed that the appearance of the insertion is limited to individuals (genotypes) that respond to growth conditions by modifying their genome (Chen *et al*, 2009). Single-copy insertion LIS-1 is assembled from short sequences scattered throughout the flax genome in a short period of time before flowering (Cullis, 1976). The LIS-1 is a 5.7kb nucleotide sequence that inserts at a specific site in the flax genome. This specific site contains two genes, inhibitor of growth-1, and kip-related cyclin-dependent kinase inhibitor-2. The target sequence is also modified when LIS-1 is inserted. A total of 129 single nucleotide polymorphisms and InDels were found in the target sequence when comparing lines with and without LIS-1 (Bickel *et al*, 2012). Contrasting conditions of mineral nutrition of seedlings caused the appearance on the plastic (or responsive) genotypes (line Pl) of two types of stable genotrophs: L-genotrophs and S-genotrophs (Durrant and Nicholas, 1970; Durrant and Jones, 1971). Treatment with low or imbalanced nutrients (different concentrations of nitrogen, phosphate and potassium in soil, or lack of water) gives rise to the small, or S, genotroph. Morphologically the S-genotrophs are shorter than the Pl line, have a non-branching stem, hairy capsule septa, and contain single-copy insertion LIS-1. A high nutrient and water treatment results in the large, or L-genotrophs, which are much taller than the S-genotrophs, have more branching stems than Pl or S, hairless capsule septa, and do not contain LIS-1. Both L- and S-genotrophs have been shown to be better adapted to the nutrient environment in which they were induced. In the responsive flax genotype, which did not form stable genotrophs, LIS-1 was lost in the absence of inducing conditions (Bickel *et al*, 2012; Chen *et al*, 2009, 2005). The characteristics altered in the genotrophs include height, weight, number of branches, the presence of hairs on seed capsule septa, total nuclear DNA contents, the number of genes coding for the large ribosomal RNAs and the 5S ribosomal RNAs as well as a number of other repetitive sequence families (Chen *et al*, 2005).

Chen *et al* (2005) have shown that the insertion LIS-1 could occur also naturally in many flax and linseed varieties, and concluded that the external environment can act both as an inducer of directed genetic variability and as a selective factor for beneficial mutations. Flax is a self-pollinator, and the ability to modify the genome may be an adaptive property (Cullis, 1986).

To sum up, in responsive flax genotypes many genomic rearrangements occur in response to environmental challenges, including the occurrence of LIS-1. Using LIS-1 as a marker of genome plasticity is a fast and cost-efficient tool for primary screening of genotypes as the insertion could be detected by polymerase chain reaction. Thus, the LIS-1 sequence is a promising molecular marker for identifying flax forms with genome plasticity and, accordingly, adaptive capacity.

Artificial intelligence and machine learning algorithms are used in different fields of science to solve problems of classification or regression. By learning from existing data, supervised, semi-supervised or unsupervised techniques could be applied in chemistry (Raghunathan and Priyakumar, 2022) for predicting properties and designing molecules and materials; in the pharmaceutical industry (Volkamer *et al*, 2023) for the predictions of bio-activity and physical properties; and in active learning (Bajorath, 2022) for drug discovery. It was shown that random forest provides comparable performance and easier interpretation for many applications (Volkamer *et al*, 2023) or outperforms other models (Yang *et al*, 2022), like support vector machine, decision tree, and extreme gradient boosting tree algorithms. Random forest is a supervised ensemble method that randomly builds and integrates multiple decision trees to create a forest structure. The choice of the machine learning algorithm depends on the data structure, data types and questions you want an answer to (Hu and Xing, 2021).

The aim of this study was to investigate local ancient Belarusian flax varieties with the LIS-1 insertion as a marker of genome plasticity and, based on available morphological features, to construct a classification model to predict the presence or absence of the LIS-1 insertion using a random forest classifier.

Materials and methods

Plant material

For the analysis, 21 ancient local varieties of flax *Linum usitatissimum* convar. *elongatum* Vav. et Ell. were used. Seeds were obtained in 1998 from the N.I. Vavilov Institute of Plant Genetic Resources (VIR) genebank, where they were collected during Vavilov expeditions from 1923 to 1958 in Belarus. Plant material origin is shown in Table 1. Since 1998, these varieties have been cultivated and studied at the Biological Experimental Station of the Institute of Genetics and Cytology of the National Academy of Sciences of Belarus. In 2000, the National Bank of Seeds of Plant Genetic Resources of Belarus was established for the conservation, investigation and use of plant genetic resources (Privalov *et al*, 2021), where investigated varieties are included and conserved. Since the seed material was obtained from the VIR genebank, the accession numbers are given according to VIR.

Experimental conditions

The 21 studied local ancient varieties of flax were sown and cultivated in the Biological Experimental Station of

Table 1. Flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) accessions included in the study. Information includes accession numbers in the VIR collection, dates of inclusion in the collection, origin and LIS-1 group. Information about the exact location of some sampling has not been preserved.

Accession number (VIR code)	Year added	Place of origin	Belarus region	LIS-1 group
624_595	1923	Homielskaja vobłasć	South-East	R2
624_596	1923	Homielskaja vobłasć	South-East	NR
624_776	1923	Viciebskaja vobłasć	Nord	R1
624_781	1923	Minskaja vobłasć, Červieński district	Center	NR
624_784	1923	Belarus	Unknown	NR
624_786	1923	Belarus	Unknown	R1
624_787	1923	Belarus	Unknown	R1
624_789	1923	Belarus	Unknown	R1
624_791	1923	Mahiloŭskaja vobłasć, Čerykaŭski district	East	R0
624_1043	1924	Viciebskaja vobłasć, Połacki district	Nord-East	R0
624_1044	1924	Viciebskaja vobłasć, Haradocki district	Nord	R1
624_5462	1939	Minskaja vobłasć, Dokšycki district	Center	R1
624_5463	1939	Viciebskaja vobłasć, Pastaŭski district	Nord-West	R2
624_5464	1958	Viciebskaja vobłasć, raka Dzisna	Nord	R2
624_6213	1958	Viciebskaja vobłasć, Šarkaŭščynski district	Nord-West	R2
624_6214	1958	Viciebskaja vobłasć, Hłybocki district	Nord-West	R1
624_6215	1958	Hrodzienskaja vobłasć, Navahrudski district	West	R1
624_6216	1958	Bresckaja vobłasć	South-West	R0
624_6220	1958	Bresckaja vobłasć	South-West	R2
624_6219	1958	Bresckaja vobłasć, Ivanaŭski district	South-West	R2
624_6222	1958	Hrodzienskaja vobłasć, Karelicki district	West	R1

the Institute of Genetics and Cytology in Minsk from 2017 to 2022. The studied varieties are self-pollinating, and no crosses were carried out with them. About 50 plants per variety were grown each year, 10 plants were taken to test the phenotype, and 5–10 plants per variety were tested by polymerase chain reaction (PCR) for the presence or absence of the LIS-1 insertion. No further studies were conducted for those varieties for which LIS-1 was not found. Seeds were collected from each plant found to have LIS-1, planted again in the following year, and the presence or absence of LIS-1 was checked again. For each variety, there were at least two generations of plants with traceable LIS-1 presence or loss.

Nitrogen fertilizer was applied to the soil in accordance with the field fertilization schedule, both when preparing the soil for seeding and during plant growth. Plants were watered as needed.

Weather conditions (rainfall, mm and temperature, °C) during the flax vegetation period from May to August for the study years are shown in Figure 1.

The conditions of the growing season varied depending on the year. The average temperatures in May, June, July and August were 12.83°C, 18.51°C, 18.48°C, and 18.52°C, respectively. The coldest years were 2017 and 2020 (mean growing season temperatures were 16.48°C and 16.20°C, respectively), and the warmest was 2018 (average temperature 18.70°C), with a 6-year average growing season temperature of 17.09°C. The wettest years were 2018 and 2019 (total rainfall during the growing season was 328 and 333mm), and the driest year was 2020 (total rainfall 240mm), with an average rainfall during the growing season over six years of 307.17mm.

The optimal temperature for flax development is from 15–18°C. The phenological stages of flax are: germination (May), leaf development (end of May), active growth (June), budding and flowering (June, July), development of seed capsules (July), ripening of seed (end of July, August). Flax continues growing until the end of flowering.

Phenotypic characterization

The morphological features of the flowers and plants of the varieties were evaluated according to the flax descriptors (Maggioni *et al*, 2001; Nôžková *et al*, 2016). For the flower, the following were detected: shape of the flower, shape of the corolla, size of the corolla, shape of the petals, longitudinal folding of the petals, colour of corolla (the petals colour, when fully developed), colour of the veins of the petals, anther colour, seed colour. For the plant: foliation, plant height, technical length of the stem; number of productive seed capsules per plant; and number of seeds in the capsule. For the modelling, morphological features included an extended list of characteristics, which were not investigated before: the ciliation of seed capsule septa and the presence of anthocyanin pigmentation in the hypocotyl in addition to the plant height, the technical length of

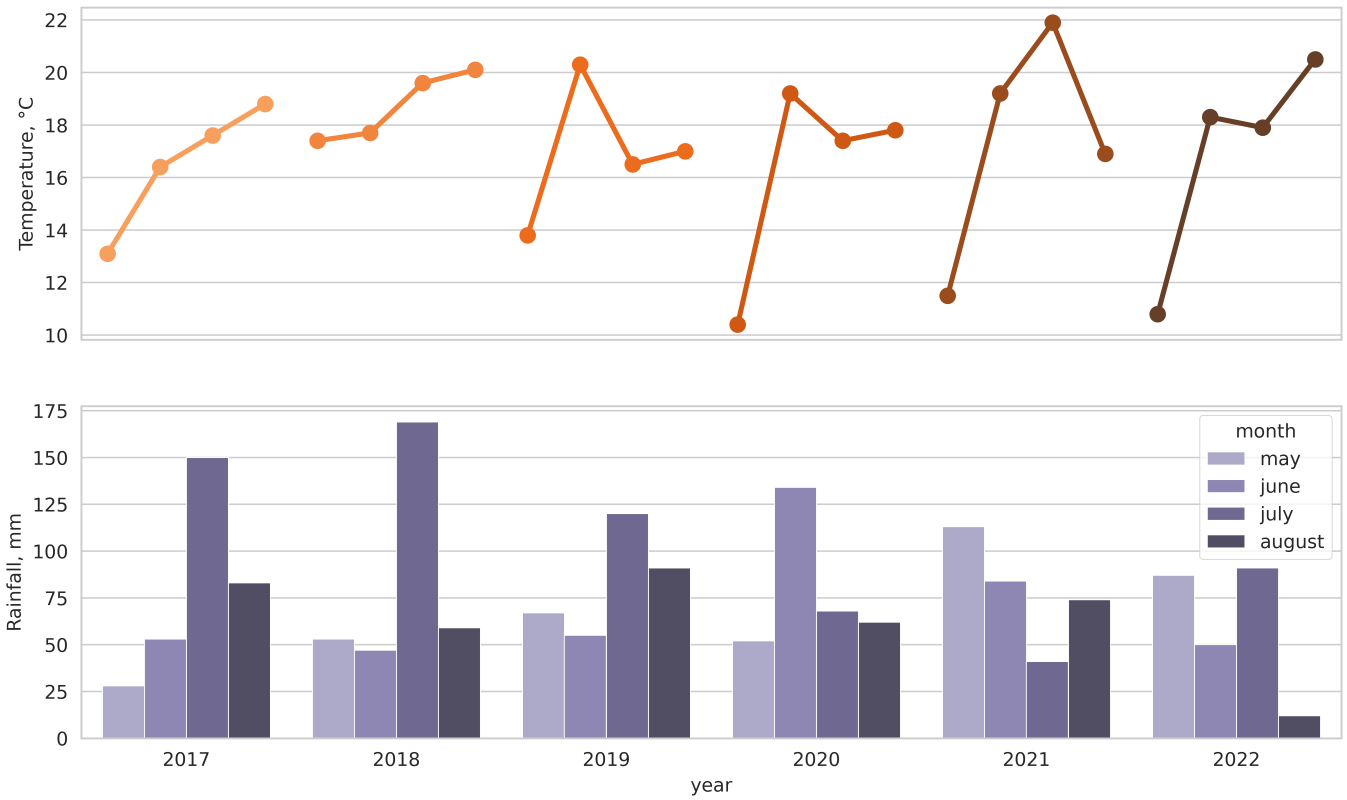

Figure 1. Rainfall (mm) and temperature (°C) in May, June, July and August for each year of the study.

the stem, the number of productive seed capsules per plant and the number of seeds in the capsule. For the modelling, seven flax varieties which were grown in 2022 were used, for which these characteristics were available: 624_6222, 624_6219, 624_1044, 624_791, 624_786, 624_6215, and 624_789. Ciliation of seed capsule septa was shown to be associated with LIS-1 presence/absence (Chen *et al*, 2005), and therefore this trait could be important for the machine learning model. We included the presence of anthocyanin pigmentation in the hypocotyl characteristic in our list of traits as one of the possible responses of flax genotypes to abiotic stress. This information is available in Supplemental Table 1.

Molecular genetic analysis

For molecular analysis, upper leaves were collected at the budding stage, since in the experiment of Cullis (1976) it was shown that the LIS-1 insertion is assembled from short fragments of DNA distributed throughout the genome in the short time preceding flowering. Leaves (50–100mg) of 5–10 individual plants per accession were taken for DNA isolation according to Sambrook and Russell (2006). Briefly, plant material was placed in 2.0ml microcentrifuge tubes, 15μl of TE buffer (10mM Tris-HCl, pH 7.5; 1mM EDTA) was added and the tissue ground using a Tissue Lyser tissue homogenizer (Qiagen). Then, 400μl of lysis solution (TrisHCl 1M, pH = 8.0, NaCl 5M, EDTA 0.5M, SDS 10%) was added to each tube and incubated for 10 minutes, with periodical shaking. Then, 600μl of phenol-chloroform mixture (1:1 v/v) was immediately added, mixed gently and centrifuged at 10,000rpm for 10 minutes. The supernatant was transferred to new tubes, and 600μl of a mixture of chloroform-isoamyl alcohol (24:1 v/v) was added, mixed and centrifuged for 2 minutes at 10,000rpm. The upper phase was transferred into clean tubes, and 800μl of ice-cold 96% ethanol was added, mixed with gentle rocking, and placed for 15 minutes at -20°C, then centrifuged for 7 minutes at 10,000rpm, and the supernatant was completely removed. The next step was to dissolve the DNA on a vortex at low speed in 100μl of a 1.2M NaCl solution. Then 300μl of ice-cold 96% ethanol was added, and the DNA was allowed to precipitate (for up to 2 hours at -20°C) and then centrifuged for 4 minutes at 12,000rpm. The precipitate was washed in 300μl of 70% cold ethanol, and the alcohol was removed with a pipette. DNA was dissolved in 100μl of double-distilled water. The DNA concentration in the resulting solution was measured using an Ultrospec 3300 pro spectrophotometer (Amersham Biosciences, USA) at a wavelength of 260nm (ultraviolet spectrum).

The LIS-1 insertion was detected by sequence-specific polymerase chain reaction (PCR) as described by Chen *et al* (2009). The primers used for amplification of LIS-1 at the insertion site were: 2 and 3' (5'-ggtttcagaactgtaacgaa-3' and 5'-gaggatggaagatgaagaagg-3'), 18 and 19' (5'-cataaattcagtcctatcgac-3' and 5' taacagctcggatctaggc 3'); the absence of the insertion was

detected by amplification with primers: 2 and Pl9' 5'-ggtttcagaactgtaacgaa-3' and 5'-gcttggatttagacttggcaac-3'. The sizes of the amplified fragments were 416, 398 and 417bp, respectively (Chen *et al*, 2009). Electrophoretic separation was carried out in 1.5% agarose gel, with further detection in ultraviolet light.

Statistical analysis

Correlation analysis using Python libraries: NumPy (Harris *et al*, 2020), SciPy (Virtanen *et al*, 2020) was performed to measure the strength of the relationship between the examined features and to calculate their association.

Generalized linear models were used to identify whether there were significant contributions of interaction between factors genotype (LIS-1 groups) and weather conditions (temperature and rainfall by month, average temperature and total rainfall) to predict the dependent variables (plant height, technical length of the stem, number of seed capsules per plant, and number of seeds in the capsule). The formula with interaction was: 'dependent variable ~ LIS_group*weather_condition'. The Shapiro test was applied to check the normality of models residuals with statistical significant effect. Analysis was run in Python version 3.10.9 using the statsmodels (Seabold and Perktold, 2010).

Analysis of variances (ANOVA) was performed to identify if factor LIS-1 groups had a significant effect on dependent variables: plant height, technical length of the stem, number of seed capsules per plant, and number of seeds in the capsule. The Shapiro-Wilk test (Wickham *et al*, 2022) was used to determine the normality and Levene's test (Wickham *et al*, 2022) was used to determine homoscedastisity to check ANOVA assumptions. Shapiro test was applied to check the normality of models residuals distribution. For the 'technical length of the stem' and 'number of seed capsules per plant', non-parametric Kruskal-Wallis ANOVA test was applied as the normality assumptions failed. Post-hoc Tukey's test for parametric and Dunn's test for non-parametric distributed characteristics were used for LIS-1 groups pairwise comparisons for models with significant effect of LIS-1 group factor on morphological characteristics of plants. P-values were taken into account under Bonferroni correction. Analysis was run in R version 4.1.2 using the lme4 (v1.1-35.1; Bates *et al* (2015)), rstatix (v0.7.2; Kassambara (2023)), dplyr (v1.1.4; Wickham *et al* (2022)) packages.

Machine learning model

For the analysis, a sample of 56 plants with known morphologic characteristics were taken from seven flax varieties: 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), 624_6219 (R2 group), and 624_791 (R0 group). Before building the model, the sample of plants were tripled using the function pandas.DataFrame.sample() with replacement (pandas.pydata.org, The pandas development team (2020)). Analysis was run in Jupyter notebook (Kluyver, 2016).

The basis of the method is the use of a set of single classifiers (a decision tree) to make the final decision on the classification of objects. The selection of hyperparameters for the calculation was carried out using the GridSearchCV algorithm, which checks each combination of hyperparameters from a given range of values and selects the most successful one. The number of iterations was set to seven for better cross-validation results. The range of values was as follows: param_grid = {'bootstrap': [True, False], 'max_depth': [3, 5, 7, 10], 'max_features': ['auto', None], 'criterion': ['gini', 'entropy'], 'n_estimators': [600, 800, 1000]}. The best combination of hyperparameters was calculated as follows: {'bootstrap': True, 'criterion': 'gini', 'max_depth': 7, 'n_estimators': 1000}. Random forest classifier algorithm was run with the following options: {'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 7, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 1000, 'n_jobs': None, 'oob_score': False, 'random_state': 13, 'verbose': 0, 'warm_start': False}. The number of features used at nodes for split was defined by grid search as sqrt(n_features). The criterion for feature importance estimation is gini impurity, which shows how each feature decreases the impurity of the split.

In total, nine features were used: four numeric (plant height; technical length of the stem; number of productive seed capsules per plant; number of seeds in the capsule), and two categorical (ciliation of septa and anthocyanin pigmentation in the hypocotyl). Categorical features were encoded to numeric with the pandas library method get_dummies(), and resulted in five features (ciliation of septa as: hairless hh, heterozygous Hh, hairy HH); the presence of anthocyanin pigmentation in the hypocotyl as: week or missing). Target variable was the presence of LIS-1 insertion ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote', when three PCR products were revealed).

The dataset was divided into training and test sets, with the proportion of test size being 0.3. At first, the machine learning algorithm will use only the training part of data (70% from all dataset) including features (characteristics) and the target variable (what we want to predict). On this data, the algorithm will learn about the structure and relationships between features, and finally make a prediction about the target variable. Then the algorithm takes a new 30% of test data (only features) and makes a prediction again. When comparing the real values of the test target variable with the predicted values, we can assess the degree of success achieved. As a criterion for the success of the classification, the accuracy score was calculated (the ratio of correctly classified objects to the total number of operations performed). A confusion matrix provided a visualization of additional information about

classification results: precision, recall and f1-score. Abbreviations in formulas to calculate these metrics were: TP (true positive, when both the actual and predicted values were 1), TN (true negative, when both the actual and predicted values were 0), FP (false positive, when the actual value was 0, but the predicted value was 1), FN (false negative, when the actual value was 1, but the predicted value was 0). Classification metrics were the following: precision, the proportion of positive identifications that were actually correct: TP/(TP+FP); recall, the proportion of actual positives correctly identified: TP/(TP+FN); f1-score, the weighted average of precision and recall – this score takes both false positives and false negatives into account: 2*(Recall * Precision) / (Recall + Precision).

Figures 1, 5, 6 and 7 were produced using Python with the 'matplotlib' and 'seaborn' (Hunter, 2007; Waskom, 2021) packages. Figure 3 was produced using R with the ggplot2 (v3.5.0; Wickham (2016)) package.

Results

Morphological characteristics of the studied flax varieties are shown in Table 2 and Table 3.

Varieties with the largest mean plant height were 624_6216, 624_6219, 624_6222, whereas varieties with the largest mean technical stem length were also 624_6216, 624_6219 and 624_5462. Varieties with the smallest mean plant height and technical stem length were 624_595, 624_6215, 624_781, and 624_6215, 624_786, 624_781.

