



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente (DAFNAE)

CORSO DI DOTTORATO DI RICERCA IN: CROP SCIENCE
CICLO XXXIV

**MULTIPLE GENOMIC TECHNOLOGIES APPLIED TO
GERMPLASM CHARACTERIZATION AND PLANT VARIETY
BREEDING AND PROTECTION**

Coordinatore: Ch.mo Prof. Claudio Bonghi

Supervisore: Ch.mo Prof. Gianni Barcaccia

Co-Supervisore: Ch.mo Dott. Fabio Palumbo

Dottorando: Francesco Scariolo

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

(signature/name/date)

Francesco Scariolo

29/10/2021



A copy of the thesis will be available at <http://paduaresearch.cab.unipd.it/>

Dichiarazione

Con la presente affermo che questa tesi è frutto del mio lavoro e che, per quanto io ne sia a conoscenza, non contiene materiale precedentemente pubblicato o scritto da un'altra persona né materiale che è stato utilizzato per l'ottenimento di qualunque altro titolo o diploma dell'università o altro istituto di apprendimento, a eccezione del caso in cui ciò venga riconosciuto nel testo.

(firma/nome/data)

Francesco Scariolo

29/10/2021



Una copia della tesi sarà disponibile presso <http://paduaresearch.cab.unipd.it/>

Index

Riassunto generale.....	1
General abstract	5
Molecular Hallmarks to Exploit Neglected Genetic Resources of Common Bean	7
0. <i>Abstract</i>	9
1. <i>Introduction</i>	11
2. <i>Materials and Methods</i>	13
3. <i>Results</i>	18
4. <i>Discussion</i>	25
5. <i>References</i>	30
6. <i>Supplementary material</i>	34
Molecular relationships and phylogenies of Venetian Radicchio (leaf chicory, <i>Cichorium intybus</i> subsp. <i>intybus</i> var. <i>foliosum</i>, $2n=2x=18$) varietal groups.....	41
0. <i>Abstract</i>	43
1. <i>Introduction</i>	45
2. <i>Materials and Methods</i>	48
3. <i>Results</i>	51
4. <i>Discussion</i>	59
5. <i>Conclusions</i>	60
6. <i>References</i>	62
7. <i>Supplementary materials</i>	65
Molecular characterization and genetic structure evaluation of breeding populations of fennel through microsatellite genotyping	67
0. <i>Abstract</i>	69
1. <i>Introduction</i>	71
2. <i>Materials and methods</i>	72
3. <i>Results and Discussion</i>	74
4. <i>Conclusions</i>	85
5. <i>References</i>	87
Genotyping Analysis by RAD-Seq Reads is Useful to Assess the Genetic Identity and Relationships of Breeding Lines in Lavender Species Aimed at Managing Plant Variety Protection	89
0. <i>Abstract</i>	91
1. <i>Introduction</i>	93

2.	<i>Materials and Methods</i>	94
3.	<i>Results</i>	97
4.	<i>Discussion</i>	107
5.	<i>References</i>	111
6.	<i>Supplementary materials</i>	117
7.	<i>Informatic material</i>	127
	Publications :.....	129
	Congresses :.....	130

Riassunto generale

Al giorno d'oggi, la genomica gioca un ruolo importante nel miglioramento delle colture, nella tracciabilità delle cultivar e nella protezione delle varietà vegetali. La disponibilità di strumenti genomici per caratterizzare molecularmente le specie coltivate è in continuo aumento, e dimostra la sua utilità per molti scopi, come la selezione delle migliori linee parentali per la costituzione di nuovi ibridi F1, la caratterizzazione di varietà locali di una specie per studiarne il germoplasma locale, la valutazione del pool genico di origine di piante coltivate, la tracciabilità delle cultivar per proteggere i diritti dei coltivatori, o la determinazione della distinguibilità genetica e delle relazioni genetiche di diversi biotipi coltivati. I principali strumenti genomici utilizzati in questo campo sono i marcatori molecolari, che consistono in brevi regioni genomiche polimorfiche utili per determinare le differenze o le somiglianze genetiche tra individui, o per stimare la loro omozigosi e altre importanti statistiche genetiche. Queste informazioni possono poi essere utilizzate per diversi fini, come è già stato menzionato sopra, che possono migliorare l'agricoltura e la tracciabilità dei suoi prodotti, aspetti importanti che vengono considerati non solo dai coltivatori, ma anche dai consumatori. Questo progetto di ricerca si propone di illustrare le potenzialità degli strumenti e delle piattaforme genomiche per fini applicativi nel miglioramento genetico di piante coltivate e nella tracciabilità delle colture. Gli argomenti presentati riguarderanno il miglioramento genetico assistito da marcatori (MAB), che consiste nel migliorare le varietà di colture selezionando le migliori linee parentali da utilizzare nello sviluppo di ibridi F1 destinati al commercio, la caratterizzazione del germoplasma finalizzata a indagare la genetica delle varietà o dei biotipi locali, e la protezione delle varietà vegetali (PVP) incentrata sulla tracciabilità delle cultivar per proteggere i diritti dei coltivatori dalle frodi e dalle appropriazioni indebite da parte dei loro concorrenti.

Il primo capitolo di questo progetto di ricerca è incentrato sull'uso di marcatori molecolari microsatelliti e *Single Nucleotide Polymorphism* (SNP) per la genotipizzazione e la caratterizzazione aplo-tipica gene-specifica di linee d'élite italiane e di varietà di nicchia locali del Veneto di fagiolo comune (*Phaseolus vulgaris* L.). Le analisi effettuate sono state condotte per calcolare la similarità genetica (GS) e l'omozigosi delle suddette tipologie di varietà coltivate con fini di miglioramento genetico. Inoltre, è stata stimata la struttura genetica della popolazione nel suo insieme, oltre al pool genico di origine delle accessioni analizzate. Questo lavoro evidenzia l'importanza degli strumenti genomici per la caratterizzazione del germoplasma attraverso la ricostruzione degli antenati delle popolazioni in questa specie proponendo un metodo combinato basato su loci marcatori neutri e gene-specifici.

La seconda parte di questo lavoro si è concentrata sull'analisi di molteplici popolazioni coltivate di "Radicchio" (*Cichorium intybus* L.), tipiche della regione Veneto (Italia), utilizzando marcatori molecolari dominanti, rispettivamente *Random Amplification of Polymorphic DNA* (RAPD) e *Amplified Fragment Length Polymorphism* (AFLP), per studiare le relazioni genetiche e le origini dei più importanti biotipi di questa coltura orticola. Questa ricerca è stata eseguita utilizzando vecchie popolazioni di agricoltori che sono state ottenute dalla selezione fenotipica nell'era pre-icedente al miglioramento genetico di questa coltura. Come comunemente fatto dagli agricoltori, prima che le aziende sementiere iniziassero a pianificare strategie di selezione in radicchio, i semi venivano raccolti dalle piante che mostravano i migliori fenotipi e venivano poi utilizzati per la semina dell'anno successivo. In questo modo, a causa dell'auto-incompatibilità di questa specie, si è perpetuata la variazione genetica che rende le popolazioni localmente coltivate un importante deposito di germoplasma da utilizzare per la selezione di nuove varietà che siano più adattabili all'agricoltura locale sostenibile. L'uso di marcatori dominanti ha limitato le informazioni ottenibili sulle popolazioni analizzate in termini di omozigosi, ma ha permesso di stimare la struttura genetica e le relazioni dei biotipi studiati, dimostrando così la loro discendenza e distinguibilità.

La terza parte del lavoro è incentrata sul mostrare le potenzialità delle tecnologie di genotipizzazione attraverso marcatori molecolari co-dominanti per caratterizzare più popolazioni di finocchio (*Foeniculum vulgare* Mill.). A causa della sua elevata auto-compatibilità, le strategie di miglioramento in questa coltura sono comunemente eseguite utilizzando tre linee parentali. La prima è il mantenitore (M), che viene selezionato per il suo fenotipo e viene propagato attraverso autoimpollinazione, la seconda è la linea maschio-sterile (CMS) che viene utilizzata per trasferire la maschio-sterilità citoplasmatica nell'ideotipo del genitore materno incrociandolo con la linea del mantenitore, e la terza linea parentale è l'impollinatore (P), ancora una volta riprodotto per autoimpollinazione, che viene incrociato con la progenie CMS-mantenitore per ottenere ibridi F1 maschio-sterili (H). Nel primo capitolo di questa tesi, sono state analizzate otto popolazioni di finocchio e le loro tre linee parentali (CMS, M e P), insieme alla relativa progenie F1 (H), sono state genotipizzate utilizzando 12 marcatori *Simple Sequence Repeats* (SSR), o microsatelliti, per calcolare diverse statistiche genetiche e ricostruire la composizione genetica della macro-popolazione. Questa ricerca dimostra l'informatività del pannello di marcatori SSR utilizzato nelle analisi di genotipizzazione per i progetti di miglioramento genetico assistito da marcatori molecolari in finocchio.

Il quarto e ultimo capitolo di questo studio era volto a studiare la possibile applicazione del *genotyping-by-sequencing* (GBS) attraverso il sequenziamento del DNA associato a siti di restrizione (RADseq) nelle specie di lavanda. Questa ricerca si è basata sull'analisi di quindici campioni di lavanda appartenenti a due diverse specie, *Lavandula stoechas* L. e *Lavandula pedunculata* (Mill.) Cav.. La presente ricerca ha lo scopo di stimare la somiglianza genetica e l'omozigosi delle linee da miglioramento genetico di lavanda, ma ha anche cercato di identificare dei loci marcatori da poter utilizzare nella tracciabilità delle varietà e nella protezione dei diritti dei costitutori, oltre alla possibilità di identificare incroci interspecifici tra le due specie analizzate. Un aspetto importante di questo studio è stato il consistente approccio bioinformatico utilizzato e l'enorme quantità di loci marcatori analizzati (oltre 16 mila).

General abstract

Nowadays, genomics plays an important role in crops' breeding, cultivars' traceability, and plants variety protection. The availability of genomic tools to molecularly characterize crop species is continually increasing, and demonstrates its usefulness for many purposes, like selecting the best breeding parental lines for the constitution of new F1 hybrids, characterizing multiple landraces to investigate a species local germplasm, assessing the gene pool of origin of cultivated plants, tracing cultivars in order to protect the plant breeders' rights, or determining the genetic distinctiveness and genetic relationships of different cultivated biotypes. The main genomic tools used in this field are DNA molecular markers, which consist in short polymorphic genomic regions useful in determining the genetic differences or similarities among individuals or in estimating their homozygosity and other important genetic statistics. These information can then be used for different purposes, as it has already been mentioned above, which can improve agriculture and its products' traceability, which are important aspects considered not only by cultivators, but also by costumers. This research project aims at illustrating the potentials of genomic tools and platforms for applicative purposes in plant breeding and crops traceability. The topics presented will concern markers-assisted breeding (MAB), which consists in improve crops varieties by selecting the best parental lines to be used in the development of commercially destined F1 hybrids, germplasm characterization aimed at investigating the genetics of landraces or local biotypes, and plant variety protection (PVP) focused on the traceability of cultivars to protect breeders' rights from frauds and embezzlements from their competitors.

The first chapter of this research project is focused on the use of microsatellites and Single Nucleotide Polymorphism (SNP) molecular markers for genotyping and gene-specific haplotyping characterizations of Italian elite lines and Venetian niche landraces of common bean (*Phaseolus vulgaris* L.). The analyses performed were conducted to calculate the genetic similarity (GS) and homozygosity of the above-mentioned typologies of cultivated varieties for breeding purposes. Moreover, the genetic structure of the core collection was estimated, together with the gene pool of origin of the analysed accessions. This work highlights the importance of genomic tools for germplasm characterizations through the reconstruction of the populations' ancestors in this species proposing a combined method based on neutral and gene-specific marker loci.

The second part of this work focused on the analysis of multiple locally cultivated populations of "Radicchio" (*Cichorium intybus* L.), typical from the Veneto region (Italy), using dominant molecular markers, Random Amplification of Polymorphic DNA (RAPD) and Amplified Fragment Length Polymorphism (AFLP) respectively, to study the genetic relationships and origins of the most important biotypes of this horticultural crop. This research was performed using old

farmers' populations that were obtained by phenotypic selection in the pre-breeding era of this crop. As commonly done by farmers, before seed companies started planning breeding strategies in chicory, seeds were collected from plants showing the best phenotypes that were then used for the next year seedling. In this way, due to the self-incompatibility of this species, genetic variation was perpetuated that makes local cultivated populations an important germplasm repository to be used for breeding new varieties that will be more adaptable and suitable for local sustainable agriculture. The use of dominant markers limited the obtainable information on the analysed populations in terms of homozygosity, but it enabled the estimation of the genetic structure and relationships of the studied biotypes, thus demonstrating their ancestry and distinctiveness.

The third part of the work is focused on showing the potentials of genotyping technologies through codominant molecular markers to characterize multiple populations of fennel (*Foeniculum vulgare* Mill.). Due to its high self-compatibility, breeding strategies in this crop is commonly performed by using three parental lines. The first one is the maintainer (M), which is selected for its phenotype and is propagated through self-pollination, the second one is the male-sterile line (CMS) that is used to transfer cytoplasmic male-sterility into the maternal parent ideotype by crossing it with the maintainer line, and the third parental line is the pollinator (P), once again reproduced by self-pollination, that is crossed with the CMS-maintainer progeny to obtain male-sterile F1 hybrids (H). In the first chapter of this thesis, eight breeding populations of fennel were analysed and the three parental lines (CMS, M and P), along with their progenies (H), were genotyped using 12 Simple Sequence Repeats (SSR) markers, or microsatellites, to calculate several genetic statistics and to reconstruct the genetic composition of the core collection. This research demonstrates the informativeness of the used SSR markers panel in genotyping analyses for MAB projects in fennel.

The fourth and last chapter of this study was aimed at study the possible application of genotyping-by-sequencing (GBS) through Restriction-site Associated DNA sequencing (RADseq) in lavender species. This research was based on the analyses of fifteen samples of lavender belonging to two different species, *Lavandula stoechas* L. and *Lavandula pedunculata* (Mill.) Cav., respectively. The present research has its purpose at estimating the genetic similarity and homozygosity of breeding lines of lavender, but it also looked forward to identifying putative marker loci to be used in variety traceability and breeders' rights protection, or PVP, plus the possibility of identifying interspecific crosses between the two analysed species. An important aspect of this study was the consistent bioinformatics approach used and the massive amount of marker loci analysed (over 16 thousand).

Chapter I

Molecular Hallmarks to Exploit Neglected Genetic Resources of Common Bean

This chapter has been extracted from:

Sica, P., **Scariolo, F.**, Galvao, A., Battaglia, D., Nicoletto, C., Maucieri, C., ... & Barcaccia, G. (2021). Molecular Hallmarks, Agronomic Performances and Seed Nutraceutical Properties to Exploit Neglected Genetic Resources of Common Beans Grown by Organic Farming in Two Contrasting Environments. *Frontiers in plant science*, 12.

0. Abstract

Common bean (*Phaseolus vulgaris* L.) is an essential source of food proteins and an important component of sustainable agriculture systems around the world. Thus, conserving and exploiting the genetic materials of this crop species play an important role in achieving global food safety and security through the preservation of functional and serendipitous opportunities afforded by plant species diversity. Our research aimed to collect and perform molecular-genetic characterizations of common bean accessions of Venetian niche landraces (ancient farmer populations) and Italian elite lineages (old breeder selections). Molecular characterization with SSR and SNP markers grouped these accessions into two well-separated clusters that were linked to the original Andean and Mesoamerican gene pools, which was consistent with the outputs of ancestral analysis. Genetic diversity in the two main clusters was not distributed equally and the Andean gene pool was found to be much more uniform than the Mesoamerican pool. Additional subdivision resulted in subclusters, supporting the existence of six varietal groups. Accessions were selected according to preliminary investigations and historical records. On the whole, the genetic-molecular information put together for these univocal bean entries were used to select and transform the best accessions into commercial varieties (*i.e.*, pure lines) suitable for wider cultivation.

Keywords: *Phaseolus vulgaris* L.; local varieties; genetic diversity; genotyping; haplotyping; gene pool

1. Introduction

Legumes play an important role in addressing issues related to the environment, health, and food security and are also important due to their health benefits, such as preventing and helping manage hypercholesterolemia, hypertension [1], obesity, diabetes, and coronary conditions [2]. They are also a critical and affordable source of plant-based proteins, vitamins, and essential minerals such as calcium, magnesium, and zinc, contributing to the food security and nutrition of people around the world, especially subsistence smallholder farmers in developing countries [2]. In developed countries, vegetarians, vegans, and individuals following flexitarian diets tend to increase, and legumes are recommended as the main plant-based protein source [3].

The common bean (*Phaseolus vulgaris* L.) is a diploid ($2n = 2x = 22$) annual species belonging to the *Fabaceae* family grown worldwide for its edible green pods and dry seeds. Given the relative simplicity and the small dimension (650 Mb) of its genome, *P. vulgaris* provides a useful model for studying closely related species of agronomic interest. It is a predominantly self-pollinating plant, with occasional occurrence of insect-mediated cross-pollination [4]. Breeding strategies for the common bean rely on the selection of homozygous individuals for the development of pure lines of high agronomic value.

The domestication process of the common bean was a unique process that occurred in two geographically distinct regions simultaneously and in two partially isolated gene pools: Mesoamerican and Andean. Genetic evidence suggests that Mesoamerica is the centre of origin of the common bean, whereas the Andean population was derived as a consequence of a strong predomestication bottleneck. Despite having undergone independent domestication processes, both gene pools are partially sexually compatible and morphologically similar. Differences between the two gene pools have been revealed using different molecular markers, such as random amplified polymorphic DNA (RAPD) [5,6], amplified fragment length polymorphisms (AFLP) [7-9], and microsatellites or simple sequence repeat (SSR) markers [10]. More recently, single-nucleotide polymorphism (SNP) markers have been used to characterize genotype and haplotype diversity in common bean accessions, assaying both nuclear [4,11-13] and plastidial genomic regions [14].

Varieties of *P. vulgaris* are distributed worldwide and are cultivated in the tropics, subtropics, and temperate zones [15], showing great variability in terms of agronomic performance, seed size, shape and color, the relative duration of the reproductive cycle, and many other qualitative and quantitative traits [16]. This diversity enabled its cultivation in a wide range of cropping systems and environments, such as China, Eastern Africa, the Americas, the Middle East and Europe, with more than 40,000 varieties [17].

The first introduction of the common bean from Central/South America into Western Europe most likely took place in the sixteenth century [18]. A peculiarity of the European population of *P. vulgaris* is the high proportion (44%) of Mesoamerican and Andean hybrids. This could be explained by the presence of different landraces, which are traditionally cultivated in proximity to each other, facilitating occasional outcrossing and gene flow [19]. In Italy, beans made their first appearance in 1515 in a painting of Giovanni di Udine [20], whereas they were officially mentioned in historical documents by 1532, which was later recognized as the year of its introduction to the Italian Peninsula. In particular, in the Veneto region, the diffusion of the common bean occurred quickly, and currently, the cultivation of this pulse still has great economic relevance, especially in Belluno Province [21]. Bean cultivation gave rise to a long tradition that allowed the evolution of many landraces adapted to microclimates in restricted areas, representing a pastiche of cultures and traditions that provide an irremissible good for Italy that is being used in low-environmental impact agriculture [22]. However, as for other Venetian crops [23,24], local accessions have been gradually substituted by superior and genetically uniform commercial varieties [22,25]. In particular, after the 1950s, the large scale of breeding programs and the fast disappearance of landraces caused the disappearance of an unknown number of populations and the marginalization of others in private gardens. The commercial relevance of these landraces is generally limited, as the product is often sold only in local markets and appreciated and used in the preparation of local dishes [26]. This fact and the possibility of using these rustic and vigorous genotypes in breeding projects increase the importance of collecting and preserving those local accessions in germplasm, *in-situ*, or *ex-situ*, contributing to the improvement of food crops and preserving their genetic diversity [27]. Collecting, conserving, and preserving niche landraces from the Veneto region are some of the objectives of this study.

Although Veneto is not one of the primary domestication centres of the common bean, the collected material can be considered an example of serendipitous value from conserving landraces in a variety of places. Through their preservation, these niche landraces have undergone autochthony with minor adaptations, as they have been cultivated in isolation for centuries in a sort of ecotypization process. This process selected accessions, and peculiar gene combinations were able to achieve maximum adaptation to the new pedoclimatic and anthropic conditions. Long-term germplasm utilization and conservation practiced yearly by farmers, along with natural evolutionary processes, have therefore brought about the constitution of a regional multispecies germplasm that includes, among others, common beans.

High-quality and authentic agri-food products, which are mostly recognized as coming from organic farming or labeled with geographical indications, are increasing in Europe as consumers

consider the quality and its association with agro-ecological characteristics as some of the most important purchasing factors. Italian agriculture for quality food production is widely spread across the national territory, highlighting a major role in the European context that is favored due to its great variety in terms of pedoclimatic and orographic conditions, together with its cultural and traditional approach to food [28]. Organic farming systems cover approximately 15.2% of the national utilized agricultural area [29], generating almost 7.5% of the Italian agriculture value production [30]. In this scenario, interest in preserving and enhancing local genotypes to identify their agronomic and qualitative characteristics is evident. This is done to find varieties able to satisfy the main needs of organic farming that can often be considered pillars, namely, the rusticity of the genotypes and their ability to adapt to different environmental conditions and marginal contexts with reduced pedoclimatic performances. Furthermore, bean cultivation is fully part of this contest, considering its potential role not only in terms of the nutritional value of the grain produced but also in relation to the benefits that the cultivation of a legume can bring in terms of soil fertility, especially for organic agriculture, as recently stated by Regulation (EU) 2018/848.



























In this work, 26 bean accessions – 13 ancient local varieties (Venetian farmer populations) and 13 improved old lineages (Italian breeder selections) – were genetically characterized through molecular markers for assessing the population structure, genotype composition, and relationships among accessions. Overall, this information will be of great help for the socioeconomic valorisation of ancient local genetic resources that are well characterized molecularly in order to become suitable for implemented cultivation and agronomic practices in organic agricultural systems.

2. Materials and Methods

2.1. Plant material

Accessions assessed in this study were selected from the germplasm bank of the Department of Agronomy, Food, Natural Resources, Animal and Environment (DAFNAE) of the University of Padova in Legnaro, Italy. Among the species conserved in the DAFNAE germplasm, *P. vulgaris* was represented by 48 accessions (**Supplementary Table 1**), of which 26 were selected. A list of the names, identification numbers, some characteristics and pictures of the seeds of these 26 accessions is shown in **Table 1**. Of these 26 accessions, 13 were lowland and mountain-climbing Venetian local varieties (hereafter defined as farmer populations), and 13 were Italian old lineages (breeder selections), among which were seven dwarf and six climbing (Italian elite varieties) beans.

Table 1. Identification numbers, names, type, and growth habit of the 26 common bean accessions assessed in this study. (Each accession is followed by a picture of its seeds on the right.)

Id	Name	Type	Growth		Id	Name	Type	Growth	
1	Mangiatutto rampicante	Italian pre-commercial	Indeterminate (climbing)		23	Gialet	Venetian landrace	Indeterminate (climbing)	
2	Borlotto nano A	Italian elite line	Determined (Dwarf)		24	Posenati	Venetian landrace	Indeterminate (climbing)	
3	Borlotto nano B	Italian elite line	Determined (Dwarf)		25	Semi-rampicante abruzzese	Venetian landrace	Indeterminate (climbing)	
6	Fagiolo nano creso	Italian elite line	Determined (Dwarf)		26	Fasol dela nonna	Venetian landrace	Indeterminate (climbing)	
7	Blue lake sel. Gia	Italian pre-commercial	Indeterminate (climbing)		27	Maseleta rossa	Venetian landrace	Indeterminate (climbing)	
8	Anellino di Trento	Italian elite line	Determined (Dwarf)		28	Zia Orsolina	Venetian landrace	Indeterminate (climbing)	
9	Anellino giallo	Italian pre-commercial	Indeterminate (climbing)		29	Meraviglia di Venezia	Venetian landrace	Indeterminate (climbing)	
11	Bortollo lingua di fuoco 3	Italian pre-commercial	Indeterminate (climbing)		30	Secle	Venetian landrace	Indeterminate (climbing)	
12	Blue lake a grano nero	Italian pre-commercial	Indeterminate (climbing)		31	Della Clorinda	Venetian landrace	Indeterminate (climbing)	
17	Fagiolo nano valdarno	Italian elite line	Determined (Dwarf)		32	Pegaso	Venetian landrace	Indeterminate (climbing)	
18	Coco nain blanc precoce	Italian elite line	Determined (Dwarf)		33	SC-iosela	Venetian landrace	Indeterminate (climbing)	
19	Tondino abruzzese	Italian pre-commercial	Indeterminate (climbing)		34	D'oro (val di fiemme)	Venetian landrace	Indeterminate (climbing)	
20	Verdone del piave	Italian elite line	Determined (Dwarf)		36	Maron	Venetian landrace	Indeterminate (climbing)	

2.2. Genetic characterization

In total, 193 genomic DNA samples were extracted from young leaves of 26 *P. vulgaris* using the DNeasy 96 Plant kit (Qiagen, Hilden, Germany) following the instructions provided by the supplier. After extraction, the DNA quality and quantity were evaluated using a NanoDrop 2000c UV-Vis spectrophotometer (Thermo Fisher, Pittsburgh, PA, United States). The DNA

sample integrity was also checked by electrophoresis on a 2% agarose/1× TAE gel containing 1× Sybr® Safe DNA gel stain (Life Technology, Carlsbad, CA, United States).

An initial number of 24 SSR markers was chosen from the literature [19,31,32] based on their polymorphism information content (PIC), linkage map position, and sequence length. Tests on the amplification efficiency of the designed primer pairs were conducted in singleplex reactions on a subset of 8 samples of as many varieties. Amplifications were accomplished following the M13-tailed SSR method described by Schuelke [33] and modified as reported by Palumbo et al. [23,24] using 6-FAM, VIC, NED, and PET fluorophores. The 10 best SSR marker loci were selected and organized into two multiplexes based on the primer annealing temperature, amplicon size, amplification efficiency, and dimer formation tendency (**Table 2**). PCR was performed in a final volume of 20 µL containing 1x Platinum Multiplex PCR Master Mix (Thermo Scientific, Carlsbad, CA, United States), 5% GC Enhancer (Thermo Scientific), 0.25 µM of each tailed primer, 0.75 µM of each non-tailed primer, 0.5 µM of each labelled primer (Applied Biosystem, Carlsbad, CA, United States), 10 ng of DNA and sterile water. The fluorescently labelled PCR products were electrophoresed on an ABI 3730 DNA Analyzer (Applied Biosystems). Finally, the size of each fragment was determined by Peak Scanner software 1.0 (Applied Biosystems).

Table 2. List and basic information of the primer pairs for the Microsatellite (neutral) marker loci selected for DNA genotyping.

Marker	LG	Primer forward (5'-3')	Primer reverse (5'-3')	Motif	Size (bp)	PIC	Ta (°C)
BM200	1	TGGTGGTTGTTATGGGAGAAG	ATTGTCTCTGTCTATTCCTCCAC	(AG)10	221	0.89	56.0
AY1	1	ATCAGGGTCTGTCATGATCTG	CCTCCTCTCTTGTTCCT	(AT)5	203	0.70	55.0
GATS91	2	GAGTGCGGAAGCGAGTAG	TCCGTGTTCTCTGTCTGT	(GA)17	229	0.91	57.0
BM197	3	TGGACTGGTCGATACGAAG	CCCAGAAGATTGAGAACACCA	(GT)8	201	0.56	56.0
IAC52	4	TGCATGTATGTAGCGGTTTA	GTGGCTTTTGCTTTTGTAGTCA	(GA)11	203	0.64	55.0
IAC66	4	AATCACATCTTTAACCCAACAGGT	TTCCACTCCCTCCCTATCT	(GA)10	282	0.86	56.0
BM183	7	CTCAAATCTATTCCTGATCAGC	TCTTACAGCCTTGACAGACATC	(TC)14	149	0.84	55.5
BM210	7	ACCACTGCAATCCTCATCTTTG	CCCTCATCCTCCATTCTTATCG	(CT)15	166	0.88	56.4
PvM04	8	GGTTCCTCCTCTCTGC	GCGCCGTCTTTTGGTAG	(TTC)10	210	0.77	56.0
PvAG001	11	CAATCCTCTCTCTCATTCCAATC	GACCTGAAGTCGGTGTTCG	(GA)12	157	0.74	56.5

The SSR raw data were analysed with the POPGENE 32 software package v. 1.32 [34], and the following statistics were calculated: the observed homozygosity (Obs_Ho), the average number of alleles per locus (na), the effective number of observed alleles per locus (ne) and the SSR allele frequencies. All statistics were calculated for both the SSR loci used and the 48 accessions analysed. To determine the allele variability of the assessed marker loci, Nei's index was calculated

and assumed to be the polymorphism index content (PIC) [35]. Otherwise, considering the subpopulation later identified, the same index was used to express their heterozygosity grade.

Genetic similarity (GS) estimates were also calculated between individuals in all possible pairwise comparisons by applying Rohlf's simple matching (SM) coefficient using NTSYS v2.1 software [36]. The resulting similarity matrix was later used for the construction of a UPGMA dendrogram. The average similarity was also calculated within and among both the niche landraces and elite lineages, and a principal coordinates analysis (PCoA) graph was developed from the similarity matrix. Samples were then labelled on the basis of the results obtained by both STRUCTURE software, which was used for ancestry group reconstruction, and the UPGMA dendrogram.

From the initial core collection including 193 samples, a panel of 41 accessions of *P. vulgaris* belonging to 25 varieties (variety 19 was not considered due to the high number of missing loci) was selected for haplotyping analysis based on SNP variants. Samples were chosen to be representative of the internal variability of each population and with 100% homozygosity. This is preferred to obtain single haplotypes and because pure lines are used in breeding programs for the establishment of patented and stable varieties.

The nuclear target genes to be used for DNA haplotyping were preliminarily obtained from the scientific literature [37-39], choosing single-copy loci with functions related to plant and seed development and associated with abiotic stress resistance or tolerance. In particular, the final set of target genes was selected from among those reported by selected papers [40-42]. For each of these target genes, the pair of primers was designed and optimized to obtain fragments suitable for the subsequent sequencing analysis. The investigated marker regions are coding and/or noncoding fragments of eleven genes: β -1,3-endoglucanase (β -EG), heat shock transcription factor (HSF), β -glucan binding protein (β -GBP), phosphoenolpyruvate carboxylase (PEPC), late embryogenesis abundant protein (LEA), serine/threonine kinase (STK), nitrate reductase (NR), linoleate lipoxygenase (LOX), beta-amylase (β -AM) [42], histone H4 (H4) [40], and shatterproof (SHP) [41]. Whenever the indicated amplified fragment was larger than 900 bp, a new primer pair was designed (in bold in **Table 3**) to obtain a fragment shorter than 900 bp in substitution of the original fragment. Using the *P. vulgaris* reference genome hosted in the Phytozome database [43], primers were designed to match the exons of the selected genes, and where possible, introns were included to maximize the genetic variation estimate. Primer-BLAST [44] was used for primer design, and primers were purchased from Invitrogen (Invitrogen, Carlsbad, CA, United States).

Table 3. List and basic information on the Mendelian loci (expressed regions) selected for DNA haplotyping, including genomic localization (linkage group, LG), gene name, sequences of primer pairs used for PCR amplification, size of the amplified fragment (bp), temperature of annealing (T_a °C), and primer used for sequencing (highlighted in bold) in order to obtain amplicons <900 bp in length in substitution to the original ones.