The flowers differed in diameter, from small (624_782, 624_784, 624_6213, 624_6217, 624_6222) to large (624_791, 624_6216). According to the colour of the corolla and the anthers, the following combinations were noted: violet/bluish, blue/creamish, light blue/bluish, light blue/dark bluish, blue/bluish, and blue/dark bluish. Only one variety had cream-coloured anthers (624_776).

The majority of studied local ancient Belarusian varieties had a medium-sized corolla, a regular shape of the flower with circular petals, blue flower petals, anthers and veins of the petals, and brown seeds.

Presence of LIS-1

The studied flax varieties could be divided into four groups, corresponding to LIS-1 presence and preservation: R0, R1 and R2, which are responsive genotypes as they formed LIS-1 insertion, and NR group which includes non-responsive genotypes, i.e. LIS-1 insertion was not detected for them. Data is shown in Table 4. This LIS-1 group subdivision of accessions is the first attempt to generalize data.

Shared and individual morphological characteristics of the flower, seeds and stem of the studied flax varieties grouped by LIS-1 presence are shown in Figure 2. The analysis was run online using Venny (Oliveros, 2015). Based on the morphological characteristics listed in Table 3, we could not distinguish responsive and non-responsive accessions. The NR group of accessions had no distinctive individual characteristics, yet shared ten common traits with all four groups, and only one trait in common with the R1 and R2 groups (small corolla), and one with the R2 group (violet colour of petals). Groups of responsive accessions (R0, R1, R2) had three characteristics in common (blue colour of petal veins and petals, semi-star shape of flower), R0 and R1 groups had in common dark bluish colour of anthers, R0 and R2 groups were characterized by stem high foliation, and R1 and R2 groups had light brown colour of seeds coat. The R0 group had one individual characteristic (large flower); the R1 group had two individual traits (elliptical shape of petals and creamish anthers). So, we can consider that the presence or absence of LIS-1 insertion could have morphological effects, but the studied characteristics of the flower and stem are not sufficient for an exact morphological differentiation.

Average morphological characteristics of the four LIS-1 groups and two responsive groups (R0, R1, R2 are responsive, NR is non-responsive) are shown in Table 5.

Correlation analysis

We used correlation analysis to establish the relationship between quantitative characteristics, environmental conditions (rainfall, temperature) during vegetation period, and genetic group defined by LIS-1, as we want to reveal which factors impacting on morphologic characteristics depending on plant development stage and genetic group.

The LIS-1 groups exhibited significant negative correlations with both plant height and technical length of the stem (r = -0.294** and r = -0.248*, respectively, Table 6). Conversely, there was a strong positive correlation between plant height and technical length of the stem (r = 0.889***), between plant height and number of seeds per capsule (r = 0.25**), and plant height and temperature in August (r = 0.314***). Similarly, positive significant correlations were found between technical stem length and temperature in August (r = 0.427***), rainfall in June (r = 0.107*), rainfall in July (r = 0.2**), whereas negative significant correlations were observed between technical stem length and number of seed capsules per plant (r = -0.269**), temperature in June (r = -0.466***), and rainfalls in May and August (r = -0.341***, r = -0.271***). Plant height also exhibited negative significant correlations with temperatures in May and June (r = -0.08**, r = -0.264**), and rainfalls in May and August (r = -0.289***, r = -0.265***). Positive significant correlations were found between the number of seeds per capsule and the number of seeds capsules per plant (r = 0.45***), temperatures in June and the number of seeds in the capsule (r = 0.323***), rainfall in May and the number of seeds in the capsule (r = 0.368***), with additional statistically significant correlations between the number of seeds per capsule and temperatures in June (r = 0.241*) and rainfall in May (r = 0.151*). The number of seeds per

Table 2. Quantitative morphological characteristics of flax varieties. Numbers given represent the mean ± standard deviation (std) of ten plants examined.

Variety	Mean plant height ± std	Mean technical length of the stem ± std	Mean number of seed capsules per plant ± std	Mean number of seeds in the capsule ± std
624_595	53.08±1.82	40.37±1.997	7.80±1.64	7.92±1.64
624_596	47.97±1.90	36.7±2.01	7.3±1.55	7.63±1.55
624_776	53.97±2.57	42.34±2.50	6.43±1.32	7.82±1.32
624_781	48.03±2.52	36.78±2.39	7.14±1.76	7.68±1.76
624_784	55.09±2.32	43.64±2.90	7.18±1.37	7.85±1.37
624_786	53.22±2.50	39.67±2.57	8.52±2.25	8.07±2.25
624_787	55.52±2.23	42.62±2.06	6.53±1.059	8.2±1.059
624_789	56.43±2.64	45.08±2.38	7.695±1.53	8.08±1.53
624_791	55.90±3.24	43.34±3.067	7.92±2.048	8.058±2.05
624_1043	55.98±2.32	46.08±3.17	7.0017±1.45	7.985±1.45
624_1044	55.46±3.042	42.88±3.51	7.55±1.41	8.18±1.41
624_5462	57.06±4.58	48.2±4.43	6.496±1.56	7.79±1.56
624_5463	57.07±2.33	46.40±2.49	6.052±1.15	8.26±1.15
624_5464	56.42±3.15	47.22±3.08	5.64±1.107	7.705±1.11
624_6213	58.24±3.00	46.88±2.84	6.228±1.28	8.094±1.28
624_6214	53.94±1.96	43.87±2.089	6.61±1.202	7.92±1.202
624_6215	51.93±3.22	40.34±4.02	8.296±1.38	8.016±1.38
624_6216	63.86±2.99	52.76±3.22	6.53±1.27	8.32±1.27
624_6219	61.33±3.58	48.062±3.98	8.556±1.77	8.136±1.77
624_6220	54.09±3.41	45.68±2.64	6.697±1.43	7.44±1.43
624_6222	58.36±4.83	46.26±3.55	7.067±1.51	8.08±1.51

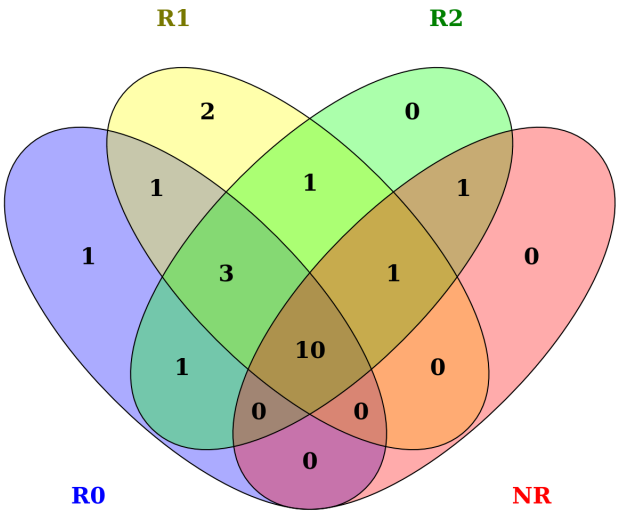

Figure 2. Venn diagram showing the number of common morphological characteristics of the flower, seeds and stem for the varieties grouped by LIS-1. R0: responsive varieties, formed LIS 1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

Table 3. Morphological characteristics of flax varieties and assignment to LIS-1 groups. The morphological characteristics of the flower, petal, foliation and seed coat colour were stable. *Anther colour in 2023, which is not included in this manuscript, for accession K-776 was not creamish. R0: responsive varieties, formed LIS-1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

Variety	Flower			Petal				Anther colour	Foliation	Seed colour	LIS-1 group
	Shape	Corolla shape	Size of corolla	Shape	Longitudinal folding	Colour of corolla	Vein colour				
624_595	Regular	Funnel	Medium	Circular	Absent	Violet	Violet	Bluish	Medium	Brown	R2
624_596	Regular	Funnel	Medium	Circular	Absent	Violet	Violet	Bluish	Medium	Brown	NR
624_776	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Creamish*	Medium	Light brown	R1
624_781	Regular	Plate like	Medium	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	NR
624_784	Regular	Funnel	Small	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	NR
624_786	Regular	Funnel	Small	Circular	Absent	Light blue	Blue	Dark bluish	Medium	Brown	R1
624_787	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_789	Semi-star	Plate like	Small	Elliptical	Absent	Blue	Violet	Bluish	Medium	Brown	R1
624_791	Regular	Plate like	Large	Circular	Absent	Light blue	Blue	Bluish	Medium	Brown	R0
624_1043	Regular	Funnel	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R0
624_1044	Regular	Plate like	Medium	Circular	Absent	Blue	Violet	Dark bluish	Medium	Brown	R1
624_5462	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_5463	Semi-star	Plate like	Medium	Circular	Absent	Light blue	Blue	Bluish	High	Brown	R2
624_5464	Semi-star	Plate like	Medium	Circular	Absent	Blue	Violet	Bluish	Medium	Light brown	R2
624_6213	Regular	Plate like	Small	Circular	Absent	Light blue	Blue	Bluish	Medium	Brown	R2
624_6214	Regular	Plate like	Medium	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	R1
624_6215	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_6216	Semi-star	Plate like	Large	Circular	Absent	Light blue	Blue	Dark bluish	High	Brown	R0
624_6219	Regular	Plate like	Medium	Circular	Absent	Blue	Violet	Bluish	Medium	Brown	R2
624_6220	Regular	Funnel	Medium	Circular	Absent	Blue	Blue	Bluish	High	Brown	R2
624_6222	Regular	Funnel	Small	Circular	Absent	Blue	Blue	Dark bluish	Medium	Brown	R1

Table 4. Accessions and the LIS-1 groups to which they could be affiliated based on the LIS-1 insertion presence and preservation.

LIS-1 group	Definition	Accessions
R0	Responsive, formed the insertion and then, over a number of generations, completely lost it.	624_6216, 624_791, 624_1043
R1	Responsive, formed the insertion and retained it over a number of generations.	624_776, 624_5462, 624_6214, 624_6222, 624_786, 624_787, 624_789, 624_1044, 624_6215
R2	Responsive, formed the insertion and then partially lost it over a number of generations.	624_595, 624_6220, 624_5463, 624_5464, 624_6213, 624_6219
NR	Non-responsive varieties, in which the insertion was not found.	624_784, 624_596, 624_781

Table 5. Quantitative morphological characteristics of the four LIS-1 groups and two responsive groups. R0: responsive varieties, formed LIS-1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

LIS-1 group	Mean plant height, cm ± std	Mean technical length of stem, cm ± std	Mean no. of seed capsules per plant ± std	Mean no. of seeds in the capsule ± std
R0	59.70±2.95	48.66±3.19	7.16±1.62	8.07±1.62
R1	54.88±3.16	43.16±3.02	7.28±1.48	8.02±1.48
R2	55.63±2.90	44.72±2.82	6.82±1.40	7.88±1.40
NR	50.18±2.15	39.07±2.43	7.24±1.60	7.73±1.60
Responsive group				
Responsive (R0+R1+R2)	55.95±3.04	44.61±2.99	7.11±1.48	7.98±1.48
Non-responsive (NR)	50.18±2.15	39.07±2.43	7.24±1.60	7.73±1.60

capsule exhibited a negative significant correlation with temperature in July (r = -0.261**).

As a result, plant height and technical stem length are correlated with the LIS-1 group, requiring a thorough analysis of their relationship; correlations between environmental conditions (rainfall, temperature) and quantitative characteristics (plant height, technical length of the stem, number of seed capsules per plant, and number of seeds per capsule) align with flax preferences across various stages of plant development.

The interaction effects of temperature in June were significant for groups R0, R2 and NR, impacting plant height ($p < 0.05$), while the effect for the R1 group was not statistically significant ($p = 0.108$) when implementing the generalized linear model. Effect on the border of significance between genotype (LIS-1 groups) and rainfall in May for the reproductive trait 'number of seeds in the capsule' was observed in the NR group ($p = 0.056$).

Analysis of variances (ANOVA)

The analysis of variances (ANOVA) showed statistically significant differences among LIS-1 groups for 'plant height' ($p = \mathbf{0.00119}$) and 'technical length of the stem' ($p = \mathbf{0.00581}$). Results are shown in Table 7.

For traits where ANOVA showed significant results, Tukey's and Dunn's tests were implemented to reveal the statistical significance of differences among LIS-1 groups in pairwise comparisons (Table 8).

A statistically significant difference was found in the plant height between groups NR and R0 ($p < 0.01$), and NR and R2 ($p < 0.05$). Similarly, for the technical length of the stem, a statistically significant difference was observed between the same groups NR and R0, and NR and R2 ($p < 0.05$). Before the sequential Bonferroni correction, significant differences were shown for groups R1 and R0, and R1 and NR ($p < 0.05$) in the technical length of the stem.

Accessions grouped as R0 and R2 were characterized by greater mean plant height and technical length of the stem compared to accessions of the NR group. The R1 group was statistically significantly taller than the NR group ($p = 0.0365$ before the sequential Bonferroni correction) for what concerns the technical length of the stem. In contrast, for reproduction-related traits such as 'number of seed capsules per plant' and 'number of seeds in the capsule', responsive genotypes (R0, R1, R2) did not significantly differ from non-responsive genotypes (NR). Figure 3 shows comparisons of all four groups by morphological characteristics. The R0, R1 and R2 groups are generalized as 'responsive group', and NR as 'non-responsive'.

Thus, from the plant's point of view, the most important thing is to leave seeds at the end of the vegetation period. Both responsive and non-responsive genotypes, revealed by LIS-1, are successful at this. But, from the point of identifying potentially useful genotypes for fibre flax selection, LIS-1 responsive

Table 6. Correlation analysis performed between the examined features (LIS-1 genetic group, plant height, technical length of the stem, number of seed capsules per plant, number of seeds per capsule, and temperature and rainfall in May, June, July and August). *, **, ***: statistically significant at $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. Bold font indicates significant correlations between quantitative characteristics, environmental conditions, and genetic group defined by LIS-1. Correlations between environmental conditions (rainfall, temperature) were out of the scope of manuscript, so are not discussed.

		Quantitative traits			
	LIS-1 groups	**Plant height**	**Technical stem length**	**No. of seed capsules per plant**	**No. of seeds per capsule**
Plant height	**-0.294****				
Technical stem length	**-0.248***	**0.889*****			
No. of seed capsules per plant	-0.036	-0.001	**-0.269****		
No. of seeds per capsule	-0.189	**0.251****	-0.012	**0.452*****	
Temp May	-0.001	**-0.08****	0.06	-0.213	-0.21
Temp June	0.011	**-0.264****	**-0.466*****	**0.323*****	**0.241***
Temp July	0.003	-0.177	-0.008	-0.072	**-0.261****
Temp August	-0.001	**0.314*****	**0.427*****	-0.029	-0.035
Rainfall May	0.009	**-0.289*****	**-0.341*****	**0.368*****	**0.151***
Rainfall June	0.004	0.157	**0.107***	-0.132	-0.117
Rainfall July	-0.006	0.098	**0.2****	-0.224	-0.11
Rainfall August	-0.005	**-0.265*****	**-0.271*****	-0.216	-0.124

Table 7. F-values, statistics and p-values from analysis of variance (ANOVA) and non-parametric Kruskal-Wallis analysis of variance of morphological characteristics for flax accessions grouped into four LIS-1 groups. Bold font indicates a significant difference.

Morphological characteristic	**ANOVA**	**Shapiro-Wilk normality test**
Plant height	F = 5.643, p-value = **0.00119**	W = 0.98964, p-value = 0.484
Number of seeds in the capsule	F = 1.549, p-value = 0.206	W = 0.9933, p-value = 0.8254
	Kruskal-Wallis	
Technical length of the stem	statistic = 12.5, p-value = **0.00581**	-
Number of seed capsules per plant	statistic = 1.41, p-value = 0.703	-

Table 8. Tukey's and Dunn's tests results and p-values. Bold font indicates a significant difference after the sequential Bonferroni correction. Numbers in brackets indicate a significant difference before the correction.

	Morphological characteristics	
	Plant height	**Technical length of the stem**
LIS-1 groups	**Tukey's test results, p-values**	**Dunn's test, p-values**
R1_R0	0.0613309	0.223 (0.0371)
R2_R0	0.1956210	1 (0.187)
NR_R0	**0.0004668**	**0.00379** (0.000632)
R2_R1	0.9618691	1 (0.403)
NR_R1	0.0710239	0.219 (0.0365)
NR_R2	**0.0420870**	**0.0580** (0.00966)

genotypes appear to have an advantage visualized by larger plant height and stem length.

Ciliation of septa

A. Durrant and O. R. Joarder (1978) described that the differences between flax genotrophs must be due to differences in gene activity and regulation induced by the environment. In particular, the capsule quantitative character H-h (hairy-hairless septa) is conditioned by multiple genes. Also, H and h are stable in homozygotes but unstable in a heterozygous state. HH and hh, and their adjacent regulators, could be controlled by genes elsewhere in the genome. The capsule character was defined as follows: HH (number of hairs more than 55), hh (hairless septa), or Hh (number of hairs from 1 to 55) genotype. Durrant and Joarder (1978) also concluded that, if more than 55 hairs are present on the seed capsule septa, then most likely, but not necessarily, this plant may have the LIS-1 insertion. The heterozygous condition Hh was shown to be unstable due to gene interaction and could be visually detected by the number of hairs on the septa, which varied from about 20 to 55 (Durrant and Joarder, 1978).

We counted the number of hairs on the capsule septa for seven flax varieties: 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), 624_6219 (R2 group),

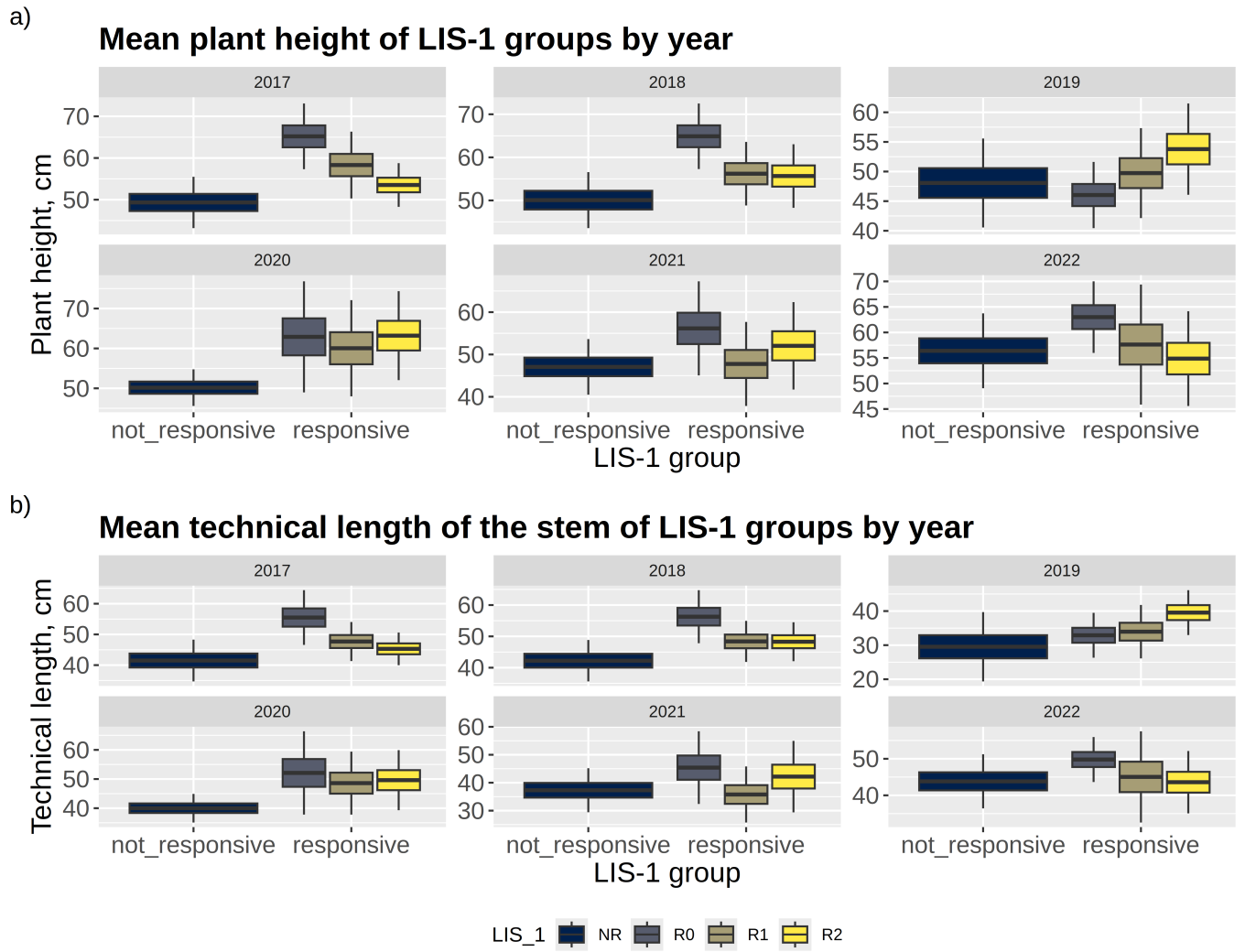

Figure 3. Comparison of quantitative morphological characteristics by four LIS-1 groups over six years where ANOVA showed significant results. a) mean height of the plant (cm); b) mean technical length of the stem (cm) for R0, R1, R2 groups ('responsive group'), and NR ('non-responsive').

and 624_791 (R0 group). The number of hairs varied from 0 to about 65 (Figure 4). According to Durrant and Joarder (1978), we assigned plants with hairless septa as recessive homozygotes (hh), plants with more than 55 hairs as dominant homozygotes (HH) and others as heterozygotes (Hh). In Figure 4, the different types of ciliation of the septa are shown.

In our study, dominant homozygotes HH were only present in varieties from the LIS-1 group R1 (624_6222, 624_1044), which have the LIS-1 insertion, whereas recessive homozygotes hh were represented in all LIS-1 groups, except varieties 624_6222 and 624_1044, that were dominant homozygotes HH or heterozygotes Hh, respectively (Figure 5).

Prediction of LIS-1 presence using the machine learning algorithm random forest classifier

Using the machine learning algorithm random forest classifier (Breiman, 2001), we tried to predict the presence, absence or heterozygosity of the LIS-1 insertion based on the characteristics of plants. This classifier also allowed us to calculate the importance of each characteristic of the object for the classification. The features used in the machine learning model were: ciliation of the seed capsule's septa (hairless hh, heterozygous Hh, hairy HH), plant height, technical length of the stem (as they are shown in literature to be associated with LIS-1), number of productive seed capsules per plant, and number of seeds in the capsule (based on correlation analysis), presence (weak or missing) of anthocyanin pigmentation in the hypocotyl.

The target variable was the presence of LIS-1 insertion ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote'). The heterozygosity of LIS-1 could be attributed to the insertion occurring in one chromosome, while being absent in the other. This condition is unstable, and due to uncertainty regarding its transmission in the next generation, we encoded it separately.

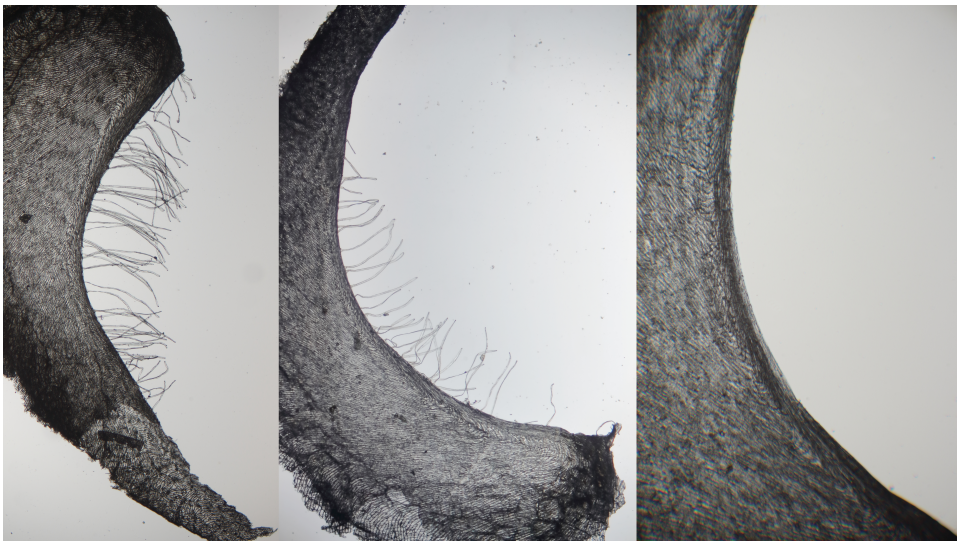

Figure 4. Ciliation of septa of flax genotypes. From left to right: dominant homozygous HH (624_6222), heterozygous Hh (624_6215) and recessive homozygous hh (624_791).

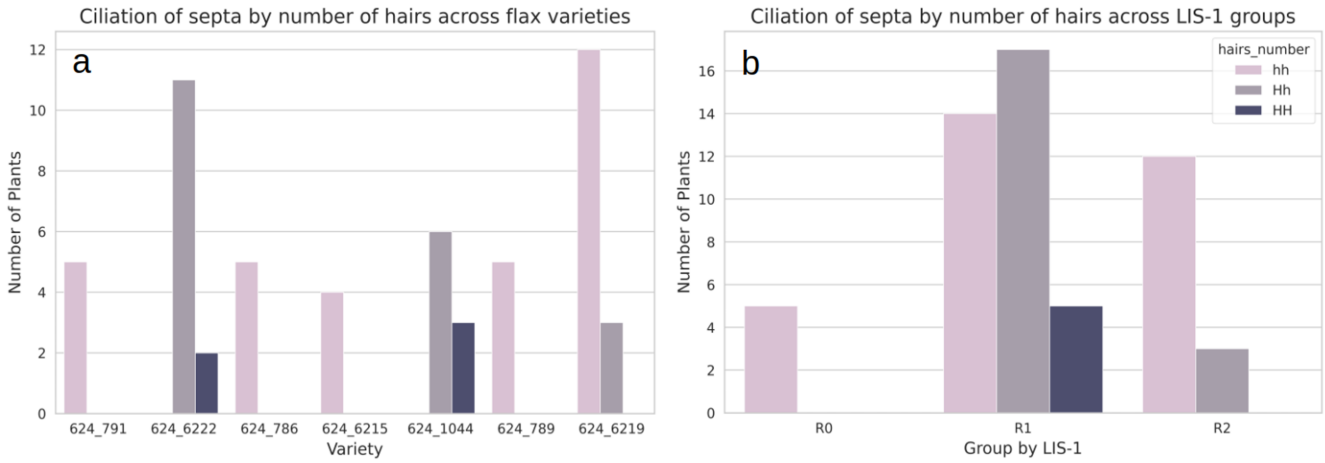

Figure 5. Distribution of hairy septa feature of plants across flax varieties (a) and LIS-1 groups (b). Accessions 624_791 (R0 group), 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), and 624_6219 (R2 group); groups by LIS-1: 0 – R0, 1 – R1, 2 – R2, N = 5, 37, 15 plants.

Overall, the classification accuracy of the predictive model was 99.14% on training data and 98.039% on test data. Of the 51 objects, 34 were correctly classified as 'LIS-1 present', 9 belonged to 'LIS-1 absent', and 7 were 'heterozygotes LIS-1'. One object was erroneously assigned to 'LIS-1 heterozygotes' (in fact, 'LIS-1 present', Figure 6). So, groups with 'LIS-1 present' and 'LIS-1 heterozygote' based on studied morphological characteristics could be misclassified.

Classification metrics that show the success of the class prediction were not equal for all studied groups of flax varieties. The average accuracy of predictions was 98%. The general classification report is given in Table 9.

The random forest classifier calculates the importance of each characteristic of the object (shown in Figure 7) for the classification task.

Five morphological features had classification importance exceeding 10%: technical stem length (19.86%), plant height (19.33%), number of capsules (18.73%), ciliation of septa hh (means hairless septa, 11.68%), and number of seeds per capsule (10.61%).

Discussion

The ancient local flax varieties investigated in this study originated from different regions of Belarus. They are adapted to certain growing conditions in which they have evolved and could therefore serve as useful sources of genetic diversity. Surveying and inventorying the pool of diversity in local varieties, including using molecular markers, is a priority to sustain future agricultural production (FAO, 2012).

Genome plasticity could be defined as a change in genome structure (mutations, genome expansion, transposable elements, etc.) associated with environmental challenges, leading to the development of new phenotypes. Meanwhile, adaptive plasticity is a phenotypic

Table 9. Classification metrics for each group of plants.

	Precision	Recall	F1-score	Number of objects
LIS-1 absent	100	100	100	9
LIS-1 present	100	97	99	35
LIS-1 heterozygotes	88	100	93	7
Accuracy			**98**	

plasticity that increases the global fitness of a genotype. Genotype fitness refers to the relative abundance and success of a species' genes over multiple generations (total biomass, seed number and growth rates of a single generation) (Nicotra *et al*, 2010).

The majority of studied flax varieties were assigned to the responsive group of genotypes, defined by presence of the LIS-1 insertion (groups R0, R1 and R2, which included 18 of the studied accessions), and only three accessions (assigned to NR group) were not responsive in terms of LIS-1.

For plant height and technical length of the stem, the groups R0 (that formed the insertion and then lost it) and R2 (that formed the insertion, and partly lost it) were statistically significant ($p < 0.05$) higher than group NR (the insertion was not found). Group R1 included the accessions that formed the LIS-1 insertion and retained it. Statistically, there was no significant difference observed between group R1 and groups R0 and R2 ($p > 0.05$) by plant height and technical length of the stem and at the same time, group R1 was not significantly higher than the NR group ($p > 0.05$). Bickel *et al* (2012) indicated that stable S- and L-genotrophs (which have retained and lost the insertion sequence) were well adapted to environmental stress (lack and excess nitrogen or water in the soil, respectively) compared to the PL line, which in normal conditions does not form the insertion, but under stressful conditions could produce

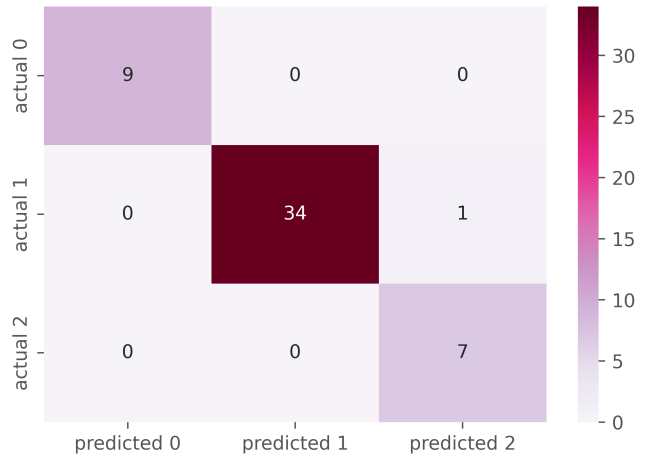

Figure 6. Confusion matrix showing classification accuracy of prediction of LIS-1 presence by machine learning. Of the 51 objects, 50 are correctly classified. One object was misclassified. ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote').

two types of stable S- and L-genotrophs, as well as retain the ability to be genetically plastic. Flax varieties that stably inherit the LIS-1 insertion (S-genotrophs) are characterized by a shorter plant height than L-genotrophs and the PL-line (Bickel *et al*, 2012). This could be attributed to the LIS-1 insertion being one of multiple genomic rearrangements occupying a region with two genes involved in growth processes, inhibitor of growth-1, and kip-related cyclin-dependent kinase inhibitor-2 (Bickel *et al*, 2012). Thus, it could affect both the plant height and technical stem length. The weather conditions in the years 2017 to 2022 differed in terms of rainfall and temperatures during the flax vegetation period. The lack of a statistically significant interaction effect between genotype (R1 group) and June temperatures on plant height ($p = 0.108$) indicates that these genotypes may be phenotypically stable, and would not modify their height in response to temperatures in the stage of active growth. An effect on the border of significance was revealed for genotype (NR group) × Rainfall_May interaction for the reproduction-related trait 'number of seeds in the capsule' ($p = 0.056$). This indicated that while this group of accessions did not exhibit genetic responsiveness, they displayed phenotypic plasticity. Significant negative correlations between growth-related traits (plant height and technical stem length) and LIS-1 groups showed that the absence of the insertion (NR group) is associated with shorter plant height and stem length.

For reproduction-related traits ('number of seed capsules per plant' and 'number of seeds in the capsule'), responsive genotypes (R0, R1 and R2) did not significantly differ from non-responsive genotypes (NR, $p > 0.05$).

Flax is mostly inbred and, therefore, under selection by the environment for particular combinations of alleles, it could become homozygous at all loci with little variation (Cullis, 2019), making it vulnerable to any environmental changes. Therefore, the ability to modify the genome in response to growth challenges could be an evolutionary advantage. The uniqueness of the genome plasticity mechanism in fibre flax is that modification occurs not in a single gene but in different genome regions, resulting in potential phenotypic and biochemical variability (Cullis, 2019). Thus, the LIS-1 sequence is a promising molecular marker for identifying flax forms with genome plasticity.

A complex of phenotypic changes in the stable L- and S-genotrophs are described in the literature (Bickel *et al*, 2012; Cullis, 2019). Among them are the height of the plant, the hairy septa, and the number of

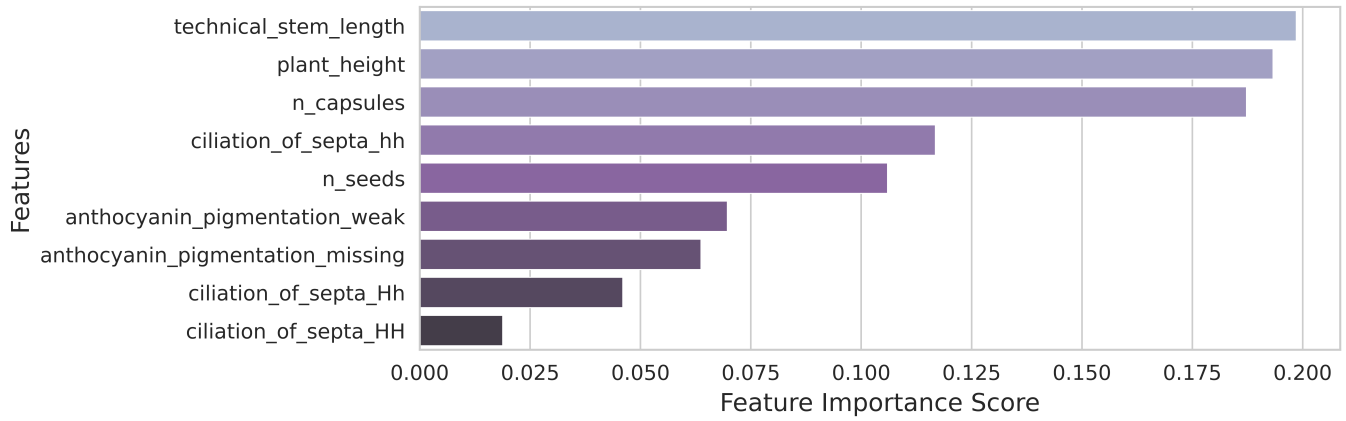

Figure 7. Feature importance score calculated by random forest classifier. The features were: technical length of the stem, plant height, number of seed capsules per plant (n_capsules), ciliation of the seed capsule's septa (hairless hh, heterozygous Hh, hairy HH), and number of seeds in the capsule (n_seeds), anthocyanin pigmentation in the hypocotyl (weak or missing).

capsules, which were the most important for our random forest classification model. The overall accuracy of LIS-1 status (presence, absence or heterozygous condition) prediction based on nine studied morphological features was 98.039%.

From the evolutionary point of view, both LIS-1 responsive and non-responsive flax varieties are successful in transmitting their genes over generations as they do not differ by the number of seed capsules and seeds. The study of plant genetic resources using LIS-1 as a molecular marker of genome plasticity will provide us with knowledge about flax genotypes that are potentially valuable for fibre flax breeding and biodiversity conservation.

Conclusion

Among the 21 local varieties studied, four groups were identified based on their ability to modify their genome using LIS-1 as a molecular marker of genome plasticity. The most promising in terms of sources for the selection of fibre flax varieties adaptive to environmental challenges is the group of responsive varieties that have formed LIS-1 insertion (R0, R1 and R2 groups). Existing associations between patterns of LIS-1 sequence presence and morphological traits of flax allowed us to classify them correctly with 98% accuracy.

Supplemental data

Supplemental Table 1. Characteristics of flax varieties for machine learning modelling

Acknowledgements

We thank Viachaslau N. Kipen for providing photos of the hairy and hairless septa of flax capsules.

The work was carried out within the framework of the activity 'Genetically identify collection samples of agricultural crops to form a new gene pool of donors of economically valuable traits for use in breeding', subprogramme 'Study, identification and rational use of collections of plant genetic resources' of the state programme 'Scientific and innovative activities of the National Academy of Sciences of Belarus' for 2021–2025.

Conflict of interest statement

The authors have no conflicts of interest to report.

Author contributions

MP contributed to the study conception, writing of the draft and final manuscript, statistical analysis, modelling, visualization and interpretation of the results, laboratory experiments conduction, and final manuscript revision. VL contributed to the study conception and design, interpretation of the results, final manuscript revision, resource provision and supervision of the research on field and laboratory experiments. EL contributed to the study conception and design, laboratory experiments conduction, writing of the final manuscript and interpretation of the results. VS contributed to the study conception and design, field experiment design and conduction, made field measurements, and worked on the draft. AB contributed to field experiment design and conduction, and worked on the draft, study conception and design. EG contributed to the laboratory experiments design and conduction, worked on the draft, and to the study conception and design. LK contributed to the study conception and design, and the final manuscript revision. All authors discussed the results and commented on the manuscript, and have read and agreed to the published version of the manuscript.

References

Bajorath, J. (2022). Revisiting active learning in drug discovery through open science. *Artificial Intelligence in the Life Sciences* 2, 100051–100051. doi: https://doi.org/10.2144/fsoa-2022-0010

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. *Jour-*

nal of Statistical Software (1), 67–67. doi: https://doi.org/10.18637/jss.v067.i01

Bickel, C., Lukacs, and Cullis, C. (2012). The loci controlling plasticity in flax. *Research and Reports in Biology* 3, 1–11. doi: https://doi.org/10.2147/RRB.S27198

Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. doi: https://doi.org/10.1023/A:1010933404324

Chen, Y., Lowenfeld, R., and Cullis, C. A. (2009). An environmentally induced adaptive (?) insertion event in flax. *Journal of Genetics and Molecular Biology* 1, 38–047.

Chen, Y., Schneeberger, R. G., and Cullis, C. A. (2005). A site-specific insertion sequence in flax genotrophs induced by environment. *The New Phytologist* 167, 171–80. doi: https://doi.org/10.1111/j.1469-8137.2005.01398.x

Cullis, C. (1976). Environmentally induced changes in ribosomal RNA cistron number in flax. *Heredity* 36, 73–79. doi: https://doi.org/10.1038/hdy.1976.8

Cullis, C. A. (1981). DNA sequence organization in the flax genome. *Biochim Biophys Acta* 652(1), 1–15. doi: https://doi.org/10.1016/0005-2787(81)90203-3

Cullis, C. A. (1986). Phenotypic consequences of environmentally induced changes in plant DNA. *Trends in Genetics* 2(86), 90285–90289. doi: https://doi.org/10.1016/0168-9525(86)90285-4

Cullis, C. A. (2019). Origin and Induction of the Flax Genotrophs. *Genetics and Genomics of Linum* 227-234. doi: https://doi.org/10.1007/978-3-030-23964-0_14

Diederichsen, A. (2019). A Taxonomic View on Genetic Resources in the Genus *Linum* L. for Flax Breeding. *Genetics and Genomics of Linum* 227-234. doi: https://doi.org/10.1007/978-3-030-23964-0_1

Durrant, A. and Joarder, O. I. (1978). Regulation of hairless septa in flax genotrophs. *Genetica* 48, 171–183. doi: https://doi.org/10.1007/BF00155567

Durrant, A. and Jones, T. (1971). Reversion of induced changes in amount of nuclear DNA in *Linum*. *Heredity* 27, 431–439. doi: https://doi.org/10.1038/hdy.1971.106

Durrant, A. and Nicholas, D. (1970). An unstable gene in flax. *Heredity* 25, 513–527. doi: https://doi.org/10.1038/hdy.1970.60

Ehrensing, D. (2008). Oilseed Crops: Flax (EM 8952-E) (Oregon State University Extension Service).

Evans, G., Durrant, A., and Rees, H. (1966). Associated Nuclear Changes in the Induction of Flax Genotrophs. *Nature* 212, 697–699. doi: https://doi.org/10.1038/212697a0

FAO (2012). Synthetic account of the Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture. url: https://www.fao.org/3/i2650e/i2650e.pdf.

Goldsbrough, P. B., Ellis, T. H., and Cullis, C. A. (1981). Organisation of the 5S RNA genes in flax. *Nucleic Acids Res* 9(22), 5895–904. doi: https://doi.org/10.1093/nar/9.22.5895

Harris, C. R., Millman, K. J., and Van Der Walt, S. J. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: https://doi.org/10.1038/s41586-020-2649-2

Hu, Z. and Xing, E. P. (2021). Toward a 'Standard Model' of Machine Learning. *Harvard Data Science Review* . doi: https://doi.org/10.1162/99608f92.1d34757b

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9(3), 90–95. doi: https://doi.org/10.1109/MCSE.2007.55

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. url: https://rpkgs.datanovia.com/rstatix/.

Kluyver, T. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas,* ed. Loizides, F. and Schmidt, B. 87-90.

Maggioni, L., Pavelek, M., Van Soest, L. J. M., and Lipman, E. (2001). Flax Genetic Resources in Europe Ad hoc meeting (Rome, Italy: International Plant Genetic Resources Institute), 72-73. url: https://www.ecpgr.cgiar.org/fileadmin/bioversity/publications/pdfs/Ad_Hoc_Fibre_Crops_WG_ad_hoc_meeting_Flax_genetic_resources_in_Europe_Czech_Rep_2001.pdf.

Nicotra, A. B., Atkin, O. K., Bonser, S. P., Davidson, A. M., Finnegan, E. J., Mathesius, U., Poot, P., Purugganan, M. D., Richards, C. L., Valladares, F., and Van Kleunen, M. (2010). Plant phenotypic plasticity in a changing climate. *Trends in plant science* 15(12), 684–692. doi: https://doi.org/10.1016/j.tplants.2010.09.008

Nôžková, J., Pavelek, M., Bjelková, M., Brutch, N., Tejklová, E., Porokhovinova, E., and Brindza, J. (2016). Descriptor list for flax (*Linum usitatissimum* L.) . doi: https://doi.org/10.15414/2016.9788055214849

Oliveros, J. C. (2007-2015). Venny. An interactive tool for comparing lists with Venn's diagrams. url: https://bioinfogp.cnb.csic.es/tools/venny/index.html.