LG	Gene	Primer forward	Primer reverse	Size	T_c	Sequencing
LG01	Beta-1,3-endoglucanase ¹	CAAACAAATGGGTGCAAGACAA	TCATGCTCTGGATGCTTCTG	670	59°	FW
LG02	Heat shock transcription factor ¹	CTTGTTGGGTATTGGGGTTA	AACCGGCTTCCGTCTATG	794	57°	FW
LG03	Histone H4 ²	ATCAGCCATGTCTGGAAGAGGAAA	TGCTTGAAAATGTCCAAATCATTG	502	57°	RV
LG04	Beta-glucan binding protein ¹	TGAGCCTTGTA CTTCC TACCC	AAGTTAGTTCTT GTT TACCCCGTG	683	57°	FW
LG05	Phosphoenolpyruvate carboxylase ¹	GCTGCAAGAGATGTACAACCAA	ATAACGAAAGGAAGATGGGTGA	618	57°	FW
LG06	Shatterproof ³	TTTGCTGATGTCGAGTT CATGC	ATTGTTGTTGTTGTTGTTGAGCTC	531	57°	FW
LG07	Late embryogenesis abundant protein ¹	GAGATGAAGGATGCGGCGAA	TCCTCCAGTTTCTCCTTGC	753	63°	FW
LG08	Serine/Threonine kinase ¹	TCCTGAACTCAGCCCCAAG	CCTGAGATCCGTCAACACCC	640	57°	RV
LG09	Nitrate reductase ¹	TGTGGAGCGTCTGGAGAAAC	TGCACACGTTCGTCTTCACT	889	57°	FW
LG10	Linoleate 13S-lipoxygenase ¹	TTAGCCCCATA CCAG TGCTT	TGAACAATCTGTTACCATGCAGT	628	59°	RV
LG11	Beta-amylase ¹	TGGTCCACTTTTGGCATCT	CCACAAAATCAAGGATGGGAAT	685	57°	RV

¹,Goretti et al. [42]; ²,McConnell et al. [40]; ³,Nanni et al. [41].

The PCR mixes were composed of 10 µl Mango mix [45], 0.25 µM of both primers, 20 ng genomic DNA, and sterile distilled H₂O to reach a final volume of 20 µl per reaction. Amplifications were carried out using the following conditions: initial denaturation for 2 min at 95°C, then 40 cycles of 30 s at 95°C, 30 s at the annealing temperature, and 1 min at 72°C, followed by one last final extension for 10 min at 72°C. Annealing temperatures specific to each primer pair can be found in **Table 2**. Amplification products were purified using exonuclease I and FastAP thermosensitive alkaline phosphatase (Thermo Fisher) according to the manufacturer's protocol for PCR product clean-up before sequencing. After purification, gene fragments were analyzed using Sanger sequencing by performing a single-strand reaction on ABI 3730XL with PHRED20. Sequencing data were analysed using Geneious 3.6.1 [46], and the resulting sequences were aligned using the Muscle algorithm implemented in MEGA-X [47]. The detected SNPs were then classified as synonymous or nonsynonymous mutations. Haplotype reconstruction was performed by identifying all the unique combinations of genes existing in the core collection. In addition, the haplotype number (Hn) and haplotype diversity (Hd) were estimated according to Nei [48]. Polymorphic loci were analysed using NTSYS v2.1 [36] software to obtain the genetic similarity (GS) estimate in a pairwise comparison using the simple matching coefficient. The resulting similarity matrix was used for the development of a UPGMA dendrogram using hierarchical clustering. To reconduct the two main clusters highlighted by the UPGMA tree within the Andean or Mesoamerican gene pool, sequences of accessions with known geographical origins were

retrieved from studies by Nanni et al. [41] and Goretti et al. [42]. Information was not available for all the analyzed markers and was available only for β -EG, β -GBP, and SHP. The genetic sequences were aligned with MEGA-X using the Muscle algorithm, and the UPGMA tree was then generated using 1000 bootstrap and Kimura 2-parameter models.

Population structure analysis of the core collection was performed using STRUCTURE software, which exploits a systematic Bayesian clustering approach by applying Markov chain Monte Carlo (MCMC) estimation [49], which compares the molecular marker data belonging to each accession among themselves to infer their membership in a series of putative clusters. The simulation was performed assuming the admixture model, with no *a priori* population information. The SNP data were analysed with 10^6 iterations and a burning period of $2 \cdot 10^5$, and ten replicate runs were executed with the value of K ranging between 1 and 16. The most likely K value was estimated using ΔK [50] and was considered for the assignment to an ancestral group.

3. Results

3.1. SSR Marker-Based Genotyping

The first part of this study aimed to identify the most suitable SSR markers for the characterization of the common bean core collection conserved in the DAFNAE germplasm bank. Ten of the 24 initially selected markers were chosen for further analyses since preliminary tests exhibited easy scorability, a marked attitude to be amplified in multiplex PCRs and the highest polymorphism information content (PIC) coefficients.

Descriptive statistics of genetic diversity calculated for the 10 marker loci exploited for the analyses are reported in **Table 4**. The mean number of observed alleles per locus (n_a) was 5.8, with values ranging from 2 (AY1) to 11 (GATS91). The effective number of alleles (n_e) for the analysed accessions ranged between 0.52 (AY1) and 6.69 (GATS91), with a mean value equal to 3.55. Each microsatellite locus scored high levels of observed homozygosity, ranging from 0.98 (IAC52 and IAC66) to 1.00, with an average observed homozygosity of 0.99. The PIC coefficients were also calculated using the marker allele frequencies at each locus to determine the discriminant ability of each marker locus among the different genotypes. With a minimum value of 0.56 (BM197) and a maximum value of 0.91 (BM183), the panel of selected microsatellite markers proved to be hypervariable and highly informative.

Table 4. Descriptive statistics of genetic diversity calculated for each of the 10 SSR marker loci analysed in the common bean core collection. (The number of observed alleles (n_a), the number of effective alleles (n_e), observed homozygosity (H_o), and the polymorphic information content (PIC) are reported for each SSR marker locus investigated.)

Locus	n_a	n_e	H_o	PIC
IAC52	6	3.84	0.98	0.89
AY1	2	1.50	0.99	0.70
BM183	4	1.77	0.99	0.91
BM197	3	1.68	0.99	0.56
IAC66	7	2.26	0.98	0.64
BM210	6	4.89	1.00	0.86
PvM04	8	4.83	1.00	0.84
BM200	7	4.50	0.99	0.88
Pvag001	4	3.52	1.00	0.77
GATS91	11	6.69	1.00	0.74
Mean	5.8	3.55	0.99	0.78

Supplementary Figure 1 shows the mean genetic similarity matrix calculated within and among the 25 populations. Values on the diagonal representing the genetic similarity of each population showed an overall very high uniformity – 15 populations scored 100% identity, and only two populations showed a genetic similarity below 0.95 (accessions “Blue lake a grano nero” and “Pegaso”).

Genetic relationships among common bean accessions were further studied using principal coordinate analysis (Figure 1A). Individual unique genotypes were differently classified and discriminated, depending on whether they were assigned through the subsequent haplotyping analysis to the Andean or Mesoamerican gene pools and to Venetian farmer’s varieties or Italian breeder’s lineages (please note that symbol dimensions are proportional to the number of individuals sharing 100% similarity and hence are characterized by the same coordinates). The vast majority of populations represented by two or more subgroups are located in proximity to each other, forming almost isolated subgroups (for example, accession 12 “Blue lake a grano nero” and 36 “Maron”). Considering the first discriminant coordinate, there is a clear separation between the elite lineages and the niche landraces (with the only exception of genotype 1.1 from population “Mangiatutto rampicante”). Taking into consideration the second discriminant coordinate, the distinction is less clear (Figure 1A). It is, however, possible to highlight a trend in which the genotypes with Mesoamerican origin, as subsequently determined from the SNP variant analysis, possess a generally higher coordinate value for the second dimension compared to the genotypes of Andean origin, even with a large and shared area of centroids (Figure 1A). A total of 41 genotypically different individuals was then chosen as representative of the genetic diversity of the

entire core collection and used in the subsequent haplotyping analyses, as they showed 100% homozygosity and similarities to each other lower than 100%.

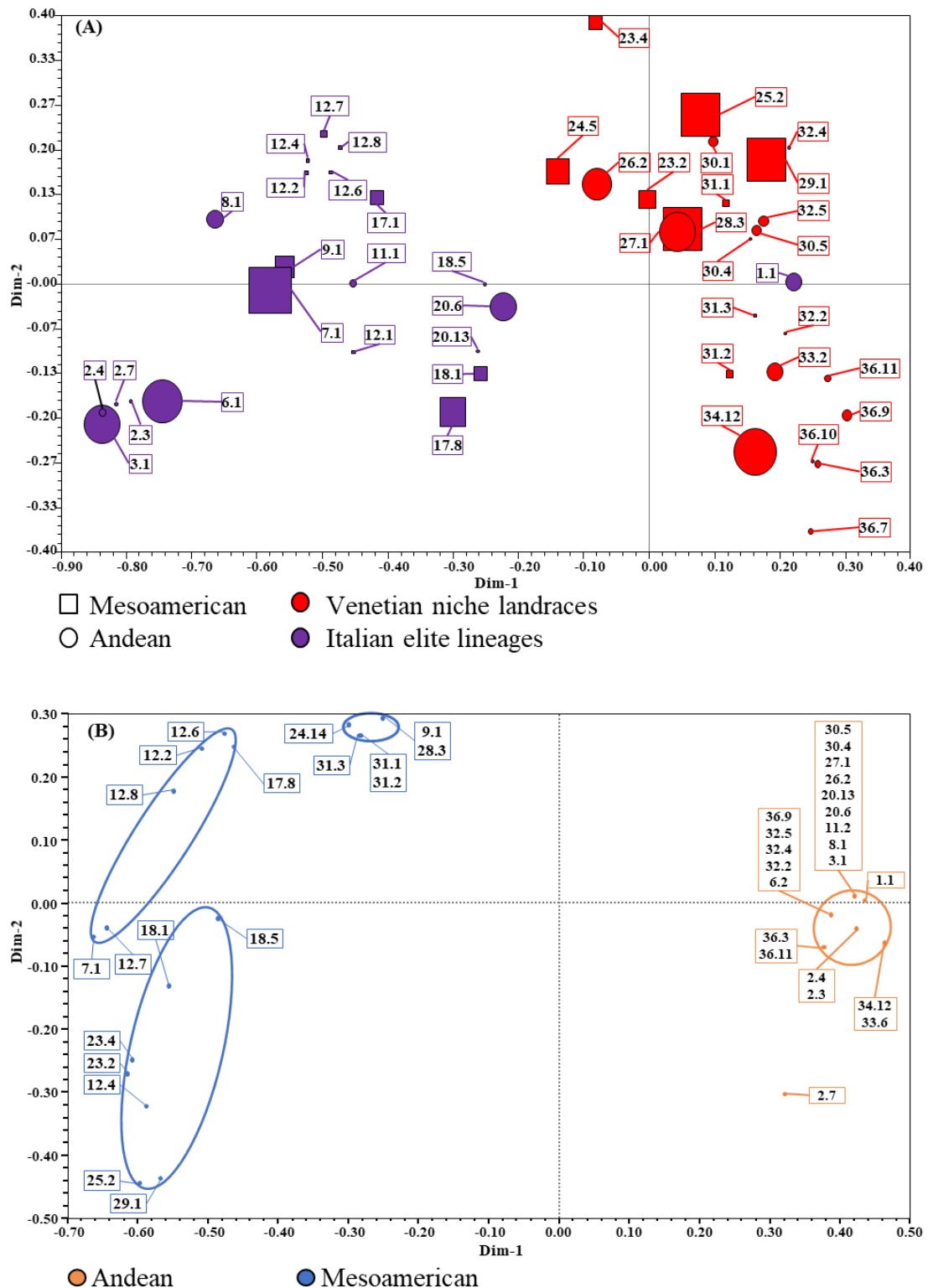


Figure 1. Principal Coordinate Analysis of the common bean accessions based on SSR and SNP markers: **(A)** Bidimensional centroids of the total 193 analyzed samples (symbol size is proportional to the number of samples scoring full genetic identity). **(B)** Bidimensional clusters of the 41 common bean unique genotype accessions identified through SSR markers and analysed using SNP markers for the eleven chosen genes.

3.2. SNP Variant-Derived Haplotyping

The subset of 41 unique genotypes belonging to 25 of the chosen 26 populations (accession 19 exhibited a significant amount of missing data, so its haplotype characterization was not reported in this study) was analysed for 11 target genes using Sanger sequencing to investigate and characterize their haplotypes. Each gene is a single copy located on a distinct linkage group and is related to traits of agronomic interest or linked to loci known to be under selective pressure during the domestication process. The primer pairs that were used in this study revealed good PCR efficiency and were able to work with a 100% success rate in *Phaseolus* genomic DNA samples and resulted in the amplification of single genomic regions (**Supplementary Figure 2**).

Sequenced fragments ranged between 409 and 818 bp, for a cumulative length of 6,533 bp composed of 55.5% exonic and 44.5% intronic regions. Based on sequence analysis, all the gene regions revealed 100% homozygosity, which was in agreement with the results obtained from microsatellite marker analysis. By aligning the DNA sequences of these target genes, in total, 48 SNP variants were detected, as well as 8 INDELs, which were distributed unequally across the DNA fragments. All the polymorphic sites found were biallelic, with the exception of the third single point variant of the LOX gene, where an adenine was either substituted with a thymine or deleted. The marker with the highest variability was the intronic region of the shatterproof (SHP) gene, with 10 SNP variants and 3 INDELs (1, 4, and 7 nucleotides long, respectively), whereas the β -AM gene was monomorphic in the investigated core collection. The occurrence of polymorphisms varied significantly based on the region taken into consideration – in exons, 0.47% of the nucleotides were polymorphic, while in introns, the occurrence of variants was three times higher, equalling 1.55%. This discrepancy was expected since noncoding regions are known to be neutral and thus retain point mutations, and they usually show higher diversity. Considering only the SNP variants detected in exon regions, it is worth mentioning that 8 out of 17 proved to be nonsynonymous mutations. Additional information on the SNP distribution and classification can be found in **Supplementary Table 2**.

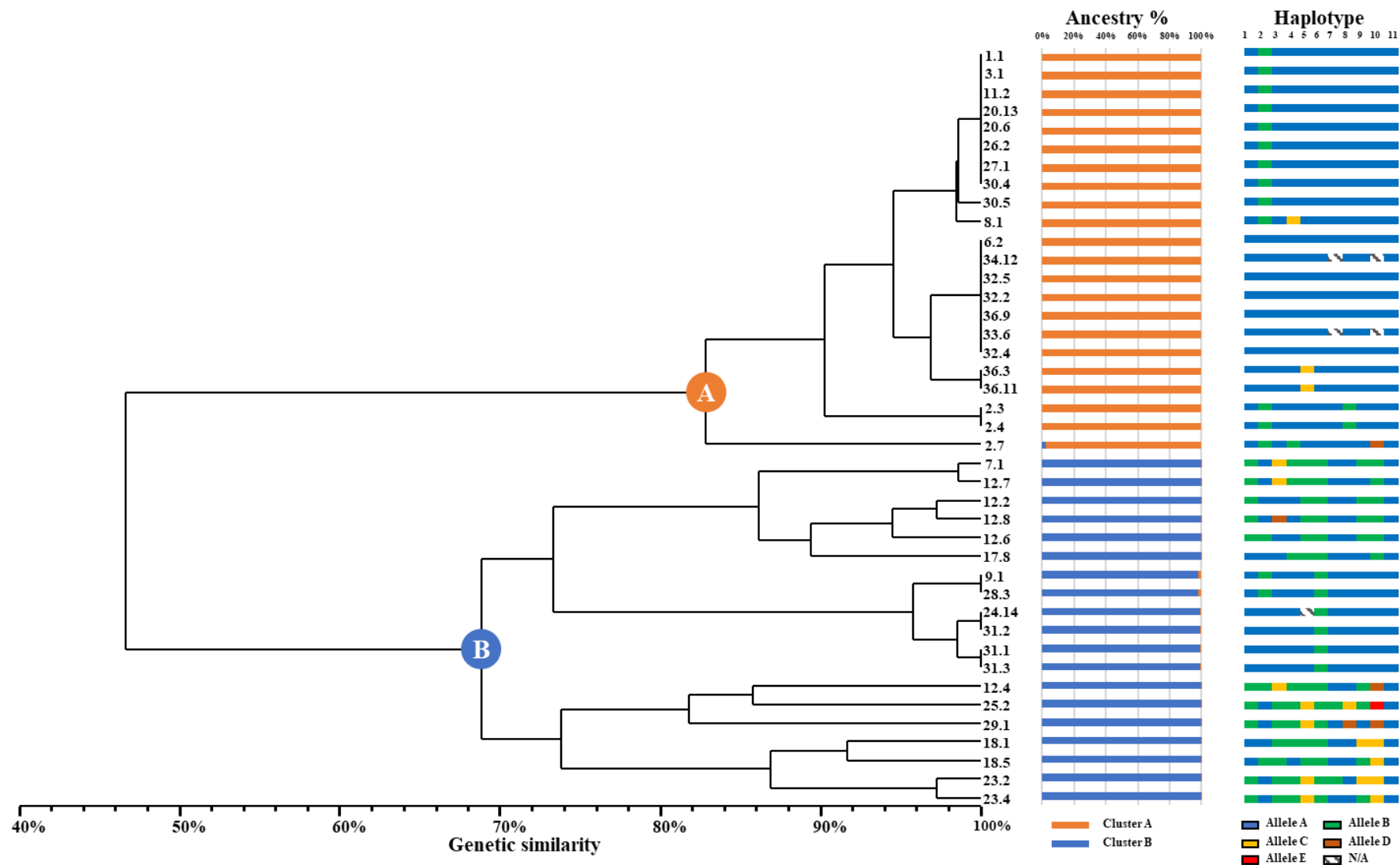


Figure 2. UPGMA dendrogram with all the common bean accessions analysed in this study grouped in two main clusters (A,B) and several subclusters. The ancestry coloured bars refer to the percentage of ancestral membership estimated by STRUCTURE analysis (for $K = 2$). The haplotype coloured bars refer to the allele combinations found for each of the samples: individual bars are split into eleven blocks as the number of sequenced genes and each bar as a whole represents a multi-locus haplotype including 11 specific alleles (dashed blocks indicate missing data).

All the accessions were grouped into two distinct and well-separated branches labelled A and B, respectively. Each of these two groups scored high mean genetic similarity, especially in cluster A, wherein all 22 genotypes were divided into seven subgroups. Moreover, 16 subgroups were found in cluster B, wherein the 19 genotypes showed lower mean genetic similarity. Notably, genotypes belonging to the same population were consistently assigned to only one of the two main clusters, even if placed in different minor branches. In total, 14 varieties were assigned to cluster A, and 11 varieties were associated with cluster B.

The least homogeneous population, accession 12 “Blue lake a grano nero,” was represented by five genotypes and was assigned to cluster B, contributing to the higher variability in this group. It is worth mentioning that in three cases, in both clusters A and B, individuals were present belonging to both Venetian niche landraces and national lineages grouped together with full haplotype genetic identity (e.g., samples 1.1 and 20.13 in cluster A and samples 9.1 and 28.3 in cluster B). The ancestral membership of the core collection was investigated using STRUCTURE software by the estimation of ΔK . This result suggested that the core collection most likely originated from two genetically distinct ancestors ($K = 2$ as the most likely value), as shown in **Figure 2**, where each genotype is represented by a histogram divided into two segments that are proportional to the membership to ancestor 1 or 2. This resulted in a clear division into two distinct and highly uniform groups composed of 22 and 19 accessions each, which was in agreement with the current evolutionary model of modern common beans derived from two distinct gene pools, Andean and Mesoamerican.

Then, we demonstrated that the two main clusters highlighted by the UPGMA tree analysis and supported by the STRUCTURE ancestry analysis do correspond to Andean or Mesoamerican origins. To interpret the results, the genetic sequences were aligned with accessions of known geographical origin and gene pool identity among those available on the NCBI database for the analysed loci. Samples assigned to cluster A univocally belonged to the Andean gene pool, similar to those in cluster B to the Mesoamerican pool, supporting the hypothesis of the two gene pools of origin.

For 10 of the 11 regions sequenced, numbers of alleles (meaning specific combinations of polymorphic positions) ranging from 2 to 5 were identified (**Table 3**). Since recombination is unlikely to occur in closely linked loci, SNP variants are instead inherited together as a single unit, and the unique combinations found were limited: 2 alleles for β -EG, HSF, SHP, and LEA; 3 alleles for β -GBP, PEPC, and NR; 4 alleles for H4; and 5 alleles for STK and LOX. β -AM was monoallelic. In H4, β -GBP, LEA, STK, NR, and LOX SNP combinations were uncommon among

the examined genotypes, with an incidence below 5%. The relative values are reported in **Table 5** and **Supplementary Table 4**.

Table 5. Relative frequency (%) of the SNP-derived allele variants found in the common bean core collection for each of the 11 target genes. (The number of polymorphic sites at each locus (S), the relative number of identified alleles (Hp n), and Nei's haplotype distance (Hp d) [48] are reported)

Genes	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	75.60	53.70	75.60	73.20	61.00	53.70	90.20	90.30	75.60	61.00	100.00
Allele B	24.40	41.50	14.60	24.40	22.00	46.30	4.90	4.90	19.50	14.60	
Allele C			7.30	2.40	14.60			2.40	4.90	9.80	
Allele D			2.40					2.40		7.30	
Allele E										2.40	
S	3	3	7	3	3	22	1	8	2	19	0
Hp n	2	2	4	3	3	2	2	4	3	5	1
Hp d	0.68	0.77	0.85	0.80	0.85	0.75	0.59	0.80	0.80	0.92	0.00

The various SNP combinations were arranged in 21 different multi-locus haplotypes, as reported in **Table 6**. In agreement with the previous analysis, cluster A was characterized by higher genetic uniformity, and most of the accessions were represented by a single haplotype (Haplo_01), while cluster B showed wider genetic variability with 14 different haplotypes (Haplo_08-21). Particularly relevant in terms of protein functionality are the 8 SNP variants identified as responsible for amino acid substitutions (please note that all missense mutations are marked with an asterisk in the consensus sequence, see **Supplementary Table 3**).

Table 6. Relative haplotype number (Hp n) and haplotype distance (Hp d) [48] of the SNP-derived allele variants found in the core collection for each of the 11 target genes and for each of the four identified clusters, based on accession identity (farmer populations and breeder selections) or ancestry identity (Andean and Mesoamerican), and in total.

Resources/Genes		β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Venetian niche populations	Hp n	2	2	2	2	2	2	2	3	3	4	1
	Hp d	0.65	0.68	0.65	0.65	0.75	0.74	0.67	0.73	0.75	0.87	0.00
Italian elite lineages	Hp n	2	2	4	3	2	2	1	2	3	4	1
	Hp d	0.71	0.73	0.87	0.83	0.75	0.75	0.00	0.59	0.83	0.91	0.00
Andean lineages	Hp n	1	2	1	3	2	1	1	2	1	2	1
	Hp d	0.00	0.74	0.00	0.72	0.58	0.00	0.17	0.58	0.00	0.63	0.00
Mesoamerican lineages	Hp n	2	2	4	2	3	1	2	3	3	5	1
	Hp d	0.75	0.69	0.92	0.75	0.89	0.00	0.59	0.73	0.86	0.95	0.00
Total	Hp n	2	2	4	3	3	2	2	4	3	5	1
	Hp d	0.68	0.77	0.85	0.80	0.85	0.75	0.59	0.80	0.80	0.92	0.00

4. Discussion

Available historical records suggest that the common bean core collection we have studied is formed by neglected Venetian niche landraces locally maintained by farmers for several decades (*i.e.*, Venetian farmer's local varieties) and old varieties genetically improved by breeders several decades ago using local materials (Italian breeder's elite lineages). As a main finding, this core collection is represented by a large number of highly homozygous individuals and within-population homogeneous varieties. A considerable range of variation among populations is phenotypically detectable and genotypically verifiable, but between-population differentiation is also particularly evident and measurable for several distinctive plant and seed traits, likely as a consequence of both natural and human selection pressure.

4.1. Molecular Characterization of Venetian Niche Landraces and Italian Elite Lineages and the Genetic Structure of the Core Collection as a Whole

The region of Veneto covers an area of 18,364 km², of which 57% is a vast plain and 29% is a mountainous area composed of the Carnic Alps, Eastern Dolomites, and Venetian Prealps [51]. This specific geographic formation allowed farmers or small farms and rural communities to grow beans in isolation for centuries. This study shows that over the years, new introductions and exchanges of different accessions have occurred from different domestication centres and origins. However, currently, based on visual characteristics, it is possible to identify dozens of landraces typical of that region. In addition to agronomic and nutraceutical categorization, one of the first goals of this study was the genetic and molecular characterization of a *P. vulgaris* core collection composed of local farmer varieties and elite breeder lineages using DNA markers. This approach was considered to be crucial for the genetic diversity estimation of potentially valuable bean germplasm resources to avoid any loss of genotypes/biotypes, to promote long-term conservation programs, and to allow commercial valorisation of ancient bean varieties typical of the Veneto region, Italy.

The first SSR-based approach applied to a consistent number of samples (193 individuals belonging to 25 populations) highlighted a very high extent of homozygosity, which was in accordance with the autogamous reproduction system of this species characterized by a very low rate of occasional hybridization. Two important considerations are as follows: first, although homozygosity is linkage group-independent, its estimate was nearly equal to 100% and was found to be constant in the sequenced genes and amplified markers throughout the genome, and second,

the set of expressed and neutral regions investigated was shown to be representative of all the basic chromosomes, as confirmed by their genome and/or linkage map localization analysis.

The mean genetic similarity within each population was calculated based on all the pairwise comparisons among the accessions. The vast majority of populations scored a very high genetic similarity (>95%) and, hence, genetic uniformity, with 14 out of 25 varieties showing full genetic identity (*i.e.*, 100% genetic similarity estimates). This homogeneity can have different explanations, considering the different origins of the accessions. For the landraces (populations numbered from 23 to 36), the lack of variation within populations may be ascribed to the production system locally adopted by these niche plant materials in isolated geographical areas, which are often mountainous, limiting gene flow events among populations and preventing seed exchange among farmers. For the elite lineages (populations numbered from 1 to 20), the lack of variation within populations may be expected, as they are likely derived from single pure line selection methods to meet the genetic uniformity and stability requirements for commercial varieties. From the comparison among populations, four accessions (7, 23, 25, and 29) were highly differentiated from the rest, with a genetic similarity almost always below 70%. Interestingly, all the outliers previously mentioned have a Mesoamerican origin, and the majority (23, 25, and 29) are Venetian niche landraces. The Mesoamerican centre is considered the first domestication centre for the common bean, from which the Andean gene pool originated as a consequence of a strong bottleneck, and thus, the Mesoamerican gene pool is characterized by higher variability with the presence of uncommon genotypes. This peculiarity is also reflected in the core collection even after many generations of adaptation to the Italian environment. Overall, 56% of accessions were from the Andean center of origin, and 44% were from the Mesoamerican center. The proportion of Andean and Mesoamerican accessions in Veneto is close to that found by Angioi et al. [19] in Europe: 67% Andean and 33% Mesoamerican. However, this proportion is slightly different when compared to their findings for Italian landraces: 75% Andean and 25% Mesoamerican. The data obtained in this study are also in agreement with Angioi et al. [19] in terms of hybridization since their data showed that 44% of the European *P. vulgaris* landraces were derived from hybridization between Andean and Mesoamerican gene pools; however, in Spain and Italy, the distribution of hybrids was very low.

The SSR marker data were very useful for clustering the 193 samples initially selected into 41 distinct genotypes based on their genetic structures (allele/genotype diversity) and their genetic similarity estimates. This allowed us to construct a PCoA that led to some considerations. Generally, in populations represented by more than one genotype, it is possible to group the samples within narrow areas of the PCoA, meaning that even if not genetically identical, these populations possess a high degree of genetic homogeneity. Looking at the first dimension, a major

division was clear between the elite lineages and the local varieties. This can be a result of the convergent evolution that led the local varieties to better adapt to the climatic and environmental conditions of the Veneto region and farmers' cultivation methods, while the improved varieties are the results of breeders' selection for their high yields and satisfying consumers' preferences. The only misplacement is represented by population 1, possibly because this pure line was derived from local germplasm in recent times. Such a well-defined separation is not present in the second dimension, but samples of Andean origin showed the lowest values, while those of Mesoamerican origin showed the highest values. While such classification by ancestry origin was evident with the SNP marker data, upon using microsatellites, the result is not the same, thus highlighting the limits of this methodology based on neutral markers in reconstructing genetic relationships of more genetically distant accessions.

Microsatellite markers are known to be extremely useful for assessing the genetic similarity between individuals and populations of the same species and testing the genetic identity of varieties. However, they are not reliable for describing phylogenetic relationships. Based on the SSR marker-based genotyping results, 41 unique genotypes were selected to represent the genetic variability existing in the core collection under study and were further characterized through SNP variant-based haplotyping using specific genes as target regions.

The genetic variability of the core collection was evaluated using ten marker genes spanning a total of 6,533 nucleotides, where 48 SNPs and 8 INDELS were found. The SNP frequency was 1 every 136 bp, which is considerably high when compared to the average found in other legumes (1 every 233 bp in *Medicago trunculata* [52] and 1 every 588 bp in *Glycine max* [53]), meaning that the target regions chosen for this analysis are characterized by a particularly high polymorphism rate. As expected, introns were found to be less conserved, since mutations in these regions are silent and neutral, with no association with any phenotypic variation.

Considering only the SNP variants detected in exons, it is interesting to note that in 8 out of 17 cases, the mutation was nonsynonymous, determining an amino acid substitution. In these cases, the altered protein can be responsible for phenotypic variation. Our finding is particularly relevant since the genes chosen for the haplotyping analysis were selected based on their putative association with traits of agronomic interest.

Polymorphic nucleotides were actually arranged in a few unique combinations. A total of 21 multi-locus haplotypes was detected for the marker genes used, with 7 in cluster A and 14 in cluster B. Fifteen of these SNP-derived haplotypes were encountered in less than 5% of the accessions, mainly in those attributable to Mesoamerican ancestry (cluster B). The presence of uncommon gene variants caused an increase in the genetic diversification of this subgroup, which may be exploited

as a useful resource for the development of new varieties with unique characteristics and adapted to local conditions. These results are in agreement with the previous SSR marker-based analysis, which also highlighted the higher variability in the group with a Mesoamerican origin. Another aspect to take into account is the functionality of the allele associated with a specific haplotype, which implies that genetic mutations are translated into protein modifications. Nonfunctional haplotypes are useful to reconstruct the evolutionary history of an organism but will not be reflected in phenotypic variations. In total, 9 distinct functional haplotypes were present in the core collection – 3 were in cluster A and 8 in cluster B, with 2 shared between the two groups. It is thus possible that in cluster B, not only the genetic diversity but also the phenotypic diversity is higher. Selection within a wide gene pool has a higher probability of finding individuals with particular traits, conferring adaptability to mutable or extreme environments or resistance to pests and diseases. These characteristics are becoming increasingly valuable considering climate change and the growing demand for food that the world will face in the upcoming years.

The abundance of different genotypes of Mesoamerican origin was also evident in the UPGMA dendrogram, which was composed of branches grouping only 1 or 2 populations. One of the experimental lines (population 12), which were also selected based on uniformity criteria, was even composed of 5 genetically distinct genotypes. The subgroup with Andean origin was markedly different and much more uniform. In cluster A, 15 of the 22 genotypes were grouped and shared the same haplotype. Experimental lines and local varieties in some cases cannot be distinguished, and this may indicate that they are related and that pure lines were developed from those landraces or closely related materials.