Privalov, F. I., Grib, S. I., and Matys, I. S. (2021). National seed bank of genetic economically useful plant resources is a scientific object of a National property of the Republic of Belarus. *Crop Farming and Plant Growing* 2, 10–14.

Rachinskaya, O. A., Lemesh, V. A., Muravenko, O. V., Yurkevich, O., Yu, Guzenko, E. V., Bol'sheva, N. L., Bogdanova, M. V., Samatadze, T. E., Popov, K. V., Malyshev, S. V., Shostak, N. G., Heller, K., Hotyleva, L. V., and Zelenin, A. V. (2011). Genetic polymorphism of flax *Linum usitatissimum* based on the use of molecular cytogenetic markers. *Russ J Genet* 47, 56–65. doi: https://link.springer.com/article/10.1134/S1022795411010108

Raghunathan, S. and Priyakumar, U. D. (2022). Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chem-*

istry 122(7), 26870–26870. doi: https://doi.org/10.1002/qua.26870

Sa, R., Yi, L., Siqin, B., An, M., Bao, H., Song, X., Wang, S., Li, Z., Zhang, Z., Hazaisi, H., Guo, J., Su, S., Li, J., Zhao, X., and Lu, Z. (2021). Chromosome-Level Genome Assembly and Annotation of the Fiber Flax (*Linum usitatissimum*) Genome. *Front Genet* 12, 735690–735690. doi: https://doi.org/10.3389/fgene.2021.735690

Sambrook, J. and Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol:chloroform. *CSH protocols* . doi: https://doi.org/10.1101/pdb.prot4455

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference.

The pandas development team (2020). pandas-dev/pandas: Pandas 1.0.0 (v1.0.0). Zenodo. url: https://doi.org/10.5281/zenodo.3630805.

Vavilov, N. I. (1926). Studies on the origin of cultivated plants. *Bull Appl Botany* 16(2), 3–248.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., Vanderplas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and Van Mulbregt (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17(3), 261–272. doi: https://doi.org/10.1038/s41592-019-0686-2

Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine Learning for Small Molecule Drug Discovery in Academia and Industry. *Artificial Intelligence in the Life Sciences* . doi: https://doi.org/10.1016/j.ailsci.2022.100056

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software* 6(60), 3021–3021. doi: https://doi.org/10.21105/joss.03021

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis (Springer International Publishing), 2nd edition. url: https://ggplot2.tidyverse.org.

Wickham, H., Franc üller, K. (2022). dplyr: A Grammar of Data Manipulation. url: https://dplyr.tidyverse.org.

Yang, Z., Tian, Y., Kong, Y., Zhu, Y., and Yan, A. (2022). Classification of JAK1 Inhibitors and SAR Research by Machine Learning Methods. *Artificial Intelligence in the Life Sciences* 2, 100039–100039. doi: https://doi.org/10.1016/j.ailsci.2022.100039

Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 61–71
DOI: 10.46265/genresj.RYAJ6068
https://www.genresj.org
ISSN: 2708-3764

Combined cytogenetic and molecular methods for taxonomic verification and description of *Brassica* populations deriving from different origins

Cyril Falentin[*,a,†], Houria Hadj-Arab[b,†], Fella Aissiou[b], Claudia Bartoli[a], Giuseppe Bazan[c], Matéo Boudet[a], Lydia Bousset-Vaslin[a], Marwa Chouikhi[d], Olivier Coriton[a], Gwenaëlle Deniot[a], Julie Ferreira De Carvalho[a], Laurène Gay[e], Anna Geraci[c], Pascal Glory[a], Virginie Huteau[a], Riadh Ilahy[d], Vincenzo Ilardi[c], José A Jarillo[f], Vladimir Meglič[g], Elisabetta Oddo[c], Mónica Pernas[f], Manuel Piñeiro[f], Barbara Pipan[g], Thouraya Rhim[d], Vincent Richer[a], Fulvia Rizza[h], Joëlle Ronfort[e], Mathieu Rousseau-Gueutin[a], Rosario Schicchi[i], Lovro Sinkovič[g], Maryse Taburel[a], Valeria Terzi[h], Sylvain Thérèné[a], Mathieu Tiret[a], Imen Tlili[d], Marie-Hélène Wagner[j], Franz Werner Badeck[h] and Anne-Marie Chèvre[a]

[a] *IGEPP, INRAE, Institut Agro, Université de Rennes, 35650, Le Rheu, France*
[b] *Faculty of Biological Sciences FSB, University of Sciences and Technology Houari Boumediene USTHB, BP 32, 16111, Bab-Ezzouar, El-Alia, Algiers, Algeria*
[c] *Department of Biological, Chemical and Pharmaceutical Sciences and Technologies (STEBICEF), Università degli Studi di Palermo, via Archirafi 38, 90123, Palermo, Italy*
[d] *Laboratory of Horticulture, National Agricultural Research Institute of Tunisia (INRAT), University of Carthage, Menzah 1, 1004, Tunis, Tunisia*
[e] *UMR AGAP Institut, Université de Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France*
[f] *Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/CSIC), Campus Montegancedo UPM, Madrid, Spain*
[g] *Crop Science Department, Agricultural Institute of Slovenia, Hacquetova ulica 17, SI-1000, Ljubljana, Slovenia*
[h] *Research centre for Genomics & Bioinformatics, Council for Agricultural Research and Economics (CREA), I -29017, Fiorenzuola d'Arda (PC), Italy*
[i] *Department of Agricultural, Food and Forest Sciences (SAAF), Università degli Studi di Palermo, viale delle Scienze ed. 4, 90128, Palermo, Italy*
[j] *GEVES, Station Nationale d'Essais de Semences, 49071, Beaucouzé, France*

Abstract: Agriculture faces great challenges to overcome global warming and improve system sustainability, requiring access to novel genetic diversity. So far, wild populations and local landraces remain poorly explored. This is notably the case for the two diploid species, *Brassica oleracea* L. (CC, 2n=2x=18) and *B. rapa* L. (AA, 2n=2x=20). In order to explore the genetic diversity in both species, we have collected populations in their centre of origin, the Mediterranean basin, on a large contrasting climatic and soil gradient from northern Europe to southern sub-Saharan regions. In these areas, we also collected 14 populations belonging to five *B. oleracea* closely related species. Our objective was to ensure the absence of species misidentification at the seedling stage among the populations collected and to describe thereafter their origins. We combined flow cytometry, sequencing of a species-specific chloroplast genomic region, as well as cytogenetic analyses in case of unexpected results for taxonomic verification. Out of the 112 *B. oleracea* and 154 *B. rapa* populations collected, 103 and 146, respectively, presented a good germination rate and eighteen populations were misidentified. The most frequent mistake was the confusion of these diploid species with *B. napus*. Additionally for *B. rapa*, two autotetraploid populations were observed. Habitats of the collected and confirmed wild populations and landraces are described in this study. The unique plant material described here will serve to investigate the genomic regions involved in adaptation to climate and microbiota within the framework of the H2020 Prima project 'BrasExplor'.

Citation: Falentin, C., Hadj-Arab, H., Aissiou, F., Bartoli, C., Bazan, G., Boudet, M., Bousset-Vaslin, L., Chouikhi, M., Coriton, O., Deniot, G., Ferreira De Carvalho, J., Gay, L., Geraci, A., Glory, P., Huteau, V., Ilahy, R., Ilardi, V., Jarillo, J. A., Meglič, V., Oddo, E., Pernas, M., Piñeiro, M., Pipan, B., Rhim, T., Richer, V., Rizza, F., Ronfort, J., Rousseau-Gueutin, M., Schicchi, R., Sinkovič, L., Taburel, M., Terzi, V., Thérèné, S., Tiret, M., Tlili, I., Wagner, M., Badeck, F. W., Chèvre, A. (2024). Combined cytogenetic and molecular methods for taxonomic verification and description of *Brassica* populations deriving from different origins. *Genetic Resources* 5 (9), 61–71. doi: 10.46265/genresj.RYAJ6068.

Received: 14.09.2023 | Accepted: 12.04.2024 | Published online: 10.05.2024

Introduction

Agriculture has to face great challenges to overcome global climate change and improve the sustainability of agricultural systems while maintaining crop production and quality. Regarding crop improvement, there are at least two main questions to consider: (i) which type of genetic diversity should we promote in breeding programmes to withstand the new climatic regime and (ii) which material to select for the development of new relevant varieties in this erratic context. Intensive farming systems and particularly modern breeding techniques have led to a drastic reduction in crop genetic diversity. On the other hand, local landraces and wild plant populations are a great source of genetic diversity. However, for many crop species such plant material has either never been collected, is not available, or has been poorly analyzed and/or characterized.

The two diploid species that we focused on in this study, *Brassica oleracea* L. (CC, 2n=2x=18) and *B. rapa* L. (AA, 2n=2x=20), are native to the Mediterranean basin (Cheng *et al*, 2016; Bird *et al*, 2017; Qi *et al*, 2017; Cai *et al*, 2021; Mabry *et al*, 2021; McAlvay *et al*, 2021; Cai *et al*, 2022), in which they grow as wild populations or as local landraces selected over several generations by farmers. They encounter a large gradient of contrasted climate, soils and biotic factors from northern Europe to southern sub-Saharan regions, which makes these species particularly relevant for the analysis of diversity in relation to adaptation to the climate. The new populations will be complementary to accessions available in Biological Resource Centres (BRC) as they continue to evolve under current climatic constraints. Indeed, exploring these wild populations and local varieties represents a unique opportunity to identify locally adapted material for which genetic diversity and adaptive traits could be relevant to face upcoming climatic changes and disease emergences correlated to global change in the Mediterranean area, thus contributing to biodiversity-based agriculture.

Convergent evolution has led to similar morphotypes in these two economically important vegetable species that were locally selected for a long time by farmers all over the Mediterranean basin, mainly for their inflorescence at budding stage (cauliflower or broccoli for *B. oleracea*, broccoletto for *B. rapa*), leaves (cabbage, kale for *B. oleracea*; fodder turnip for *B. rapa*) or epicotyls/roots (kohlrabi for *B. oleracea*, turnip for *B. rapa*). This morphological convergence between the two species is linked to their recent common ancestor (Cheng *et al*, 2016) as they diverged only 2–4 million years ago (Cheng *et al*, 2014). The morphological similarity between them is one of the reasons for some confusion when identifying the species. Additionally, a third species widely cultivated for seeds, resulting from the hybridization and genome doubling of the two diploid species, *B. napus* L. (AACC, 2n=4x=38), can also produce edible roots in swede cultivars, or leaves as forage or vegetable. As both species share many morphological characteristics with *B. napus*, species identification remains difficult at the seedling stage and controls are required before further analyses.

In this paper, we describe the collection, along a broad climatic gradient, of more than 100 populations each of *B. oleracea* and *B. rapa* species, including both landraces and wild populations, which co-evolve under current climatic constraints. To ensure the absence of species misidentification or potential interspecific hybrids at the seedling stage before sequencing, plants of each population were assessed using different methods sequentially from the easiest to the most time-consuming: (1) flow cytometry on all the plants based on different genome size and chromosome number (630Mb for 18 chromosomes in *B. oleracea*, 529Mb for 20 chromosomes in *B. rapa*) (Belser *et al*, 2018), (2) Sanger sequencing of a species-specific chloroplast genomic region on a sub-sample per population (Li *et al*, 2017), and (3) cytogenetic approaches in the event of unexpected results from the previous analyses. After these controls, the geographical distribution and ecological environment of each population were described. This unique plant material will support further analyses from our consortium investigating the genomic regions involved in local adaptation to climate and microbiota.

Materials and methods

Plant material

Wild populations of both *B. oleracea* and *B. rapa* species were collected in France based on information in the National Inventory of Natural Heritage database INPN (2024) and Maggioni *et al* (2020). In addition, *B. rapa* wild populations were gathered in Italy, Algeria, Slovenia and *B. oleracea* in Spain (Gomez-Campo *et al*, 2005) based on local flora and long field experiences. Siliques were collected from 30 plants per population (when available), depending on the size and accessibility of populations. Some wild populations of *B. oleracea* closely related species were identified and added to the analysis: eight *B. montana* Pourr. populations (six from France and two from Italy), as well as two *B. rupestris* Raf. (subsp. *rupestris*), two *B. villosa* Biv. [subsp. *drepanensis* (Caruel) Raimondo & Mazzola and subsp. *tineoi* (Lojac.) Raimondo & Mazzola], one *B. macrocarpa* Guss., and one *B. incana* Ten., all from Sicily, Italy. *B. oleracea* and *B. rapa* landraces were collected in five different countries either through direct collects on farms in Algeria, Tunisia and Italy or in BRC maintaining old landraces in France (BRC BrACySol) and Slovenia (Slovene Plant Gene Bank in Slovenia, SRGB KIS). In agreement with each country's policy, the Nagoya Protocol will be applied, pending the introduction of the relevant collected material into the Multilateral System of the FAO's International Treaty

*Corresponding author: Cyril Falentin (cyril.falentin@inrae.fr.)
†These authors contributed equally to this work

on Plant Genetic Resources for Food and Agriculture. Thus, during this transition period, the material will be available after seed production upon request, either in BRC BrACySol and SRGB for French and Slovenian populations, respectively, or by contacting the partner in each country, as reported in Supplemental Tables 1 and 2.

Each collected population was named following a specific code. It starts with (1) two letters representing the species (BO for *B. oleracea*, BR for *B. rapa*, BM for *B. montana*, BU for *B. rupestris*, BV for *B. villosa*, BA for *B. macrocarpa*, and BI for *B. incana*), followed (2) by a letter for the country of origin (F for France, I for Italy, S for Slovenia, E for Spain, A for Algeria, or T for Tunisia), (3) then four letters indicating the location of the collecting site, (4) either a W for a wild population or an L for a landrace, (5) and an additional letter (A, B, C, etc.) in case of several collecting sites at the same location (i.e. BR_I_CAST_W_A and BR_I_CAST_W_B). For all these populations, a common sheet was filled for wild populations to describe the environment (Supplemental Table 1) and another one for landrace collects at the farm or when seeds were acquired from Genetic Resource Centres (GRC BrACySol in France, KIS in Slovenia) (Supplemental Table 2).

Thirty plants per population were grown in the greenhouse for taxonomy assessments. For wild populations, we planted one seed of each of the 30 collected mother plants. When seeds were collected from fewer than 30 plants, we sowed several seeds per mother plant, equally represented, to reach a total of 30 seeds. For landraces, 30 seeds were sown.

As controls for the different experiments, we used a known representative of *B. oleracea, B. rapa* and *B. napus* species: doubled haploid lines of *B. oleracea* subsp. *italica* (HDEM) and *B. rapa* subsp. *trilocularis* (Z1) (Belser *et al*, 2018) and a pure line of *B. napus* subsp. *oleifera*, 'Darmor'.

Cytogenetic control and chromosome counts

Flow cytometry was performed on all plants to assess the chromosome number of each plant using leaves as described by Leflon *et al* (2006). Briefly, approximately 0.5cm^2 of fresh leaves were harvested and transferred to a Petri dish. This material was chopped using a sharp razor blade in 300μl of nuclei extraction staining buffer (from kit Cystain^{TM} UV Presice P-Sysmex) and incubated at room temperature for 30 to 90sec. 1.2ml of DAPI staining buffer was added per sample and the solution was then filtered through a 50μm nylon mesh. Estimation for each accession was obtained with FlowMax software using a CyFlow space cytometer (Sysmex Inc.). For the screening of *B. oleracea* and independently of *B. rapa* populations, the control variety, HDEM for *B. oleracea* and Z1 for *B. rapa*, was adjusted to a fluorescence intensity value of 300 for nuclei at G1 stage. Coincidence or deviation was compared with these controls.

For populations for which flow cytometer and chloroplast sequencing data were not congruent, the chromosome number was also determined from mitotic chromosomes observed on metaphasic cells isolated from root tips. Root tips of 0.5–1.5cm in length were treated in the dark with 0.04% 8-hydroxiquinoline for 2h at 4°C followed by 2h at room temperature to accumulate metaphases. They were then fixed in 3:1 ethanol:glacial acetic acid for 48h at 4°C and stored in 70% ethanol at -20°C until use. After being washed in distilled water for 10min, in HCl 0.25 N for 10min, then treated for 15min with a 0.01M citric acid-sodium citrate buffer (pH 4.5), root tips were incubated at 37°C for 30min in an enzymatic mixture (5% Onozuka R-10 cellulase (Sigma), 1% Y23 pectolyase (Sigma)). The enzymatic solution was removed and the digested root tips were then carefully washed with distilled water for 30min. One root tip was transferred to a slide and macerated with a drop of 3:1 fixation solution. Dried slides were then stained by a drop of 4',6-diamidino-2-phenylindole (DAPI). Cells were viewed with an ORCA-Flash4 (Hamamatsu, Japan) on Axio Imager Z.2 (Zeiss, Oberkochen, Germany) and analyzed using Zen software (Carl Zeiss, Germany).

Fluorescence *in situ* hybridization (FISH)

The BoB014O06 BAC clone from *B. oleracea* BAC library (Howell *et al*, 2008) was used as probe for 'genomic *in situ* hybridization (GISH)-like' to distinguish specifically all C-genome chromosomes in *B. napus* (Suay *et al*, 2014). The BoB014O06 clone was labelled by random priming with Alexa-594 dUTP (red) (Thermo Fisher Scientific). The ribosomal probe 45S rDNA used in this study was pTa71 (Gerlach and Bedbrook, 1979) which contained a 9-kb EcoRI fragment of rDNA repeat unit (18S-5.8S-26S genes and spacers) isolated from *Triticum aestivum* L. pTa71 was labelled by random priming with biotin-14-dUTP (Invitrogen, Life Technologies). Biotinylated probes were immunodetected by Fluorescein avidin DN (green) (Vector Laboratories). The chromosomes were mounted and counterstained in Vectashield (Vector Laboratories) containing 2.5μg/mL 4',6-diamidino-2-phenylindole (DAPI) (grey). Fluorescence images were captured using an ORCA-Flash4 (Hamamatsu, Japon) on an Axio Imager Z.2 (Zeiss, Oberkochen, Germany) and analyzed using Zen software (Carl Zeiss, Germany).

Species identification by sequencing of a chloroplast region

The aim was to amplify a chloroplast genomic region containing diagnostic single nucleotide polymorphisms (SNPs) or indels for *B. oleracea*, *B. rapa* or *B. napus*. To that purpose, we first retrieved and aligned the *Brassica* chloroplast genome sequences available for the three species from Li *et al* (2017) using Geneious Prime 2022.2.2 (https://www.geneious.com). We then identified a genomic region and designed consensus primers enabling us to discriminate each species. The

consensus primers allowed amplification o f 1,118pb for *B. oleracea*, 1,088pb for *B. rapa* or 1,084pb for *B. napus*. DNA of one to three plants per population and of control lines was extracted using 50mg of fresh leaf tissue, which had previously been freeze-dried, and the Nucleospin Plant II kit (Macherey Nagel). The consensus primers used were trnK-rps16_F (5' CATAAACAGGTAGACTGCTAACTGG 3') and trnK-rps16_R (5' GTATTCTTCCTAAAGGTATGAAAACTAAC 3') with following PCR reagents: 1X buffer, MgCl2 2mM, dNTPs 0.25mM, Primers 0.5μM each, Taq Promega 1.5U and 5ng DNA of the sample analyzed. The PCR conditions were a denaturation 94°C 2min, then 35 cycles 94°C 30sec - 59°C 30sec - 72°C 1min 30sec, with a final elongation 72°C 10min. The amplified region was then sequenced by Sanger (Genoscreen) and analyzed using Geneious software (https://www.geneious.com). All amplified chloroplast sequence data have been deposited into NCBI/GenBank as PopSet 2716368500.: PP619885 - PP620127).

Results

Taxonomic verification of the collected populations

Among the collected populations (Table 1), the first limiting factor encountered was the germination of the collected seeds, even under favourable controlled conditions applied on automated germination tools for *B. rapa*, in spite of seed viability confirmed by tetrazolium staining. Specifically, 6.8% of the collected populations showed a poor emergence in the greenhouse with less than 30 plants per population and were not considered for further analyses. This low germination rate may be attributed to two different factors: the high level of seed dormancy (observed here in 18.2% of *B. rapa* wild populations) and the seed conservation of landraces collected on farms (5.2% and 14.9% of seeds showed very poor germination for *B. rapa* and *B. oleracea* landraces, respectively).

To validate the correct species identification of each collected population and to verify the absence of contamination in the collected seeds, we performed flow cytometry on all the plants grown representing a population. As the investigated species have different profiles linked to their differences in DNA content (630Mb for 18 chromosomes in *B. oleracea*, 529Mb for 20 chromosomes in *B. rapa*) (Figure 1A), it was possible to determine with +/-2 chromosomes the genomic structure of each plant.

Due to possible contamination with species having a close chromosome number, this analysis was complemented by sequencing a chloroplast genomic region that showed species-specific differences. We chose a genomic region with a sequence specific to each species according to Li *et al* (2017). The size of the amplified regions was 1,118bp, 1,088bp and 1,084bp for *B. oleracea*, *B. rapa* and *B. napus*, respectively (all these chloroplast sequences are available on NCBI/GenBank as PopSet 2716368500.: PP619885 - PP620127). After aligning the sequences, we compared the sequences obtained in the sampled populations with those of the controls for the three species. We observed four SNPs and six Indels specific to *B. oleracea*, four SNPs and four Indels specific to *B. rapa* and three SNPs and five Indels specific to *B. napus* (examples provided in Figure 2A). *B. montana* (2n=18) differed from *B. oleracea* at only three SNPs and one Indel whereas *B. villosa* and *B. macrocarpa* differed from *B. oleracea* at 17 SNPs and six Indels. *B. rupestris* showed exactly the same sequence as the two latter species except for one SNP, indicating that these three species (*B. villosa*, *B. macrocarpa* and *B. rupestris*) are highly related to each other whereas *B. montana* seems closer to *B. oleracea* (Figure 2B).

When flow cytometer and sequencing data were not congruent, chromosome counting was performed during mitosis to identify the species. This observation was combined with GISH-like allowing identification of the C chromosomes and of rDNA locus number, specific to each species with four, ten and 12 rDNA loci for *B. oleracea*, *B. rapa* and *B. napus*, respectively (Figure 3).

By FISH, in *B. rapa* (A genome) (Figure 3B), the 45S rDNA probe (green) marks five different chromosomes. The strongest FISH signal located on the A03 chromosomes reflects a large number of genes. The second gene-rich locus is located on A01 chromosome proximal to the centromere. The remaining sites are located on cytogenetically undistinguishable A05, A06 and A09 chromosomes. *B. oleracea* (C genome)(Figure 3D) had two pairs of chromosomes (C07, C08) containing 45S rDNA loci. The sites localized on chromosome C08 show extensive decondensation while loci on C07 are fully condensed. In natural *B. napus*, we observed twelve 45S rDNA signals and the BOB014O06 staining revealed that eight signals were located on A genome and four on the C genome (Książczyk *et al*, 2011) (Figure 3C).

All the misidentified populations were listed in Supplemental Table 3. The most frequent mistake was confusing *B. oleracea* or *B. rapa* with *B. napus*. Among the 103 *B. oleracea* populations analyzed, only three were misidentified (one wild and two landraces) and were thereafter confirmed to belong to *B. napus* using flow cytometry (Figure 1D). This misidentification was also validated by chloroplast sequencing (Figure 2A) and chromosome counting. Among the 146 analyzed *B. rapa* populations, 15 were misidentified, out of which 12 were identified as *B. napus*. Nine of these 12 populations were sampled in the wild and are probably volunteers of *B. napus*, i.e. escaped from the fields. All these data were confirmed by the sequencing of a chloroplast genomic region (Figure 2A) revealing that all carried *B. napus* chloroplasts except for one wild Tunisian population (BR_T_ARIA_W_A), which had a *B. rapa* type chloroplast. The *B. napus* origin of this population was confirmed by cytogenetic analyses, revealing the presence of nine C chromosomes and 12 45S rDNA signals by FISH, eight on A genome and four on C genome

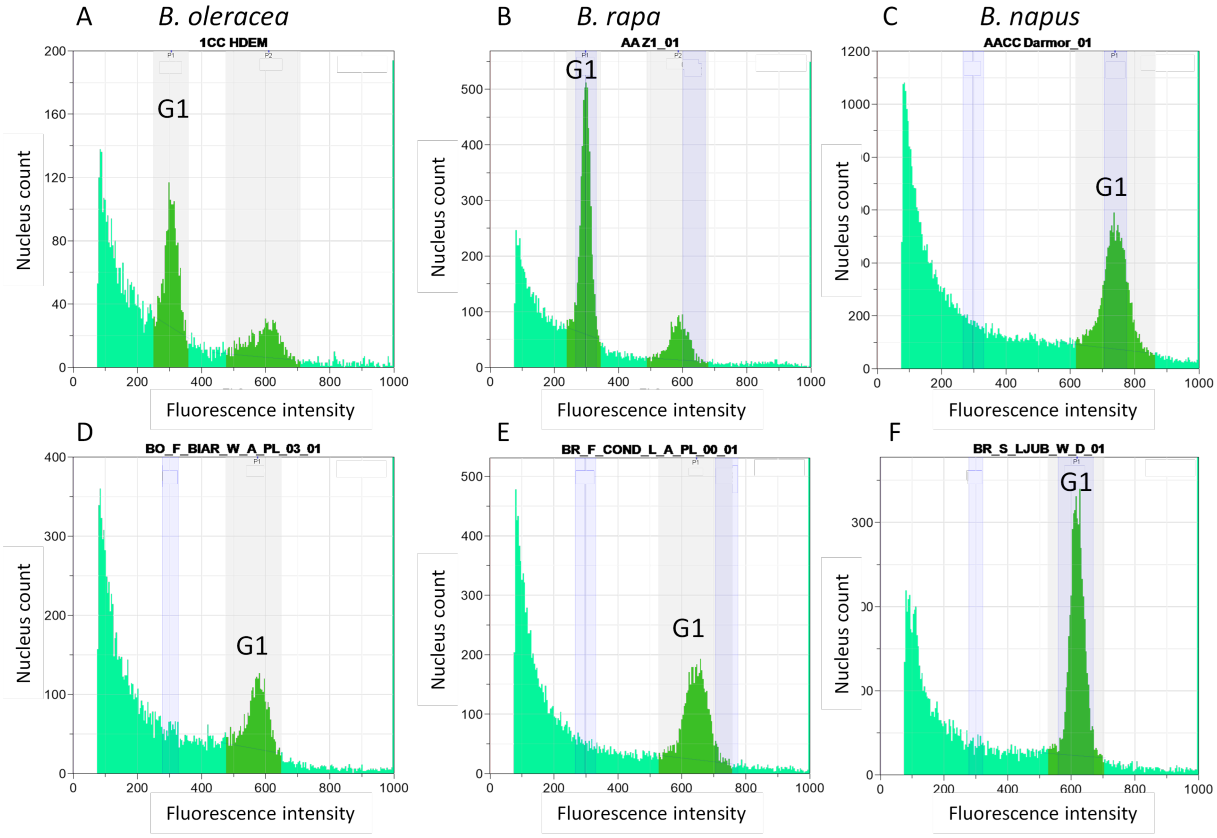

Figure 1. Flow cytometry profiles of *Brassica* controls and selected populations harbouring an unexpected profile: A) *Brassica oleracea*, B) *B. rapa*, C) *B. napus*. For the screening of *B. oleracea* and independently of *B. rapa* populations, the control variety was adjusted to 300 for fluorescence value of nuclei at G1 stage. Coincidence or deviation was compared with these controls. Three examples of populations misidentified (D, E, F) are presented with a fluorescence intensity of G1 nuclei close to the one of *B. napus*. Further analyses revealed that D) was a *B. napus* population whereas E) and F) were *B. rapa* autotetraploids.

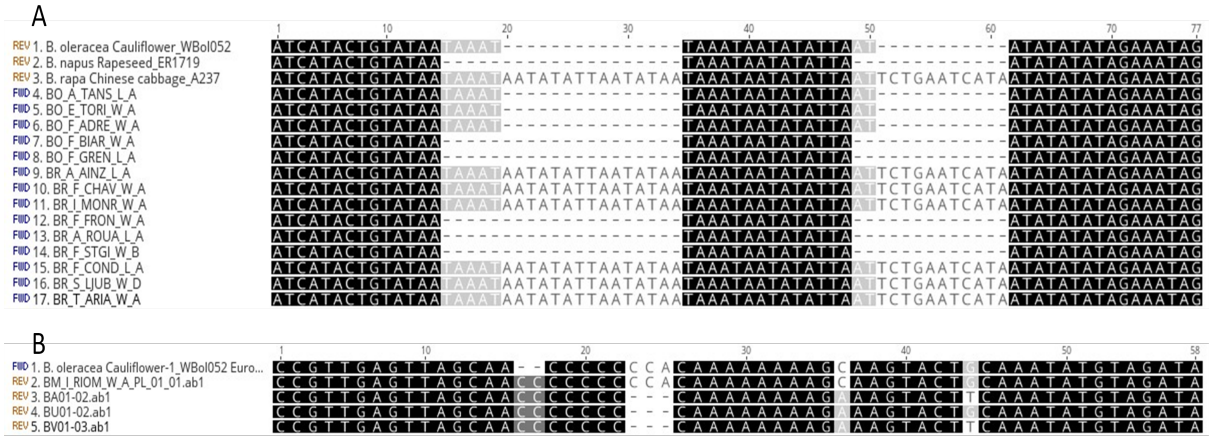

Figure 2. Alignments of chloroplast regions showing differences between the *Brassica* species: (A) comparison between the controls and different *Brassica oleracea* and *B. rapa* populations. The lines 7 (BO_F_BIAR_W_A), 8 (BO_F_GREN_L_A), 12 (BR_F_FRON_W_A), 13 (BR_A_ROUA_L_A) and 14 (BR_F_STGI_W_B) were misidentified populations with a *B. napus* chloroplast, (B) comparison between the controls and different *B. oleracea* related species, *B. montana* (BM), *B. macrocarpa* (BA), *B. rupestris* (BU) and *B. villosa* (BV), highlighting polymorphisms between the different species.

Table 1. Origin and number of collected *Brassica oleracea* and *B. rapa* wild and landrace populations, as well as five *B. oleracea*-related species. The number of populations, for which we obtained a germination sufficient for their multiplication, is indicated. For these latter, the number of populations for which the species was validated using flow cytometry, chloroplast sequencing, plus cytogenetic controls when required, is also given in the last column.

Expected species	Expected subspecies	Collected populations	Populations with a satisfying germination	Validated populations/ species-subspecies
Wild populations				
Brassica oleracea	*oleracea*	45	45	44
Brassica incana		1	1	1
Brassica macrocarpa		1	1	1
Brassica montana		8	8	8
Brassica rupestris	*rupestris*	2	2	2
Brassica villosa	*drepanensis*	1	1	1
Brassica villosa	*tineoi*	1	1	1
Brassica rapa	*sylvestris/campestris*	77	73	63
Landraces				
Brassica oleracea	*acephala*	9	9	9
Brassica oleracea	*botrytis*	6	6	6
Brassica oleracea	*capitata*	19	19	19
Brassica oleracea	*gemmifera*	1	1	1
Brassica oleracea	*gongylodes*	1	1	1
Brassica oleracea	*italica*	6	5	5
Brassica oleracea	*medullosa*	6	6	6
Brassica oleracea	*ramosa*	2	2	2
Brassica oleracea	*sabauda*	1	1	1
Brassica oleracea	unknown	16	8	6
Brassica rapa	*rapa*	71	68	63
Brassica rapa	*sylvestris* var. *esculenta*	6	5	5

(Figure 3E). Among the three remaining misidentified *B. rapa* populations, one wild population from Tunisia had a cytometry value close to *B. rapa* but no chloroplast gene amplification was detected; further morphological observations of this population revealed that it probably belongs to the genus *Sinapis*. The two last cases observed were *B. rapa* populations (one Slovenian wild population BR_S_LJUB_W_D and one French landrace BR_F_COND_L_A) having a flow cytometry value close to the one of *B. napus* (Figure 1E and Figure 1F) but a *B. rapa* chloroplast genomic sequence. Using cytogenetics, we detected no C chromosomes after a GISH-like experiment and 20 45S rDNA were counted, i.e. five rDNA loci per A genome (Figure 3F), which led us to the conclusion that these populations were in fact *B. rapa* autotetraploids (AAAA, 2n=4x=40).

Most of the populations confirmed as belonging to a specific species had an identical chloroplast sequence. Nevertheless, we observed a few SNPs specific to some populations. In *B. oleracea*, two SNPs were specific to only seven populations (BO_F_JOUY_L_A, BO_S_LJUB_L_G, BO_S_LJUB_L_H, BO_S_LJUB_L_L, BO_S_LJUB_L_M, BO_S_LJUB_L_N and BO_S_LJUB_L_O) and one allele at a different SNP was specific to BO_F_MERS_W_A. In *B. rapa*, three variations differentiated a few populations, one SNP in BR_A_DELL_W_A, one base deletion in BR_A_SEBA_W_A and BR_A_BOME_W_A and one SNP in BR_A_BLID_W_A, BR_A_BOUF_W_A, BR_A_CHLE_W_A, BR_A_BARA_W_A. These differences were observed in all the individuals tested per population.

For *B. oleracea* related species (*B. montana*, *B. rupestris*, *B. villosa*, *B. macrocarpa* and *B. incana*), all collected populations per species had the same flow cytometry value and the same chloroplast sequence.

Description of the populations

After discarding the few populations that did not germinate or were misidentified (Supplemental Table 3), we further characterized the remaining populations and their respective data collected during harvest.

Wild *B. oleracea* populations were collected on cliffs on the Atlantic coast in France and Spain (Figure 4), whereas its related species (*B. montana*, *B. rupestris*, *B. villosa*, *B. macrocarpa* and *B. incana*) were growing more in southern regions, on the Mediterranean coast. Their locations and the characteristics of each environment are described in Supplemental Table 1.

B. oleracea landraces were selected by farmers in each country, even in very warm regions such as the south of Algeria (Figure 4; Supplemental Table 2). Selection of different organs for crop production (flowers, leaves, stems or roots) has led to the divergence of highly diverse phenotypes. It is worth mentioning that some

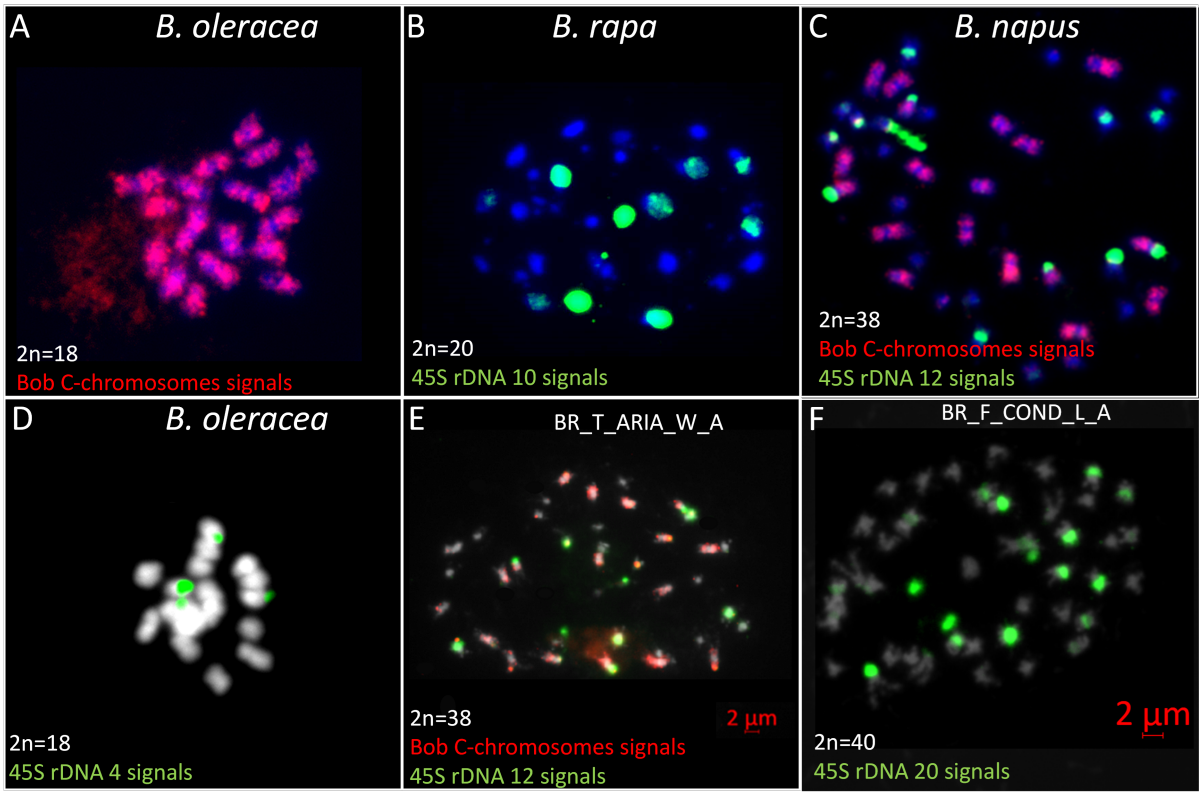

Figure 3. *Brassica* chromosomes stained by Fluorescence *in situ* hybridization (FISH). Chromosome number counted in mitosis with the three controls A and D) *B. oleracea*, B) *B. rapa* and C) *B. napus* and two populations showing an unexpected structure: E) BR_T_ARIA_W_A with *B. napus* genomic structure with 18 C chromosomes and 12 rDNA signals and F) BR_F_COND_L_A, an autotetraploid of *B. rapa* with 40 A chromosomes and 20 rDNA signals. The BoB014O06 BAC clone (red) is specific to C chromosomes allowing to distinguish A and C genomes.

morphotypes were difficult to classify in one subspecies as some of them were domesticated at the same time for leaf production (such as subsp. *acephala*) and for head cabbage (such as subsp. *capitata*, e.g. BO_A_TAZL_L_A). Additionally, even within the same morphotype, different developmental traits can be observed such as in Mugnuli populations (south of Italy) with several floral heads compared to common broccoli (Laghetti *et al*, 2005).

Wild *B. rapa* populations (Figure 5; Supplemental Table 1) were found in locations where competition with other species is lower, such as vineyards, orchards or field margins. Thus, regardless of the country, the populations were generally large.

The majority of the collected *B. rapa* local landraces (Figure 5; Supplemental Table 2) were turnips (subsp. *rapa*) with the exception of few broccoletto (subsp. *sylvestris* var. *esculenta*) selected by Italian farmers.

Discussion

In this paper, we described the sampling of wild populations and local landraces of *B. oleracea* and *B. rapa* along a large climatic and soil gradient from the north of France to the Sub-Saharan regions. Our objective was to validate at the early stage of plant development before sequencing that the seeds collected from plants of 112 and 154 of *B. oleracea* and *B. rapa* populations (both wild and local landraces), respectively, belonged to the expected botanical species. Then the origin of each population is described as a preliminary material for future botanical determination and plant adaptation genetic studies.

The first limiting factor was germination. Seed dormancy was only detected among *B. rapa* populations. In spite of seed viability confirmed by tetrazolium staining and of cold treatment, we did not succeed in getting enough seedlings per mother for four *B. rapa* wild populations to keep the initial genetic diversity of the populations. This trait, described in *Brassica* as primary physiological dormancy (Finch-Savage and Leubner-Metzger, 2006), seems to be a characteristic of some wild *B. rapa* populations. In our case, some populations met problems of imbibition as the seed coat was impermeable. Puncturing the seed coat before adding gibberellic acid improved germination for Sicilian and some Algerian wild populations. These results indicated a seed coat imposed dormancy in *B. rapa* which has not been described for *Brassica* (Baskin and Baskin, 1998). The conditions of seed conservation on the other hand is a likely explanation for the low germination rate in landraces of both species. This observation highlights the importance

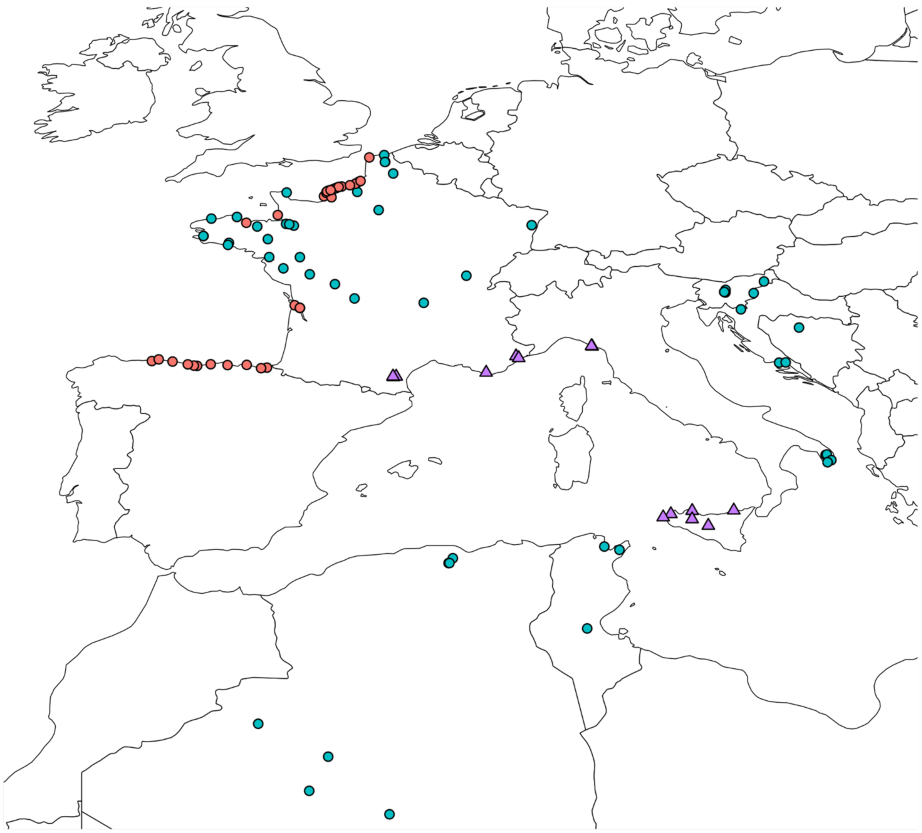

Figure 4. Distribution of the *Brassica oleracea* populations collected: 44 wild populations indicated with red dots, 56 landraces with green dots and 14 related species' populations with pink triangles.

of seed quality and storage conditions, especially in BRCs (Subramanian *et al*, 2023).