Comparing the molecular tools used in this study, noncoding region-based SSR markers (neutral regions) revealed more variability in the Andean gene pool, whereas the coding region-based SNP markers (expressed regions) detected higher polymorphism in the Mesoamerican gene pool. One example is population 36 (“Maron”), for which the SSR analysis result was one of the most variable, while the subsequent haplotyping defined it as almost uniform. Moreover, SSR markers showed greater informativeness about the common bean typology – Venetian niche landraces or Italian lineage – compared to SNP markers, which otherwise demonstrated higher suitability for ancestral gene pool reconstruction (for details, see **Figure 1**).

The clear distinction between the Andean and Mesoamerican gene pools was not ensured since the genetic material used for this study can be geographically ascribed to north-eastern Italy. Europe, in fact, is considered a secondary diversification centre where gene flow occurred by spontaneous events of hybridization and introgression, but still, the distinction into two well-separated clusters calculated for the ancestry reconstruction of the analysed samples is noteworthy.

Remarkable are the results obtained from the analysis of the intronic region of the shatterproof gene, where an extremely high abundance of polymorphisms was localized – totals of 10 SNPs and 3 INDELS were located in a 440 bp-long sequence. This marker gene has proven to be extremely predictive in distinguishing between Andean and Mesoamerican origin. The presence of intraspecific indels is also a particularly relevant and peculiar feature, which could be exploited for the development of cheap and fast molecular tools able to discriminate the gene pool of origin based only on the length of this fragment. It will be necessary to perform additional investigations with individuals of different geographical origins to confirm this hypothesis.

Thus, by using neutral SSR markers in combination with functional SNP markers, it was possible to unambiguously group the 25 *P. vulgaris* populations (“Semirampicante abruzzese”) into two clusters based on their gene pool of origin – either Andean or Mesoamerican. Genetic variability was distributed unequally across the two subgroups, as highlighted by haplotyping analysis, in which 16 out of the 23 multi-locus haplotypes detected in the core collection were found within the Mesoamerican gene pool, while the Andean was much more homogeneous.

The molecular characterization provided useful information for the selection of pure lines among the analysed populations, which can be combined with the nutraceutical characterization and the agronomic performance of these accessions in different Venetian environments. These are fundamental steps for the development of new varieties highly adaptable to local agronomic and climatic conditions and valuable for organic cultivation systems. Italian citizens value local food with a strong connection to the production area, increasing their value and demand in a microregion [54], and another factor that can add value to these varieties is represented by the possibility of their genetic traceability along the whole supply chain using the described molecular tools to guarantee high-quality standards and to protect consumers and producers.

Therefore, this study provides original information that allows not only the conservation of this genetic material from Venetian niche landraces and Italian elite lineages, but also the commercial valorization of this genetic material with a strong connection to the region where they have been cultivated for centuries, thereby allowing the selection of landraces with good agronomic performance and high nutraceutical value, which may become commercial varieties with high added value. Italian citizens value local food with a strong connection to the production area, thus increasing their value and demand in a microregion. Another factor that can add value to these varieties is represented by the possibility of their genetic traceability along the whole supply chain using the described molecular tools to guarantee high-quality standards and to protect consumers and producers.

5. References

1. Arnoldi, A.; Zanoni, C.; Lammi, C.; Boschini, G. The Role of Grain Legumes in the Prevention of Hypercholesterolemia and Hypertension. *Critical Reviews in Plant Sciences* **2015**, *34*, 144-168, doi:10.1080/07352689.2014.897908.
2. Calles, T. Preface to special issue on leguminous pulses. *Plant Cell Tissue and Organ Culture* **2016**, *127*, 541-542, doi:10.1007/s11240-016-1146-7.
3. Nelson, M.E.; Hamm, M.W.; Hu, F.B.; Abrams, S.A.; Griffin, T.S. Alignment of Healthy Dietary Patterns and Environmental Sustainability: A Systematic Review. *Adv Nutr* **2016**, *7*, 1005-1025, doi:10.3945/an.116.012567.
4. Rendon-Anaya, M.; Montero-Vargas, J.M.; Saburido-Alvarez, S.; Vlasova, A.; Capella-Gutierrez, S.; Ordaz-Ortiz, J.J.; Aguilar, O.M.; Vianello-Brondani, R.P.; Santalla, M.; Delaye, L., et al. Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biol* **2017**, *18*, 60, doi:10.1186/s13059-017-1190-6.
5. Johns, M.A.; Skroch, P.W.; Nienhuis, J.; Hinrichsen, P.; Bascur, G.; MunozSchick, C. Gene pool classification of common bean landraces from Chile based on RAPD and morphological data. *Crop Science* **1997**, *37*, 605-613, doi:DOI 10.2135/cropsci1997.0011183X003700020049x.
6. Beebe, S.; Skroch, P.W.; Tohme, J.; Duque, M.C.; Pedraza, F.; Nienhuis, J. Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Science* **2000**, *40*, 264-273, doi:DOI 10.2135/cropsci2000.401264x.
7. Tohme, J.; Gonzalez, D.O.; Beebe, S.; Duque, M.C. AFLP analysis of gene pools of a wild bean core collection. *Crop Science* **1996**, *36*, 1375-1384, doi:DOI 10.2135/cropsci1996.0011183X003600050048x.
8. Beebe, S.; Rengifo, J.; Gaitan, E.; Duque, M.C.; Tohme, J. Diversity and origin of Andean landraces of common bean. *Crop Science* **2001**, *41*, 854-862, doi:DOI 10.2135/cropsci2001.413854x.
9. Pallottini, L.; Garcia, E.; Kami, J.; Barcaccia, G.; Gepts, P. The genetic anatomy of a patented yellow bean. *Crop Science* **2004**, *44*, 968-977, doi:DOI 10.2135/cropsci2004.0968.
10. Diaz, L.M.; Blair, M.W. Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor Appl Genet* **2006**, *114*, 143-154, doi:10.1007/s00122-006-0417-9.
11. Ariani, A.; Teran, J.C.B.M.Y.; Gepts, P. Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS). *Molecular Breeding* **2016**, *36*, doi: 10.1007/s11032-016-0512-9.
12. Ariani, A.; Berny Mier, Y.T.J.C.; Gepts, P. Spatial and Temporal Scales of Range Expansion in Wild *Phaseolus vulgaris*. *Mol Biol Evol* **2018**, *35*, 119-131, doi:10.1093/molbev/msx273.

13. Kuzay, S.; Hamilton-Conaty, P.; Palkovic, A.; Gepts, P. Is the USDA core collection of common bean representative of genetic diversity of the species, as assessed by SNP diversity? *Crop Science* **2020**, *60*, 1398-1414, doi:10.1002/csc2.20032.
14. Nicole, S.; Erickson, D.L.; Ambrosi, D.; Bellucci, E.; Lucchin, M.; Papa, R.; Kress, W.J.; Barcaccia, G. Biodiversity studies in Phaseolus species by DNA barcoding. *Genome* **2011**, *54*, 529-545, doi:10.1139/G11-018.
15. Hidalgo, R. The Phaseolus world collection. *Genetic Resources of Phaseolus Beans* **1988**, 67-90.
16. Rodino, A.P.; Santalla, M.; De Ron, A.M.; Singh, S.P. A core collection of common bean from the Iberian peninsula. *Euphytica* **2003**, *131*, 165-175, doi:Doi 10.1023/A:1023973309788.
17. Jones, A.L. Phaseolus bean: Post-harvest Operations. *Phaseolus Bean: Post-Harvest Operations* **1999**.
18. Zeven, A.C. The introduction of the common bean (*Phaseolus vulgaris* L) into Western Europe and the phenotypic variation of dry beans collected in the Netherlands in 1946. *Euphytica* **1997**, *94*, 319-328, doi:Doi 10.1023/A:1002940220241.
19. Angioi, S.A.; Rau, D.; Attene, G.; Nanni, L.; Bellucci, E.; Logozzo, G.; Negri, V.; Spagnoletti Zeuli, P.L.; Papa, R. Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. *Theor Appl Genet* **2010**, *121*, 829-843, doi:10.1007/s00122-010-1353-2.
20. Albala, K. *Phaseolus vulgaris*: Mexico and the World. *Beans. a History* **2007**, 127-190.
21. Piergiovanni, A.R.; Lioi, L. Italian Common Bean Landraces: History, Genetic Diversity and Seed Quality. *Diversity* **2010**, *2*, 837-862, doi:10.3390/d2060837.
22. Bianco, M.L.; Grillo, O.; Cremonini, R.; Sarigu, M.; Venora, V.J.A.J.o.C.S. Characterisation of Italian bean landraces ('*Phaseolus vulgaris*' L.) using seed image analysis and texture descriptors. **2015**, *9*.
23. Palumbo, F.; Galla, G.; Barcaccia, G. Developing a Molecular Identification Assay of Old Landraces for the Genetic Authentication of Typical Agro-Food Products: The Case Study of the Barley 'Agordino'. *Food Technol Biotechnol* **2017**, *55*, 29-39, doi:10.17113/ftb.55.01.17.4858.
24. Palumbo, F.; Galla, G.; Martinez-Bello, L.; Barcaccia, G. Venetian Local Corn (*Zea mays* L.) Germplasm: Disclosing the Genetic Anatomy of Old Landraces Suited for Typical Cornmeal Mush Production. *Diversity-Basel* **2017**, *9*, doi: 10.3390/d9030032.
25. Spagnoletti Zeuli, P.L.; Baser, N.; Riluca, M.; Laghetti, G.; Logozzo, G.; Masi, P.; Molinari, S.; Negri, V.; Olita, G.; Tiranti, B. Valorisation and certification of Italian bean agro-ecotypes (*Phaseolus vulgaris*). *Proceedings of the Ecotipi Vegetali Italiani: Una Preziosa Risorsa di Variabilità Genetica* **2004**, 19.

26. Piergiovanni, A.R.; Cerbino, D.; Brandi, M. The common bean populations from Basilicata (Southern Italy). An evaluation of their variation. *Genetic Resources and Crop Evolution* **2000**, *47*, 489-495, doi:Doi 10.1023/A:1008719105895.
27. Piergiovanni, A.R.; Laghetti, G. The common bean landraces from Basilicata (Southern Italy): an example of integrated approach applied to genetic resources management. *Genetic Resources and Crop Evolution* **1999**, *46*, 47-52, doi:Doi 10.1023/A:1008641731573.
28. Dal Ferro, N.; Borin, M. Environment, agro-system and quality of food production in Italy. *Italian Journal of Agronomy* **2017**, *11*, 133-143, doi:10.4081/ija.2017.793.
29. Eurostat. Available online: <https://ec.europa.eu/eurostat> (accessed on 2020)
30. Istat. Available online: <https://www.istat.it/> (accessed on 2020)
31. Yu, K.; Park, S.J.; Poysa, V.; Gepts, P. Integration of simple sequence repeat (SSR) markers into a molecular linkage map of common bean (*Phaseolus vulgaris* L.). *Journal of Heredity* **2000**, *91*, 429-434, doi:DOI 10.1093/jhered/91.6.429.
32. Blair, M.W.; Torres, M.M.; Giraldo, M.C.; Pedraza, F. Development and diversity of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). *BMC Plant Biol* **2009**, *9*, 100, doi:10.1186/1471-2229-9-100.
33. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* **2000**, *18*, 233-234, doi:10.1038/72708.
34. Yeh, F.C.; Yang, R.; Boyle, T.B.; Ye, Z.; Mao, J.X. POPGENE, the user-friendly shareware for population genetic analysis. *Molecular biology biotechnology centre, University of Alberta, Canada* **1997**, *10*, 295-301.
35. Palumbo, F.; Barcaccia, G. Critical aspects on the use of microsatellite markers for assessing genetic identity of crop plant varieties and authenticity of their food derivatives. *Rediscovery of Landraces as a Resource for the Future* **2018**, 129-160.
36. Rohlf, F.J. NTSYS: numerical taxonomy and multivariate analysis system version 2.02. *Applied Biostatistics Inc., Setauket, NY* **1998**.
37. Schlotterer, C.; Pemberton, J. The use of microsatellites for genetic analysis of natural populations. *EXS* **1994**, *69*, 203-214, doi:10.1007/978-3-0348-7527-1_11.
38. Rakoczy-Trojanowska, M.; Bolibok, H. Characteristics and a comparison of three classes of microsatellite-based markers and their application in plants. *Cellular & Molecular Biology Letters* **2004**, *9*, 221-238.
39. Kumar, S.; Banks, T.W.; Cloutier, S. SNP Discovery through Next-Generation Sequencing and Its Applications. *Int J Plant Genomics* **2012**, *2012*, 831460, doi:10.1155/2012/831460.
40. McConnell, M.; Mamidi, S.; Lee, R.; Chikara, S.; Rossi, M.; Papa, R.; McClean, P. Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* **2010**, *121*, 1103-1116, doi:10.1007/s00122-010-1375-9.

41. Nanni, L.; Bitocchi, E.; Bellucci, E.; Rossi, M.; Rau, D.; Attene, G.; Gepts, P.; Papa, R. Nucleotide diversity of a genomic sequence similar to SHATTERPROOF (PvSHP1) in domesticated and wild common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* **2011**, *123*, 1341-1357, doi:10.1007/s00122-011-1671-z.
42. Goretti, D.; Bitocchi, E.; Bellucci, E.; Rodriguez, M.; Rau, D.; Gioia, T.; Attene, G.; McClean, P.; Nanni, L.; Papa, R. Development of single nucleotide polymorphisms in *Phaseolus vulgaris* and related *Phaseolus* spp. *Molecular Breeding* **2014**, *33*, 531-544, doi:10.1007/s11032-013-9970-5.
43. Phytozome. Available online: <https://phytozome.jgi.doe.gov> (accessed on 2020)
44. Primer-BLAST. Available online: <https://www.ncbi.nlm.nih.gov/tools/primer-blast> (accessed on 2020)
45. Bioline, L.U. *PCR, qPCR & NGS Reagents | Bioline | Meridian Bioscience* **2020**.
46. Geneious Bioinformatics Software for Sequence Data Analysis. Available online: <https://www.geneious.com> (accessed on 2020)
47. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **2018**, *35*, 1547-1549, doi:10.1093/molbev/msy096.
48. Nei, M. *Molecular Evolutionary Genetics* **1987**.
49. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945-959.
50. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **2005**, *14*, 2611-2620, doi:10.1111/j.1365-294X.2005.02553.x.
51. Venetian Geography. Available online: <https://www.venetoinside.com/it/scopri-il-veneto/geografia/> (accessed on 2020)
52. Branca, A.; Paape, T.D.; Zhou, P.; Briskine, R.; Farmer, A.D.; Mudge, J.; Bharti, A.K.; Woodward, J.E.; May, G.D.; Gentzittel, L., et al. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc Natl Acad Sci U S A* **2011**, *108*, E864-870, doi:10.1073/pnas.1104032108.
53. Lam, H.M.; Xu, X.; Liu, X.; Chen, W.; Yang, G.; Wong, F.L.; Li, M.W.; He, W.; Qin, N.; Wang, B., et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* **2010**, *42*, 1053-1059, doi:10.1038/ng.715.
54. Belletti, G.; Marescotti, A.; Brazzini, A. OldWorld Case Study: The Role of Protected Geographical Indications to Foster Rural Development Dynamics: The Case of Sorana Bean PGI. *The Importance of Place: Geographical Indications as a Tool for Local and Regional Development* **2017**, 253-276.

6. Supplementary material

Table S1. List of accessions from DAFNAE's bean germplasm. The 26 accessions used in this study are highlighted in grey. Accessions that are not highlighted were not assessed in this study due to low germination and/or low availability of seeds.

Accession Id	Name	Type	Growth
1	Fagiolino mangiatutto rampicante	Italian elite lineages	Indeterminate (climbing)
2	Borlotto nano	Italian elite lineages	Determined (dwarf)
3	Borlotto nano	Italian elite lineages	Determined (dwarf)
4	Fagiolo bianco nano	Italian elite lineages	Determined (dwarf)
5	Fagiolo rampicante	Italian elite lineages	Indeterminate (climbing)
6	Fagiolo nano creso	Italian elite lineages	Determined (dwarf)
7	Fagiolo rampicante mangiatutto (blue lake sel. Gia)	Italian elite lineages	Indeterminate (climbing)
8	Fagiolo nano mangiatutto anellino di trento	Italian elite lineages	Determined (dwarf)
9	Fagiolo rampicante mangiatutto anellino giallo	Italian elite lineages	Indeterminate (climbing)
10	Fagiolo rampicante stortino di trento	Italian elite lineages	Indeterminate (climbing)
11	Fagiolo borlotto rampicante (bortollo lingua di fuoco 3)	Italian elite lineages	Indeterminate (climbing)
12	Fagiolo rampicante (blue lake a grano nero sel. Tom)	Italian elite lineages	Indeterminate (climbing)
13	Fagiolo dolico (nano dall'occhio)	Italian elite lineages	Determined (dwarf)
14	Fagiolo nano mangiatutto (OR arno)	Italian elite lineages	Determined (dwarf)
15	Fagiolo nano (montalbano)	Italian elite lineages	Determined (dwarf)
16	Fagiolo rampicante (dolico del metro)	Italian elite lineages	Indeterminate (climbing)
17	Fagiolo nano valdarno	Italian elite lineages	Determined (dwarf)
18	Fabiolo nano coco nain blanc precoce (lotto verdone)	Italian elite lineages	Determined (dwarf)
19	Fagiolo rampicante tondino abruzzese	Italian elite lineages	Indeterminate (climbing)
20	Fagiolo verdone del piave (professional seed)	Italian elite lineages	Determined (dwarf)
21	Fagiolo dolico rampicante mangiatutto o stringa	Italian elite lineages	Indeterminate (climbing)
22	Fasole del diavolo	Venetians niche populations	Indeterminate (climbing)
23	Gialet	Venetians niche populations	Indeterminate (climbing)
24	Posenati	Venetians niche populations	Indeterminate (climbing)
25	Semi-rampicante abruzzese	Venetians niche populations	Indeterminate (climbing)
26	Fasol dela nonna	Venetians niche populations	Indeterminate (climbing)
27	Maseleta rossa	Venetians niche populations	Indeterminate (climbing)
28	Zia Orsolina	Venetians niche populations	Indeterminate (climbing)

Accession Id	Name	Type	Growth
29	Meraviglia di Venezia	Venetians niche populations	Indeterminate (climbing)
30	Secle	Venetians niche populations	Indeterminate (climbing)
31	Della Clorinda	Venetians niche populations	Indeterminate (climbing)
32	Pegaso	Venetians niche populations	Indeterminate (climbing)
33	SC-iosela	Venetians niche populations	Indeterminate (climbing)
34	D'oro (val di fiemme)	Venetians niche populations	Indeterminate (climbing)
35	Meso e Meso	Venetians niche populations	Indeterminate (climbing)
36	Maron	Venetians niche populations	Indeterminate (climbing)
37	Righetti 1	Venetians niche populations	Indeterminate (climbing)
38	Verdine	Venetians niche populations	Indeterminate (climbing)
39	Cuna	Venetians niche populations	Indeterminate (climbing)
40	Oci Della Madona	Venetians niche populations	Indeterminate (climbing)
41	Sciosele	Venetians niche populations	Indeterminate (climbing)
42	Monachelle	Venetians niche populations	Indeterminate (climbing)
43	Righetti 2	Venetians niche populations	Indeterminate (climbing)
44	Mamme bianche di Bassano - Prod 2018 Azienda	Venetians niche populations	Indeterminate (climbing)
45	Mame Bianche B1 - Prod Azienda 2018	Venetians niche populations	Indeterminate (climbing)
46	Banel fonzaso - Prod Azienda 2018	Venetians niche populations	Indeterminate (climbing)
47	Zolferini Rovizetti - Prod Azienda 2018	Venetians niche populations	Indeterminate (climbing)
48	Bala rossa	Venetians niche populations	Indeterminate (climbing)

Table S2. Information on the regions analyzed and SNPs characteristics, including amplicons length and composition (exon length and intron length), SNPs number, typology (synonymous, nonsynonymous, indel), and incidence of SNPs in the different genetic portions. Data are reported for each marker and the total analyzed region.

	β-EG	HSF	H4	β-GBP	PEPC	SHP	LEA	STK	NR	LOX	β-AM	TOT
Fragment length (bp)	597	720	409	576	529	447	680	571	818	574	612	6533
Exon length (bp)	597	299	295	576	420	54	388	188	139	244	428	3628
Intron length (bp)	0	421	114	0	109	393	292	383	679	330	184	2905
SNP number	3	3	7	3	3	10	1	3	2	13	0	48
SNPs in introns	0	2	3	0	0	10	0	2	2	12	-	31
SNPs in exons	3	1	4	3	3	0	1	1	0	1	-	17
Synonymous	2	1	0	1	3	-	0	1	-	1	-	9
Nonsynonymous	1	0	4	2	0	-	1	0	-	0	-	8
Indels	0	0	0	0	0	3	0	1	0	4	0	8

Table S3. Variable sites in the 10 loci considered, corresponding to 51 SNPs and 4 indels. Haplotypes (Haplo01-Haplo21) are ordered based on the UPGMA clustering, and the number of entries for each haplogroup is reported. The consensus sequence and the alternative nucleotide are reported for each position. Mutations that caused an amino acid substitution are marked with an asterisk on the consensus sequence.

		Entries	β -EG	HSF	LEA	STK	NR	LOX	
Cluster A	Hap01	45		T T C					
	Hap02	21							
	Hap03	5		T T C					
	Hap04	5		T T C		- - - - -			
	Hap05	4							
	Hap06	3		T T C					
	Hap07	1		T T C				T C C C	
Cluster B	Hap08	18		T T C					
	Hap09	13	C G C				G	- - - C T C T T C A	
	Hap10	12	C G C		G		T	G - - - - C C C G	
	Hap11	12	C G C			- - - - - C T		T C - - - - C C	
	Hap12	8						- - - C T C T T C A	
	Hap13	5	C G C		G		G A	T T T G C G	
	Hap14	4					G A	T T T G C G	
	Hap15	4	C G C				G	T T T G C G	
	Hap16	3							
	Hap17	2	C G C					- - - C T C T T C A	
	Hap18	1	C G C				G	- - - C T C T T C A	
	Hap19	1	C G C	T T C			G	- - - C T C T T C A	
	Hap20	1	C G C				G	- - - C T C T T C A	
	Hap21	1		T T C			G	- T T T G C G	
CONSENSUS			A A G*	G C T	T*	C A A A C T C G	T C	C A T T G C C T T A A C T A A T G T	
		Entries	H4		β -GBP	PEPC	SHP		
Cluster A	Hap01	45							
	Hap02	21							
	Hap03	5							
	Hap04	5							
	Hap05	4				T	G		
	Hap06	3					A		
	Hap07	1			A A				
Cluster B	Haplo_08	18						T A T - G - - - - G C C A T G T C A T T A C	
	Haplo_09	13	G A A	A A G A	A A	A		T A T - G - - - - G C C A T G T C A T T A C	
	Haplo_10	12	G A A	A G A	A A	T G		T A T - G - - - - G C C A T G T C A T T A C	
	Haplo_11	12	G A A	A G A	A A	T G		T A T - G - - - - G C C A T G T C A T T A C	
	Haplo_12	8			A A	A		T A T - G - - - - G C C A T G T C A T T A C	

Haplo_13	5	G	A	A	A	G	A	A	A	T	G	T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_14	4	G	A	A	A	G	A	A	A		A	T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_15	4	G	A	A	A	G	A	A	A	T	G	T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_16	3											T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_17	2	G	A	A	A	A	G	A	A	A		T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_18	1									A		T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_19	1									A		T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_20	1					A	G			A		T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
Haplo_21	1	G	A	A	A	G	A			A		T	A	T	-	G	-	-	-	-	G	C	C	A	T	G	T	C	A	T	T	A	C	
CONSENSUS		T*	G*	G*	G*	G	A	G	T*	G	C*	C	T	A	C	G	C	T	T	G	A	T	A	T	G	A	T	C	-	-	-	-	-	T

Table S4. Relative frequency (%) of the SNP-derived allele variants found in the core collection for each of the 11 target genes, number of polymorphic sites (S) at each locus, relative number of identified alleles (Hp n°), Nei's haplotype distance (Hp d.) (Nei, 1987). Information are calculated for each of the four identified clusters, based on variety identity (Italian elite lineages and Venetians niche populations) or ancestry identity (Andean and Mesoamerican), and in total

Total	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	75.60	53.70	75.60	73.20	61.00	53.70	90.20	90.30	75.60	61.00	100.00
Allele B	24.40	41.50	14.60	24.40	22.00	46.30	4.90	4.90	19.50	14.60	
Allele C			7.30	2.40	14.60			2.40	4.90	9.80	
Allele D			2.40					2.40		7.30	
Allele E										2.40	
S	3	3	7	3	3	22	1	8	2	19	0
Hp n°	2	2	4	3	3	2	2	4	3	5	1
Hp d.	0.68	0.77	0.85	0.80	0.85	0.75	0.59	0.80	0.80	0.92	0.00
Italian elite lineages	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	70.00	35.00	70.00	65.00	55.00	50.00	100.00	90.00	65.00	50.00	100.00
Allele B	30.00	65.00	10.00	30.00	45.00	50.00		10.00	30.00	30.00	
Allele C			15.00	5.00					5.00	10.00	
Allele D			5.00							10.00	
Allele E											
Hp n°	2	2	4	3	2	2	1	2	3	4	1
Hp d.	0.71	0.73	0.87	0.83	0.75	0.75	0.00	0.59	0.83	0.91	0.00
Venetians niche populations	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	80.95	76.19	80.95	80.95	66.67	57.14	80.95	90.48	85.71	71.43	100.00
Allele B	19.05	23.81	19.05	19.05		42.86	9.52		9.52		
Allele C					23.81			4.76	4.76	9.52	
Allele D								4.76		4.76	
Allele E										4.76	
Hp n°	2	2	2	2	2	2	2	3	3	4	1
Hp d.	0.65	0.68	0.65	0.65	0.75	0.74	0.67	0.73	0.75	0.87	0.00
Andean	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	100.00	40.91	100.00	90.91	90.91	100.00	90.91	90.91	100.00	86.36	100.00
Allele B		59.09		4.55				9.09			
Allele C				4.55	9.09						
Allele D										4.55	
Allele E											
Hp n°	1	2	1	3	2	1	1	2	1	2	1
Hp d.	0.00	0.74	0.00	0.72	0.58	0.00	0.17	0.58	0.00	0.63	0.00
Mesoamerican	β -EG	HSF	H4	β -GBP	PEPC	SHP	LEA	STK	NR	LOX	β -AM
Allele A	47.37	73.68	42.11	52.63	26.32		89.47	89.47	47.37	31.58	100.00
Allele B	52.63	26.32	31.58	47.37	47.37	100.00	10.53		42.11	31.58	
Allele C			15.79		21.05			5.26	10.53	21.05	
Allele D			5.26					5.26		10.53	
Allele E										5.26	
Hp n°	2	2	4	2	3	1	2	3	3	5	1
Hp d.	0.75	0.69	0.92	0.75	0.89	0.00	0.59	0.73	0.86	0.95	0.00

Chapter II

Molecular relationships and phylogenies of Venetian Radicchio (leaf chicory, *Cichorium intybus* subsp. *intybus* var. *foliosum*, $2n=2x=18$) varietal groups

(This chapter has been submitted to *MDPI Diversity* journal and is under revision)

0. Abstract

Chicory (*Cichorium intybus* L., $2n=2x=18$), is an important leafy vegetable cultivated worldwide. In Italy, this horticultural crop is known as Radicchio, and different biotypes of this crop are cultivated, especially in the northern-east of the Italian peninsula. Known to be introduced and cultivated since the 17th century in the Venice area, the original biotype, still cultivated and named “Late Red of Treviso”, differentiated among the centuries and it also hybridised with endive (*C. endivia*) giving origin to many other biotypes. Several studies, based on morphological characterizations and historical reports, describe the relationships between the most popular cultivated local varieties of this species, but this work, focused on the use of molecular markers information obtained through DNA fingerprinting, presents validations and new insights on the genetic relatedness and phylogenesis of these biotypes. By means of Random Amplified Polymorphic DNA (RAPD) and Amplified Fragment Length Polymorphism (AFLP) molecular markers, this study provides insights into the genetic relationship that intercourses among the five most important local biotypes historically cultivated in the Veneto region, which is also the geographic centre of differentiation of this cultivated leafy vegetable. Through the construction of a Maximum-likelihood tree and the reconstruction of the genetic structure of a core collection, constituted by 652 samples belonging to 5 biotypes of Radicchio divided into 22 old farmers populations, original data on their genetic origin, distinctiveness, relatedness and differentiation are reported and discussed.

Keywords: Radicchio; DNA fingerprinting; population genetics; phylogenesis; local varieties.

1. Introduction

Chicory (*Cichorium intybus* L., $2n=2x=18$) is among the most popular leafy vegetables in the world [1]. It belongs to the Asteraceae, a very large botanical family with approximately 23,000 species subdivided into 1,535 genera grouped into three subfamilies: Asteroideae, Barnadesioideae and Cichorioideae [2]. In the subfamily Cichorioideae, the tribe Lactuceae includes the genus *Cichorium*, with different horticultural species recognized according to their origin and utilization [3].

Integrating data collected from the investigation of morphological descriptors and molecular markers with geographical dispersal area and commercial indicators, *C. intybus* appeared as the most known cultivated species along with *C. endivia* [3-5]. Considering its taxonomy, distinct subspecies were established for *C. intybus* including *intybus* L. (C. Presl) Arcang. [3,5]. Moreover, several botanical varieties and cultivar groups of *C. intybus* subsp. *intybus* were recognized and they could be classified as follows: var. *foliosum* (Witloof chicory), var. *porphyreum* (Pain de Sucre), var. *latifolium* (Radicchio), var. *sylvestre* (Catalogne), and var. *sativum* (Root chicory) [3]. Within *C. intybus* subsp. *intybus*, cultivated chicory types are biennial whereas wild chicory types are perennial plants.

Most likely known by the Egyptians as a medicinal plant and used as a vegetable crop by ancient Greeks and the Romans, chicory gradually underwent a process of naturalization in Europe [5]. Currently, wild *C. intybus* covers all regions of the Italian peninsula and it is also widespread in the entire European continent. Although there are large differences in cultivation techniques and cultural uses, leafy products from chicory landraces have traditionally become a part of the diet of local populations as an important ingredient of typical local dishes [3,5]. In horticulture markets, leaf chicory traditionally includes all the cultivar groups whose commercial products are the leaves and are used in the short food supply chain (for preparation of both cooked and fresh salads) whereas, the other types of commercial products derived from the roots, are destined to either industrial transformation (inulin extracts) or human consumption (coffee substitutes), and are classified as root chicory [3].

Chicory is commonly an allogamous species due to an efficient sporophytic self-incompatibility system and a consistent entomophilous pollination that favours outcrossing [5-7]. Furthermore, hybridization among plants is also promoted by floral morphological barriers that hamper selfing and physiological mechanisms that boost germination and growth of pollen grains and tubes in case of outcrossing [3,5]. Within leaf chicory, cultivated populations of Radicchio adopted for large-scale farming systems are nowadays represented by commercial seeds of OP

varieties, synthetic varieties, and F₁ hybrids that are available on the global chicory market [1]. However, a great proportion of Radicchio is planted in many small farming units, using seed of local varieties selected and maintained through mass selection by individual farmers [7,8].