Because of the morphological similarity between the species at the seedling stage, our controls have revealed the importance of performing molecular and cytogenetic analyses before undertaking genetic sequencing and agronomic studies. We decided to combine a straightforward method, flow cytometry for assessment of chromosome number with a more expensive one, sequencing of a species-specific chloroplast region to validate the taxonomy. We applied more difficult and time-consuming cytogenetic methods for populations showing incongruent results with the two first methods. Flow cytometry is a high throughput technique allowing DNA content assessment of all plants, here 30 plants per population. Yet, as several species of the Brassiceae tribe have a similar DNA content, this technique might not be precise enough (Leflon *et al*, 2006) to validate the species. That is the reason why we complemented this analysis by sequencing a species-specific chloroplast genomic region taking advantage of the whole chloroplast genome sequences of many *Brassica* species/populations published by Li *et al* (2017). The combination with the analysis of chloroplast sequences allowed the confirmation of a misidentification for one Tunisian population presenting a flow cytometry value similar to *B. rapa* but no chloroplast amplification as it probably belongs to the genus *Sinapis*. However, the most frequent mistake was a confusion with *B. napus*, showing a higher DNA content, detectable by flow cytometry. Yet, among the 17 populations identified as *B. napus* by flow cytometry (three populations in the *B. oleracea* and 14 in the *B. rapa* collections), three had a chloroplast sequence similar to *B. rapa*. This conflicting result called for further cytogenetic experiments for these three populations, using GISH-like on mitotic chromosomes with a BAC specific to *B. oleracea* chromosomes (Suay *et al*, 2014) and 45S rDNA probes revealing the number of rDNA loci (Książczyk *et al*, 2011). From this data, we concluded that one Tunisian population was indeed a *B. napus* population. It could be interesting to precisely compare after chloroplast assembly with the results reported by Li *et al* (2017). These authors reported that *B. napus* chloroplasts can be classified into two different clades identified from different *B. rapa* morphotypes. The two other populations were *B. rapa* autotetraploids, with 40 A chromosomes and 20 45S rDNA loci as expected when doubling the A genome. Such autopolyploid populations were previously reported for the production of new forage varieties (Olsson and Ellerström, 1980).

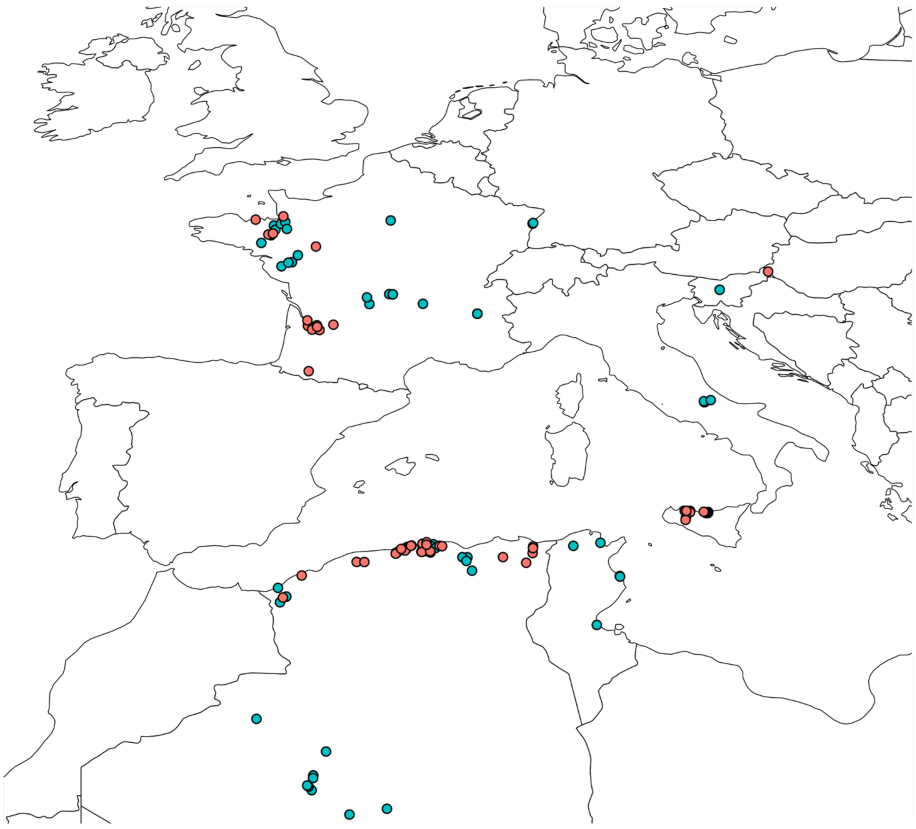

Figure 5. Distribution of the *Brassica* rapa populations collected: 63 wild populations indicated with red dots, 68 landraces with green dots.

Among the 100 and 131 confirmed diploid populations for *B. oleracea* and *B. rapa* respectively, chloroplast sequences revealed only a few variants SNV (Li *et al*, 2017) for some accessions in both species. The low mutation rate of the chloroplast DNA in most flowering plant families can explain these variations as already reported from global chloroplast assembly. Interestingly, Li *et al* (2017) observed more SNVs in the *B. rapa* than in the *B. oleracea* genotypes that they investigated, with 343 and 16 SNV, respectively. By investigating an enlarged *B. oleracea* diversity, Perumal *et al* (2021) described more SNVs with clustering of different cultigroups. In our collected wild and landrace populations, we observed that a common variation is shared by seven populations belonging to capitata and acephala groups originating from Slovenia with the exception of one French landrace. For *B. rapa*, SNV were only observed in some wild Algerian populations. Further studies are in progress in order to compare the genetic diversity from chloroplast assembly and nuclear SNP, taking into account the different cultigroups and their geographic origins.

A large morphological diversity was observed among the *B. oleracea* landraces whereas wild populations were morphologically similar to forage kales. For Mugnoli belonging to the same group as broccoli (subsp. *italica*), Biancolillo *et al* (2023) developed a non-destructive tool based on Multivariate Image analysis and agro-morphological descriptors for the characterization and authentication of these local varieties. For *B. rapa*, landraces selected by farmers are mainly turnips, with the exception of five populations of Broccoletto. In this paper, we describe the different environments in which these different populations were collected.

This well-characterized material collected on a very large climatic and soil gradient opens the prospect of identifying genomic regions involved in adaptation to climatic constraints and microbiota descriptors (fungus and bacterial composition). To do so, seeds were produced at the same geographic location in order to avoid the environmental effects of the collecting site on seed quality. High-throughput sequencing for bulks of 30 plants per population is currently ongoing to capture the maximum diversity existing within the population. Mapping the reference genome of each species and SNP calling will allow the description of genetic diversity and the design of nested core collections. Genome-wide association (GWAS) and genotype-environment association (GEA) analyses will be possible from the project consortium to identify genomic regions involved in climate adaptation. Functional analyses will be

performed on the most contrasted populations to finely investigate their responses to cold and warm temperatures. Field experiences of core collections in five countries will allow the validation of favourable alleles under different environmental conditions. All these data will be used (1) to promote local landraces, as several are endangered, and (2) to design crosses that could be relevant to produce pre-breeding populations, each adapted to the climatic evolution of each country.

Supplemental data

Supplemental Table 1. Description of *B. oleracea* and *B. rapa* wild populations
Supplemental Table 2. Description of *B. oleracea* and *B. rapa* landraces
Supplemental Table 3. Populations that did not germinate or were misidentified

Acknowledgements

We thank the Genetic Resource Centers BrACySol (https://igepp.rennes.hub.inrae.fr/l-igepp/plateformes/bracysol) and the Agricultural Institute of Slovenia (https://www.kis.si/en/) for providing seeds from different landraces. We thank Biogenouest (the western French network of technology core facilities in life sciences and the environment, supported by the Conseil Regional des Pays de la Loire) for access to molecular cytogenetics (https://www6.rennes.inrae.fr/igepp_eng/About-IGEPP/Platforms/Molecular-Cytogenetics-Platform-PCMV) and GenOuest bioinformatic platforms (https://www.genouest.org/). We thank Plant imaging platform PHENOTIC in Angers (INRAE-IRHS, Angers University, Institut Agro, GEVES, France) for experiments on seed germination and V. Blouin for tetrazolium staining. We also thank all the staff who took care of our plant material (especially L. Charlon, J-P. Constantin and F. Letertre).

All the research is funded by H2020 Prima, project no. 1425, BrasExplor (https://brasexplor.hub.inrae.fr/) for 'Wide exploration of genetic diversity in *Brassica* species for sustainable crop production' and by INRAE through TSARA Initiative (Transforming food systems and agriculture through a partnership research with Africa) promoting a specific French-Algerian collaboration.

Author contributions

CF, HH, and AMC designed and managed all the experiments. CF, HH, FA, CB, GB, LB, MC, GD, JFC, LG, AG, RI, VI, JAJ, VM, EO, MP, MP, BP, TR, FR, JR, RS, LS, VT, ST, IT, FWB, AMC participated to the collects and the local description of the populations. VM, BP, VR, ST provided landraces and their description from BRC. MT performed flow cytometer analyses. GD and MRG designed chloroplast markers and performed experiments. OC and VH performed all molecular cytogenetic experiments. MB managed the database for population description. CF, HH, MRG and MT contributed to writing the manuscript, which was finalized by AMC.

Conflict of interest statement

The authors declare that they have no financial or competing interests.

References

Baskin, C. and Baskin, J. M. (1998). Seeds. Ecology, Biogeography and evolution of dormancy and germination. US: Academic Press, San Diego, pp. 666

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A. M., Delourme, R., Deniot, G., Denoeud, F., Duffé, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., Vekemans, X., D'hont, A., Rousseau-Gueutin, M., Barbe, V., Cruaud, C., Wincker, P., and Aury, J. M. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* 4, 879–887. doi: https://doi.org/10.1038/s41477-018-0289-4

Biancolillo, A., Ferretti, R., Scappaticci, C., Foschi, M., D'archivio, A. A., Santo, M. D., and Martino, L. D. (2023). Development of a Non-Destructive Tool Based on E-Eye and Agro-Morphological Descriptors for the Characterization and Classification of Different Brassicaceae Landraces. *Applied Sciences* 13. doi: https://doi.org/10.3390/app13116591

Bird, K. A., An, H., Gazave, E., Gore, M. A., Pires, J. C., Robertson, L. D., and Labate, J. A. (2017). Population Structure and Phylogenetic Relationships in a Diverse Panel of Brassica rapa L. . *Frontiers in Plant Science* 8. doi: https://doi.org/10.3389/fpls.2017.00321

Cai, C., Bucher, J., Bakker, F. T., and Bonnema, G. (2022). Evidence for two domestication lineages supporting a middle-eastern origin for Brassica oleracea crops from diversified kale populations. *Horticulture Research* 9(166). doi: https://doi.org/10.1093/hr/uhac033

Cai, X., Chang, L., Zhang, T., Chen, H., Zhang, L., Lin, R., Liang, J., Wu, J., Freeling, M., and Wang, X. (2021). Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biology* 22, 166–166. doi: https://doi.org/10.1186/s13059-021-02383-2

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., Liu, B., Liang, J., Zhuang, M., Liu, Y., Lin, K., Bucher, J., Zhang, N., Wang, Y., Wang, H., Deng, J., Liao, Y., Wei, K., Zhang, X., Fu, L., Hu, Y., Liu, J., Cai, C., Zhang, S., Zhang, S., Li, F., Zhang, H., Zhang, J., Guo, N., Liu, Z., Liu, J., Sun, C., Ma, Y., Zhang, H., Cui, Y., Freeling, M. R., Borm, T., Bonnema, G., Wu, J., and Wang, X. (2016). Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nature Genetics* 48, 1218–1224. doi: https://doi.org/10.1038/ng.3634

Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of Brassica plants

1(14024). doi: https://doi.org/10.1038/hortres.2014.24

Finch-Savage, W. E. and Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. *New Phytologist* 171, 501–523. doi: https://doi.org/10.1111/j.1469-8137.2006.01787.x

Gerlach, W. L. and Bedbrook, J. R. (1979). Cloning and characterization of ribosomal RNA genes from wheat and barley. *Nucleic Acids Research* 7. doi: https://doi.org/10.1093/nar/7.7.1869

Gomez-Campo, C., Aguinagalde, I., Ceresuela, J. L., Lazaro, A., Martınez-Laborde, J. B., Parra-Quijano, M., Simonetti, E., Torres, E., Tortosa, E., and E, M. (2005). An exploration of wild *Brassica oleracea* L. germplasm in Northern Spain. *Genetic Resources and Crop Evolution* 52, 7–13.

Howell, E. C., Kearsey, M. J., Jones, G. H., King, G. J., and Armstrong, S. J. (2008). A and C Genome Distinction and Chromosome Identification in Brassica napus by Sequential Fluorescence in Situ Hybridization and Genomic in Situ Hybridization. *Genetics* 180, 1849–1857. doi: https://doi.org/10.1534/genetics.108.095893

INPN (2024). National Inventory of Natural Heritage in France. url: https://inpn.mnhn.fr/accueil/index?lg=en.

Książczyk, T., Kovarik, A., Eber, F., Huteau, V., Khaitova, L., Tesarikova, Z., Coriton, O., and Chèvre, A. M. (2011). Immediate unidirectional epigenetic reprogramming of NORs occurs independently of rDNA rearrangements in synthetic and natural forms of a polyploid species *Brassica napus*. *Chromosoma* 120, 557–571. doi: https://doi.org/10.1007/s00412-011-0331-z

Laghetti, G., Martignano, F., Falco, V., Cifarelli, S., Gladis, T., and Hammer, K. (2005). "Mugnoli": a Neglected Race of *Brassica oleracea* L. from Salento (Italy). *Genetic Resources and Crop Evolution* 52, 635–639. doi: https://doi.org/10.1007/s10722-005-8511-4

Leflon, M., Eber, F., Letanneur, J. C., Chelysheva, L., Coriton, O., Huteau, V., Ryder, C. D., Barker, G., Jenczewski, E., and Chèvre, A. M. (2006). Pairing and recombination at meiosis of *Brassica rapa* (AA) × *Brassica napus* (AACC) hybrids. *Theoretical and Applied Genetics* 113. doi: https://doi.org/10.1007/s00122-006-0393-0

Li, P., Zhang, S., Li, F., Zhang, S., Zhang, H., Wang, X., Sun, R., Bonnema, G., and Borm, T. J. A. (2017). A Phylogenetic Analysis of Chloroplast Genomes Elucidates the Relationships of the Six Economically Important *Brassica* Species Comprising the Triangle of U. *Frontiers in Plant Science* 8. doi: https://doi.org/10.3389/fpls.2017.00111

Mabry, M. E., Turner-Hissong, S. D., Gallagher, E. Y., Mcalvay, A. C., An, H., Edger, P. P., Moore, J. D., Pink, D. A. C., Teakle, G. R., Stevens, C. J., Barker, G., Labate, J., Fuller, D. Q., Allaby, R. G., Beissinger, T., Decker, J. E., Gore, M. A., and Pires, J. C. (2021). The Evolutionary History of Wild, Domesticated, and Feral *Brassica oleracea* (Brassicaceae). *Molecular Biology and Evolution* 38, 4419–4434. doi: https://doi.org/10.1093/molbev/msab183

Maggioni, L., Bothmer, R., Poulsen, G., and Aloisi, K. H. (2020). Survey and genetic diversity of wild *Brassica oleracea* L. germplasm on the Atlantic coast of France. *Genetic Resources and Crop Evolution* 67, 1853–1866. doi: https://doi.org/10.1007/s10722-020-00945-0

McAlvay, A. C., Ragsdale, A. P., Mabry, M. E., Qi, X., Bird, K. A., Velasco, P., An, H., Pires, J. C., and Emshwiller, E. (2021). Brassica rapa Domestication: Untangling Wild and Feral Forms and Convergence of Crop Morphotypes. *Molecular Biology and Evolution* 38, 3358–3372. doi: https://doi.org/10.1093/molbev/msab108

Olsson, G. and Ellerström, S. (1980). Polyploidy breeding in Europe. In *Brassica crops and wild allies; biology and breeding*, ed. Tsunoda, S., Hinata, K., and Grimez-campo, C., (Tokyo: Japan Scientific Soc. Press), 167-190.

Perumal, S., Waminal, N. E., Lee, J., Koo, H. J., Choi, B., Park, J. Y., Ahn, K., and Yang, T. J. (2021). Nuclear and chloroplast genome diversity revealed by low-coverage whole-genome shotgun sequence in 44 *Brassica oleracea* breeding lines. *Horticultural Plant Journal* 7, 539–551. doi: https://doi.org/10.1016/j.hpj.2021.02.004

Qi, X., An, H., Ragsdale, A. P., Hall, T. E., Gutenkunst, R. N., Pires, J. C., and Barker, M. S. (2017). Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Molecular Ecology* 26, 3373–3388. doi: https://doi.org/10.1111/mec.14131

Suay, L., Zhang, D., Eber, F., Jouy, H., Lodé, M., Huteau, V., Coriton, O., Szadkowski, E., Leflon, M., Martin, O. C., Falque, M., Jenczewski, E., Paillard, S., and Chèvre, A. M. (2014). Crossover rate between homologous chromosomes and interference are regulated by the addition of specific unpaired chromosomes in *Brassica*. *New Phytologist* 201, 645–656. doi: https://doi.org/10.1111/nph.12534

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. *Frontiers in Plant Science* 14. doi: https://doi.org/10.3389/fpls.2023.1220134

Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 72–82
DOI: 10.46265/genresj.WETA7514
https://www.genresj.org
ISSN: 2708-3764

Morphological and molecular characterization of 'Saragolla' wheats (*Triticum turgidum* subsp. *durum* from Abruzzo, Italy)

Agata Rascio *,[a], **Vanessa De Simone**[a], **Lorenzo Goglia**[b], **Silvana Paone**[a], **Maria Pellegrino**[a] and **Giuseppe Sorrentino**[b]

[a] *Council for Agricultural Research and Economics, Research Centre for Cereal and Industrial Crops S.S, 673 Km 25, 200 71122, Foggia, Italy*
[b] *Institute for Sustainable Plant Protection - Italian National Research Council (IPSP-CNR), Piazzale Enrico Fermi, 1, 80055, Portici (NA), Italy*

Abstract: A morphological and genetic characterization of autochthonous 'Saragolla' wheats, currently cultivated in Abruzzo Region (Italy), was carried out. Using 15 simple sequence repeat (SSR) markers and 24 UPOV morphological traits we compared: (a) 13 'Saragolla' genotypes with traits of the *italicum/apulicum* botanical varieties (Saragolla (Sar.) *italicum*), (b) 26 'Saragolla' genotypes with traits of *leucurum/affine* botanical varieties (Sar. *leucurum*), (c) 8 breeding varieties (pure lines), and (d) 5 Italian autochthonous wheats and 1 *turanicum* line (old wheats). One hundred twenty-six (126) alleles were identified. The number of alleles per locus spanned from 4 to 15 and the number of alleles per genotype varied between 12 and 21. Values of gene diversity (Nei) across the 53 genotypes was 0.17. The groups of Sar. *leucurum* and Sar. *italicum* genotypes were morphologically distinguishable from the groups of old wheats and pure lines. Likewise, the analysis of molecular data using the discriminant analysis revealed that genotypes with the Sar. *italicum* phenotype displayed distinct genetic differences from Sar. *leucurum,* pure lines and old wheats. These results make Sar. *italicum* genotypes distinguishable and eligible as a conservation variety. Ward's clustering analysis of the 53-genotype pool showed that the 'Saragolla' landrace is a valuable repository of genetic diversity.

Keywords: Abruzzo Region, 'Saragolla' landrace, durum wheat diversity, genetic characterization

Citation: Rascio, A., De Simone, V., Goglia, L., Paone, S., Pellegrino, M., Sorrentino, G. (2024). Morphological and molecular characterization of 'Saragolla' wheats (*Triticum turgidum* subsp. *durum* from Abruzzo, Italy). *Genetic Resources* 5 (9), 72–82. doi: 10.46265/genresj.WETA7514.

Introduction

Traditional crops are generally viewed by consumers as more natural and healthier options (Rascio *et al*, 2015, 2016). Their reintroduction into cultivation and reinforcing their value chain can play a vital role in bolstering conservation efforts and elevating the value of their cultivation environment. This, in turn, can substantially increase the economic benefits for rural communities (Pallante *et al*, 2016). Before the onset of the breeding era in Italy, initiated by Strampelli in the early twentieth century (Scarascia-Mugnozza, 2005), cultivated wheat primarily consisted of autochthonous varieties. Although there is not a worldwide consensus on this definition, these wheats are often referred to as old/ancient wheats, or landraces (Negri *et al*, 2009).