In Italy, where Radicchio is widely cultivated, especially in the north-eastern regions, for a long time plant materials grown by farmers were mainly represented by local varieties known to possess variation and adaptation to the natural and anthropological environment where they were originated and still are widely cultivated [3,5]. Such cultivated populations were conserved and multiplied by farmers as local varieties (*i.e.*, farmer populations) via phenotypic selection, and thus they were highly heterozygous and heterogeneous. Although a considerable range of phenotypic variation within each population was present across all cultivated types, a clear genetic differentiation was also noticeable among populations for various traits and molecular markers [3,5].

Currently, there are five main varietal groups of Radicchio cultivated in the Italian territory: “Late Red of Treviso”, “Early Red of Treviso”, “Red of Verona”, “Variegated of Castelfranco” and “Red of Chioggia” [5]. The last of these biotypes is the most widespread and well known, while all the others represent locally valuable high-quality crops. Although a clear-cut morphological differentiation among the five biotypes does exist, their genetic distinctiveness, relationships and phylogenies are becoming increasingly important for breeders, producers and consumers.

There is no documented history about the origin of coloured leaf chicory in Italy. All red types of Radicchio currently cultivated appear to derive from red-leaved individuals first introduced in the 15th century. According to Bianchedi [9], the cultivation of red chicory dates back to the first half of the 16th century. It is largely accepted that the original type corresponds to the “Late Red of Treviso” since it was for a long time the only cultivated Radicchio in the Venetian territories surrounding the ancient town of Treviso [5]. After spreading to nearby lands, the original type underwent strong morphological and agronomic selection according to very different criteria adopted by individual farmers, but at least partially due to or dependent on the various environmental conditions of cultivation. Thus, after many years of repeated hybridization and selection carried out by farmers within their own populations, a heavy head with imbricated leaves was bred and this new type called “Early Red of Treviso” becoming locally popular in 1965-70 [5]. Meanwhile, crosses between red-leaved plants of *C. intybus* and plants of *C. endivia* - occurred spontaneously or intentionally performed by farmers back in the 18th century [10,11] - enabled to obtain a new type with red spotted or variegated leaves, currently known as “Variegated of Castelfranco”, related to a small medieval town in the province of Treviso. Later on, in the area of Chioggia, a traditional horticultural area established on sandy soils extending southward from this small sea-side town just south of Venice, new types with variegated- and red-leaved traits able to

form rather conical or spherical, and tightly closed heads were originally generated around 1930 and 1950, respectively [5]. Similarly, in the agricultural area of the town of Verona, a small winter hardy type forming a rosette of deep-red coloured and egg-shaped leaves was initially selected from the “Late Red of Treviso” and then, in 1950-60, populations of “Red of Verona” were obtained and started to be cultivated locally [5].

As a matter of fact, the biotype “Red of Chioggia” is by far the most widely grown among the various cultivar groups of Radicchio and it presents the highest within-type differentiation among cultivars in terms of earliness able to guarantee production almost year-round. Indeed, this biotype of Radicchio has exhibited great adaptability to very different environmental situations worldwide, becoming the most grown one outside the Italian territory and thus the most known at the international level [3,5]. In Italy, the Radicchio of Chioggia is cultivated on a total area of approximately 18,000 ha, half of which is in the Veneto region, with a total production of approximately 270,000 tons (more than 60% obtained using professional seeds), reaching an overall turnover of approximately 10,000,000 euro per year [3].

During the past two decades, the agricultural scenery in the Mediterranean countries has profoundly changed for chicory cultivations, including Radicchio biotypes, where subsistence mixed farming units have been transformed into extensive farming systems growing mainly modern improved varieties instead of local varieties. In recent years, professional breeders have developed protocols based on controlled hybridizations among chosen individual plants to obtain genetically improved synthetic varieties showing higher distinctiveness, uniformity, and stability for both agronomic and esthetical traits [7,8]. The modern breeding programs aim to isolate individuals within the best local populations for the selection of inbred lines suitable for the production of commercial F₁ hybrids [7,8]. These programs are increasingly assisted by the use of molecular markers in order to breed genetically distinguishable, uniform and stable varieties [8,12,13]. Radicchio materials, grown in the second half of the last century, not only provide a valuable source for potentially useful traits, but they are also an irreplaceable bank of co-adapted genotypes. In fact, the Radicchio germplasm was represented by local populations, yearly maintained by farmers through mass selection and known to possess a high variation and adaptation to the natural and anthropological environment where they have originated and have been cultivated for a long time.

This research deals with the use of molecular markers for fingerprinting genomic DNA of Venetian Radicchio biotypes that belong to the five main varietal groups and that correspond to old farmer populations cultivated locally in the 1980-90s. Overall results highlighting the genetic structure and distinctiveness of single populations and biotypes, along with the phylogenies and

genetic relationships among these varietal groups of Venetian Radicchio are presented and critically discussed.

2. Materials and Methods

2.1. Plants materials

An overall number of 797 samples of Radicchio, provided by “Veneto Agricoltura”, and belonging to the five biotypes named hereafter TvT, TvP, Vr, Cf and Ch (respectively Late Red of Treviso”, “Early Red of Treviso”, “Red of Verona”, “Variegated of Castelfranco” and “Red of Chioggia”), that represented 23 populations in total, were selected for DNA fingerprinting analyses through RAPD and AFLP markers. Samples used belonged to local populations phenotypically selected by farmers.

2.2. Genomic DNA isolation

DNA extraction of each sample was performed by means of the procedure described by Barcaccia and Rossellini [14]. DNA quality and quantity of the obtained extracts were evaluated using spectrophotometry and gDNA integrity was verified through agarose gel electrophoresis in a 1% agarose/1× TAE gel containing 1× Sybr® Safe DNA gel stain (Life Technology, Carlsbad, CA, United States). After these evaluations, good quality gDNA samples were used for PCR amplification.

2.3. Molecular markers

PCR parameters and gel electrophoresis in RAPD markers analysis were those described by Barcaccia et al. [15]. Primers' sequences (Operon Technologies, Inc.) used are reported in **Table 1**. PCRs were conducted by means of *Taq* DNA polymerase (EURx Ltd. Poland) in a total volume of 25 µL using the following conditions: 1× POL Buffer A, 2,5 mM of 25 mM MgCl₂, 0,2 µL of dNTP Mix (5 mM), 10 µM primers (1:10), 1,25 U of *Taq* DNA polymerase, 2 µL of DNA (20 ng/µL) and dH₂O to reach the final volume. PCR parameters were chosen following the procedure described by Barcaccia et al. [7]. Amplification products were separated by electrophoresis in 2% agarose/ 1× TAE containing 1× Sybr® Safe DNA gel stain (Life Technology, Carlsbad, CA, United States). On the other hand, AFLP analyses were performed according to Barcaccia et al. [15] with some modifications. The analysis of AFLP loci was based on the detection of *PstI/MseI* genomic restriction fragments by PCR amplification of three different primer combinations having 2 and 3 selective nucleotides for *PstI* and *MseI*, respectively (**Table 1**).

Table 1. RAPD and AFLP (with restriction enzymes) primers used

Primer name	Primer sequence
RAPD	
OP-P01	GTAGCACTCC
OP-Q17	GAAGCCCTTG
OP-Q03	GGTCACCTCA
OP-A08	GTGACGTAGG
OP-M10	TCTGGCGCAC
AFLP	
<i>Pst</i> I+AA/ <i>Mse</i> I+CAA	GACTGCGTACATGCAGAA GACGATGAGTCCTGAGAGTAACAA
<i>Pst</i> I+AT/ <i>Mse</i> I+CAA	GACTGCGTACATGCAGAT GACGATGAGTCCTGAGAGTAACAA
<i>Pst</i> I+AG/ <i>Mse</i> I+CAG	GACTGCGTACATGCAGAG GACGATGAGTCCTGAGAGTAACAG

2.4. Genetic relationships analysis

After the gel electrophoresis and the molecular fingerprints screening, two initial datasets were recovered, comprising 35 molecular traits from RAPD and 92 from AFLP, for each sample. In a few cases, missing data were present that could have invalidated the forthcoming bioinformatical analyses. For this reason, and following the preliminary analyses on the complete datasets of single markers, a threshold of 2% missing data for filtering the datasets was considered to reduce the possibility of misleading results. An initial genetic distance (GD) analysis was performed on NTSYS v2.21 software [16] that was based on the default NEI72 algorithm. The same protocol was adopted for both the RAPD and the AFLP datasets. The two GD matrixes were then used for the construction of two neighbour-joining (NJ) trees. After this, the two datasets were also combined into a bigger one that comprehended 652 samples that had, for both the kinds of molecular markers used (127 traits), a percentage of missing data < 2%. Using the combined dataset, the same genetic distance analysis was performed, and the resulting NJ tree was then used as the initial tree in the subsequent phylogenetic analysis.

The phylogenetic analysis was performed according to the maximum-likelihood method (ML) implemented in the IQ-Tree v1.6.12 software [17]. RAPD and AFPL matrices were analyses as morphological data using the best fitting model identified by the ModelFinder algorithm available in IQ-Tree [18]. The BIC values were used to identify the best fitting model (MK+FQ+I+G4) for both datasets [18]. Initially, the RAPD and AFPL datasets were analysed independently. Successively, they were merged in single set, and studied simultaneously. Ten independent runs were performed in each phylogenetic analysis to minimise the possibility to be entrapped in suboptimal trees.

The discrepancies among the obtained topologies were identified with the Phylo.io application [19] and further assessed by visual inspection. The statistical supports to the phylogenetic trees were computed by running 10,000 replicates and until convergence for the standard bootstrap (BT) [20], the ultrafast bootstrap (UFB) [21,22] and the SH-like approximate likelihood ratio tests (SH-aLRT) [23]. Significant values: bootstrap ≥ 75 , UFB ≥ 90 , and SH-aLRT ≥ 75 were indicated in **Supplementary Figure 1**. According to literature, the TvT clade was set as root.

2.5. Genetic structure analyses of the core collection

Parallely, a Bayesian clustering algorithm implemented in STRUCTURE v.2.2 [24] was used to model the genetic structure of the Radicchio core collection. The number of founding groups ranged from 1 to 20, and 10 replicate simulations were conducted for each value of K based on a burn-in of 20,000 and a final run of 100,000 Markov chain Monte Carlo (MCMC) steps. STRUCTURE HARVESTER [25] was used to estimate the most likely value of K, and the estimates of membership were plotted as a histogram using an Excel spreadsheet.

2.6. Genetic statistics

The average number of alleles (n_o), the effective number of alleles (n_e) according to Kimura and Crow [26] and the polymorphic loci expressed in number (n_{pl}) and percentage ($\%_{pl}$) were calculated for each population considered in the combined markers dataset and for the clusters identified through the ML supported tree obtained from IQ-Tree software. Also, Nei's [27] genetic diversity statistics were computed and averaged for the 22 populations and the 5 biotypes over all RAPD and AFLP loci of the combined dataset to evaluate the total diversity of the entire core collection (H'_T), the within population diversity (H'_s), the among genetic differentiation (D_{ST}) and the proportion expressed between populations (G_{ST}) parameters. From G_{ST} , Gene flow (Nm) was estimated as follows: $Nm = 0.5(1 - G_{ST})/G_{ST}$ [28]. Moreover, the polymorphism degree was calculated over all the 22 populations and the 5 clusters using Shannon's information index (I) of phenotypic diversity [29]. All statistics were computed using POPGENE version 1.32 [30].

2.7. Genetic similarity estimates

Genetic similarity (GS) based on Dice's coefficient [31] was computed in all pairwise comparisons using the combined markers dataset. Moreover, mean GS within and among single populations was calculated. Also, from the entire genetic similarity matrix, GS and standard error were calculated for each of the five clusters previously identified. Genetic similarity was calculated using NTSYS software v2.21 [16]

3. Results

3.1. Genetic relationships among Radicchio populations and biotypes

Phylogenetic investigations performed on singles datasets produced a series of incongruent and unstable trees, *i.e.* each independent run (see Methods) produced a different topology, probably in consequence of the molecular markers used, their heritability, numerosity and variability (data not shown). Conversely, the obtained results from the combined matrix showed a relatively more stable topology both in the backbone and in the clusters within varieties (**Figure 1**). The best tree had a topology where the higher statistical corroboration was associated to terminal or sub-terminal nodes, but some notable exceptions occurred in basal and sub-basal nodes (supplementary figure).

The Vr, TvP and Ch biotype samples formed monophyletic groups. The Vr and TvP received statistical support by all tests, while the Ch clade was corroborated by the BT and SH-aLRT values (supplementary figure). The TvT samples were almost all grouped together (**Figure 1**). However, four of them (TvT2_15, TvT2_16, TvT2_34, TvT2_32) clustered with the Vr clade. Finally, the Cf samples formed globally a cluster that was paraphyletic with respect to the Ch clade (**Supplementary Figure 1**).

The node splitting the TvP clade and the Cf+Ch group was supported only by UFB. Within each of the main para/monophyletic biotypes were present several subclades. Most of them were supported by one or more statistical tests at different ranks, probably mirroring the AFLP clustering.

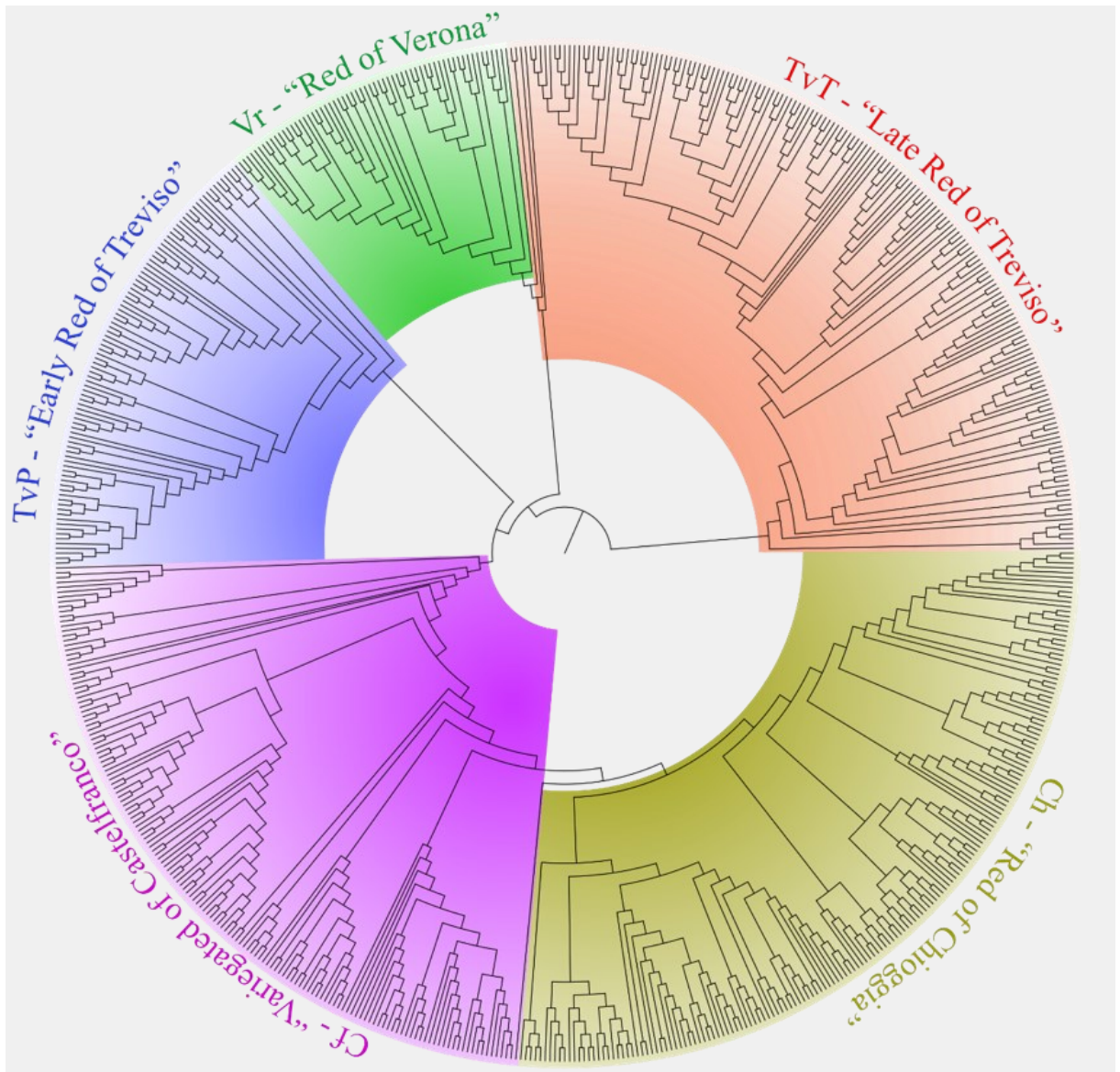


Figure 1. Maximum likelihood tree topology ($-\ln = 5698.207$) depicting the phylogenetic relationships among the different biotypes.

3.2. Genetic structure of the Radicchio core collection and biotypes clustering

Regarding the investigation of the genetic structure of the Radicchio core collection, the STRUCTURE harvester software estimated the best value of K equal to 5, ($\Delta K = 238.6$ in **Figure 2**) and the memberships of 652 samples grouped them in accordance with the biotypes they belonged to. Each group was labelled using the same colours of the ML tree previously described, with 634 samples showing strong memberships with the respective cluster ($>90\%$), and 18 samples presenting admixtures between the five groups identified. The vast majority of admixed samples belonged to the Red of Chioggia and the Variegated of Castelfranco biotypes (5 and 13 samples with main membership below 90%, respectively), but two samples belonging to the “Late Red of Treviso” were present with membership to the respective group equal to 79.8% and 79.4%, so slightly lower than the considered threshold (**Figure 3**)

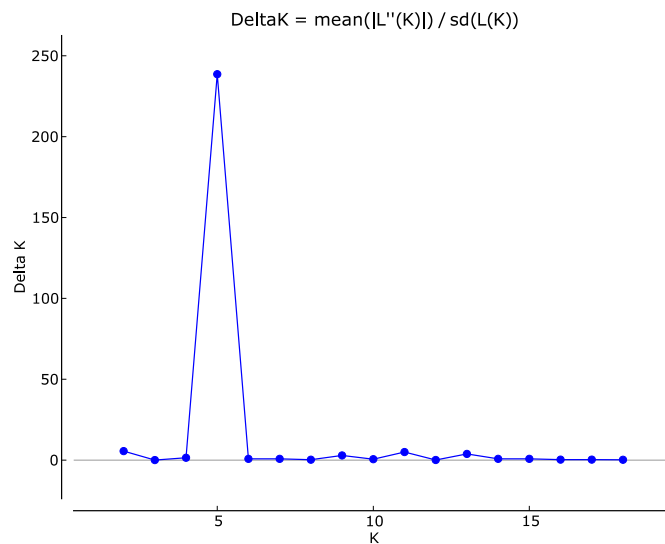


Figure 2. Graph representing the ΔK values resulting from Structure Harvester software [25]

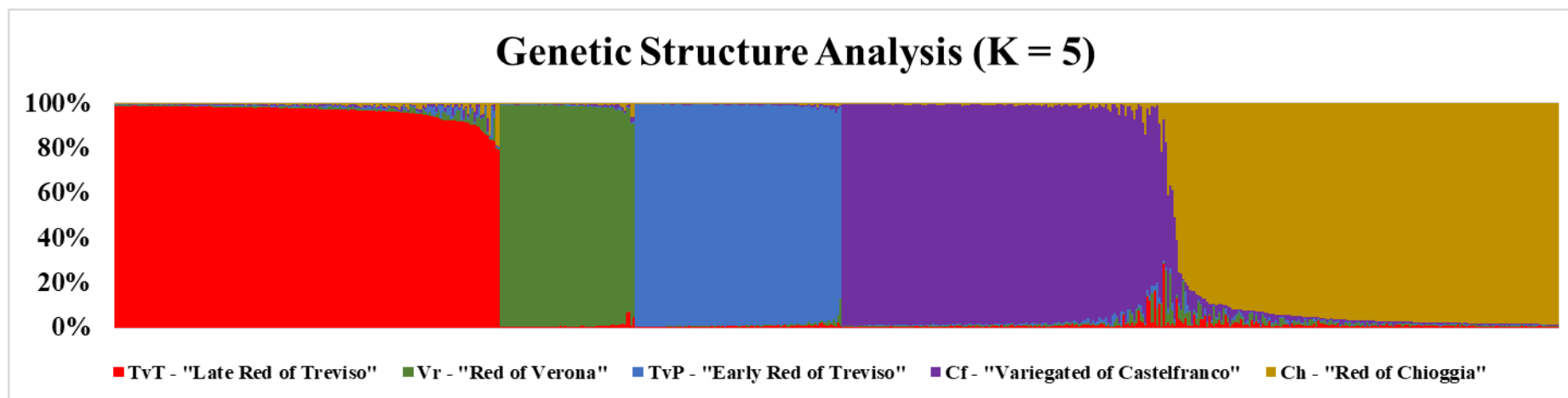


Figure 3. Genetic structure analysis of the Radicchio core collection. Identified most probable value of K = 5. Clusters and samples' memberships agree with the biotype of Radicchio they belong to.

3.3. Genetic statistics of populations and biotypes

Descriptive statistics overall RAPD and AFLP markers for single and grouped populations, based on their biotype, and the whole Radicchio core collection are reported, together with Nei's diversity statistics and gene flow estimates, in **Table 2**. The average number of observed alleles per locus (n_a) calculated among all biotypes was equal to 1.480, while the average effective number of alleles (n_e) per locus was equal to 1.216 (**Table 2**). The percentage of polymorphic loci on the whole core collection was 48%, ranging from 21.3% in the Vr biotype to 36.2% in the Ch one. The Nei's unbiased genetic diversity, calculated among the entire dataset was $H' = 0.130$, and it varied between 0.070 (Overall Vr) and 0.110 (Overall Ch). Shannon's information index (I) of phenotypic diversity overall biotypes was $I = 0.202$, with the minimum value calculated for Vr (0.105) and the maximum one for Ch (0.170) (**Table 2**). Dice's genetic similarity (GS) was also computed that ranged between 0.946, within Chioggia biotype, and 0.962, within Verona biotype. The average GS calculated among the 652 samples of the Radicchio core collection was equal to 0.931 (**Table 2**). Nei's diversity statistics were calculated for each Radicchio population, biotype and overall. The total genetic diversity assessed with the molecular marker dataset was $H'_T = 0.132$, and it was higher in the Ch biotype (0.109), while lower in the Vr one (0.070). The average within genetic diversity calculated overall was $H'_S = 0.068$, with values been between 0.002 of Ch1 and 0.053 of Ch3. Among biotypes, the H'_S value was the highest in Ch ($H'_S = 0.079$) and the lowest in both TvP and Cf ($H'_S = 0.060$). The genetic differentiation (D_{ST}) among the core collection was equal to 0.064, with the highest values calculated for Ch1 (0.107) and the lowest one for Vr1 (0.028). The biotypes genetic differentiation estimates ranged from 0.006 (Red of Verona) and 0.030 (Red of Chioggia). The proportion of the overall genetic diversity among the core collection was $G_{ST} = 0.488$, while the gene flow estimate was $N_m = 0.524$. Specifically, the biotypes' gene flow was never below 1 and ranged from 1.068, in the Variegated of Castelfranco, to 5.032, in the Red of Verona, while the G_{ST} values were between 0.090 (Verona) and 0.319 (Castelfranco) (**Table 2**). Noteworthy, the two Red of Verona populations had N_m values of 0.731 and 0.422, in Vr1 and Vr2 respectively.

Table 2. Descriptive statistics over all SSR loci including the number of individuals, number (n_{pl}) and percentage ($\%_{pl}$) of polymorphic loci, mean number of observed (n_a) and effective (n_e) alleles per locus, Nei's genetic diversity (H') Shannon's information index (I), Dice's genetic similarity (GS) coefficient, total genetic diversity per biotypes and overall (H'_T), expected heterozygosity (H'_s) within biotypes and overall, genetic differentiation (D_{ST}) and proportional genetic diversity (G_{ST}), and gene flow estimates (N_m)

Population ID	N° of Individuals	n_{pl}	$\%_{pl}$	n_a	n_e	H'	I	GS	H'_T	H'_s	D_{ST}	G_{ST}	N_m
TvT1	18	28	22.1%	1.221	1.144	0.080	0.120	0.951		0.007	0.080	0.924	0.041
TvT2	30	23	18.1%	1.181	1.124	0.065	0.096	0.962		0.044	0.043	0.494	0.512
TvT3	30	25	19.7%	1.197	1.120	0.067	0.101	0.961		0.030	0.057	0.656	0.262
TvT4	36	19	15.0%	1.150	1.094	0.049	0.074	0.971		0.025	0.062	0.715	0.199
TvT5	28	17	13.4%	1.134	1.096	0.049	0.072	0.970		0.027	0.059	0.684	0.231
TvT6	32	25	19.7%	1.197	1.126	0.065	0.098	0.962		0.024	0.063	0.723	0.191
Overall TvT	174	44	34.7%	1.347	1.136	0.084	0.133	0.958	0.086	0.067	0.019	0.225	1.727
s.d.				0.478	0.274	0.152	0.223	0.014	0.024	0.016			
Vr1	35	23	18.1%	1.181	1.107	0.061	0.092	0.965		0.041	0.028	0.406	0.731
Vr2	26	25	19.7%	1.197	1.111	0.060	0.093	0.964		0.032	0.038	0.542	0.422
Overall Vr	61	27	21.3%	1.213	1.119	0.070	0.105	0.962	0.070	0.063	0.006	0.090	5.032
s.d.				0.411	0.277	0.151	0.220	0.013	0.023	0.019			
TvP1	30	22	17.3%	1.173	1.099	0.053	0.080	0.968		0.030	0.048	0.614	0.314
TvP2	35	21	16.5%	1.165	1.099	0.059	0.089	0.965		0.003	0.075	0.959	0.021
TvP3	28	24	18.9%	1.189	1.113	0.068	0.102	0.959		0.050	0.029	0.368	0.859
Overall TvP	93	32	25.2%	1.252	1.130	0.078	0.120	0.953	0.078	0.060	0.018	0.229	1.687
s.d.				0.436	0.274	0.154	0.226	0.015	0.024	0.015			
Cf1	22	24	18.9%	1.189	1.126	0.066	0.099	0.960		0.045	0.043	0.489	0.523
Cf2	36	16	12.6%	1.126	1.075	0.043	0.065	0.975		0.024	0.064	0.724	0.190
Cf3	29	21	16.5%	1.165	1.112	0.060	0.089	0.966		0.042	0.046	0.523	0.456
Cf4	31	25	19.7%	1.197	1.111	0.076	0.111	0.956		0.025	0.063	0.715	0.199
Cf5	33	20	15.8%	1.158	1.094	0.054	0.080	0.970		0.016	0.072	0.814	0.114
Overall Cf	151	37	29.1%	1.291	1.142	0.085	0.131	0.952	0.088	0.060	0.028	0.319	1.068
s.d.				0.456	0.282	0.159	0.233	0.014	0.026	0.015			
Ch1	23	21	16.5%	1.167	1.101	0.057	0.087	0.966		0.002	0.107	0.982	0.009
Ch3	35	22	17.3%	1.175	1.109	0.065	0.096	0.963		0.053	0.057	0.519	0.463
Ch4	35	27	21.3%	1.213	1.130	0.073	0.109	0.960		0.056	0.054	0.493	0.515
Ch5	23	26	20.5%	1.205	1.124	0.067	0.101	0.963		0.047	0.062	0.569	0.379
Ch6	26	32	25.2%	1.205	1.124	0.079	0.121	0.954		0.050	0.059	0.543	0.421
Ch7	31	32	25.2%	1.252	1.196	0.084	0.127	0.952		0.048	0.061	0.561	0.391
Overall Ch	173	46	36.2%	1.362	1.181	0.110	0.170	0.946	0.109	0.079	0.030	0.274	1.323
s.d.				0.483	0.304	0.170	0.249	0.017	0.029	0.016			
Mean Overall	652	61	48.0%	1.480	1.216	0.130	0.202	0.931	0.132	0.068	0.064	0.488	0.524
s.d.				0.502	0.328	0.178	0.257	0.020	0.032	0.011			

3.4. Genetic similarity analysis within and between populations and biotypes

Dice's genetic similarity was computed to create a GS matrix in all pair-wise comparisons. Average GS values within and among each population and biotype were calculated to create two GS matrices (**Table 3** and **Table 4**, standard errors were computed that were below 0.001 overall). Average genetic similarity within populations is reported in **Table 2**, while that calculated among them is reported in **Table 3** and ranged from 0.891 (TvP3 vs. Ch6) to 97.1 (TvT4 vs. TvT5).

Similarly, the average genetic similarity calculated among and within biotypes is reported in **Table 4**, with the within values also reported in **Table 2**. The average GS values calculated by comparing the different biotypes ranged from 0.905 (TvP vs. Ch) and 0.934 (Cf vs. Ch).

4. Discussion

4.1. Molecular characterization of the core collection of Radicchio: novel genetic insights integrate and corroborate available historical information on the origin of biotypes

From the analyses of the genetic variability and relationships, including genetic diversity and similarity statistics, and gene flow estimates, it was observed that the five Radicchio biotypes analysed are distinguishable from each other and that they are uniform within each of the biotypes. Specifically, the genetic differentiation (D_{ST}) detected within biotypes was lower than 3.0%, thus meaning that the populations they were constituted by were highly similar, while the core collection's value was above 6%. This result was also obtained through Dice's GS analysis, from which within GS estimates were above 94.6% and the among ones were below 93.4%. Once again, these findings were also supported by the low N_m values calculated for the entire core collection (Overall $N_m < 1$), and those observed within groups (within biotypes $N_m > 1$), thus demonstrating that genotypes had low gene migration among them, while that of their populations was higher, hence gene flow occurs among populations of the same biotype, but it is low or absent among different ones.

The ML analysis reflects the distinctiveness of the biotypes but not that of the populations within them (**Figure 1**). Selecting TvT as the ML-tree root following Bianchedi's work [9], the selection events described by historical literature was supported. Thus, their relationship could be linked and also supported by the chronological events. The paraphyly of Late Red of Treviso is likely related to the abundance of fixed marker loci into the combined dataset, such as the instability of the sub-basal and sub-terminal nodes within each clade. Nevertheless, TvT and Vr forms a sister group of the other biotypes (TvP, Cf and Ch). Considering the historical report indicating that the Vr biotype was selected using phenotypic traits from a group of TvT [5], probably located in the related province of Verona, we can infer the likelihood of the results. The positioning on the tree of the "Early Red of Treviso", placed later on of the group of TvT and Vr, is justified by its less genetic similarity to its ancestor even if it is still better related to it than the other two biotypes, noteworthy, being most likely interspecific crosses (Cf and Ch) between TvT and endive (*C. endive*). Considering some typical morphological traits (*i.e.*, leaf shape, white ribbing and percentage of red-leaf area) shared between TvT and TvP biotypes, this result supports the recent formation of the biotype, which was reasonably constituted through more professional skills and thorough methods than what occurred in the decade before to the Vr biotype. Furthermore, the clear separation of the Variegated of Castelfranco and the Red of Chioggia from the other biotypes was

attributed to their ancestral interspecific origin. The fact that these two lineages were clustered subsequently, with Ch placed more distant to TvT (ML-tree root) was, in our hypothesis, due to the fact that firstly Cf was obtained (during the XVIII century) and, from it, Ch was differentiated in the first half of the last century from variegated sub-populations. The tree topology supports this hypothesis disposing all Cf accessions basal to those of Ch, thus clustering them as a monophyletic group. In addition to this, sample CF5_27 placed basal to Ch cluster, and statistically supported (SH-aLRT and UFB), indicates the solid relatedness reported into the historical reports [10,11].

Furthermore, the genetic structure analysis agreed in the clustering of samples and biotypes with the ML analysis. The obtained results were able to group the accessions depending on their belonging biotype in both cases, and in it was observed that Ch cluster was the one presenting the higher variability, findings that were supported also by the GS and genetic variability results. This said, Ch and Cf biotypes showed in different analyses to be more related among them than with the others (GS = 93.4% (**Table 4**), 18 admixed samples (**Figure 2**), ML-tree clustering (**Figure 1**).

Summarising, the DNA fingerprinting approach used to molecularly characterise these 22 old farmers populations of Radicchio enabled to verify not only the genetic distinctiveness of the five most common biotypes cultivated in the Veneto region, but it also allowed the reconstruction of the genetic relationships among these locally cultivated varieties and to corroborate the historical information available on their origin and chronological order of appearance, despite overall supporting statistics (BT, SH-aLRT, UFB) among the tree nodes. Further investigations, performed using a significant number of discriminant co-dominant marker loci widespread into the genome or adding information retrieved from other analyses of mapped or expressed molecular markers, could increase the statistical supports and the resolution of the tree.

5. Conclusions

5.1. Marker-assisted characterization and genomic selection of Radicchio populations

Historically, most cultivated varieties of Venetian Radicchio have been developed using mass selection to obtain uniform populations characterized by high yield and suitable commercial standards [3]. Currently, two genetically distinct types of chicory cultivars are on the market: OP or synthetics and F₁ hybrids [7,8]. Newly released cultivars are mostly synthetics, developed through inter-crossing or poly-crossing among many selected parental individuals or clonal lines, followed by progeny testing to assess general combining ability [3,7]. By their nature, synthetics have a wide genetic base represented by a mixture of highly heterogeneous and heterozygous individuals, yet showing rather similar phenotypes. In recent years, however, developing F₁ hybrid cultivars has

become more common, mainly done in the private sector [3,8]. Experimental data on how these hybrids are developed are currently scarce, and presumably, each company employs its own protocol depending on genetic materials used and the system(s) of pollination control during inbred line development and F₁ hybrid seed production [3]. In general, the strong self-incompatibility (SI) system in chicory has been a great barrier to the development of parental inbred lines or clones used to produce single-cross hybrids [3,5,8]. However, there has been an increased interest in the production of F₁ hybrids due to the discovery of male-sterility genes [32,33]. For instance, an increasing number of cultivars of the Witloof and Radicchio types are commercialized as true F₁ hybrids. Further, owing to the economic benefits, most newly released varieties of leaf chicory are F₁ hybrids, mainly developed by European seed companies. Moreover, most commercial breeding programs have improved their efficiency during the past several years due to the use of genomic tools. Various types of molecular tools, including SSR, EST and SNP markers, have been implemented for genotyping elite breeding stocks of leaf chicory [8,12,33]. The available data show that markers have been reliable for assessing multi-locus genotypes of individual plants, breeding stocks and lineages, including assessing the degree of homozygosity of inbred lines and their genetic stability. Moreover, markers have also been used to accurately estimate the specific combining ability between parental lines, as judged based on their genetic diversity and predicted degree of heterozygosity in their F₁ hybrid progeny. Such information could be utilized for planning 2-way crosses and predict heterosis of the experimental F₁ hybrids on the basis of genetic distance and allelic divergence between parental inbred lines. Information on the parental genotypes would also allow protection of newly registered cultivars assessment of genetic purity and identity of the seed stocks of commercial F₁ hybrids.

In conclusion, local farmer varieties of Venetian Radicchio represent invaluable and irreplaceable genetic resources, which should be collected and preserved in gene banks for characterization and future exploitation by breeding programs. Most of these germplasms are local farmer-derived varieties and a few of them include also professional breeder-improved varieties that typically exhibit a great deal of genetic diversity in morphological and physiological characteristics, highly desirable to breeding programs. However, it appears that variations in traits related to biotic and abiotic stresses are scarce in these local populations. The reason could be that farmers selections were traditionally focused on morphological and esthetical characteristics important to the market instead of selecting for disease resistance, abiotic stress tolerance, or post-harvest quality traits.

6. References

1. Simko, I.; Jia , M.; Venkatesh, J.; Kang, B.-C.; Weng, Y.; Barcaccia, G.; Lanteri, S.; Bhattarai, G.; Foolad, M.R. Genomics and marker-assisted improvement of vegetable crops. *Critical Reviews in Plant Sciences* **2021**, 10.1080/07352689.2021.1941605, doi:10.1080/07352689.2021.1941605.
2. BREMER, K.J.C.; classification. Asteraceae. **1994**.
3. Barcaccia, G.; Ghedina, A.; Lucchin, M. Current Advances in Genomics and Breeding of Leaf Chicory (*Cichorium intybus* L.). *Agriculture-Basel* **2016**, 6, 50, doi:10.3390/agriculture6040050.
4. M Kiers, A. Endive, chicory, and their wild relatives. A systematic and phylogenetic study of *Cichorium* (Asteraceae). *Gorteria-Supplement* **2000**, 5, 1-77.
5. Lucchin, M.; Varotto, S.; Barcaccia, G.; Parrini, P. Chicory and Endive. Handbook of plant breeding, vegetables I: asteraceae, brassicaceae chenopodi-caceae. New York: Springer: 2008.
6. Barcaccia, G.; Varotto, S.; Soattin, M.; Lucchin, M.; Parrini, P. Genetic and molecular studies of sporophytic self-incompatibility in *Cichorium intybus* L. In Proceedings of Proceedings of the Eucarpia meeting on Leafy Vegetables Genetics and Breeding.
7. Barcaccia, G.; Pallottini, L.; Soattin, M.; Lazzarin, R.; Parrini, P.; Lucchin, M. Genomic DNA fingerprints as a tool for identifying cultivated types of radicchio (*Cichorium intybus* L.) from Veneto, Italy. *Plant Breeding* **2003b**, 122, 178-183, doi:DOI 10.1046/j.1439-0523.2003.00786.x.
8. Patella, A.; Scariolo, F.; Palumbo, F.; Barcaccia, G. Genetic Structure of Cultivated Varieties of Radicchio (*Cichorium intybus* L.): A Comparison between F1 Hybrids and Synthetics. *Plants-Basel* **2019**, 8, 213, doi:10.3390/plants8070213.
9. Bianchedi, A.J.I.A. I radicchi di Treviso: storia, coltivazione, forzatura e commercio. **1961**, 98, 37-51.
10. Pimpini, F.; Chillemi, G. Evoluzione delle tecniche colturali e prospettive di sviluppo dei radicchi veneti. *Atti Convegno "I radicchi veneti* **1993**, 20, 9-28.
11. Pimpini, F.; CHILLEMI, G.; LAZZARIN, R.; BERTOLINI, P.; MARCHETTI, C. Il Radicchio Rosso di Chioggia. Aspetti tecnici ed economici di produzione e conservazione. *Edizione in sito interet di Veneto Agricoltura, anno* **2001**.
12. Ghedina, A.; Galla, G.; Cadalen, T.; Hilbert, J.L.; Caenazzo, S.T.; Barcaccia, G. A method for genotyping elite breeding stocks of leaf chicory (*Cichorium intybus* L.) by assaying mapped microsatellite marker loci. *BMC Res Notes* **2015**, 8, 831, doi:10.1186/s13104-015-1819-z.

13. Galla, G.; Ghedina, A.; Tiozzo, S.C.; Barcaccia, G. Toward a first high-quality genome draft for marker-assisted breeding in leaf chicory, Radicchio (*Cichorium intybus* L.). In *Plant genomics*, IntechOpen: 2016.
14. Barcaccia, G.; Rossellini, D. A quick method for the isolation of plant DNA suitable for RAPD analysis. *Journal of Genetics Breeding* **1996**, *50*, 177-180.
15. Barcaccia, G.; Mazzucato, A.; Albertini, E.; Zethof, J.; Gerats, A.; Pezzotti, M.; Falcinelli, M.J.T.; Genetics, A. Inheritance of parthenogenesis in *Poa pratensis* L.: auxin test and AFLP linkage analyses support monogenic control. **1998**, *97*, 74-82.
16. Rohlf, F.J. *NTSYS-pc Numerical Taxonomy and Multivariate Analysis System* **1993**.
17. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **2015**, *32*, 268-274, doi:10.1093/molbev/msu300.
18. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **2017**, *14*, 587-589, doi:10.1038/nmeth.4285.
19. Robinson, O.; Dylus, D.; Dessimoz, C. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Mol Biol Evol* **2016**, *33*, 2163-2166, doi:10.1093/molbev/msw080.
20. Felsenstein, J. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **1985**, *39*, 783-791, doi:Doi 10.2307/2408678.
21. Minh, B.Q.; Nguyen, M.A.; von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **2013**, *30*, 1188-1195, doi:10.1093/molbev/mst024.
22. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **2018**, *35*, 518-522, doi:10.1093/molbev/msx281.
23. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **2010**, *59*, 307-321, doi:10.1093/sysbio/syq010.
24. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945-959.
25. Earl, D.A.; Vonholdt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **2012**, *4*, 359-361, doi:10.1007/s12686-011-9548-7.
26. Kimura, M.; Crow, J.F.J.G. The number of alleles that can be maintained in a finite population. **1964**, *49*, 725.
27. Nei, M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **1973**, *70*, 3321-3323, doi:10.1073/pnas.70.12.3321.

28. McDermott, J.M.; McDonald, B.A. Gene Flow in Plant Pathosystems. *Annual Review of Phytopathology* **1993**, *31*, 353-373, doi:DOI 10.1146/annurev.py.31.090193.002033.
29. Lewontin, R.C. The apportionment of human diversity. In *Evolutionary biology*, Springer: 1972; pp. 381-398.
30. Yeh, F.C.; Yang, R.; Boyle, T.B.; Ye, Z.; Mao, J.X.J.M.b.; biotechnology centre, U.o.A., Canada. POPGENE, the user-friendly shareware for population genetic analysis. **1997**, *10*, 295-301.
31. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297-302.
32. Gonthier, L.; Blassiau, C.; Morchen, M.; Cadalen, T.; Poiret, M.; Hendriks, T.; Quillet, M.C. High-density genetic maps for loci involved in nuclear male sterility (NMS1) and sporophytic self-incompatibility (S-locus) in chicory (*Cichorium intybus* L., Asteraceae). *Theor Appl Genet* **2013**, *126*, 2103-2121, doi:10.1007/s00122-013-2122-9.
33. Palumbo, F.; Qi, P.; Pinto, V.B.; Devos, K.M.; Barcaccia, G. Construction of the First SNP-Based Linkage Map Using Genotyping-by-Sequencing and Mapping of the Male-Sterility Gene in Leaf Chicory. *Frontiers in Plant Science* **2019**, *10*, 276, doi:10.3389/fpls.2019.00276.

7. Supplementary materials

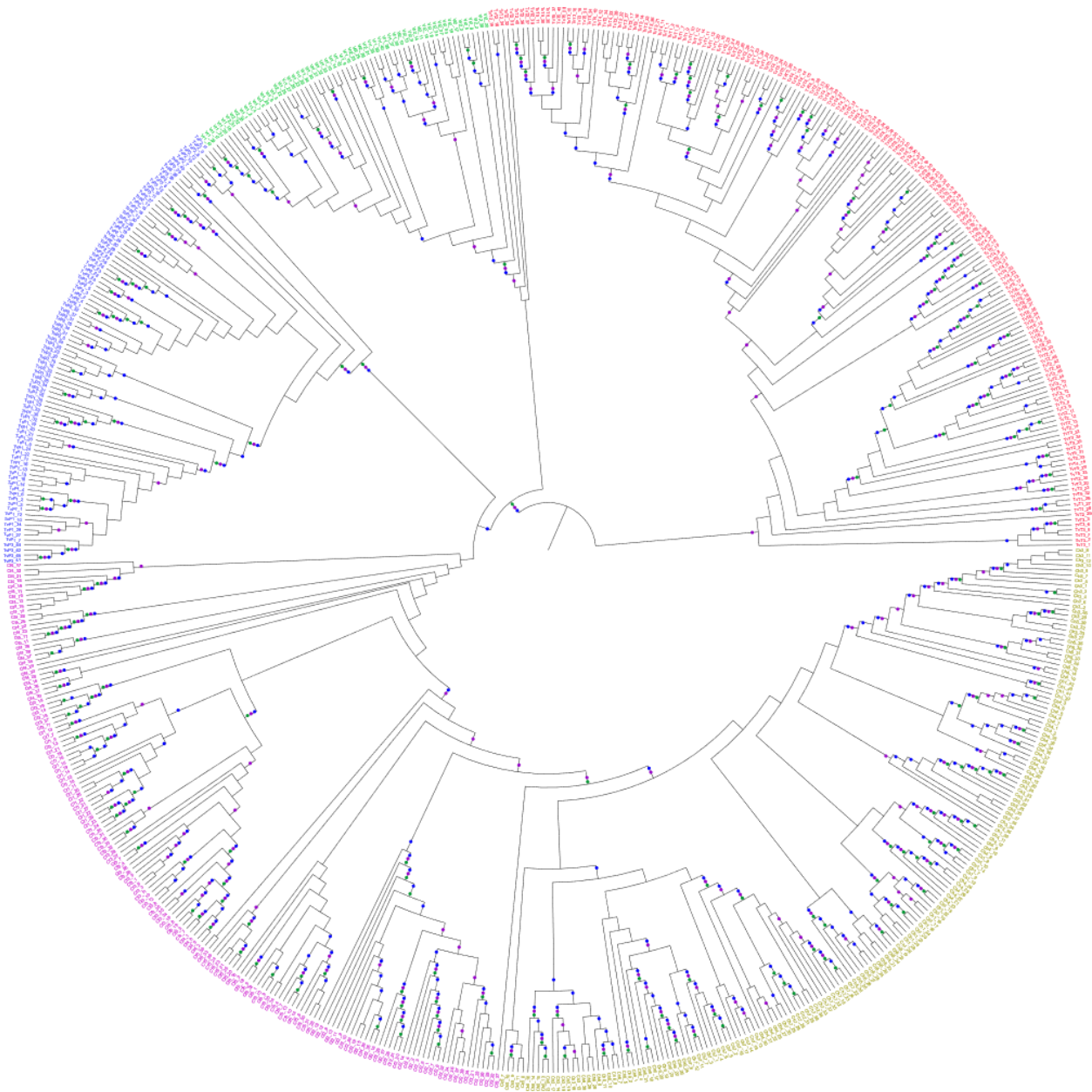


Figure S1. Maximum Likelihood tree ($-\ln = 5698.207$), obtained analyzing the combined dataset of molecular markers (RAPD+AFLP). Three-alphabetic codes show the biotypes: TvT, "Late Red of Treviso"; Vr, "Red of Verona"; TvP, "Early Red of Treviso"; Cf, "Variegated of Castelfranco" and Ch, "Red of Chioggia". Numbers report the original population and the progressive number of the samples within the population. Coloured dots on the nodes indicate the support: ●, Bootstrap (≥ 75); ●, SH-aLRT (≥ 75); ●, UFB (≥ 90).

Chapter III

Molecular characterization and genetic structure evaluation of breeding populations of fennel through microsatellite genotyping

(This chapter has been submitted to *MDPI Horticulturae* journal and is under revision)

0. Abstract

Fennel, or *Foeniculum vulgare* Mill., is an important horticultural crop belonging to the Apiaceae family, that is cultivated worldwide and used in the agri-food sector and for pharmaceutical preparations. Breeding strategies in this species usually involve 3 parental lines, two maternal (one cytoplasmatic male-sterile line and an ideotype representative maintainer line) that are crossed to obtain an ideotype representative cytoplasmatic male-sterile line, and one paternal, used as pollinator in crosses with the derived maternal lines' progeny. From this cross, F1 hybrid progenies are obtained, that are characterized by high levels of heterozygosity, and so, hybrid vigour. In this study, over 450 plants, representing 8 breeding populations and their respective three parental and one progeny lines, were genotyped by means of co-dominant molecular markers. The twelve highly polymorphic microsatellites used enabled the analyses of the genetic variability, distinctiveness and stability of each breeding line. Moreover, the genetic structure of the core collection was investigated, that, together with the homozygosity, gene flow and genetic similarity results, allowed the identification of unsuitable lines to be used in breeding plans due to their low homozygosity (10.4% in the pollinator line of population 7). Moreover, the Bayesian reconstruction of the core collection's genetic structure, based on the co-dominant markers used, allowed to confirm the distinctiveness results obtained from the genetic similarity investigation and the gene flow estimates computed. Among these, it was also observed a trend in hybrids' heterozygosity that increased when the genetic similarity between their respective parental lines decreases.

This research proposes a suitable method for genotyping fennel populations in pre- and post-breeding approaches, like marker-assisted breeding or breeding lines distinctiveness and stability verifications.

Keywords: *SSR makers; genotyping; breeding population genetics, marker-assisted breeding; genetic distinctiveness; genetic stability; populations' genetic structure*

1. Introduction

Fennel (*Foeniculum vulgare* Mill., $2n=2x=22$) is a diploid horticultural crop characterised by a biennial or perennial developmental cycle. This species belongs to the Apiaceae family and it originated in the southern Mediterranean regions [1]. After its domestication, it spread all over the world and began an important crop used for food and pharmaceutical purposes [2-5]. Due to the economic value of this species (nearly 2 million tons around the world, and over 42 thousand tons in Europe in 2019 [6]) and to the increasing market demand, breeders are called upon to respond by developing ever more performing varieties.

Breeding strategies in fennel are mainly based on the constitution and production of F1 hybrids, greatly facilitated by the prevalently allogamous and proterandrous behaviour (i.e. anthers mature before pistils) [7]. Although, self-fertilization usually does not occur within the same flower, it is still possible within the same umbel or between two umbels of the same plant. For this reason, a male sterility system is required. In this regard, hybrid seeds in fennel are commonly obtained by exploiting a three lines-based system characterized by a cytoplasmic male-sterile CMS seed line (strain A), a male fertile sister line (also known as maintainer line, strain B) and a pollinator line (strain C) with a general combining ability (GCA) with the CMS line. Strain C and B are initially obtained through several generations of selfing or sibling to achieve high uniformity and high homozygosity. Strain A is instead developed through backcross, by using strain B as a recurrent parent and a CMS genotype as a non-recurrent parent: after several cycles of backcrossing (usually 6–7), the resulting progeny will be isogenic to strain B except for the cytoplasm, which will be CMS. The newly obtained line (strain A) will be then used as a mother plant (or seed plant) and crossed with strain C for F1 hybrids production whilst strain B will be used to maintain strain A. By crossing highly dissimilar strain C and A, the resulting offspring is expected to exhibit high level of heterozygosity, maximizing the heterotic vigour. [7-9]. Following F1 hybrids development, the registration process of the new plant variety is subject to compliance with rigorous and specific requirements concerning the distinctness (D), uniformity (U), and stability (S). Specifically, the new variety must be distinguishable from those already registered, must be phenotypically uniform and must be stable during subsequent propagation cycles.

The entire process, from the constitution of the parental lines to the evaluation of the resulting F1 hybrids, to the variety registration, is greatly facilitated by the use of molecular markers. In particular, SNPs and SSR are the two most attractive classes of markers because of their reproducibility, co-dominant nature, locus-specificity and random genome-wide distribution. SSR and SNP can be in fact used to genotype the parents in order to select and to cross those ones

genetically more dissimilar; they are pivotal to estimate the homozygosity of the parental lines as well the heterozygosity of the resulting offspring; they are exploited to determine the stability of a new variety and any possible similarity with registered cultivars; they represent an effective tool for addressing legal disputes related to improper use of registered varieties.

In this study, based on the SSR panel set developed by Palumbo et al. [5], 8 breeding populations each represented by parental lines (cytoplasmic male sterile seed plants CMS, maintainers M and pollinators P) and F1 hybrid progenies (H) were genotyped by means of 12 highly polymorphic Simple Sequence Repeats (SSR) markers. These analyses aimed at determining the uniformity of each population in terms of genetic similarity and homozygosity in order to identify any possible correlation between the stability of the F1 hybrids and the genomic background of their parents. The effectiveness of the SSR panel will also be discussed in broader terms for marker assisted breeding (MAB) analyses, DUS testing and varietal registration.

2. Materials and methods

2.1. Plant material

In this study, 451 samples belonging to eight breeding populations of fennel were considered. Each population (numbered from 1 to 8) was composed of four different lines: cytoplasmic male-sterile (CMS), maintainer (M), pollinator (P) and F1 hybrid (H). Considering that populations 5 and 6 shared the same CMS and M lines, overall the study involved 30 lines, each composed of 4-28 individuals.

2.2. DNA isolation and SSR primers

451 genomic DNA (gDNA) were extracted from young leaves using the DNeasy 96 Plant kit (Qiagen, Hilden, Germany) following the protocols provided by the supplier. DNA quality and quantity were estimated by means of a NanoDrop 2000c UV-Vis spectrophotometer (Thermo Fisher, Pittsburgh, PA, United States). The gDNA integrity of the extracted samples was evaluated by electrophoresis on a 1% agarose/1× TAE gel containing 1× Sybr® Safe DNA gel stain (Life Technology, Carlsbad, CA, United States). For the genotyping analyses, 12 SSR markers were retrieved from Palumbo et al. [5] (Table 1) by selecting those ones with high Polymorphism Information Content (PIC). After an initial phase of testing to verify the presence of polymorphic alleles for each marker, primers were organised into 2 multiplexes. Amplifications were performed using the M13-tailed SSR method described by Schuelke [10] and modified as reported by Palumbo et al. [11,12] using four different fluorophores (6-FAM, VIC, NED and PET, respectively). PCRs

were performed in a final volume of 20 μ L containing 1x Platinum Multiplex PCR Master Mix (Thermo Scientific, Carlsbad, CA, United States), 5% GC Enhancer (Thermo Scientific), 0.25 μ M of each tailed primer, 0.75 μ M of each non-tailed primer, 0.5 μ M of each labelled primer (Applied Biosystem, Carlsbad, CA, United States), 30 ng of gDNA and sterile water to volume. PCR products were then analysed through capillary electrophoresis using an ABI 3730 DNA Analyser (Applied Biosystem) and the resulting chromatograms were screened to determine the fragment size at each locus using Peak Scanner software 2.0 (Applied Biosystem).

Table 1. SSR markers list reporting locus name, primer pair sequences, microsatellite motif, minimum and maximum sizes, anchor [5].

Locus name	Primer forward	Primer reverse	Motif	Min Size	Max Size	Anchor
FV_2	CAAAGAATGGAAAACATGCTG	CAAAGAATGGAAAACATGCTG	CAA	129	152	PAN1
FV_6	TATGTTCTCAGATTCGGGTTA	TATGTTCTCAGATTCGGGTTA	TC	214	226	M13
FV_253	TTGTAGAGATACAGGGTCGAA	TTGTAGAGATACAGGGTCGAA	TC	196	252	PAN1
FV_9919	AGTAAAGGCATAATCTGTTGGTGG	AGTAAAGGCATAATCTGTTGGTGG	GT	231	248	PAN3
FV_11537	TTCATGTATCAACTACGCACAC	TTCATGTATCAACTACGCACAC	AG	152	166	M13
FV_15981	CTAGCGTTTCCATCTCGTCTC	CTAGCGTTTCCATCTCGTCTC	TC	235	245	PAN1
FV_18902	GTTTGAAGCTCGAATGACCACCT	GTTTGAAGCTCGAATGACCACCT	TC	410	424	PAN2
FV_179837	ATTCACCATGACATCACCTC	ATTCACCATGACATCACCTC	TC	320	336	M13
FV_217218	ACAAACGTACCTCTGTACGAA	ACAAACGTACCTCTGTACGAA	AG	345	360	M13
FV_217225	AAAGAATGGAGAGAAGAATGG	AAAGAATGGAGAGAAGAATGG	AG	309	344	PAN1
FV_290063	TGATTTCTCAAAGGCATTCTA	TGATTTCTCAAAGGCATTCTA	GA	294	324	PAN3
FV_290202	AGGGCTGAGATTAGTTTCTAGTT	AGGGCTGAGATTAGTTTCTAGTT	TA	139	210	PAN2

* PIC value was calculated as Nei's genetic diversity index [5,13]

2.3. Genetic statistics, genetic variability estimates, and genetic structure reconstruction

Raw SSR data were analysed using POPGENE software package v. 1.32 [14] and the following statistics were calculated for each locus: number of alleles, frequency of the most abundant allele and PIC [5,13]. For each line, the number of observed (n_o) and effective (n_e) alleles [15], the number (n_{pl}) and the percentages ($\%_{pl}$) of polymorphic alleles, the Shannon's information index (I), the observed (H_o) and the expected (H_e) homozygosity, the Nei's genetic diversity (H, [16]), and the gene flow (Nm, [17]) were computed. The same analyses were repeated also considering together the CMS and M lines of each population. In this later case, the populations differentiation (D_{ST}), the total genetic diversity (H_T), the genetic diversity within each population (H_S), and the proportion of genetic diversity between populations (G_{ST}) calculated as $G_{ST} = 1 - H_S/H_T$ [18] were estimated too.

Raw SSR data were also used to calculate the genetic similarity (GS) estimates between individuals in all possible pairwise comparisons using Rohlf's simple matching (SM) coefficient

implemented in NTSYS v2.1 software [19]. Results were summarized in a GS matrix later used to calculate the average GS within and among each line.

Finally, the genetic structure of the core collection was investigated by means of a Bayesian clustering algorithm using STRUCTURE v. 2.2 software [20]. The set number of possible groups ranged from 1 to 30 and 10 replicates were conducted for each value of K based on a burn-in of 200,000 and a final run of 1,000,000 Markov chain Monte Carlo (MCMC) steps. The obtained results were analysed using STRUCTURE HARVESTER [21] web software to calculate the most likely value of K, and to determine the individuals' memberships that were then plotted as a histogram using an Excel spreadsheet.

3. Results and Discussion

3.1. SSR markers descriptive statistics and genetic variability

Descriptive statistics for all microsatellite markers are available in **Table 2**.

Table 2. Descriptive statistics of SSR markers reporting the polymorphic information content (PIC), number of alleles per locus and the highest allele frequency observed per locus

Locus name	PIC*	N° alleles	Highest allele frequency
FV_2	0.73	7	0.366
FV_6	0.65	7	0.514
FV_253	0.86	13	0.234
FV_9919	0.69	5	0.343
FV_11537	0.80	6	0.290
FV_15981	0.63	5	0.472
FV_18902	0.80	8	0.279
FV_179837	0.79	9	0.346
FV_217218	0.75	8	0.405
FV_217225	0.85	11	0.194
FV_290063	0.77	8	0.288
FV_290202	0.89	15	0.186
Mean	0.77	8.5	0.326

*PIC is calculated as Nei's diversity in agreement with Palumbo and Serrote [5,13]

Descriptive statistics on the SSR markers used, selected from Palumbo et al. [5] study for being highly polymorphic (with initial PIC > 0.5), demonstrated their informativeness in relation to the high number of alleles observed among the core collection, and the PIC values, always above 0.63. The polymorphism degree resulted fully comparable between this study and that of Palumbo et al: in both case FV_290202, FV_253 and FV_217225 resulted the loci exhibiting the highest PIC. According to Botstein et al. [22], marker loci with PIC>0.5 are considered as highly informative, those with 0.5>PIC>0.25 as reasonably informative, and those with PIC<0.25 as slightly

informative. Thus, the results observed for the SSR markers used in this study are all highly informative and suitable for comparative genotyping analyses.

The statistics calculated for each line are reported in **Table 3**.

Table 3. Descriptive statistics over all SSR loci including number of individuals, number (n_{pi}) and percentage ($\%_{pi}$) of polymorphic loci, mean number of observed (n_a) and effective (n_e) alleles per locus, Shannon's information index (I), observed (H_o) and expected (H_e) homozygosity, Nei's genetic diversity (H), Rohlfs Simple Matching genetic similarity (GS) coefficient and gene flow estimates (Nm)

Population ID	N° of individuals	n_{pi}	$\%_{pi}$	n_a	n_e	I	H_o	H_e	H	GS	Nm
CMS1	24	10	83.3%	2.33	1.39	0.38	0.72	0.77	0.23	95.6%	0.32
CMS2	24	4	33.3%	1.67	1.11	0.14	0.91	0.92	0.08	98.3%	0.33
CMS3	16	7	58.3%	1.58	1.21	0.21	0.83	0.86	0.13	97.8%	0.44
CMS4	16	5	41.7%	1.42	1.15	0.15	0.90	0.91	0.09	98.3%	0.33
CMS5-6	27	4	33.3%	1.33	1.10	0.09	0.94	0.94	0.06	98.9%	0.24
CMS7	24	5	41.7%	1.50	1.09	0.10	0.93	0.94	0.06	98.9%	0.34
CMS8	14	11	91.7%	2.25	1.47	0.46	0.64	0.72	0.28	94.9%	0.43
Average CMS	20.7	6.6	54.8%	1.73	1.22	0.22	0.84	0.87	0.13	97.5%	0.35
M1	23	7	58.3%	1.75	1.16	0.19	0.87	0.89	0.11	97.9%	0.32
M2	24	4	33.3%	1.33	1.03	0.06	0.97	0.97	0.03	99.5%	0.14
M3	13	4	33.3%	1.33	1.03	0.06	0.97	0.97	0.03	99.3%	0.28
M4	14	3	25.0%	1.42	1.19	0.15	0.92	0.91	0.09	97.9%	0.21
M5-6	28	5	41.7%	1.42	1.10	0.09	0.95	0.95	0.05	98.9%	0.23
M7	19	5	41.7%	1.75	1.09	0.14	0.93	0.93	0.07	98.4%	0.32
M8	13	5	41.7%	1.50	1.21	0.20	0.86	0.87	0.13	97.4%	0.30
Average M	19.1	4.7	39.3%	1.50	1.12	0.13	0.92	0.93	0.07	98.5%	0.26
P1	22	8	66.7%	1.83	1.18	0.18	0.92	0.88	0.11	98.4%	0.34
P2	24	5	41.7%	1.50	1.11	0.11	0.98	0.92	0.07	99.4%	0.05
P3	8	9	75.0%	2.00	1.23	0.29	0.81	0.84	0.15	96.3%	0.39
P4	9	8	66.7%	1.90	1.49	0.38	0.69	0.75	0.24	95.3%	0.42
P5	12	3	25.0%	1.50	1.27	0.20	0.92	0.87	0.13	96.8%	0.10
P6	12	2	16.7%	1.17	1.05	0.06	0.96	0.96	0.03	99.3%	0.39
P7	4	11	91.7%	2.08	1.98	0.68	0.10	0.46	0.47	97.5%	4.78
P8	12	6	50.0%	1.50	1.23	0.20	0.86	0.86	0.13	97.3%	0.18
Average P	11.6	6.3	52.4%	1.67	1.34	0.28	0.76	0.81	0.18	97.4%	0.90
H1	12	10	83.3%	2.75	2.16	0.78	0.22	0.50	0.48	92.7%	0.98
H2	12	11	91.7%	2.25	1.91	0.65	0.21	0.55	0.43	97.0%	2.35
H3	8	11	91.7%	2.42	1.94	0.68	0.24	0.544	0.44	95.3%	1.70
H4	7	11	91.7%	2.75	2.31	0.84	0.16	0.44	0.52	91.8%	1.08
H5	7	10	83.3%	2.17	2.01	0.67	0.17	0.51	0.45	96.9%	1.56
H6	7	10	83.3%	2.00	1.93	0.62	0.17	0.53	0.43	98.3%	3.51
H7	8	12	100.0%	2.17	1.99	0.71	0.06	0.48	0.49	98.2%	5.00
H8	8	9	75.0%	1.92	1.79	0.55	0.32	0.59	0.38	96.9%	0.73
Average H	8.1	10.6	88.1%	2.24	1.98	0.67	0.19	0.52	0.45	96.3%	2.28
Overall Mean		7.2	59.7%	1.82	1.43	0.33					
Among overall	451						0.78	0.23	0.77	78.9%	0.09

Considering each line separately, the number of observed alleles ranged from 2 to 12, on average 7.2. The highest numbers of polymorphic loci were observed in the F1 hybrids (H) group (on average 10.6 polymorphic loci), although few CMS lines (i.e. CMS1 and CMS8) exhibited a considerable amount of polymorphic loci too. The average number of observed alleles (n_a) per marker ranged from 1.17 to 2.75. On the other hand, the number of effective alleles ranged from 1.03 to 2.31, and it was observed to be lower than 1.5 in the parental lines and higher than 1.7 in hybrid lines, with the only exception for P7 (1.98). In agreement with the previous results.