Autochthonous wheats are regarded as valuable sources of alleles for breeding programmes (Terzi *et al*, 2005). Indeed, over time, environmental conditions and, to a certain extent, purposeful farmer selection have resulted in the cultivation of plants exhibiting high adaptability and superior performance within their respective cultivation regions (Zeven, 1998). For

*Corresponding author: Agata Rascio (email:agata.rascio@crea.gov.it)

Received: 18.10.2023 | Accepted: 10.04.2024 | Published online: 14.05.2024

these reasons, the EU (EU, 2009) defines landraces or varieties which are naturally adapted to local and regional conditions as conservation varieties (Spataro and Negri, 2013). They are frequently composed of mixed genotypes (populations) and exhibit a high degree of genetic diversity (Zeven, 1998), in contrast to modern breeding varieties (cultivars or pure lines).

Following the ratification of the International Treaty on Plant Genetic Resources for Food and Agriculture by the UN Food and Agriculture Organization (FAO, 2009) in 2001, most Italian regions implemented laws aimed at protecting and promoting indigenous crops. They also provided funding for projects dedicated to studying the conservation varieties' distinctive characteristics. Their registration involves a formal process where these varieties are officially recognized, documented, and often included in seed catalogues or databases. Once registered, conservation varieties are often conserved in community seedbanks or similar facilities where they can be accessed by farmers. Enhancing access to these varieties is important to support agricultural biodiversity, improve resilience in the face of environmental challenges, and contribute to the long-term sustainability of food production systems. Some countries, by officially recognizing and documenting these varieties, provide legal protection to conservation varieties to prevent unauthorized use or commercial exploitation (EU, 2009). This protection is designed to encourage the continued conservation and sustainable use of these valuable genetic resources.

In Southern Italy, the cultivation of a wheat type known as 'Saragollo forte' which is particularly well-suited for pasta production, has been extensively documented in various commercial agreements dating back to the seventeenth century (Fiore, 2013). As introduced by old botanists (Draghetti, 1927; De Cillis, 1927), the use of the plural noun 'Saragolle,' highlights the presence of multiple forms of 'Saragolla' wheat, all falling under the Saragolla (Sar.) *leucurum* botanical variety, which was one of the 22 botanical varieties of *Triticum turgidum* documented at the start of the 1900s (Percival, 1921). In 2004, the 'Produttori Sementi Bologna' company registered a variety also named 'Saragolla'. This variety resulted from crosses between the 'Iride' cultivar and the '0114' elite line. Therefore, this 'Saragolla' pure line is an enhanced variety and not a local one.

Today, a significant number of farmers in Central and Southern Italy are cultivating the old 'Saragolla' wheat, either for personal consumption or to establish short food supply chains of autochthonous wheat. The resurgence of interest in 'Saragolla' can be attributed to its adaptability to low-fertility soils and its suitability for cultivation with minimal input methods, making it especially attractive for agriculturally marginal regions. Moreover, this revived interest in 'Saragolla' has been magnified by online sources (Eccellenze D'abruzzo, 2024), which suggest that it can be an Italian alternative to Khorasan wheat, marketed under the name 'Kamut' (Piergiovanni, 2013). Kamut is a registered variety belonging to the tetraploid species *Triticum turanicum* Jacubz, initially described by Percival (1921) as *T. orientale*. This hypothesis is reinforced by the morphological resemblance between the elongated seeds of 'Saragolla' cultivated in the Abruzzo Region and the seeds of 'Kamut' *T. orientale* Percival.

Considering the complexity of the situation, in the years 2018–2020, the local authorities of the Italian Region Abruzzo funded the SARAB project, 'Characterization of local Saragolla durum wheat populations', to characterize the 'Saragolla' wheat currently cultivated. Through intensive cataloguing based on morphology, the project focused on the main species and botanical varieties of 'Saragolla' cultivated in 11 different sites within the Abruzzo Region (Rascio *et al*, 2021). The results revealed a heterogeneous botanical composition both within and among these sites (Rascio *et al*, 2022). Nine botanical varieties of durum wheat were observed, the majority belonging to the *italicum* or *apulicum* botanical varieties, primarily differing in glume pigmentation intensity. There was also a smaller number of genotypes falling into the *leucurum* or *affine* botanical varities (referred to here as Saragolla (Sar.) *leucurum*), which exhibited variations in seed pigmentation intensity.

The main goals of this study were to conduct genetic and morphological characterizations of representative plants belonging to the most prevalent botanical varieties collected in 11 farms that participated in the SARAB project. Additionally, these varieties were compared with a collection of both pure lines (modern) and traditional Italian (old) wheat varieties. The results here shown indicate that, in contrast to the Sar. *leucurum*, the *italicum* genotypes shared a close genetic similarity among themselves. They are widespread in the Abruzzo region and genetically distinct from the groups of modern and old genotypes examined in this study, hence they are eligible for registration as a conservation variety. The study also explores the degree of diversity of the four groups of genotypes.

Material and methods

Plant material

Eleven samples of 'Saragolla' wheats from 11 locations in Abruzzo (shown in Supplemental Table 1) were grown and characterized at the Council for Agricultural Research and Economics - Research Centre for Cereal and Industrial Crops (CREA-CI) in Foggia (Rascio *et al*, 2022). About 1,000 individual plants of these heterogeneous wheats were morphologically examined and 434 durum wheat plants were found. A total of 39 representative spikes of this subset, selected to be indicative of the prevailing botanical varieties *italicum/apulicum* and *leucurum/affine* were used for further genetic and morphological characterization in this study. For comparison, eight modern varieties/durum cultivars and six samples of old autochthonous durum wheat samples

belonging to the CREA-CI working collection were also included (Supplemental Table 2). They comprised four genotype groups:

1) **Sar. *italicum***: 13 'head to row' of 'Saragolla' genotypes exhibiting at least four out of five traits of the *italicum/apulicum* botanical varieties

2) **Sar. *leucurum***: 22 'head to row' of 'Saragolla' genotypes from Abruzzo and four from Puglia displaying at least four out of five traits of the *leucurum/affine* botanical varieties

3) **Modern pure lines :** Eight seed samples of durum wheat cultivars: 'Ciccio', 'Cappelli', 'Capeiti', 'Colosseo', 'Duilio', 'Simeto', 'Svevo' and 'Saragolla'

4) **Old wheats:** Six seed samples of autochthonous durum wheats, primarily sourced from Sicily (Fiore *et al*, 2019) and belonging to the CREA-CI working collection. These include: 'Realforte', 'Russello', 'Sammartinara', 'Scorsonera', 'Vallelunga pubescent', and the *T. turanicum* pure line (PI166959), selected at CREA-CI.

Phenotypic assessment

In 2021, a total of 53 rows, each 1m in length and spaced 30cm apart, were sown according to the usual agronomic practices (Rascio *et al*, 2016). Throughout the growth stage, each row was carefully examined to ensure its purity, and one plant was selected for DNA extraction and morphological characterization. The assessment was performed on 24 traits with value scales employed for evaluation in part adhering to the guidelines outlined by the International Union for the Protection of New Varieties of Plants (UPOV, 2012) (Table 1).

Molecular marker analysis

For each genotype, the extraction of DNA was performed according to the protocol used by Marone *et al* (2009). Twenty-eight microsatellite single sequence repeat (SSR) markers were selected based on published map data (Marone *et al*, 2009, 2012), according to the following criteria: locus-specific amplification, high level of polymorphism, and good genome coverage (one marker per chromosome arm). The sequences of the SSR are available in the GrainGenes database (http://wheat.pw.usda.gov). The PCR reactions were performed in 25μl volume in Applied Biosystems 2720 Thermal Cyclers. The reaction mixture contained 60ng of template DNA, 0.2mM of dNTPs, 1X Buffer (10mM Tris–HCl—pH 8.3, 50mM KCl, 1.5mM MgCl2), 0.4μM labelled reverse primer (FAM or HEX or NED or TET), 0.4μM unlabelled forward primer and 0.2U of Taq DNA polymerase (5U/μl) (Kapa). Thermal cycling conditions were as follows: 94°C for 3min, followed by 45 cycles of 94°C for 30s, the specific annealing temperature (Ta) for each primer for 30s, 72°C for 30s, with a final extension at 72°C for 2min. The amplification products were analyzed by means of capillary electrophoresis (ABI3130), multiplexing different fluorescent dyes. Electropherograms were analyzed with GeneMapper version 4.0. The internal molecular weight standard was 500-ROX (Life Technologies).

Statistical analysis

Genotypic characterization was performed with 15 SSR markers which gave a clear electrophoretic pattern. To this aim, the genotypic data were transformed into a binomial matrix as present (1) or absent (0) for each marker and this matrix was used to construct Ward's dendrogram tree to assess genetic diversity.

Nei's gene diversity, percentage of polymorphic loci and Shannon's information index were determined using the PopGen 1.31 software (Yeh *et al*, 1999).

For assessing marker polymorphism and informativeness, the average polymorphic information content (PIC) was calculated using the following formula introduced by Anderson *et al* (1993):

$PIC = 1\text{-}\Sigma(P_i)^2$

where P_i is the number of polymorphic loci/all the number loci.

The distances among the four groups of genotypes (Sar. *leucurum*, Sar. *italicum*, pure lines and old wheats) were examined by multivariate discriminant analysis and cluster analysis, using the STATISTICA (StatSoft Inc.) software.

Results

Morphological characterization

As shown in Figure 1 and in Supplemental Table 3, the Sar. *italicum/apulicum*-like genotypes have rather compact, hairy glumes, lightly pigmented spikes, long red or brown-red awns, yellow-amber and elongated grains. The Sar. *leucurum/affine* genotypes have white and glabrous glumes, elongated, compact spikes, and white or red seeds.

The mean values of each morphological and phenological trait for genotypes belonging to the four groups (Sar. *leucurum*, Sar. *italicum*, modern cultivars or old wheats) show that the Sar. *italicum/apulicum* genotypes registered the highest values for glume hairiness and 1,000 seed weight (Table 2). In contrast, the Sar. *leucurum/affine* genotypes exhibited the highest values for the shape of the lower glume beak. Modern varieties displayed smaller height and earlier heading dates, a result of the extensive breeding efforts they underwent.

Results of stepwise discriminant analysis performed using the four groups of genotypes as classification categories and visualized through the biplot of canonical variables (Figure 2) showed that the model had a high discriminatory power (Lambda Wilks: 0,0075245; approx. F (48,90) = 7,8312; $p < 0,0000$), with two discriminant functions that accounted for 92.9% of the explained variance (Table 3).

Based on the absolute values of standardized coefficients of the canonical variables (Table 3) the main traits that horizontally contributed to the 4-group separation were the glume hairiness, which was absent in Sar. *leucurum* and the upper neck glaucosity, which

Table 1. The 24 traits used for the morphological characterization of wheat genotypes, and the value scale employed for evaluation (UPOV, 2012).

	Trait	Measure units/score	Assessment scale
1	1,000 seed weight	g	
2	Curvature of lower glume beak	1–7	1 = absent, 7 = strong
3	Lower glume length of beak	1–9	1 = very short, 7 = long
4	Lower glume hairiness	1–9	1 = absent, 9 = very present
5	Straw: pith in cross section	1–7	1 = thin, 7 = thick
6	Grain shape	3–7	3 = slightly, 7 = strongly elongated
7	Grain: length of brush hair	3–7	3 = short, 7 = long
8	Grain weight/plant	g	
9	Awn colour	1–5	1 = white, 3 = medium dark, 5 = very dark
10	Spike colour	1–5	1 = white, 3 = medium dark, 5 = very dark
11	Awn tip/ear length ratio	1–3	1 = low ratio, 3 = high ratio
12	Spike length	cm	
13	Spike shape in profile	1–5	1 = tapering, 2 = parallel sided, 3 = slightly clavate, 4 = strongly clavate, 5 = fusiform
14	Ear glaucosity	1–9	1 = absent, 9 = very strong
15	Awn divergence	1–2	1 = adherent, 2 = divergent
16	Spike density	3–7	3 = lax, 5 = medium, 7 = dense
17	Plant height	cm	
18	Growth habit	1–9	1 = erect, 9 = prostrate (score 9)
19	Recurved flag leaves	1–9	1 = absent, 9 = very high
20	Heading date from April 1st	days	
21	Flag leaf: glaucosity of sheath	1–9	1 = absent, 9 = very strong
22	Flag leaf lower side glaucosity	1–9	1 = absent, 9 = very strong
23	Upper node hairiness	1–9	1 = absent, 9 = very strong
24	Upper neck glaucosity	1–9	1 = absent, 9 = very strong

Figure 1. Comparison of distinguishing characteristics of glume (A), and spike and seed (B) of Saragolla botanical types Sar. *italicum* and Sar. *apulicum*: hairy with red glumes and red awns; Sar. *leucurum* and Sar. *affine*: glabrous with white glumes and white awns.

Table 2. Mean values of morphological traits of the four groups of durum wheat genotypes. The qualitative traits, for which no specific unit of measurement is provided, were assessed using the value scale established by the International Union for the Protection of New Varieties of Plants (UPOV, 2012).

Trait		Sar. *leucurum*	Sar . *italicum*	Modern cultivars	Old wheats
Shape of lower glume beak	Mean	6,9	2,0	2,7	2,1
	SD	0,6	1,3	2,0	1,6
Glume length of beak	Mean	4,0	4,9	4,3	3,6
	SD	1,8	1,6	3,0	1,9
Glume hairiness	Mean	1,3	8,2	1,3	1,0
	SD	1,5	1,7	0,8	0,0
Straw: pith in cross section	Mean	5,5	6,7	6,7	6,4
	SD	1,9	0,6	0,8	1,0
Ear length (cm)	Mean	9,2	9,4	8,5	8,4
	SD	1,2	1,0	2,3	2,2
Grain shape	Mean	4,6	6,8	3,7	3,6
	SD	1,7	0,6	1,0	1,5
Grain: length of hair	Mean	3,3	4,1	3,1	3,0
	SD	0,9	1,3	0,2	0,0
Awn colour	Mean	1,8	2,3	2,2	2,6
	SD	1,1	0,9	1,6	1,4
Glume colour	Mean	1,7	2,0	1,5	2,0
	SD	0,8	0,9	0,5	1,5
Awn tip/ear length ratio	Mean	1,6	2,7	1,7	1,6
	SD	0,7	0,6	0,8	0,8
Ear shape	Mean	2,0	2,2	3,0	2,0
	SD	0,9	1,2	1,3	0,8
Awn compactness	Mean	1,4	1,5	1,5	1,9
	SD	0,6	0,8	0,5	1,5
Ear density	Mean	5,2	4,7	4,7	6,1
	SD	1,9	1,4	1,5	1,1
Plant height (cm)	Mean	108,7	120,0	79,2	101,4
	SD	16,2	7,9	17,7	21,7
Grain weight/plant (g)	Mean	19,7	15,2	19,0	15,3
	SD	6,3	4,2	1,9	3,4
1,000 seed weight (g)	Mean	46,5	73,3	48,7	46,4
	SD	8,5	3,6	5,3	7,0
Growth habit	Mean	3,7	4,4	2,3	4,4
	SD	1,4	1,0	2,1	1,0
% recurved flag leaves	Mean	6,1	6,7	2,7	5,0
	SD	1,6	0,8	1,5	2,0
Heading date	Mean	28,3	29,4	18,1	20,7
(days from April 1st)	SD	4,1	1,6	6,6	6,3
Flag leaf glaucosity	Mean	5,1	5,8	4,7	4,4
	SD	1,5	1,3	2,0	2,2
Flag leaf lower glaucosity	Mean	3,4	3,0	3,7	3,3
	SD	1,4	0,0	1,0	1,4
Upper node hairiness	Mean	3,3	3,0	2,3	3,0
	SD	1,3	0,0	1,0	1,2
Upper neck glaucosity	Mean	5,1	3,3	5,0	4,4
	SD	1,5	0,8	1,3	1,5
Spike glaucosity	Mean	4,8	6,7	5,7	5,3
	SD	1,4	1,1	1,0	1,4

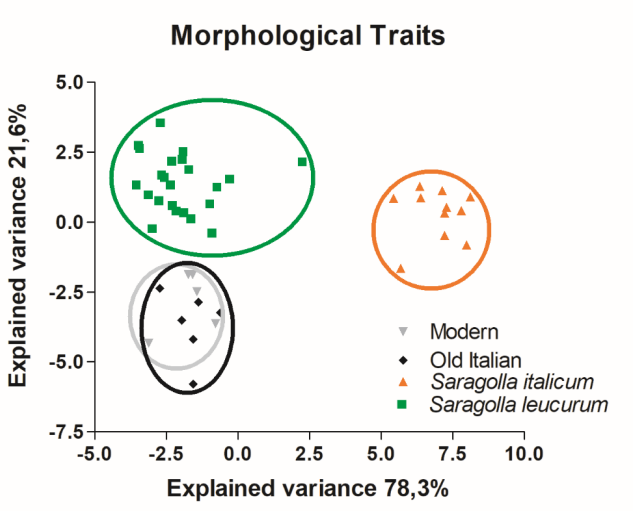

Figure 2. The biplot of canonical variables obtained using a stepwise discriminant analysis based on 24 UPOV morphological descriptors and four groups of genotypes. The percentages of explained variance by the two roots are shown.

was lacking in Sar. *italicum* (Table 2). The differences in heading date along with the upper neck green glaucosity degree (Table 2) mainly contributed to the vertical separation of Sar. *leucurum* and *italicum* from the pool of new and old genotypes.

The Mahalanobis distances between groups were all highly significant (Table 4), except between pure lines and old wheats; Sar. *leucurum* and Sar. *italicum* were the two most distant and hence morphologically different groups.

SSR patterns

The molecular analysis utilized a set of 15 SSR markers out of initially 28 tested, each characterized by a distinct electrophoretic pattern for all genotypes. In total, 126 alleles were identified, with the number of alleles per locus spanning from 4 to 15. The Polymorphic Information Content (PIC) values of the SSR markers ranged from 0.52 (for gwm60) to 0.91 (for wmc606 and gwm459), resulting in an average PIC value of 0.77 per locus (Table 5).

Excluding from the analysis the alleles that occurred at a low frequency ($p < 0.05$), the number of alleles per genotype varied between 12 and 21, with the most frequent value being 16 (Table 6). This occurrence was four times higher than what was observed in a study where 104 Ethiopian durum wheat genotypes, representing 13 populations, three regions, and four altitudinal classes, were analyzed using 14 SSR markers (Dagnaw *et al*, 2023).

Diversity

The diversity analysis for all cultivars based on SSR markers (Table 7) yielded low mean values (0.28±0.20) of Shannon's index. The values (0.176±0.5) of Nei's gene diversity were lower than the minimum observed in 40 winter wheat genotypes (Petrović *et al*, 2017) coming from European countries (Croatia, Austria, France, Italy, and Russia) and lower than that (0.56) resulting for 124 Ethiopian genotypes (Dagnaw *et al*, 2023).

The among-groups comparison indicated that old wheats and Sar. *leucurum* exhibited the highest percentage of polymorphic loci, followed by pure lines and old wheats. The measurement of gene diversity, estimated by both Nei's gene diversity and Shannon's information index, yielded similar values for Sar. *leucurum*, breeding lines, and old wheats, and the lowest values for Sar. *italicum*.

In the case of Sar. *leucurum*, Sar. *italicum*, old wheats and pure lines, the highest average number of amplified alleles per locus was observed in 1B (long arm), 6B (short arm) and 6B (long arm), respectively, with average values of 9.0, 6.5, 2.5 and 2.3, respectively. It's worth noting that old wheats and modern pure lines exhibited the highest percentage of polymorphisms detected by SSR markers in the B genome (Table 8), likely originating from a species, or several species closely related to *Aegilops speltoides* Tausch, a cross-pollinating species; while the Sar. *leucurum* and Sar. *italicum* sets had the highest percentage of polymorphisms in the A genome, which can be traced back to diploids like *T. urartu* Thumanjan ex Gandilyan (Wang *et al*, 2007).

A similar clustering pattern was observed when analyzing both the morphological traits (Figure 2) and molecular marker profiles (Figure 3) of all 53 genotypes. This analysis employed a hierarchical grouping method, without missing data in the dataset. The resulting dendrogram revealed four major clusters (Figure 3). Cluster 1A comprised all 13 Sar. *italicum* genotypes, 9 out of 22 Sar. *leucurum* genotypes of Abruzzo and 1 Sar. *leucurum* genotype from Puglia. The second major cluster, 1B, could be further subdivided into two subclusters: 1B1 and 1B2. The 1B1 cluster included two subgroups: the first subgroup contained four closely related breeding lines ('Colosseo', 'Simeto', 'Ciccio' and 'Capeiti'), two Sar. *leucurum* genotypes, and the old wheat 'Vallelunga pubescent'; the second subgroup was larger, consisting of modern varieties ('Duilio', 'Realforte', 'Svevo' and the 'Saragolla' pure line), some old wheats ('Russello', 'Scorsonera', 'Sammartinara', 'Cappelli' and the *T. turanicum* line), along with eight Sar. *leucurum* genotypes from Puglia or Abruzzo. Cluster 1B2 included six strongly related Sar. *leucurum* genotypes: one was from Puglia and five from Abruzzo.

Discussion

The morphological and genetic characterization of autochthonous wheats serves the dual purpose of safeguarding the economic interests of farmers and increasing consumers' trust in the origin and quality of food products entering the market (Terzi *et al*, 2005).