Shannon's information index (I) was observed to be higher in H lines (>0.5) than in parental lines, and the observed homozygosity H_o was above 60% in parental lines and lower than 35% in hybrid lines, once again with the only exception for P7 line ($I = 0.68$; $H_o = 10.4\%$) (**Table 3** and **Figure 1**).

Mean internal Homozygosity

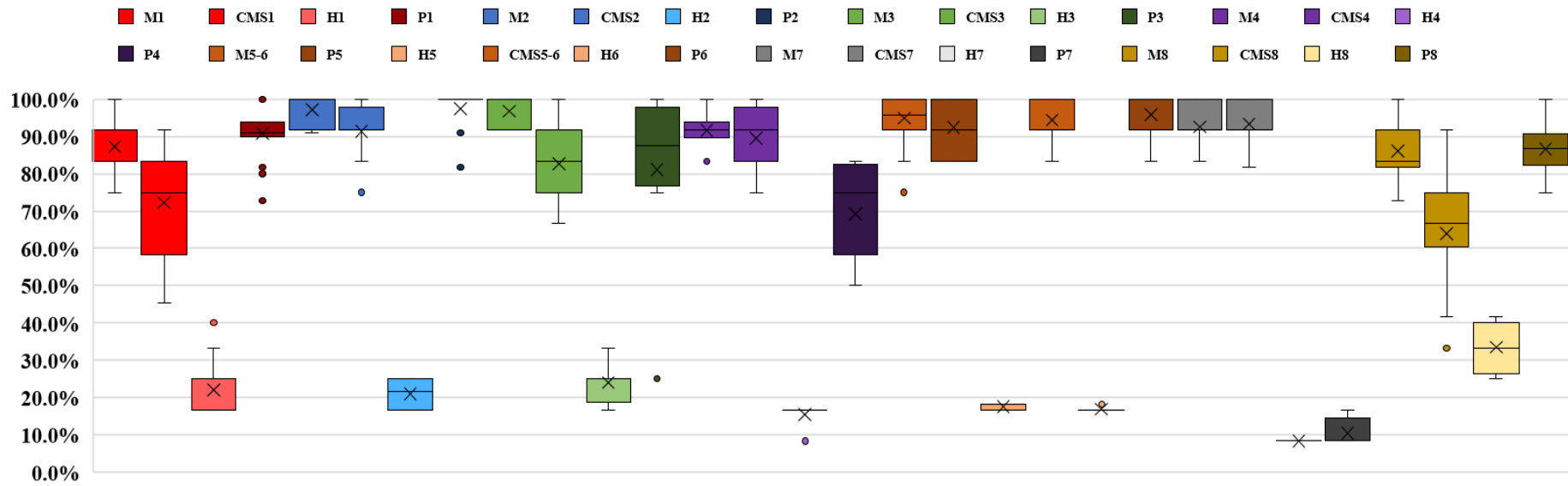


Figure 1. Box plot of the median observed homozygosity within each fennel breeding line (in percentage). Each population is labelled with different colours, CMS and M lines of each population have the same colour tone, H is lighter, and P is darker. The second and third quartiles are marked inside the square and are divided by a bar (median). The cross (×) within each box represents the mean value. Dots show outlier samples.

Nei's genetic diversity estimates were on average 0.28 in parental lines and above 0.38 in H lines (P7 was an exception, $H = 0.47$). The average Rohlfs genetic similarity (GS) (**Table 3** and **Table 4**) calculated within each line was always higher than 90% with an overall mean value equal to 97.3%. Gene flow estimates were lower than 0.5 in parental lines (only P7 was higher, $N_m = 4.78$), and higher than 0.70 in H lines, with a mean value of 2.11 and a range between 0.73 and 5.00. As it can be observed from these results, parental lines present the lower values in terms of number of effective alleles (n_e), Shannon's index (I), genetic diversity (H), and gene flow (N_m), while the same parameters in hybrid lines were consistently higher. Overall, parental lines demonstrated the expected results in terms of uniformity (due to the high levels of genetic similarity calculated within lines) and homozygosity (being derived from multiple cycles of sibling and selection). The genotyping analyses allowed the identification of an undesired event related to the P7 line. This pollinator line proved high values of genetic similarity, gene flow, effective alleles, and Shannon's index, and a low degree of homozygosity. This scenario, perfectly comparable with that of hybrid lines, would suggest a possible origin of this line from a recent crossing (e.g. $P_x \times P_y$ or $M \times P$) between highly dissimilar and homozygous lines.

On the contrary, hybrid lines resulted all characterized by high uniformity and low homozygosity, as a consequence of crosses occurring between parental lines highly homozygous for different alleles.

3.2. Parental lines uniformity and distinctiveness

In the constitution of F1 hybrids, the only possibility to sexually propagate the CMS seed plant is through the exploitation of an isogenic fertile ideotype known as maintainer. In order to keep the seed plants (and, therefore, the resulting hybrids) uniform and stable through several generations, it is of crucial importance that 1) the maintainer line is, in turn, genetically uniform and 2) the CMS seed plants are produced exclusively by crossing CMS seed plants \times isogenic fertile maintainers. For this reason, the same statistics calculated singularly for each line were also computed for 7 groups, each including the maintainer and the related CMS line of each population. These statistics are reported in **Table 5**.

Table 5. Descriptive statistics over all SSR loci for CMS-M groups. The number of individuals, number (n_{pl}) and percentage ($\%_{pl}$) of polymorphic loci, mean number of observed (n_a) and effective (n_e) alleles per locus, Shannon's information index (I), total (H_T) and within (H_S) genetic variability, genetic differentiation (D_{ST}) and proportional genetic diversity (G_{ST}), observed (H_o) and expected (H_e) homozygosity, and Rohlf's Simple Matching genetic similarity (GS) coefficient and gene flow estimates (N_m) are reported. Mean values within population (Mean CMS-M) and the mean values calculated by considering all the possible pair wise comparison among the CMS-M groups were estimated too.

Population ID	N° individuals	n_{pl}	$\%_{pl}$	n_a	n_e	I	H_T	H_S	D_{ST}	G_{ST}	H_o	H_e	GS	N_m
CMS1_M1	47	10	83.3%	2.58	1.28	0.33		0.19	0.52	0.73	0.80	0.81	0.96	2.38
CMS2_M2	48	6	50.0%	1.83	1.07	0.12		0.06	0.65	0.92	0.94	0.94	0.99	3.53
CMS3_M3	29	10	83.3%	1.92	1.13	0.19		0.10	0.61	0.86	0.89	0.90	0.98	1.78
CMS4_M4	30	6	50.0%	1.67	1.17	0.16		0.10	0.62	0.87	0.91	0.90	0.98	8.34
CMS5-6_M5-6	55	7	58.3%	1.58	1.10	0.09		0.06	0.66	0.92	0.95	0.95	0.99	69.19
CMS7_M7	43	7	58.3%	1.92	1.09	0.13		0.07	0.65	0.91	0.93	0.94	0.98	4.39
CMS8_M8	27	11	91.7%	2.42	1.39	0.39		0.23	0.48	0.67	0.74	0.77	0.95	1.98
Mean CMS-M	279	8.1	67.9%	1.99	1.18	1.40	0.71							
st. dev.		2.1	17.6%	2.68	1.15	0.28								
Among CMS-M									0.60	0.84	0.89	0.29	0.79	0.05
st. dev.									0.07	0.10	0.07	0.07	0.02	

Within male-sterile/maintainer groups (CMS-M) the number of polymorphic loci (n_{pl}) ranged from 6 to 11 and the number of observed alleles (n_a) between 1.58 and 2.58. Noteworthy, the number of effective alleles (n_e), ranging from 1.07 to 1.39, was always consistently lower than the number of observed alleles, proving a high uniformity (i.e. high genetic similarity) within each CMS-M cluster. The high uniformity of most of the CMS-M groups along with the robust differentiation from each other is also evident by comparing the within and among GS estimates (on average 98% and 79%, respectively) and by relating the gene flow (N_m) estimates calculated within and among the CMS-M groups (always higher than 1 in the first case and on average 0.05 in the second). In fact, in both cases the low GS values among the CMS-M groups and the total absence of

gene flow among them demonstrates the clear differentiation of each CMS-M and the lack of crossings between maternal lines of different populations [23].

Another aspect to consider when constituting and maintaining a CMS-M group is the degree of homozygosity. The aim is in fact to develop a highly homozygous seed plant be crossed with a dissimilar and highly homozygous pollinator (P) line. The estimate of the observed homozygosity (H_o) for each CMS-M group provided promising results, with values between 0.74 and 0.95. Specifically, all the CMS-M groups proved H_o values higher than 90%, except for CMS1-M1, CMS8-M8 and CMS3-M3, where further cycles of sibling are suggested to increase the uniformity of the resulting hybrids.

While the high levels of uniformity and homozygosity of each P line are discussed in Section 3.1, (**Table 3**), a certain degree of distinctiveness among the pollinator lines was observed, ranging from 69.3% (P4 vs. P7), to 95.8% (P5 vs. P6) (**Table 4**). It could be hypothesised a genetic relationship among the most similar P lines used in these breeding populations (e.g., P5 vs P6), putatively derived from common ancestors.

3.3. Dissimilarity estimates among parental lines and F1 hybrids evaluation

GS was calculated also among parental lines (P and CMS). GS between CMS and P lines is important to be as low as possible, in order to maximize the heterotic effect and to obtain highly heterozygous F1 hybrids. From our findings, it ranged from 67.2% (CMS7 vs. P7) to 80.1% (CMS3 vs P3, **Table 4**). By plotting and organizing the populations data based on the increment of hybrids' heterozygosity (**Figure 2**), it was possible to appreciate how those hybrids (e.g., H3) exhibiting the lower heterozygosity values (yellow area) were the results of highly similar parental lines (e.g., CMS3 and P3, green dashed lines). Vice versa, highly dissimilar parental lines produced F1 hybrids characterized by high heterozygosity. In terms of uniformity **Figure 2** clearly demonstrates how the genetic similarity of the parental lines (blue and orange bars) deeply affects the uniformity of the resulting offspring (grey bar). As a matter of fact, parent lines with low uniformity values gave rise to low uniform hybrids populations. This is especially true if parents were characterized by suboptimal homozygosity values (POP1, POP3, and POP8). In POP7, where CMS and P lines were highly uniform and highly dissimilar (67.2%), H line resulted the most heterozygous hybrid population and one of the most uniform, despite the homozygosity of its pollinator was among the lowest (H_o of P7 = 10.4%). A possible explanation for this could be the small number of hybrids and pollinators analysed for POP7 and thus a lack of representativeness. Nevertheless, the results obtained in the other analysed populations show a trend for the latter to originate uniform and

heterozygous hybrids in relation to the uniformity of the parental lines, their genetic dissimilarity, and homozygosity.

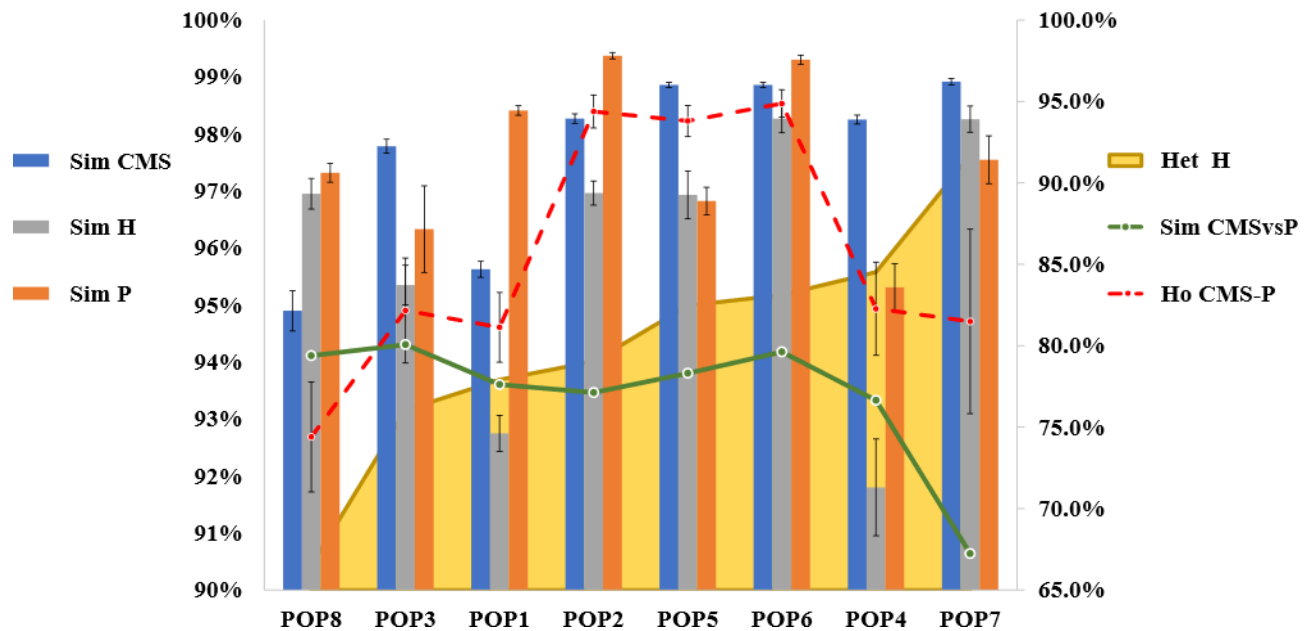


Figure 2. Graphical representation of the genetic similarity (Sim) calculated within CMS (blue bar), H (grey bar) and P (orange bar) line of each population (*vertical scale on the left*), and H line heterozygosity (yellow area), genetic similarity among CMS and P lines (green dashed line) and their average homozygosity (red dashed line) (*vertical scale on the right*)

3.4. Genetic structure of the core collection and genetic distinctiveness

Following the hypothesis of putative relationships between some breeding lines belonging to different populations, the genetic structure of the fennel core collection was investigated. Using STRUCTURE software [20], 11 clusters were identified that grouped samples in agreement with the breeding line they belonged to ($\Delta K = 39.72$) (**Figure 3** and **Figure 4**). Specifically, each of the 7 CMS-M groups was represented by a specific clusters, showing an average membership percentages of 97.6%. According to the G_S , D_{ST} and G_{ST} estimates, these findings confirmed the distinctiveness of each CMS-M group.

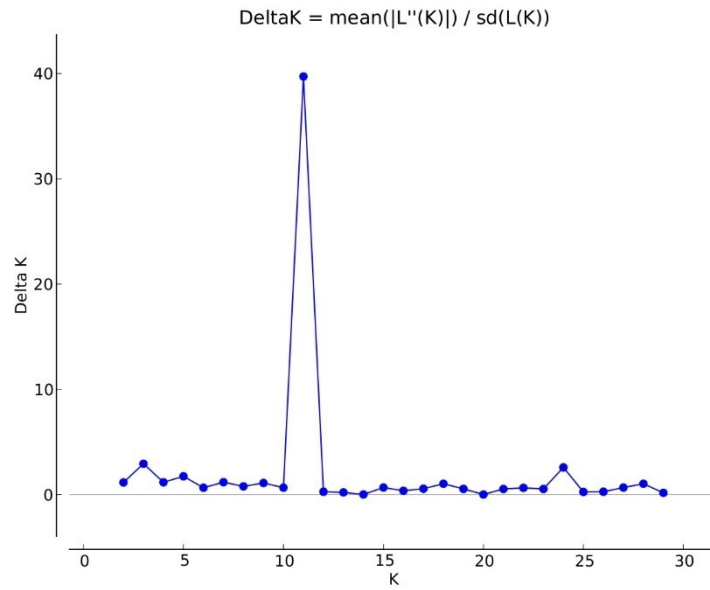


Figure 3. ΔK estimation from STRUCTURE Harvester web software elaboration of the STRUCTURE software results. Most probable value of K was equal to 11

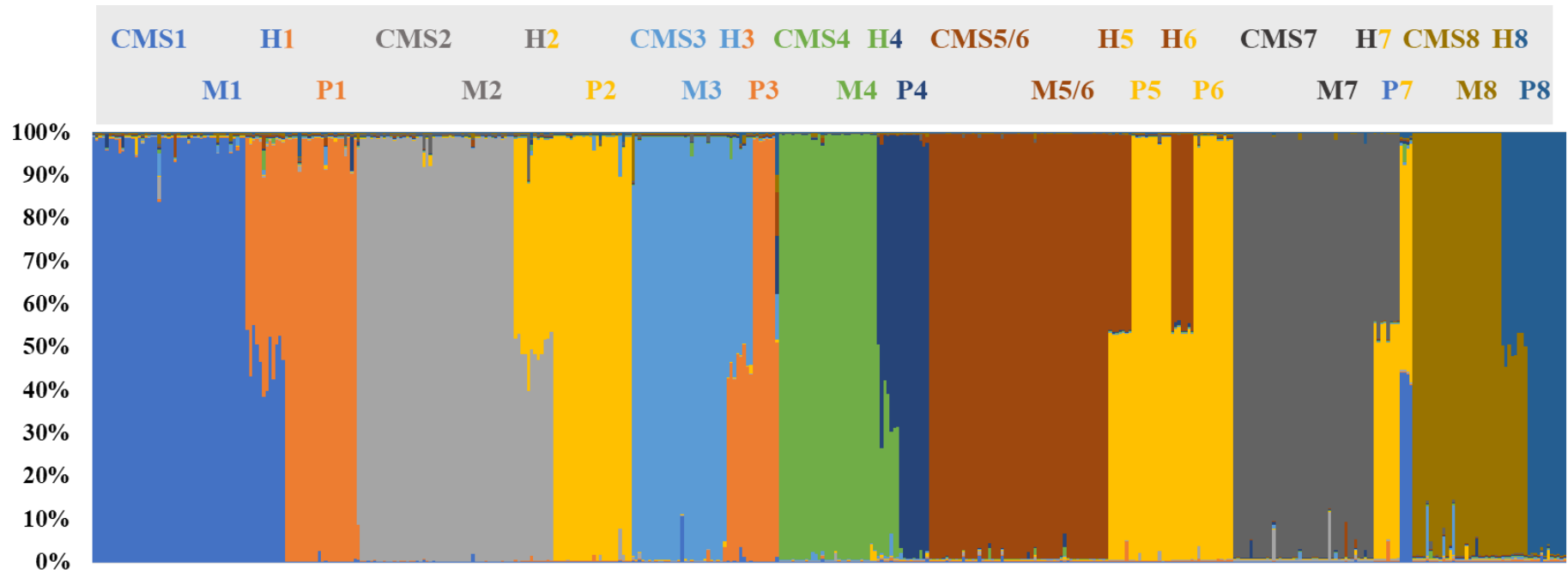


Figure 4. Histogram representing the membership of each sample to one of the 11 identified clusters. Names of each line is reported above bars and is labelled with the same colour of the respective cluster

Pollinator lines were all represented by 4 ancestors (different from the 7 observed for the CMS-M groups) where P1 and P3 (GS = 90.2%) were ascribed to the same cluster (coloured in orange in **Figure 4**), P2 was grouped with P5 and P6 (average GS = 81.3%) (coloured in yellow in **Figure 4**), while P4 and P8 constituted instead two separate clusters. Also in this case, all samples scored memberships values to their respective cluster, always higher than 95%, with few exceptions. A separate case is instead represented by P7 that resulted (~40%/60%) admixed between the P2-P5-P6 and the CMS1-M1 cluster. This finding, in addition to the considerations made around the suboptimal homozygosity and the abnormal gene flow values observed for P7, would corroborate the hypothesis that this line is the result of a recent crossing between a maintainer (M1) and a pollinator line (P2, P3 or P6).

Finally, as regards the memberships of the F1 hybrid lines, these always showed memberships ~50%/50% to the respective maternal and paternal clusters, thus demonstrating the reliability of the clustering method. Once again, results obtained from the genetic structure reconstruction agreed with the genetic variability results described in **Table 3**.

4. Conclusions

The genotyping analysis of the fennels breeding core collection described in this work, along with its genetic statistics estimates and its genetic structure reconstruction, enabled not only to determine the reliability of the method proposed based on microsatellite molecular markers, but it also provided a suitable method for plant variety traceability and post-breeding controls. Indeed, the obtained results were able to discriminate or cluster samples depending on the breeding line they belonged to and to identify unsuitable parental genotypes to be used in breeding plans (e.g.: P7), although the obtained progenies were observed to be uniform and highly heterozygous. Certainly, the obtained results highlight the impossibility of the SSR markers panel used in this study to univocally discriminate pollinator lines, thus suggesting the necessity of increasing the number of SSR markers for future analysis. Also, a certain correlation between the uniformity and heterozygosity of the hybrids with the respective parentals' uniformity, within genetic similarity and among dissimilarity was observed, demonstrating the suitability of molecular markers in helping breeders to partially predict the genetic background of F1 hybrids when planning crosses.

In conclusion, the genotyping method described in this study can be used for different purposes related to fennel's breeding and genetic traceability. Further implementations will be carried out in the future to investigate other putative SSR loci to be used for MAB or plant variety protection related analyses in this species, although these previous results demonstrated the

suitability of the molecular marker loci used in determining the distinctiveness and stability of the analysed breeding populations.

5. References

1. Badgajar, S.B.; Patel, V.V.; Bandivdekar, A.H. *Foeniculum vulgare* Mill: a review of its botany, phytochemistry, pharmacology, contemporary application, and toxicology. *Biomed Res Int* **2014**, *2014*, 842674, doi:10.1155/2014/842674.
2. Tognolini, M.; Ballabeni, V.; Bertoni, S.; Bruni, R.; Impicciatore, M.; Barocelli, E. Protective effect of *Foeniculum vulgare* essential oil and anethole in an experimental model of thrombosis. *Pharmacol Res* **2007**, *56*, 254-260, doi:10.1016/j.phrs.2007.07.002.
3. Senatore, F.; Oliviero, F.; Scandolera, E.; Tagliatalata-Scafati, O.; Roscigno, G.; Zaccardelli, M.; De Falco, E. Chemical composition, antimicrobial and antioxidant activities of anethole-rich oil from leaves of selected varieties of fennel [*Foeniculum vulgare* Mill. ssp. *vulgare* var. *azoricum* (Mill.) Thell]. *Fitoterapia* **2013**, *90*, 214-219, doi:10.1016/j.fitote.2013.07.021.
4. Diaz-Maroto, M.C.; Perez-Coello, M.S.; Esteban, J.; Sanz, J. Comparison of the volatile composition of wild fennel samples (*Foeniculum vulgare* Mill.) from central Spain. *Journal of Agricultural and Food Chemistry* **2006**, *54*, 6814-6818, doi:10.1021/jf0609532.
5. Palumbo, F.; Galla, G.; Vitulo, N.; Barcaccia, G. First draft genome sequencing of fennel (*Foeniculum vulgare* Mill.): identification of simple sequence repeats and their application in marker-assisted breeding. *Mol Breeding* **2018**, *38*, 1-17, doi:10.1007/s11032-018-0884-0.
6. FAOstat. Available online: <http://www.fao.org/home/en> (accessed on 2021)
7. Palumbo, F.; Vannozi, A.; Barcaccia, G.J.I.J.o.M.S. Impact of Genomic and Transcriptomic Resources on Apiaceae Crop Breeding Strategies. **2021**, *22*, 9713.
8. Pank, F. Three approaches to the development of high performance cultivars considering the differing biological background of the starting material. In Proceedings of International Conference on Medicinal and Aromatic Plants. Possibilities and Limitations of Medicinal and Aromatic Plant 576; pp. 129-137.
9. Palumbo, F.; Vitulo, N.; Vannozi, A.; Magon, G.; Barcaccia, G. The Mitochondrial Genome Assembly of Fennel (*Foeniculum vulgare*) Reveals Two Different atp6 Gene Sequences in Cytoplasmic Male Sterile Accessions. *Int J Mol Sci* **2020**, *21*, 4664, doi:10.3390/ijms21134664.
10. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* **2000**, *18*, 233-234, doi:10.1038/72708.
11. Palumbo, F.; Galla, G.; Barcaccia, G. Developing a Molecular Identification Assay of Old Landraces for the Genetic Authentication of Typical Agro-Food Products: The Case Study of the Barley 'Agordino'. *Food Technol Biotechnol* **2017**, *55*, 29-39, doi:10.17113/ftb.55.01.17.4858.
12. Palumbo, F.; Galla, G.; Martinez-Bello, L.; Barcaccia, G. Venetian Local Corn (*Zea mays* L.) Germplasm: Disclosing the Genetic Anatomy of Old Landraces Suited for Typical Cornmeal Mush Production. *Diversity-Basel* **2017**, *9*, doi:10.3390/d9030032.
13. Serrote, C.M.L.; Reiniger, L.R.S.; Silva, K.B.; Rabaiolli, S.; Stefanel, C.M. Determining the Polymorphism Information Content of a molecular marker. *Gene* **2020**, *726*, 144175, doi:10.1016/j.gene.2019.144175.

14. Yeh, F.C.; Yang, R.; Boyle, T.B.; Ye, Z.; Mao, J.X. POPGENE, the user-friendly shareware for population genetic analysis. *Molecular biology biotechnology centre, University of Alberta, Canada* **1997**, *10*, 295-301.
15. Kimura, M. *Population genetics, molecular evolution, and the neutral theory: selected papers*; University of Chicago Press: 1994.
16. Nei, M. *Molecular Evolutionary Genetics* **1987**.
17. McDermott, J.M.; McDonald, B.A. Gene Flow in Plant Pathosystems. *Annu Rev Phytopathol* **1993**, *31*, 353-373, doi:DOI 10.1146/annurev.py.31.090193.002033.
18. Barcaccia, G.; Lucchin, M.; Parrini, P. Characterization of a flint maize (*Zea mays* var. *indurata*) Italian landrace, II. Genetic diversity and relatedness assessed by SSR and Inter-SSR molecular markers. *Genet Resour Crop Ev* **2003**, *50*, 253-271, doi:Doi 10.1023/A:1023539901316.
19. Rohlf, F.J. NTSYS-pc, Version 2.10z. **2000**.
20. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945-959.
21. Earl, D.A.; Vonholdt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **2012**, *4*, 359-361, doi:10.1007/s12686-011-9548-7.
22. Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **1980**, *32*, 314-331.
23. Mitton, J. Gene flow. **2013**.

Chapter IV

Genotyping Analysis by RAD-Seq Reads is Useful to Assess the Genetic Identity and Relationships of Breeding Lines in Lavender Species Aimed at Managing Plant Variety Protection

(This chapter has been published with minor revisions in to *MDPI Genes*)

Scariolo, Francesco, Fabio Palumbo, Alessandro Vannozzi, Gio B. Sacilotto, Marco Gazzola, and Gianni Barcaccia. 2021. "Genotyping Analysis by RAD-Seq Reads Is Useful to Assess the Genetic Identity and Relationships of Breeding Lines in Lavender Species Aimed at Managing Plant Variety Protection" *Genes* 12, no. 11: 1656. <https://doi.org/10.3390/genes12111656>

0. Abstract

Lavender species are widely distributed in their wild forms around the Mediterranean Basin, and they are also cultivated worldwide as improved and registered clonal varieties. The economic interest of the species belonging to the *Lavandula* genus is determined by their use as ornamental plants and important source of essential oils that are destined to the production of cosmetics, pharmaceuticals, and foodstuffs. Because of the increasing number of cases of illegal commercialization of selected varieties, the protection of plant breeders' rights has become of main relevance for the recognition of breeding companies' royalties. With this aim, genomic tools based on molecular markers have demonstrated to be very reliable and transferable among laboratories, and also much more informative than morphological descriptors. With the rising of the next-generation sequencing (NGS) technologies, several genotyping-by-sequencing approaches are now available. This study deals with a deep characterization of 15 varietal clones, belonging to two distinct *Lavandula* species, by means of restriction-site associated DNA sequencing (RAD-Seq). We demonstrated that this technology screens single nucleotide variants that enable to assess the genetic identity of individual accessions, to reconstruct genetic relationships among related breeding lines, to group them into genetically distinguishable main subclusters, and to assign their molecular lineages to distinct ancestors. Moreover, a number of polymorphic sites were identified within genes putatively involved in biosynthetic pathways related to both tissue pigmentation and terpene production, useful for breeding and/or protecting newly registered varieties. Overall results highlighted the presence of pure ancestries and interspecific hybrids for the analysed *Lavandula* species, and demonstrated that RAD-Seq analysis is very informative and highly reliable for characterizing *Lavandula* clones and managing plant variety protection.

Keywords: *Lavandula*; NGS; genotyping by RAD sequencing; flavonoids; terpenes; chloroplast DNA barcoding; ancestry reconstruction; interspecific crosses; plant breeder's rights.

1. Introduction

Lavender species *Lavandula stoechas* L. and *Lavandula pedunculata* (Mill.) Cav., belonging to the Lamiaceae family, include diploid plants (both $2n=2x=30$ [1]). The wild forms of these species are widely distributed on the coast of countries around the Mediterranean Sea and are also cultivated worldwide using registered clonal varieties. The reproductive strategies of *L. stoechas* and *L. pedunculata* are prevalently allogamous and characterized by entomophilous pollination, although self-compatibility and autogamous events have also been reported [2]. Similar to many others belonging to the *Lavandula* genus, these species are known for their ornamental use and for the production of essential oils (EOs) rich in linalyl acetate, the fragrance of which is greatly appreciated for several purposes (*i.e.*, cosmetics, lotions, soaps, room fragrances and food aromas) [3]. Moreover, lavender EOs are used in pharmacology, aromatherapy, and natural medicine given their anti-inflammatory properties [4-6].

Given the growing economic interest around these species, the necessity for plant breeders and breeding companies to adequately register their varieties and to protect them from plagiarism is becoming increasingly important. As previously demonstrated in other crops, given the limits of phenotypic characterization and morphological markers, the use of molecular markers is becoming undeniably crucial [7-10]. The use of dominant markers has been reported in several studies to be helpful in assessing the genetic distinctiveness and uniformity of species belonging to the genus *Lavandula* [11-14]. However, the low reproducibility and the difficulty in associating these markers with phenotypic traits make them unsuitable for varietal registration processes. Codominant markers are instead able to overcome these limitations, and among them, SSR and SNP markers are the most commonly used markers. For example, previous studies successfully identified SSRs [15,16] strictly associated with genomic regions involved in the synthesis of EOs [17] or single nucleotide polymorphisms (SNPs) located within genes involved in the biosynthetic pathways of the main terpenes characterizing essential oils [17]. The analysis of genotypes linked to chemotypes [18,19] would allow researchers to identify the most suitable molecular markers to be used in screening analysis for breeding selection and variety registration. The use of molecular markers is also of relevant interest for marker-assisted selection (MAS) purposes: the association among molecular markers and genomic loci involved in the biosynthesis of flavonoids and other coloring compounds would allow the correlation of specific phenotypes and genotypes.

Although different molecular approaches have been used to assess the distinctiveness of varieties of the *Lavandula* species, this genus suffers from the lack of annotated genome assemblies in international databases. However, according to Jingrui Li et al. [20], there is one genome

assembly for *Lavandula angustifolia* that is not publicly available that would simplify the identification of mapped molecular markers suitable for the above-described purposes.