A recent morphological analysis conducted by the SARAB project on wheat crops in 11 farms across

Table 3. Values of the standardized coefficients for the canonical variables included in the discriminant functions, obtained using the four groups of genotypes (Sar. *leucurum*, Sar. *italicum*, pure lines and old wheats) as classification categories.

	root 1	**root 2**
Glume hairiness	-0,99	-0,26
Heading date (days from 1/4)	0,05	-1,39
Upper neck glaucosity	0,73	-0,98
Growth habit	-0,48	0,64
Grain shape	-0,06	-0,50
1,000 seed weight	0,03	0,44
Spike glaucosity	-0,43	-0,05
Spike shape	-0,27	-0,08
Glume colour	0,18	-0,57
Awn colour	-0,40	0,38
Plant height	0,48	-0,17
Flag leaf lower side glaucosity	0,19	-0,49
Grain weight/plant	0,09	-0,36
Straw: pith in cross section	-0,23	0,20
Lower glume: length of beak	-0,36	-0,29
Spike density	0,21	0,31
Eigenvalue	12.4	3.4
Explained cumulative variance (%)	72.6	92.9

Table 4. Pairwise square Mahalanobis distances (plain text) and probability values (italics) for the contrasts between the four groups of genotypes.

	Sar. *leucurum*	**Sar. *italicum***	**Pure lines**	**Old wheats**
Sar. *leucurum*		*0,0000*	*0,00036*	*0,00003*
Sar. *italicum*	69,48		*0,00000*	*0,00000*
Pure lines	23,76	78,44		*0,01962ns*
Old wheats	27,19	68,77	20,95	

Table 5. List of SSR markers used for molecular analysis, number of alleles and Polymorphic Information Content (PIC) obtained for each marker in the 39 Saragolla wheat lines. A, A genome; B, B genome; L, long arm; S, short arm.

	Marker	**Chromosome**	**No. of alleles**	**PIC**
1	gwm311	2A(L)	11	0,83
2	gwm1042	3A(L)	6	0,55
3	gwm299	3B(L)	7	0,74
4	barc45	3A(S)	4	0,77
5	gwm495	4B(S)	8	0,82
6	gwm1093	4A(S)	14	0,83
7	gwm1084	4B(L)	8	0,75
8	gwm865	5A(L)	8	0,80
9	gwm154	5A(S)	9	0,73
10	gwm499	5B(L)	11	0,80
11	gwm1017	6A(L)	6	0,80
12	gwm459	6A(S)	12	0,91
13	gwm193	6B(L)	4	0,76
14	gwm60	7A(S)	7	0,52
15	wmc606	7B(S)	11	0,91

Table 6. Average number of SSR alleles per genotype. Only alleles occurring with a frequency higher than 11 ($p < 0.05$) are included. The codes from S1 to S11 indicate the cultivation sites (see Supplemental Table 1) of Saragolla wheats from Abruzzo and S12 indicates the cultivation site in Puglia. The extra letters and numbers differentiate the genetically characterized plants within each site. Old wheats and pure lines are described in Supplemental Table 2.

Group	Genotypes	Average allele no.	Group	Genotypes	Average allele no.
Sar. *leucurum*	S1F3	13	Sar. *italicum*	S1H3	16
	S2E	13		S2P2A	15
	S2F	14		S3P27F	20
	S3P27A	16		S4P23B	14
	S3P3B	15		S5P49B	21
	S3P3I	16		S6P6A	16
	S4P4B	16		S7P43C	16
	S4P4A	16		S8P8D	16
	S4Z	15		S8P9A	16
	S5P5C	16		S10P50C	15
	S5P31D	16		S11P11A	17
	S6P21D	15		S7A	16
	S6P41C	15		S7C	14
	S8A1	14	Old wheats	'Sammartinara'	15
	S8R1	14		'Realforte'	12
	S8	15		'Scorsonera'	17
	S9P22D	15		'Russello'	16
	S9P9D	23		'Vallelunga Pubescent'	13
	S9P22A	15		'Turanicum'	17
	S10P32A	14	Modern pure lines	'Capeiti'	14
	S11P24A	15		'Ciccio'	15
	S12P12A	14		'Simeto'	15
	S12P54B	16		'Colosseo'	15
	S12P35A	14		'Duilio'	16
	S12P54A	14		'Saragolla'	15
	S1E4	15		'Cappelli'	13
				'Svevo'	16

Table 7. Genetic diversity indices over 15 SSR loci for all 53 Italian genotypes tested in the study, as well as for the four groups categorized by 'Saragolla' botanical variety or control group (Means±SD). PIC, Polymorphic Information Content.

	Shannon's information index	Percentage of polymorphic loci	Nei's gene diversity	PIC
Sar. *leucurum*	0.19±0.19	71.4	0.10±0.13	0.902±0.08
Sar. *italicum*	0.08±0.18	22.2	0.05±0.13	0.097±0.08
Old wheats	0.17±0.22	39.7	0.10±0.14	0.929±0.10
Modern pure lines	0.19±0.19	61.1	0.11±0.12	0.764±0.11
All genotypes	0.28±0.20	97.63	0.17±0.15	0.673±0.35

Table 8. Percentage of polymorphism detected by SSR in A and B genomes of four groups of genotypes

	'Saragolla' *leucurum*	'Saragolla' *italicum*	Old Italian wheats	Modern pure lines
Genome A	54,1	54,5	46,2	48,5
Genome B	45,9	45,5	53,8	51,5

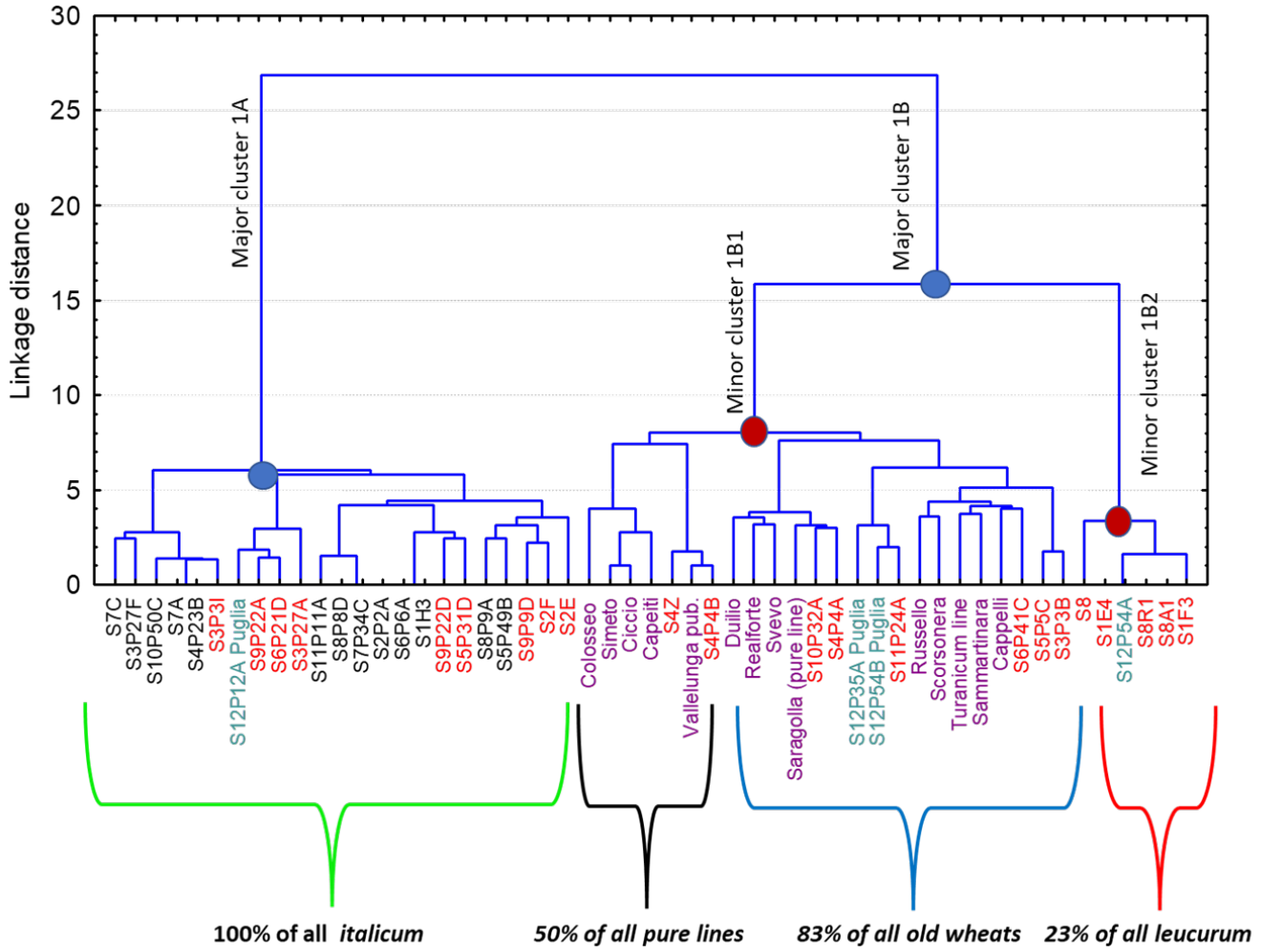

Figure 3. SSR-based genetic distances among the 53-genotype pool, through Ward's clustering. The percentage of the prevalent type of genotypes included in each cluster is indicated. Characters: red = Sar. *leucurum/affine* genotypes from Abruzzo; blue = Sar. *leucurum* from Puglia; black = Sar. *italicum/apulicum* genotypes from Abruzzo; violet = modern and old varieties. Codes are as in Table 6.

the Abruzzo Region revealed significant morphological diversity within the cultivated 'Saragolla' variety (Rascio *et al*, 2022). This diversity poses a challenge in accurately defining their distinct traits.

The morphotypes that are both quantitatively and widely spread, across most of the 11 wheat farms included in the SARAB project, belong to the Sar. *italicum/apulicum* or the Sar. *leucurum/affine* botanical varieties (Rascio *et al*, 2022). The genetic characterization described here aimed to validate whether the observed morphological similarity among the genotypes from Abruzzo corresponds to genetic similarity and to develop a tool to differentiate them.

The results presented here confirm the efficacy of SSR markers to characterize wheat genotypes (Wang *et al*, 2007; Dagnaw *et al*, 2023). In fact, the 15 SSR primers used in this experiment showed detectable polymorphisms in all the 53 genotypes and their mean polymorphic information content (PIC = 0.77) makes their use very informative. 'Saragolla' genotypes, belonging to the *italicum/apulicum* botanical varieties can be eligible as conservation varieties. These genotypes are widely cultivated across the Abruzzo Region, and they also exhibit a noteworthy genetic similarity, as indicated by the low values of Nei's gene diversity and Shannon's information index. Additionally, Sar. *italicum* genotypes display both a distinct phenotype and genotype in comparison to Sar. *leucurum*, older wheats and pure lines.

It is worth noting that Ward's clustering analysis of the 53-genotype pool revealed significant genetic diversity between Sar. *italicum* and Sar. *leucurum* genotypes. Out of the 26 Sar. *leucurum* genotypes examined, only 9 displayed a significant genetic resemblance to Sar. *italicum*. Six were categorized within the broader groups of pure lines and old wheats, while five formed a distinct group of genotypes very closely related genetically, but distinct from all others. The clustering analysis also revealed a stronger genetic similarity between most Sar. *leucurum* genotypes and three out of four 'Saragolla' genotypes from Puglia and a somewhat lesser degree of affinity with the oldest

genotypes. Expanding on the hypothesis (Zeven, 1998), that factors such as geographic distance, environmental conditions and the selection made by farmers can shape the genetic composition of local wheats, it is plausible to infer that the migration of wheat commenced from Sicily. In fact, the *leucurum* genotypes were documented in Southern Italy as early as the beginning of the 1900s (Percival, 1921; De Cillis, 1927; Draghetti, 1927) and were likely among the oldest cultivated in Sicily (Porceddu *et al*, 1981). From Sicily, it is plausible that these wheats initially spread to the nearby region of Puglia and then reached Abruzzo where the cross with indigenous wheat occurred as well as the selection of alleles improving adaptability, productivity and quality. In terms of affinities with 'Kamut', the results suggest that the Sar. *italicum* genotypes, despite having elongated and large seeds similar to *T. turanicum*, formed a distinct cluster and showed a closer genetic relationship to Sar. *leucurum* and old durum wheats.

Conclusions

The 'Saragolla' wheat presently grown in the Abruzzo Region is characterized by its rich diversity, predominantly comprising Sar. *italicum/apulicum* and Sar. *leucurum*-like durum wheats, some of which share close morphological and genetic traits. Recently, these genotypes have been officially registered as 'Saragolla' conservation varieties from Abruzzo. The genetic distance observed among 39 'Saragolla' genotypes, representative of only two out of the nine previously identified botanical varieties of durum wheat, exceeded that found among the other 13 modern, or old durum wheats used in this study, which differ for age of cultivation and origin. Consequently, at the Maiella National Park seedbank, targeted *ex situ* conservation measures will be implemented to preserve the currently cultivated populations. A more extensive genetic characterization will enable the assessment of existing variability within the 'Saragolla' landrace, for adaptive and agronomically valuable traits, useful for breeding improved varieties.

Supplemental data

Supplemental Table 1. Geographic coordinates of the cultivation sites for the 12 Saragolla wheats.
Supplemental Table 2. Passport details of old and modern wheats used in the present work.
Supplemental Table 3. Morphological traits of Saragolla *leucurum*, *italicum*, and modern and old wheats.

Acknowledgements

The authors wish to thank the farm owners who provided the studied materials, Dr Maurizio Odoardi and Dr Daniela Codoni (Department of Rural Development and Fisheries Policies – Promotion of Knowledge and Innovation in Agriculture – DPD022) of the Abruzzo Regional Authorities, for their valuable contribution to the 'SARAB project: Characterization of ancient 'Saragolla' populations from the Abruzzo Region'. The authors also wish to thank Mr Leonardo Morrone and Mr Vito De Gregorio for their valuable assistance in conducting the experimental trials.

Funding

This work was in part supported by the Abruzzo Region.

Author contributions

AR, study conception and manuscript draft; VDS, molecular analysis; LG, analysis and interpretation of results; SP, data collection; MT, field trials; GS, manuscript revision.

Conflict of interest statement

The authors declare no conflict of interest.

References

Anderson, J. A., Churchill, G. A., Sutrique, J. E., Tanksley, S. D., and Sorrells, M. E. (1993). Optimizing parental selection for genetic linkage maps. *Genome* 36, 181–186. doi: https://doi.org/10.1139/g93-024

Dagnaw, T., Mulugeta, B., Haileselassie, T., Geleta, M., Ortiz, R., and Tesfaye, K. (2023). Genetic Diversity of Durum Wheat (*Triticum turgidum* L. ssp. *durum*, Desf). *Genes* 14(1155). doi: https://doi.org/10.3390/genes14061155

De Cillis, E. (1927). I grani d'Italia volume 173. (Roma: Tipografia della Camera dei deputati).

Draghetti, A. (1927). Forme e limiti dello xerofitismo nel frumento (Forlì: Tipografie Valbonesi).

Eccellenze D'abruzzo (2024). Saragolla: il grano dei faraoni in Abruzzo. url: https://www.eccellenzedabruzzo.it/grano-saragolla/.

EU (2009). Commission Directive 2009/145/EC of 26 November 2009 providing for certain derogations, for acceptance of vegetable landraces and varieties which have been traditionally grown in particular localities and regions and are threatened by genetic erosion and of vegetable varieties with no intrinsic value for commercial crop production but developed for growing under particular conditions and for marketing of seed of those landraces and varieties. url: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02009L0145-20130828.

FAO (2009). The International Treaty on Plant Genetic Resources for Food and Agriculture (Rome, Italy: Food and Agricultural Organization of the United Nations). url: https://www.fao.org/3/a-i0510e.pdf.

Fiore, M. (2013). 'Saragollo Triticum Apulum - La produzione cerealicola nel territorio di Torremaggiore nei secoli XVI e XVII. Cooperativa Agricola Fortore, Torremaggiore.

Fiore, M. C., Mercati, F., Spina, A., Blangiforti, S., Venora, G., Dell'acqua, M., and Sunseri, F. (2019). High-throughput genotype, morphology, and quality traits evaluation for the assessment of genetic

diversity of wheat landraces from sicily. *Plants* 8(116). doi: https://doi.org/10.3390/plants8050116

Marone, D., Laido, G., Gadaleta, A., Colasuonno, P., Ficco, D. B., Giancaspro, A., Giove, S., Panio, G., Russo, M. A., De Vita, P., and Cattivelli, L. (2012). A high-density consensus map of A and B wheat genomes. *Theoretical and Applied Genetics* 125, 1619–1638. doi: https://doi.org/10.1007/s00122-012-1939-y

Marone, D., Olmo, A. I. D., Laidò, G., Sillero, J. C., Emeran, A. A., Russo, M. A., Ferragonio, P., Giovanniello, V., Mazzucotelli, E., De Leonardis, A. M., and De Vita, P. (2009). Genetic analysis of durable resistance against leaf rust in durum wheat. *Molecular Breeding* 24, 25–39. doi: https://doi.org/10.1007/s11032-009-9268-9

Negri, V., Maxted, N., and Veteläinen, M. (2009). European landrace conservation, an introduction. In *European landraces: on-farm conservation, management and use,* ed. Veteläinen, M., Negri, V., and Maxted, N., (Rome, Italy: Bioversity International), volume 15 of *Bioversity International Technical Bulletin*.

Pallante, G., Drucker, A. G., and Sthapit, S. (2016). Assessing the potential for niche market development to contribute to farmers livelihoods and agrobiodiversity conservation: Insights from the finger millet case study in Nepal. *Ecological Economics* 130, 92105–92105. doi: https://doi.org/10.1016/j.ecolecon.2016.06.017

Percival, J. (1921). The Wheat Plant: a Monograph (London: Duckworth and Co).

Petrović, S., Marić, S., Čupić, T., Rebekić, A., and Rukavina, I. (2017). Assessment of molecular and phenotypic diversity among winter wheat cultivars. *Genetika* 49, 583–598. doi: https://doi.org/10.2298/GENSR1702583P

Piergiovanni, A. R. (2013). Capillary electrophoresis: a useful tool for the management of plant genetic resources. *Analytical and Bioanalytical Chemistry* 405, 481–491. doi: https://doi.org/10.1007/s00216-012-6127-z

Porceddu, E., Vannella, S., and Perrino, P. (1981). Character analysis and numerical classification in a wheat collection from Sicily. *Die Kulturpflanze* 29, 251–266. doi: https://www.cabidigitallibrary.org/doi/full/10.5555/19831622160

Rascio, A., Beleggia, R., Platani, C., Nigro, F., Codianni, P., De Santis, and Fragasso, M. (2016). Metabolomic diversity for biochemical traits of *Triticum* sub-species. *Journal of Cereal Science* 71, 224–229. doi: https://doi.org/10.1016/j.jcs.2016.08.009

Rascio, A., Codianni, P., Paone, S., Fiorillo, F., and Marone, D. (2021). SARAB project, characterization of ancient Saragolla populations from Abruzzo Region. *Tecnica Molitoria* 72, 37–42.

Rascio, A., Fiorillo, F., Paone, S., De Santis, G., and Sorrentino, G. (2022). Quantitative botanical characterization of Saragolla wheat landraces from Abruzzo and Puglia Regions of Italy. *Plant Genetic Resources* 20, 434–441. doi: https://doi.org/10.1017/S1479262123000345

Rascio, A., Picchi, V., Naldi, J. P., Colecchia, S., De Santis, G., Gallo, A., and De Gara, L. (2015). Effects of temperature increase., through spring sowing, on antioxidant power and health-beneficial substances of old and new wheat varieties. *Journal of Cereal Science* 61, 111–118. doi: https://doi.org/10.1016/j.jcs.2014.09.010

Scarascia-Mugnozza, T. G. (2005). The contribution of Italian wheat geneticists, from Nazareno Strampelli to Francesco D'Amato. In Tuberosa, R., Phillips, R. L., and Gale, M., Proceedings of the International Congress on: The Wake of the Double Helix, 53-75.

Spataro, G. and Negri, V. (2013). The European seed legislation on conservation varieties: focus, implementation, present and future impact on landrace on farm conservation. *Genetic resources and Crop Evolution* 60, 2421–2430. doi: https://doi.org/10.1007/s10722-013-0009-x

Terzi, V., Morcia, C., Gorrini, A., Stanca, A. M., Shewry, P. R., and Faccioli, P. (2005). DNA-based methods for identification and quantification of small grain cereal mixtures and fingerprinting of varieties. *Journal of Cereal Science* 41, 213–220. doi: https://doi.org/10.1016/j.jcs.2004.08.003

UPOV (2012). Test Guideline . url: https://www.upov.int/edocs/tgdocs/en/tg120.pdf.

Wang, H., Chen, P., and Liu, D. (2007). Assessment of genetic diversity of Yunnan, Tibetan, and Xinjiang wheat using SSR markers. *Journal of Genetics and Genomics* 34, 623–633. doi: https://doi.org/10.1016/s1673-8527(07)60071-x

Yeh, F. C., Yang, R. C., and Boyle, T. (1999). Microsoft Window-based freeware for population genetic analysis (POPGENE). Version 1.31. url: https://sites.ualberta.ca/~fyeh/popgene.html.

Zeven, A. C. (1998). Landraces, a review of definitions and classifications. *Euphytica* 104, 127–139. doi: https://doi.org/10.1023/A:1018683119237