The present study is focused on the application of the Restriction site-Associated DNA (RAD) marker sequencing technology not only to assess the extent of genetic similarity and heterozygosity/homozygosity of a core collection of 15 accessions belonging to two species of the *Lavandula* genus, but also to identify genomic loci suitable for marker-assisted breeding (MAB) and for registration/protection of newly-bred varieties. These aspects are of major interest for breeding companies and plant breeders when developing new commercial clones destined to the market.

2. Materials and Methods

2.1. Plant Materials

Fifteen samples belonging to as many breeding lines of lavender were kindly granted by Gruppo Padana S.S. (Paese, TV, Italy). Specifically, 13 *L. stoechas* and 2 *L. pedunculata* (identified as 2603 and 2605) plants were analyzed. Genomic DNA (gDNA) was isolated from 200 mg of fresh leaf tissue using the DNeasy Plant mini kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol with a minor modification. Specifically, lysis and protein precipitation buffers were increased by 50% to facilitate the identification and, thus, the isolation of the supernatant phase containing oils, which was shown to deeply affect the quality of the gDNA in previous tests of DNA extraction. Both the quality and quantity of the genomic DNA samples were evaluated using a NanoDrop 2000c UV-Vis spectrophotometer (Thermo Fisher Scientific Inc., Pittsburgh, PA, USA) and by agarose gel electrophoresis (1% agarose/1× TAE gel containing 1× SybrSafe DNA stain (Life Technologies, Carlsbad, CA, USA)).

2.2. Restriction-site Associated DNA sequencing (RAD-Seq) and data analysis

The 15 gDNA samples were analyzed by means of restriction-site associated DNA sequencing (RAD-Seq) technology. One microgram of gDNA per individual sample was digested using the restriction enzyme MseI following the procedure described by Stevanato et al. [21]. For library preparation, digested DNA samples were diluted at a concentration of 3 ng/μL. Indexing, library preparation, sequencing, and bioinformatic analyses were performed according to the protocol described by Stevanato et al. [21]. Raw reads obtained through an Ion S5 sequencer (Thermo Fisher Scientific Inc., Waltham, MA, USA) were trimmed according to the restriction enzyme recognition motif. After quality assessment, all the artifacts and the Ns-containing reads

were removed. Variants were called using Stacks v2.41 software [22]. SNPs were filtered to remove those meeting the following criteria: (1) SNPs with greater than 10% missing data, (2) SNPs with a sequence depth $\times 4$, and (3) tri- and tetraallelic SNPs.

The obtained data were used for the construction of an unweighted pair group method with arithmetic mean (UPGMA) dendrogram based on Rohlf's genetic similarity simple matching coefficient and a principal coordinate analysis (PCoA) centroid using NTSYS software v2.21 [23]. Additionally, a Bayesian clustering algorithm implemented in STRUCTURE v.2.2 [24] was used to model the genetic structure of the lavender core collection. The number of founding groups ranged from 1 to 20, and 10 replicate simulations were conducted for each value of K based on a burn-in of 20,000 and a final run of 100,000 Markov chain Monte Carlo (MCMC) steps. STRUCTURE HARVESTER [25] was used to estimate the most likely value of K, and the estimates of membership were plotted as a histogram using an Excel spreadsheet.

2.3. Identification of CDS-mapping reads and reads related to terpene and anthocyanin biosynthesis pathways

Reads with no missing data in the 15 samples analyzed were used to identify those sequences most likely belonging to genomic coding sequences (CDSs). No annotated assembly is available for *Lavandula*, but Jingrui Li et al. [20] reported that an assembly was deposited in NCBI. However, a search of the accession number yields no matches, and the authors did not answer our request at the time of the submission of this article. Thus, the genomes of the two phylogenetically closest species to this genus, namely, *Sesamum indicum* (GeneBank, GCF_000512975.1) and *Salvia splendens* (GeneBank: GCA_004379255.2), were considered. While the assembly of *S. indicum* was previously annotated, all the genomic loci and the resulting proteins from *S. splendens* were "hypothetical proteins" that required an additional step of annotation prior to their usage. This step was accomplished using the KAAS platform [26], the GHOSTX aligner [27] and the KEGG database for plant organisms [28]. The RAD tags were then aligned against both the *S. indicum* and *S. splendens* CDS datasets using a local BLASTn (BLAST+ 2.11.0 package) with an E-value threshold $\leq 1.0 \times 10^{-10}$ and a percentage of identity $\geq 80\%$. The newly identified CDS-mapping reads were used for the construction of a UPGMA dendrogram and PCoA centroids as described in the previous section.

For reads matching genes involved in the biosynthetic pathways of terpenes and flavonoids, multiple Geneious alignments (Geneious software v2021.1.1, Biomatters Ltd., Auckland, New Zealand) among the 15 samples were performed to identify nonsynonymous SNPs.

2.4. DNA barcoding through Sanger sequencing for species determination

To highlight interspecific cross events between *L. stoechas* and *L. pedunculata*, DNA barcoding sequencing of all samples was accomplished using three chloroplast regions, namely, the psbA-trnH intergenic space region, the maturase K (matK) and ribonuclease large subunit (rbcL) genes. A nuclear region, namely, the internal transcribed region (ITS), was also considered. Genomic DNA amplification of the four samples considered was performed using a Veriti 96-Well Thermal Cycler (Applied Biosystems, Foster City, CA, USA) in a total volume of 25 μ L of reaction mixture including 12.5 μ L of MangoMix (Bioline, London, UK) with 1 μ L of DNA (50 ng/ μ L), 2 μ L of each primer (10 mM) and sterile water to reach the final volume. The following thermal conditions were adopted: 2 min at 95 $^{\circ}$ C; 35 cycles at 95 $^{\circ}$ C for 30 s, variable annealing temperature depending on the primer pair used (**Table 1**) for 45 s, and 72 $^{\circ}$ C for 45 s; and a final extension at 72 $^{\circ}$ C for 10 min. The PCR products were confirmed using 2% agarose/1 \times TAE gels containing 1 \times SYBR Safe DNA Gel Stain (Life Technologies), purified with ExoSAP-IT PCR Product Cleanup Reagent (Thermo Fisher) and sequenced on an ABI 3730XL Genetic Analyzer (Applied Biosystems). The obtained chromatograms were then assessed using Geneious Prime software, and sequences were trimmed at the 5' and 3' positions to remove the low-quality section were primers attached, and resulting ITS chromatograms were analyzed with “Heterozygote Plugin” version 2.0.0 (Biomatters) add-on to identify heterotic positions and then manually checked. The resulting sequences were aligned based on the barcoding region and concatenated for each sample. The resulting multiple alignment was used for the construction of a neighbor-joining tree using the Juke-Cantor algorithm, and polymorphic sites were used to create a logo graph. Bioinformatics analyses were conducted using Geneious Prime software plug-ins.

Table 1. List of primers used for each chloroplast (cpDNA) and nuclear (nuDNA) marker with their nucleotide sequence, and reference source.

Marker	Primer name	Primer sequence (5'-3')	T _a (°C)	References
rbcL gene (cpDNA)	rbcL_F	GCAGCATTYCGAGTAASTCCYCA	55	[29]
	rbcL_R	GAAACGYTCTCTCCA WCGCATAAA		[29]
matK gene (cpDNA)	matK4La	CCTTCGATACTGGGTGAAAGAT	55	[30]
	matK1932Ra	CCAGACCGGCTTACTAATGGG		[30]
trnH-psbA (cpDNA)	psbA3'f	GTTATGCATGAACGTAATGCTC	55	[31]
	trnHf	CGCATGGTGGATT CACAATCC		[32]
ITS1 (nuDNA)	ITS5	GGAAGTAAAAGTCGTAACAAGG	55	[33]
	ITS2	GCTGCGTTCTTCATCGATGC		[33]

3. Results

3.1. RAD-Seq and genetic similarity analyses

A RAD-Seq analysis was performed using 15 samples obtained from an equal number of breeding lines that belong to a core collection of the *Lavandula* genus. The sequencing produced a total of 44,219,948 raw reads with an average of 2.9 million reads per sample. After quality assessment and adapter trimming, we obtained 42,610,020 reads that were used for the creation of a catalog of 622,153 consensus loci and then used for variant calling as a reference. An initial pool of 43,271 SNPs was first identified. Then, after the filtering step, in which sequences with at least one missing value in one sample were discarded, 16,228 SNPs distributed in 14,922 RAD sequence tags were retained as all of them were shared in all samples.

The analysis of the average genetic similarity (GS), which was calculated in all pairwise comparisons among the 15 sequenced samples, is reported in **Table 2**. Overall, GS ranged from 51.6% to 93.7% (1811 vs. 2603” and “BPI vs. SD-332”, respectively), whereas the average GS among the entire pool of samples was $74.8 \pm 1.0\%$. The number of discriminative polymorphic sites among the most similar genotypes was 1966 SNPs, whereas that calculated among the most dissimilar was 9566 SNPs, both considering heterozygous loci. The UPGMA dendrogram grouped the 15 samples into 5 clusters named “Cluster A” to “Cluster E” (**Figure 1**), where the latter included the two *L. pedunculata* samples. From these findings, the mean genetic similarity was calculated among and between the identified groups. The GS calculated within the clusters ranged from 73.7% in “Cluster E” to $92.0 \pm 0.8\%$ in “Cluster C”, whereas the GS among groups ranged from $56.6 \pm 1.3\%$ (“Cluster C” vs. “Cluster E”) to $83.9 \pm 0.6\%$ (“Cluster B” vs. “Cluster C”). Moreover, due to the low genetic similarity between “Cluster E” and the other four subgroups, as shown by the UPGMA dendrogram, a comparison between this cluster and the other main group of 13 samples was also made. “Cluster A+B+C+D”, which is located in one main arm of the dendrogram with a within mean genetic similarity of $79.7 \pm 0.7\%$, exhibited an observed genetic similarity equal to $60.1 \pm 1.0\%$ when compared to “Cluster E”. Considering the number of SNPs with uncommon alleles between the *L. stoechas* and the *L. pedunculata* groups, 162 SNPs were found to have one allele in the 13 samples of “Cluster A+B+C+D” and the other allele in the 2 samples of “Cluster E”. The PCoA grouped samples in different spaces of the diagram with Dimensions 1 and 2 representing 49.2% and 19.6%, respectively, and overall, 68.8% of the molecular variation in total (**Figure 2**). From the ancestry composition reconstruction analysis, a maximum ΔK value at $K = 3$ was found ($\Delta K = 260.07$, as shown in **Supplementary Figure S1**).

Thus, an equal number of putative ancestors were hypothesized with a membership of ancestry ranging from 0 to 100%, 0 to 99.8% and from 0 to 71.3%, respectively. Notably, “Ancestor 1” had no membership in samples 2605 and 2603, for which “Ancestor 2” was greater than 40%. In contrast, “Ancestor 3” had no membership in samples BPI and ST-913 and less than 5% in samples 1811, SD-332 and 2603 (see **Figure 1**).

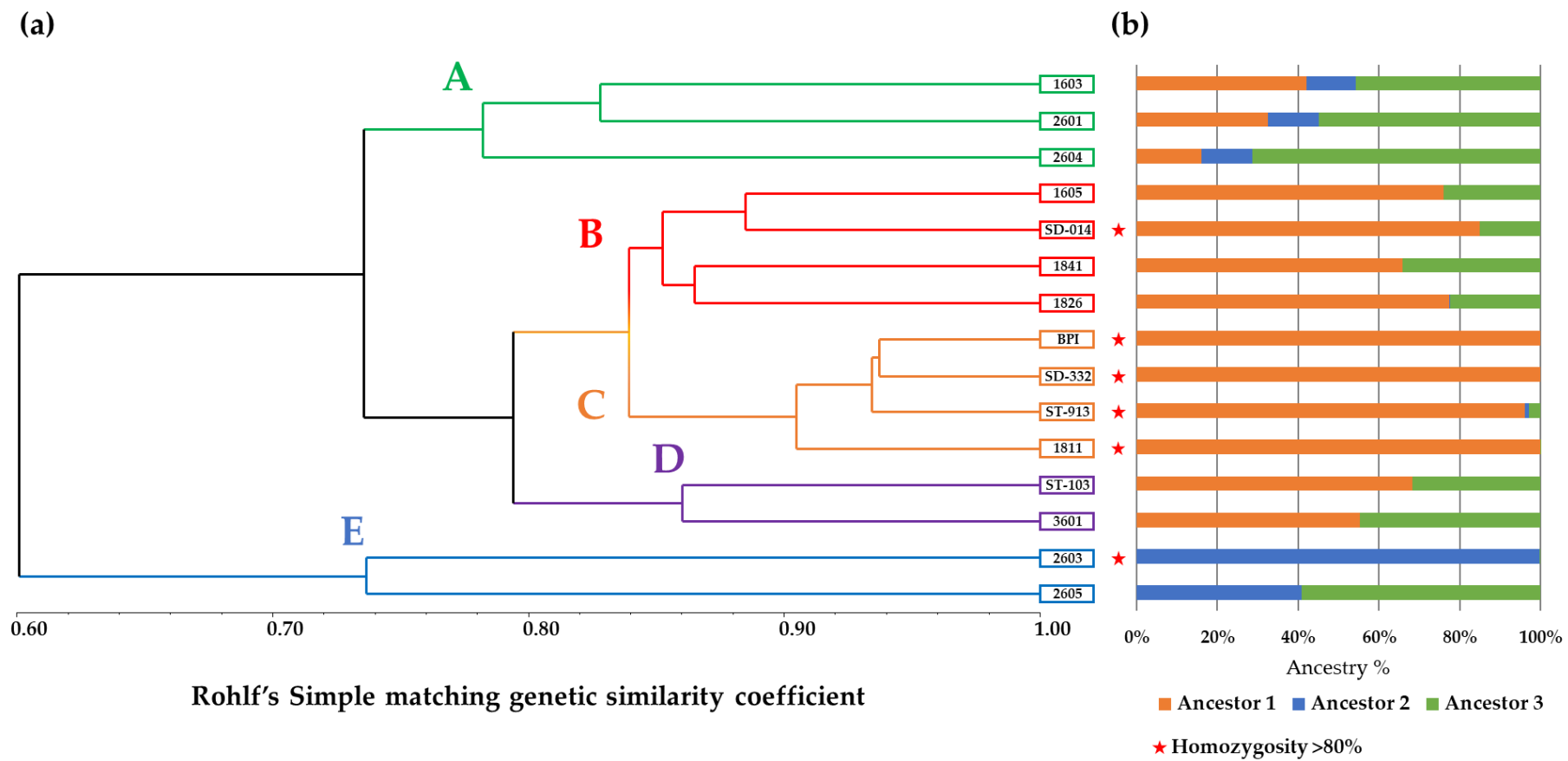


Figure 1. (a) UPGMA dendrogram based on the pair-wise genetic similarity matrix highlighting 5 main “Clusters” for the no missing values containing dataset. (b) STRUCTURE software histogram for K = 3 of 15 individuals of *Lavandula* with a no missing values containing dataset (“red star” symbol labels individuals with homozygosity >80%)

Table 2. (a) Genetic Similarity matrix of 15 *Lavandula* individuals based on 16,228 SNPs with no missing data, and relative observed homozygosity (Obs. Ho) and heterozygosity (Obs. He). (b) Average genetic similarity of Clusters identified through the construction of the UPGMA dendrogram, and average observed homozygosity (Avg. Obs. Ho)

(a)

	Obs. Ho	Obs. He	Sample	Genetic similarity (GS)																
1603	66.1%	33.9%	Cluster A	1603	100.0%															
2601	60.1%	39.9%		2601	82.8%	100.0%														
2604	72.8%	27.2%		2604	78.9%	77.6%	100.0%													
1605	76.4%	23.6%	Cluster B	1605	79.8%	77.8%	73.8%	100.0%												
1841	78.8%	21.2%		1841	77.9%	75.7%	71.1%	86.4%	100.0%											
1826	77.9%	22.1%		1826	79.2%	76.9%	74.1%	87.4%	86.5%	100.0%										
SD-014	85.5%	14.5%		SD-014	76.5%	74.3%	70.6%	88.5%	83.9%	83.3%	100.0%									
BPI	90.1%	9.9%	Cluster C	BPI	74.6%	72.3%	68.2%	82.8%	79.1%	83.7%	83.8%	100.0%								
ST-913	84.8%	15.2%		ST-913	75.0%	74.2%	70.4%	85.7%	81.1%	86.5%	83.5%	93.3%	100.0%							
SD-332	82.4%	17.6%		SD-332	75.9%	74.4%	70.5%	83.5%	79.1%	84.8%	85.4%	93.7%	93.5%	100.0%						
1811	89.7%	10.3%		1811	75.1%	72.1%	67.7%	86.0%	85.6%	85.9%	86.4%	89.7%	92.2%	89.5%	100.0%					
ST-103	77.6%	22.4%	Cluster D	ST-103	75.7%	72.9%	69.4%	80.7%	79.4%	76.9%	82.8%	83.2%	82.2%	82.3%	82.0%	100.0%				
3601	78.1%	21.9%		3601	72.2%	70.4%	67.8%	78.0%	76.5%	75.2%	80.0%	76.3%	77.2%	77.2%	80.9%	86.0%	100.0%			
2603	87.8%	12.2%	Cluster E	2603	63.0%	64.0%	65.9%	55.4%	58.8%	56.2%	54.0%	53.9%	53.9%	53.4%	51.6%	54.9%	53.1%	100.0%		
2605	71.6%	28.4%		2605	67.6%	68.9%	69.3%	62.0%	66.4%	62.9%	60.0%	58.9%	61.0%	60.3%	59.7%	64.0%	63.8%	73.7%	100.0%	
				1603	2601	2604	1605	1841	1826	SD-014	BPI	ST-913	SD-332	1811	ST-103	3601	2603	2605		
				Cluster A			Cluster B			Cluster C				Cluster D		Cluster E				

(b)

Avg. Obs. Ho	Cluster	Avg. Genetic Similarity (GS)							
66.4% ± 3.7%	Cluster A	79.8% ± 1.6%							
79.7% ± 2.0%	Cluster B	75.6% ± 0.9%	86.0% ± 0.8%						
86.7% ± 1.9%	Cluster C	72.5% ± 0.8%	83.9% ± 0.6%	92.0% ± 0.8%					
77.9% ± 0.2%	Cluster D	71.4% ± 1.1%	78.7% ± 0.9%	80.2% ± 1.0%	86.0% ± N/A				
79.7% ± 8.1%	Cluster E	66.4% ± 1.1%	59.4% ± 1.5%	56.6% ± 1.3%	58.9% ± 2.9%	73.7% ± N/A			
78.5% ± 2.4%	A+B+C+D					60.1% ± 1.0%	79.7% ± 0.7%		
		Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	A+B+C+D		

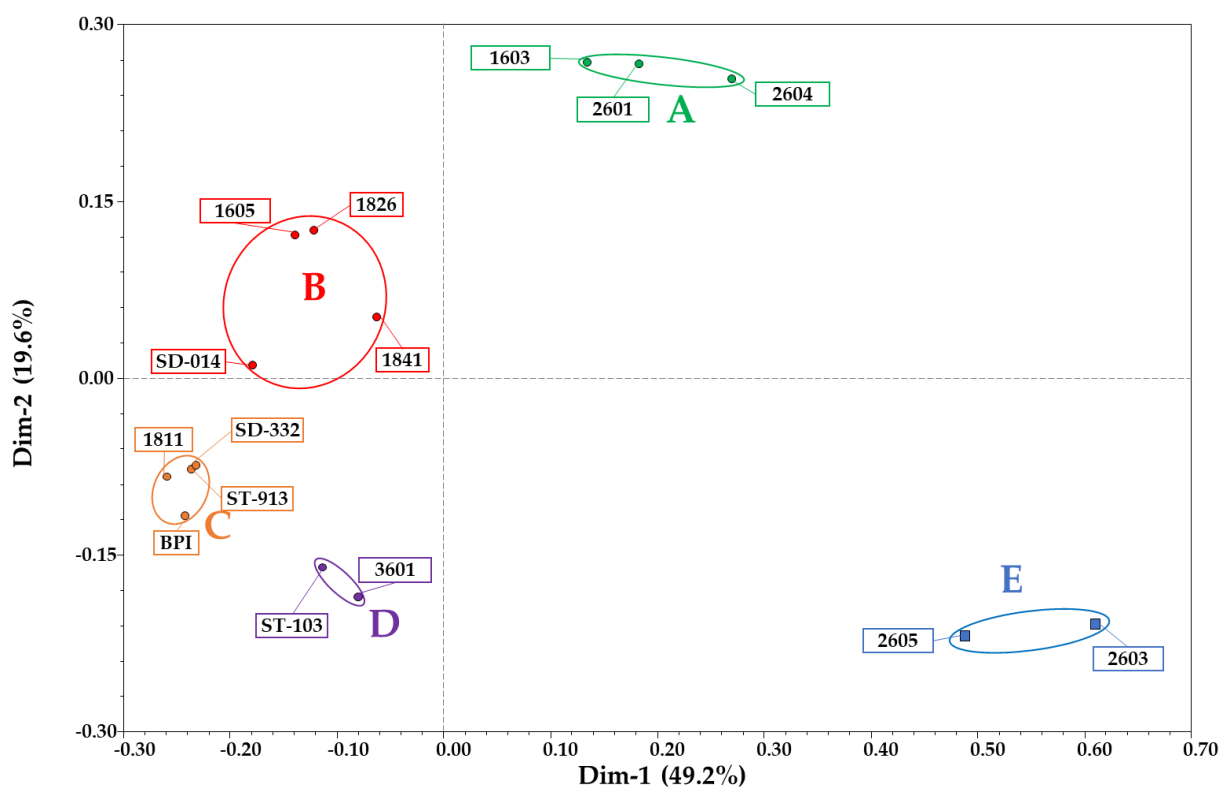


Figure 2. Principal Coordinate Analysis (PCoA), based on the eigen vectors calculated starting from the genetic similarity matrix and highlighting the 5 main “Clusters” identified for the 15 analyzed samples of *Lavandula*

Beyond the genetic similarity estimates, the observed homozygosity (Obs. Ho) of each sample was also estimated (see **Table 2**). The highest homozygosity was observed in sample “BPI” (90.1%), and the lowest (60.1%) homozygosity was observed in sample “2601”. The mean homozygosity among all samples was $78.7 \pm 2.2\%$. As for genetic similarity, homozygosity was also calculated for each of the 5 identified clusters with values ranging from $66.4 \pm 3.7\%$ to $86.7 \pm 1.9\%$ (“Cluster A” and “Cluster C”, respectively) and a mean value for group “A+B+C+D” equal to $78.5 \pm 2.4\%$.

3.2. CDS-matching reads identification

The BLASTn analysis performed aligning the 16,228 RAD tags with no missing data against the *S. indicum* exome revealed a total number of matches equal to 3,994 for 2,618 reads on 2,907 CDSs coding for 2,077 protein products, whereas that performed on the annotated *S. splendens* showed 9,107 matches for 4,239 RAD tags on 6,534 CDSs coding for 1,215 protein products. Comparing the two BLASTn analyses, 2,286 RAD tags were shared in the results from the *S. indicum* and *S. splendens* exomes. CDS-mapping reads were then used to perform a more stringent genetic similarity analysis following the procedure previously described for the entire SNP dataset (**Table 3** and **Informatic Table I1-2**).

Table 3. Summary statistics of the BLASTN analysis of the RAD-Seq reads against the exomes of *S. indicum* and *S. splendens*. Statistics information of the flavonoids and terpenes pathways involved genes is also reported.

BLASTn result	RAD-tags (n)	Accessions (n)	Protein products (n)	Avg. identity (%)	Avg. length (bp)	Avg. E-value	Avg. bitscore	Avg. score	Avg. mismatches (n)	Avg. identity (n)	Avg. positive positions
CDS <i>S.ind</i>	2,618	2,907	2,077	87.3	64.4	5.33E-12	80.2	87.5	8.2	56.2	87.3
Flavonoids	15	14	10	86.7	67.1	1.04E-12	82.1	89.6	8.9	58.2	86.7
Terpenes	20	24	19	86.0	62.9	6.20E-12	74.3	81.0	9.0	53.9	86.0
CDS <i>S.sp</i>	4,239	6,534	1,215	88.7	64.2	2.90E-12	83.8	91.5	7.3	56.9	88.7
Flavonoids	33	40	18	87.4	66.0	2.41E-12	82.5	90.1	8.3	57.6	87.4
Terpenes	61	65	28	88.9	65.6	1.45E-12	86.6	94.7	7.3	58.3	88.9

The analysis of the average genetic similarity calculated in all pairwise comparisons among the *S. indicum* and *S. splendens* exome matching reads of the 15 sequenced samples is reported in **Table 4** (see also **Supplementary Figure S2**). These estimates overall ranged from a minimum value of 56.4%/55.5% to a maximum of 94.2%/94.3% in comparisons “1811 vs. 2603” and “BPI vs. ST-913”, respectively (*S. indicum/S. splendens*), with an average genetic similarity among the entire pool of samples of $76.6 \pm 0.9\%/76.2 \pm 0.9\%$. In general, the two genetic similarity analyses performed on the datasets obtained after investigating the exome matching reads yielded highly similar results both in sample clustering and pairwise genetic similarity percentages. The only differences observed were in the UPGMA dendrogram based on the dataset containing the reads that matched the *S. splendens* exome, in which the disposition of samples “1841” and “1826” changed from those constructed using the other two datasets (see “Cluster-Bb” in the **Supplementary Figures S3-S6**). Moreover, it was observed that the genetic similarity calculated within clusters was slightly greater in the GS matrices calculated using the exome matching read datasets than in those calculated using the nonmissing data-containing dataset. The same small difference was also observed regarding the homozygosity estimations, which was generally 0.5% higher in the exome-based analyses compared with the whole 16,228 SNP dataset. The only exceptions were noticed for “Cluster D” and “Cluster E”, which were slightly lower (see **Table 4** and **Supplementary Figure S2**).

Table 4. (a) Genetic Similarity matrix of 15 *Lavandula* individuals based the BLASTN analysis against *S. indicum* exome, and relative observed homozygosity (Obs. Ho) and heterozygosity (Obs. He). (b) Average genetic similarity of Clusters identified through the construction of the UPGMA dendrogram, and average observed homozygosity (Avg. Obs. Ho) The standard error is also reported.

(a)

Obs. Ho	Obs. He	Genetic similarity (GS)																
68.3%	31.7%	Cluster A	1603	100.0%														
60.4%	39.6%		2601	83.1%	100.0%													
73.7%	26.3%		2604	79.6%	77.3%	100.0%												
78.3%	21.7%	Cluster B	1605	81.2%	78.5%	74.8%	100.0%											
86.1%	13.9%		SD-014	77.9%	75.5%	72.3%	88.9%	100.0%										
78.2%	21.8%		1841	79.7%	77.5%	73.1%	86.9%	85.1%	100.0%									
79.5%	20.5%		1826	81.0%	78.3%	75.9%	87.0%	84.5%	88.1%	100.0%								
90.4%	9.6%	Cluster C	BPI	76.4%	74.0%	70.6%	83.5%	85.8%	80.6%	84.3%	100.0%							
85.3%	14.7%		ST-913	76.9%	76.1%	72.7%	86.2%	85.5%	82.0%	87.4%	94.2%	100.0%						
83.2%	16.8%		SD-332	77.3%	76.3%	72.9%	83.9%	86.9%	80.4%	85.1%	93.9%	93.7%	100.0%					
89.7%	10.3%		1811	77.5%	74.3%	70.4%	86.8%	87.9%	86.2%	86.6%	90.4%	92.6%	90.0%	100.0%				
78.5%	21.5%	Cluster D	ST-103	77.5%	74.7%	71.3%	82.5%	84.2%	81.6%	79.0%	84.0%	83.8%	83.3%	84.1%	100.0%			
77.3%	22.7%		3601	75.0%	73.2%	70.4%	79.4%	80.9%	78.5%	76.6%	77.8%	78.9%	78.4%	82.2%	87.0%	100.0%		
87.0%	13.0%	Cluster E	2603	65.8%	67.4%	68.9%	59.0%	58.9%	63.0%	61.1%	57.9%	58.2%	58.1%	56.4%	58.7%	57.8%	100.0%	
70.2%	29.8%		2605	69.1%	70.4%	70.9%	64.1%	63.5%	69.2%	65.9%	61.5%	63.6%	63.3%	62.9%	66.5%	67.4%	74.3%	100.0%
			1603	2601	2604	1605	SD-014	1841	1826	BPI	ST-913	SD-332	1811	ST-103	3601	2603	2605	
			Cluster A			Cluster B				Cluster C				Cluster D		Cluster E		

(b)

Avg. Obs. Ho	Sample	Avg. Genetic similarity (GS)											
67.4% ± 3.9%	Cluster A	80.0% ± 1.7%											
80.5% ± 1.9%	Cluster B	77.1% ± 0.8%		86.7% ± 0.7%									
87.2% ± 1.7%	Cluster C	74.6% ± 0.7%		84.9% ± 0.6%			92.5% ± 0.8%						
77.9% ± 0.4%	Cluster D	73.7% ± 1.1%		80.3% ± 0.9%		81.6% ± 1.0%			87.0% ± N/A				
78.6% ± 8.4%	Cluster E	68.8% ± 0.8%		63.1% ± 1.2%		60.2% ± 1.0%			62.6% ± 2.5%		74.3% ± N/A		
79.1% ± 2.3%	A+B+C+D								63.4% ± 0.9%		81.0% ± 0.7%		
		Cluster A		Cluster B			Cluster C		Cluster D		Cluster E		A+B+C+D

3.3. BLASTn analysis for terpene and flavonoid pathway-related gene investigation

From the BLASTn analysis performed using the RAD tags of the 15 *Lavandula* accessions against the *S. indicum* and the *S. splendens* exomes, among the CDS-mapping reads, we selected a subgroup of sequences that aligned against genes involved in the biosynthetic pathways of terpenes and flavonoids.

In *S. indicum*, a total of 9 matches were discovered for the flavonoid biosynthetic pathway and 20 for the terpene biosynthetic pathway. From the multiple alignments of the biallelic lavender reads of the 15 samples, 6 RAD tags presented synonymous mutations, 26 were nonsynonymous and 4 coded for STOP codons that were restored in 3 cases to a coding triplet. However, in one case, it was maintained for both alleles (RAD-tag encoded 8036 matching the 1,4-dihydroxy-2-naphthoyl-CoA synthase, accession ID: XP_011071094.1). Moreover, in *S. splendens*, 33 and 61 RAD tags matched sequences related to the flavonoid and terpene biosynthetic pathways, respectively. Similar to that performed for the matches identified in sesame pathways, multiple alignments were performed only considering the lavender RAD tags. From this investigation, 16 polymorphic sites coded for synonymous mutations, 62 were nonsynonymous and 2 coded for STOP codons. One mutation was restored in some samples to an arginine coding triplet, whereas the other maintained the missense triplet in the less frequent SNP. From the two analyses performed on the sesame and scarlet sage exomes, 7 and 17 matches were common for the flavonoid and terpene pathways, respectively. Summary statistics of the BLASTn analyses for the results of the biosynthetic pathway are reported in **Table 3**, BLASTN resulting matches against *S. indicum* for the biosynthetic pathways and amino acids substitutions after multiple alignments are reported in **Table 5**, BLASTN resulting matches against *S. splendens* for the biosynthetic pathways and amino acids substitutions after multiple alignments are reported in **Supplementary Tables S1**, and complete BLASTN results are available in **Informatic Tables I3-4**.

Table 5. Multiple alignments results reporting read ID, *S. indicum* (GCF_000512975.1) accession number on NCBI database, Flavonoid/Terpenes product, KEGG ID, amino acid substitution based on the polymorphic SNP in the 15 individuals of *Lavandula*

FLAVONOIDS				
Read ID	<i>S. ind</i> CDS ID	product	KO-IDs from KEGG	SNP to AA Subs.
3043	XP_011100449.1	anthocyanidin 3-O-glucosyltransferase 2	K12930	Ile -> Met
	XP_011100453.1	anthocyanidin 3-O-glucosyltransferase 2-like		
6706	XP_011090466.1	aspartate aminotransferase and glu/asp-prephenate aminotransferase	K15849	Val -> Ala
7480	XP_011089364.1	arogenate dehydratase/prephenate dehydratase 2, chloroplastic	K05359	Glu -> Val
	XP_011089363.1			
7969	XP_011094662.1	phenylalanine ammonia-lyase	K10775	Gln -> Arg
9011	XP_011089239.2	LOW QUALITY PROTEIN: 4-coumarate--CoA ligase-like 7	K01904	Gln -> Gln
9012				Gln -> Arg
9955	XP_020554052.1	putative anthocyanidin reductase isoform X2	K08695	Uncertain
	XP_011095308.1			X -> Leu
10947	XP_011069886.1	anthocyanidin 3-O-glucosyltransferase-like	K12930	Arg -> Pro
11587	XP_011077338.1	phenylalanine ammonia-lyase	K10775	His -> Tyr
TERPENES				
Read ID	<i>S. ind</i> CDS_ID	product	KO-IDs from KEGG	SNP to AA Subs.
8036	XP_011071094.1	1,4-dihydroxy-2-naphthoyl-CoA synthase, peroxisomal	K01661	X -> X
14576	XP_011096130.1	alpha-farnesene synthase	K14173	Gly -> Glu
6208	XP_011093795.1	beta-amyrin synthase	K15813	Lys -> Glu
8386	XP_011093795.1			X -> Arg
6208	XP_011085901.1	beta-amyrin synthase-like	K15813	Lys -> Glu
8386	XP_011085901.1			X -> Arg
6276	XP_011095756.1	ent-kaur-16-ene synthase, chloroplastic	N/A	Pro -> Ala
7199	XP_011083784.1	ent-kaurene oxidase, chloroplastic-like	K04122	Val -> Met
3576	XP_020550121.1	geranylgeranyl transferase type-2 subunit alpha 1	K09833	Leu -> Ser
10802	XP_011092247.1	gibberellin 20-oxidase-like protein	K05282	Gln -> Gln
11279	XP_011096560.1	gibberellin 2-beta-dioxygenase	K04125	Phe -> Leu
10014	XP_011098626.1	gibberellin-regulated protein 4-like	N/A	Arg -> Gln
	XP_011071640.1			
4578				Uncertain
6515	XP_011084658.1	isopentenyl-diphosphate Delta-isomerase I	K01823	Phe -> Leu
13525				Pro -> Pro
9817	XP_011075409.1	probable NAD(P)H dehydrogenase subunit CRR3, chloroplastic	N/A	Trp -> Leu
14513	XP_011082816.1	probable solanesyl-diphosphate synthase 3, chloroplastic	K05356	Leu -> Phe
14513	XP_011098150.1	probable solanesyl-diphosphate synthase 3, chloroplastic isoform X2		Leu -> Phe
5640	XP_020551000.1	protein prenyltransferase alpha subunit, isoform X6	K14137	Pro -> Gln
	XP_020551002.1			
3603	XP_011078470.1	squalene monooxygenase	K00511	Asn -> Thr
9296	XP_011092466.1	squalene monooxygenase-like		Asp -> His
5280	XP_011092839.1	squalene synthase	K00801	Pro -> Ser
	XP_011092841.1			
4990	XP_011082248.1	vetispiradiene synthase 3 isoform X2	K14182	Asp -> Glu
14152	XP_020548233.1	isochorismate synthase, chloroplastic-like	K01851	Arg -> Met
14154				Gln -> Pro
14685				Val -> Leu
14687	XP_020548234.1	isochorismate synthase, chloroplastic-like	K01851	Thr -> Thr
15015				Lys -> Lys

3.4. Sanger sequencing and DNA barcoding analysis

The analysis of DNA barcoding sequences commonly used in molecular taxonomy was conducted to verify the clustering reliability of the putative interspecific crosses hypothesized after ancestor membership reconstruction. The obtained sequences were 318 bp (psbA-trnH), 644 bp (rbcL), 273 bp (ITS) and 692 bp (matK) long, and the total concatenated sequence alignment among the four samples considered was 1,926 bp long. The majority of the aligned sites were conserved, but few insertions, SNPs or heterozygous positions (ITS) were found. The different site numbers ranged from 1 (e.g., “1826” vs. “1841”) to 20 (“SD-332” vs. “2605”) among the pairwise comparisons of the aligned sequences, whereas the total number of polymorphic sites in the alignment was equal to 25. The results obtained from the neighbor-joining tree construction revealed that samples were clustered in 3 main subgroups, but no concordances were observed with the previously obtained results based on the RAD-Seq dataset (see **Figure 3**).

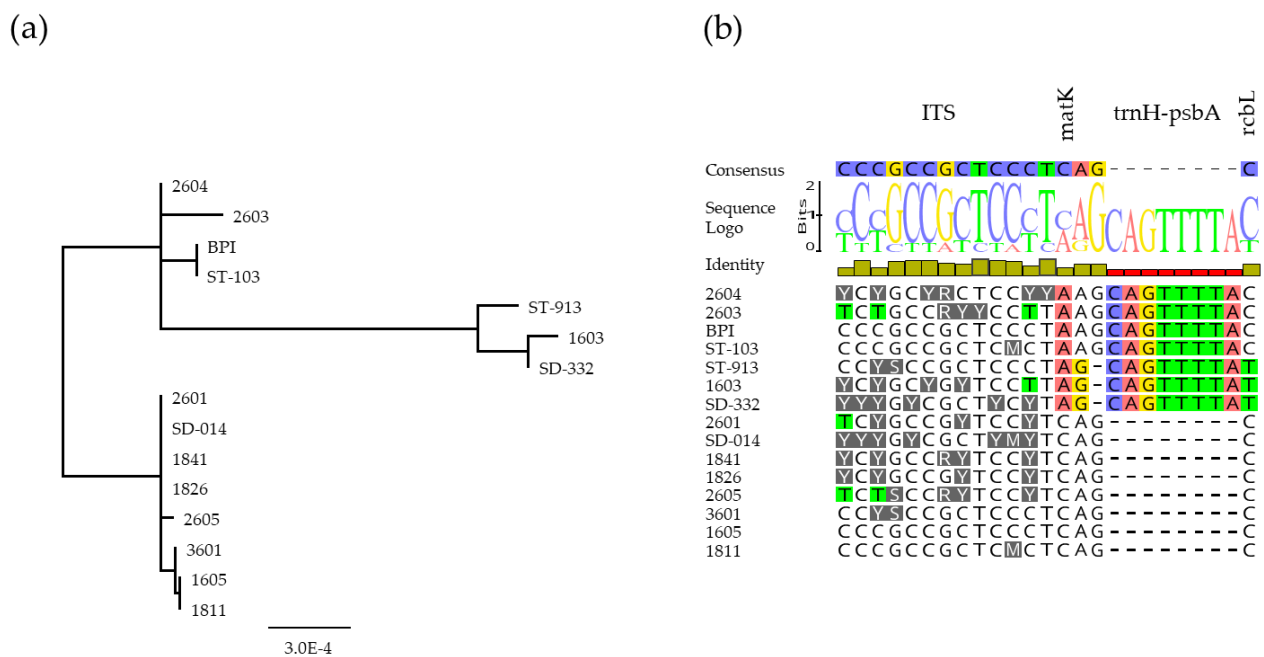


Figure 3. (a) Neighbour Joining tree based on the polymorphic sites among ITS nuclear region, and matK, trnH-psbA and rbcL chloroplast barcoding regions. (b) LOGO representation of polymorphic sites identified among the 15 *Lavandula* accessions analysed for the DNA barcoding.

4. Discussion

4.1. RAD-Seq-based genetic similarity and ancestral composition reconstruction

The use of molecular markers for genotyping analyses is currently one of the main tools in plant breeding and variety protection. Not only has this approach evolved in terms of informativeness during the late years, moving from dominant to codominant PCR-based and then to NGS-derived molecular markers, but it has also increased in the number of obtainable data and the robustness/informativeness of the resulting assays.

Indeed, RAD-Seq technology has been used for different applications in crop plant science ranging from QTL mapping in crop species [34-36] to Mendelian gene mapping [37,38] and marker-assisted breeding (MAS) [8,39-41]. This technique has also been used for crop variety identification [42] and phylogeny [43] studies, and population structure analyses [44]. In our study, we aimed to show the potential of the RAD-Seq approach in accessing the genetic identity or similarity and distinctiveness in *Lavandula* accessions, and at identifying putative genomic loci for use in breeding schemes, registering or patenting plant varieties and novelties, and protecting plant breeders' rights.

The great number of data points (42,610,020 total reads, 2,834,001 reads on average per sample) allowed us to investigate both the relatedness degree existing among the DNA samples and the SNP variants possibly linked to the biosynthesis of flavonoids and terpenes. To confer the robustness of the analysis, of the 43,271 SNP sites originally identified, only those with no missing data were retained (16,228). Notably, among the 27,043 RAD tags that were filtered and removed from the initial dataset, 1,044 had missing values in the *L. pedunculata* samples that were instead scored among the 13 individuals of *L. stoechas*. From these findings, it could be hypothesized that these loci are likely species-specific and could therefore be used for species discrimination. The filtered marker dataset used for the genetic similarity analysis allowed us to group the 15 samples into 5 main clusters. Moreover, the GS calculated within "Cluster A+B+C+D" was comparable to that calculated within "Cluster E", whereas the GS calculated between the two groups was lower, which is consistent with the fact that two different species were represented. Regarding the ancestral membership reconstruction, the number of $K = 3$ derived from the STRUCTURE software analysis was used to divide the 15 samples of the core collection of *Lavandula* into three main ancestors, showing membership percentages that were consistent with that obtained from the genetic similarity analysis. From these results, two main ancestors for accessions of *L. stoechas* were hypothesized, whereas one main ancestor mostly represented the *L. pedunculata* ancestry. The

fact that admixed memberships were present among samples belonging to different species can be explained by a few factors. In the first hypothesis, interspecific crosses can be present between the two considered species, a fact that is highly probable as they are reported to be cross-fertile and belong to the taxonomic section *Stoechas* of the genus *Lavandula* [45,46]. Notably, *L. stoechas* and *L. pedunculata* have been reported to be phylogenetically related and very close to one another. These species are so sufficiently closely related that *L. pedunculata* has been considered in the past as a subspecies of *L. stoechas* but was subsequently reassigned as a different species [45]. Then, the possibility of conserved loci among the analyzed samples is possible and could relate to common ancestral genotypes between the two species. Another consideration, excluding the possible biological explanations, is that the use of a reduced and filtered dataset based exclusively on loci that are shared among all analyzed samples and presenting no missing data could have resulted in a reduced capability of the molecular information in assessing the correct ancestry reconstruction. Specifically, missing data could be caused by missed sequencing of the genomic fragment in one or more samples or by the absence of the restricted genomic region due to a polymorphic nucleotide in the restriction site. In the first case, the missing information is not usable for genomic or statistical comparisons among the samples. In the second case, however, the absence of the data is an allele itself that could be used in species determination investigation. To address this issue, the use of an assembled genome of both or at least one of the analyzed species would be useful.

To confirm the first hypothesis, a barcoding analysis based on Sanger DNA sequencing of three cytoplasmic regions and one nuclear region was performed on the 15 samples of the core collection of *Lavandula*. The results obtained showed very few polymorphic sites among the analyzed sequences with a maximum number of 20 among 1,926 sequenced base pairs, which was approximately 1% of the total. These results were not in agreement with those obtained from the GS clustering or the ancestral reconstruction analysis performed by STRUCTURE. However, the difference can be explained by the different types of analysis performed and the nature of the molecular information used. The analyzed cytoplasmic DNA regions, including both genic and intergenic sequences, are inherited by the maternal parent, so they are not suitable for phylogenetic analyses in interspecific crosses. Thus, the ITS nuclear region was also considered and found able to discriminate the two *L. pedunculata* individuals from the other 13 accessions of *L. stoechas* (**Supplementary Figure S7**). Therefore, based on the observed data, the use of a DNA barcoding strategy in determining interspecific crosses is useless or much less informative than the RAD-Seq technology.

BLASTN analysis was also performed using the 16,228 RAD tags as queries against the *Sesamum indicum* RefSeq genome and *Salvia splendens* newly assembled genome to identify the

RAD tags most likely attributable to gene coding sequences and possibly phenotype related. A total of 16.1% of the reads matched the CDS from sesame, whereas 26.1% of the reads matched the exome regions of scarlet sage. Based on this analysis, it was possible to filter the original RAD-Seq dataset to a limited number of sequences that were subsequently used for a new and more stringent genetic similarity analysis. The resulting data used to calculate the genetic similarities and relationships among accessions and the extent of heterozygosity/homozygosity of all accessions showed no relevant differences compared with findings from the analysis of the nonfiltered dataset, with the exception of a few cases that can be explained by a higher similarity of the conserved exonic regions. In addition, the two PCoAs derived from these reduced datasets were consistently similar to the PCoAs performed using the initial 16,228 markers (**Figure 2** and **Supplementary Figures S5–6**), demonstrating once again the discriminative ability of the method used in these analyses and the relatedness of expressed and nonexpressed regions among the genomes in genotyping studies [47-49].

Regarding the heterozygosity estimates, it was observed that accessions showing a greater homozygosity were also those with the highest ancestral membership percentage to one or the other ancestors probably due to selfing or inbreeding reproductive strategies. The fact that few of the analyzed samples exhibited high levels of heterozygosity can be explained by the presence of interspecific crosses between the two species considered in this study. Notably, those samples with greater membership percentages with one of the three identified ancestors were also those with greater homozygosity (“Cluster C” and samples “SD-014” and “2603”), whereas the admixed samples showed the highest degree of heterozygosity (“Cluster A”). Consistent with the reproduction strategy of these species, autogamy rarely occurs in natural populations [2]. However, it has been reported that these species are self-compatible, so breeding lines can be obtained by increasing homozygosity levels through controlled self-pollinations. Moreover, highly heterozygous breeding lines can be maintained at their heterozygous status and can be vegetatively reproduced by cutting, thus maintaining the phenotypic characteristics of the line and their heterotic vigor and avoiding segregation after self-pollination or recombination from cross-pollination with other lines. Moreover, the use of interspecific crosses between *L. stoechas* and *L. pedunculata* is used to transfer phenotypic traits that are desired to be maintained for commercial purposes; thus, “hybrids” are reproduced by cutting to avoid loss of desired traits, which could explain the combined results of ancestry reconstruction with homozygosity. In conclusion, the results and type of data obtained through the method proposed in this study highlighted the informativeness of the approach used and showed how genotyping-by-sequencing thorough RAD-Seq is highly informative and could be considered a useful tool to be used in combination or in place of other genotyping technologies

based on PCR-based molecular markers, both dominant and codominant. Further studies are needed to confirm whether the identified SNPs are associated with phenotypic evidence.

Some findings about the STOP codons in genes involved in the synthesis of terpene precursors, including 1,4-dihydroxy-2-naphthoyl-CoA synthase, a phylloquinone precursor [50], and phosphomevalonate kinase (PMK), an inositol-diphosphate precursor [51], were particularly interesting, but further studies are needed to investigate and validate their gene function, expression, and compound synthesis to possibly correlate genotypes to chemotypes and phenotypes. This approach would be useful for MAB, including MAS approaches, and particularly for variety registration and protection.

The polymorphism information contents and molecular profiles obtained through the technology adopted in our research project would enable to guarantee the breeders' rights of the analyzed varieties and to legally protect them from any theft or embezzlement and commercialization by companies competing with the right's owner breeders. This aim would be further improved by the creation of specific molecular assays based on prebuilt arrays able to simplify and speed-up routine screenings. Most importantly, it would be helpful to legally define the genetic similarity/diversity thresholds between commercialized varieties able to consider them distinguishable or essentially derived to avoid misunderstandings or legal issues in the genus *Lavandula*, as has already been applied or suggested for other crops [52-54].

In conclusion, genotyping analysis by RAD-Seq reads was found useful to assess the genetic identity and relationships of breeding lines in lavender species aimed at managing plant variety protection.

5. References

1. Rice, A.; Glick, L.; Abadi, S.; Einhorn, M.; Kopelman, N.M.; Salman-Minkov, A.; Mayzel, J.; Chay, O.; Mayrose, I. The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. *New Phytol* **2015**, *206*, 19-26, doi:10.1111/nph.13191.
2. Munoz, A.; Devesa, J. Contribution to the knowledge of the floral biology of the genus *Lavandula* L., 2: *Lavandula stoechas* L. subsp. *stoechas*. *Anales del Jardín Botánico de Madrid* **1987**.
3. Shawl, A.S.; Kumar, S. Potential of lavender oil industry in Kashmir. *J Med Aromat Plant Sci* **2000**, *22*, 319-321.
4. Algieri, F.; Rodriguez-Nogales, A.; Vezza, T.; Garrido-Mesa, J.; Garrido-Mesa, N.; Utrilla, M.P.; Gonzalez-Tejero, M.R.; Casares-Porcel, M.; Molero-Mesa, J.; Del Mar Contreras, M., et al. Anti-inflammatory activity of hydroalcoholic extracts of *Lavandula dentata* L. and *Lavandula stoechas* L. *J Ethnopharmacol* **2016**, *190*, 142-158, doi:10.1016/j.jep.2016.05.063.
5. Zuzarte, M.d.R. Portuguese lavenders: evaluation of their potential use for health and agricultural purposes. Universidade de Coimbra, 2013.
6. Zuzarte, M.; Gonçalves, M.J.; Cavaleiro, C.; Cruz, M.T.; Benzarti, A.; Marongiu, B.; Maxia, A.; Piras, A.; Salgueiro, L. Antifungal and anti-inflammatory potential of *Lavandula stoechas* and *Thymus herba-barona* essential oils. *Industrial Crops and Products* **2013**, *44*, 97-103, doi:10.1016/j.indcrop.2012.11.002.
7. Pan, L.; Wang, N.; Wu, Z.; Guo, R.; Yu, X.; Zheng, Y.; Xia, Q.; Gui, S.; Chen, C. A High Density Genetic Map Derived from RAD Sequencing and Its Application in QTL Analysis of Yield-Related Traits in *Vigna unguiculata*. *Front Plant Sci* **2017**, *8*, 1544, doi:10.3389/fpls.2017.01544.
8. Patella, A.; Palumbo, F.; Ravi, S.; Stevanato, P.; Barcaccia, G. Genotyping by RAD Sequencing Analysis Assessed the Genetic Distinctiveness of Experimental Lines and Narrowed Down the Genomic Region Responsible for Leaf Shape in Endive (*Cichorium endivia* L.). *Genes-Basel* **2020**, *11*, 462, doi:10.3390/genes11040462.
9. Palumbo, F.; Galvao, A.C.; Nicoletto, C.; Sambo, P.; Barcaccia, G. Diversity Analysis of Sweet Potato Genetic Resources Using Morphological and Qualitative Traits and Molecular Markers. *Genes-Basel* **2019**, *10*, 840, doi:10.3390/genes10110840.
10. Barcaccia, G.; Palumbo, F.; Scariolo, F.; Vannozzi, A.; Borin, M.; Bona, S. Potentials and Challenges of Genomics for Breeding Cannabis Cultivars. *Front Plant Sci* **2020**, *11*, 573299, doi:10.3389/fpls.2020.573299.
11. Hnia, C.; Mohamed, B. Genetic diversity of *Lavandula multifida* L. (Lamiaceae) in Tunisia: implication for conservation. *African Journal of Ecology* **2011**, *49*, 10-20, doi:10.1111/j.1365-2028.2010.01223.x.
12. Prasad, A.; Shukla, S.P.; Mathur, A.; Chanotiya, C.S.; Mathur, A.K. Genetic fidelity of long-term micropropagated *Lavandula officinalis* Chaix.: an important aromatic medicinal plant.

Plant Cell, Tissue and Organ Culture (PCTOC) **2014**, *120*, 803-811, doi:10.1007/s11240-014-0637-7.

13. Ibrahim, H.M.; Salama, A.M.; Abou El-Leel, O.F. Analysis of genetic diversity of *Lavandula* species using taxonomic, essential oil and molecular genetic markers. *Sciences* **2017**, *7*, 141-154.
14. Zagorcheva, T.; Stanev, S.; Rusanov, K.; Atanassov, I. SRAP markers for genetic diversity assessment of lavender (*Lavandula angustifolia* mill.) varieties and breeding lines. *Biotechnology & Biotechnological Equipment* **2020**, *34*, 303-308, doi:10.1080/13102818.2020.1742788.
15. Adal, A.M.; Demissie, Z.A.; Mahmoud, S.S. Identification, validation and cross-species transferability of novel *Lavandula* EST-SSRs. *Planta* **2015**, *241*, 987-1004, doi:10.1007/s00425-014-2226-8.
16. Ahmed, S.M.; Alamer, K.H. Discriminating Lamiaceae Species from Saudi Arabia Using Allozyme and Specific DNA Markers. *Pak J Bot* **2018**, *50*, 969-975.
17. Adal, A.M. Development of molecular markers and cloning of genes involved in the biosynthesis of monoterpenes in *Lavandula*. University of British Columbia, 2019.
18. Angioni, A.; Barra, A.; Coroneo, V.; Dessi, S.; Cabras, P. Chemical composition, seasonal variability, and antifungal activity of *Lavandula stoechas* L. ssp. *stoechas* essential oils from stem/leaves and flowers. *J Agric Food Chem* **2006**, *54*, 4364-4370, doi:10.1021/jf0603329.
19. Tuttolomondo, T.; Dugo, G.; Ruberto, G.; Leto, C.; Napoli, E.M.; Potorti, A.G.; Fedele, M.R.; Virga, G.; Leone, R.; Anna, E.D., et al. Agronomical evaluation of Sicilian biotypes of *Lavandula stoechas* L. spp. *stoechas* and analysis of the essential oils. *Journal of Essential Oil Research* **2015**, *27*, 115-124, doi:10.1080/10412905.2014.1001527.
20. Li, J.; Wang, Y.; Dong, Y.; Zhang, W.; Wang, D.; Bai, H.; Li, K.; Li, H.; Shi, L. The chromosome-based lavender genome provides new insights into Lamiaceae evolution and terpenoid biosynthesis. *Horticulture research* **2021**, *8*, 1-14.
21. Stevanato, P.; Broccanello, C.; Biscarini, F.; Del Corvo, M.; Sablok, G.; Panella, L.; Stella, A.; Concheri, G. High-Throughput RAD-SNP Genotyping for Characterization of Sugar Beet Genotypes. *Plant Molecular Biology Reporter* **2013**, *32*, 691-696, doi:10.1007/s11105-013-0685-x.
22. Rochette, N.C.; Rivera-Colon, A.G.; Catchen, J.M. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol Ecol* **2019**, *28*, 4737-4754, doi:10.1111/mec.15253.
23. Rohlf, F. NTSYS-pc: numerical taxonomy multivariate analysis system. *Applied Biostatistics, I. & Exeter Software (Firm)* **2009**.
24. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945-959.
25. Earl, D.A.; vonHoldt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **2011**, *4*, 359-361, doi:10.1007/s12686-011-9548-7.

26. Moriya, Y.; Itoh, M.; Okuda, S.; Yoshizawa, A.C.; Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **2007**, *35*, W182-185, doi:10.1093/nar/gkm321.
27. Suzuki, S.; Kakuta, M.; Ishida, T.; Akiyama, Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One* **2014**, *9*, e103833, doi:10.1371/journal.pone.0103833.
28. Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **2010**, *38*, D355-360, doi:10.1093/nar/gkp896.
29. Nicolè, S.; Erickson, D.L.; Ambrosi, D.; Bellucci, E.; Lucchin, M.; Papa, R.; Kress, W.J.; Barcaccia, G. Biodiversity studies in Phaseolus species by DNA barcoding. *Genome* **2011**, *54*, 529-545.
30. Wojciechowski, M.F.; Lavin, M.; Sanderson, M. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *Am J Bot* **2004**, *91*, 1846-1862.
31. Sang, T.; Crawford, D.; Stuessy, T. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of Paeonia (Paeoniaceae). *Am J Bot* **1997**, *84*, 1120.
32. Tate, J.A.; Simpson, B.B. Paraphyly of Tarasa (Malvaceae) and diverse origins of the polyploid species. *Systematic Botany* **2003**, *28*, 723-737.
33. White, T.J.; Bruns, T.; Lee, S.; Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods applications* **1990**, *18*, 315-322.
34. Wang, J.; Wang, Z.; Du, X.; Yang, H.; Han, F.; Han, Y.; Yuan, F.; Zhang, L.; Peng, S.; Guo, E. A high-density genetic map and QTL analysis of agronomic traits in foxtail millet [*Setaria italica* (L.) P. Beauv.] using RAD-seq. *PLoS One* **2017**, *12*, e0179717, doi:10.1371/journal.pone.0179717.
35. Zhang, F.; Kang, J.; Long, R.; Yu, L.X.; Wang, Z.; Zhao, Z.; Zhang, T.; Yang, Q. High-density linkage map construction and mapping QTL for yield and yield components in autotetraploid alfalfa using RAD-seq. *Bmc Plant Biol* **2019**, *19*, 165, doi:10.1186/s12870-019-1770-6.
36. Wang, L.; Conteh, B.; Fang, L.; Xia, Q.; Nian, H. QTL mapping for soybean (*Glycine max* L.) leaf chlorophyll-content traits in a genotyped RIL population by using RAD-seq based high-density linkage map. *BMC Genomics* **2020**, *21*, 739, doi:10.1186/s12864-020-07150-4.
37. Wu, K.; Liu, H.; Yang, M.; Tao, Y.; Ma, H.; Wu, W.; Zuo, Y.; Zhao, Y. High-density genetic map construction and QTLs analysis of grain yield-related traits in sesame (*Sesamum indicum* L.) based on RAD-Seq technology. *Bmc Plant Biol* **2014**, *14*, 274, doi:10.1186/s12870-014-0274-7.
38. Peng, Y.; Hu, Y.; Mao, B.; Xiang, H.; Shao, Y.; Pan, Y.; Sheng, X.; Li, Y.; Ni, X.; Xia, Y., et al. Genetic analysis for rice grain quality traits in the YVB stable variant line using RAD-seq. *Mol Genet Genomics* **2016**, *291*, 297-307, doi:10.1007/s00438-015-1104-9.

39. Yang, H.; Tao, Y.; Zheng, Z.; Shao, D.; Li, Z.; Sweetingham, M.W.; Buirchell, B.J.; Li, C. Rapid development of molecular markers by next-generation sequencing linked to a gene conferring phomopsis stem blight disease resistance for marker-assisted selection in lupin (*Lupinus angustifolius* L.) breeding. *Theor Appl Genet* **2013**, *126*, 511-522, doi:10.1007/s00122-012-1997-1.
40. Fan, W.; Zong, J.; Luo, Z.; Chen, M.; Zhao, X.; Zhang, D.; Qi, Y.; Yuan, Z. Development of a RAD-Seq Based DNA Polymorphism Identification Software, AgroMarker Finder, and Its Application in Rice Marker-Assisted Breeding. *PLoS One* **2016**, *11*, e0147187, doi:10.1371/journal.pone.0147187.
41. Yamashita, H.; Uchida, T.; Tanaka, Y.; Katai, H.; Nagano, A.J.; Morita, A.; Ikka, T. Genomic predictions and genome-wide association studies based on RAD-seq of quality-related metabolites for the genomics-assisted breeding of tea plants. *Sci Rep* **2020**, *10*, 17480, doi:10.1038/s41598-020-74623-7.
42. Kawamura, K.; Kawanabe, T.; Shimizu, M.; Nagano, A.J.; Saeki, N.; Okazaki, K.; Kaji, M.; Dennis, E.S.; Osabe, K.; Fujimoto, R. Genetic distance of inbred lines of Chinese cabbage and its relationship to heterosis. *Plant Gene* **2016**, *5*, 1-7.
43. Liu, L.; Jin, X.; Chen, N.; Li, X.; Li, P.; Fu, C. Phylogeny of *Morella rubra* and Its Relatives (Myricaceae) and Genetic Resources of Chinese Bayberry Using RAD Sequencing. *PLoS One* **2015**, *10*, e0139840, doi:10.1371/journal.pone.0139840.
44. Feng, J.; Zhao, S.; Li, M.; Zhang, C.; Qu, H.; Li, Q.; Li, J.; Lin, Y.; Pu, Z. Genome-wide genetic diversity detection and population structure analysis in sweetpotato (*Ipomoea batatas*) using RAD-seq. *Genomics* **2020**, *112*, 1978-1987, doi:10.1016/j.ygeno.2019.11.010.
45. Moja, S.; Guitton, Y.; Nicolè, F.; Legendre, L.; Pasquier, B.; Upson, T.; Jullien, F. Genome size and plastid trnK-matK markers give new insights into the evolutionary history of the genus *Lavandula* L. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology* **2015**, *150*, 1216-1224, doi:10.1080/11263504.2015.1014006.
46. B, R.J.; J, C.R. Multi-bracted lavender plants. 2016.
47. Lam, H.Y.; Clark, M.J.; Chen, R.; Chen, R.; Natsoulis, G.; O'huallachain, M.; Dewey, F.E.; Habegger, L.; Ashley, E.A.; Gerstein, M.B. Performance comparison of whole-genome sequencing platforms. *Nature biotechnology* **2012**, *30*, 78-82.
48. Eklöf, H.; Bernhardsson, C.; Ingvarsson, P.K. Comparing the Effectiveness of Exome Capture Probes, Genotyping by Sequencing and Whole-Genome Re-Sequencing for Assessing Genetic Diversity in Natural and Managed Stands of *Picea abies*. *Forests* **2020**, *11*, 1185, doi:10.3390/f11111185.
49. Rabbi, I.Y.; Kulakow, P.A.; Manu-Aduening, J.A.; Dankyi, A.A.; Asibuo, J.Y.; Parkes, E.Y.; Abdoulaye, T.; Girma, G.; Gedil, M.A.; Ramu, P., et al. Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). *Bmc Genet* **2015**, *16*, 115, doi:10.1186/s12863-015-0273-1.
50. McCoy, R.M.; Utturkar, S.M.; Crook, J.W.; Thimmapuram, J.; Widhalm, J.R. The origin and biosynthesis of the naphthalenoid moiety of juglone in black walnut. *Hortic Res* **2018**, *5*, 67, doi:10.1038/s41438-018-0067-5.

51. Niu, M.; Xiong, Y.; Yan, H.; Zhang, X.; Li, Y.; da Silva, J.A.T.; Ma, G. Cloning and Expression Analysis of Mevalonate Kinase and Phosphomevalonate Kinase Genes Associated with MVA Pathway in *Santalum Album*. *Scientific Reports* **2020**.
52. Achard, F.; Butruille, M.; Madjarac, S.; Nelson, P.; Duesing, J.; Laffont, J.L.; Nelson, B.; Xiong, J.; Mikel, M.A.; Smith, J. Single nucleotide polymorphisms facilitate distinctness-uniformity-stability testing of soybean cultivars for plant variety protection. *Crop Science* **2020**, *60*, 2280-2303.
53. Jamali, S.H.; Cockram, J.; Hickey, L.T. Insights into deployment of DNA markers in plant variety protection and registration. *Theor Appl Genet* **2019**, *132*, 1911-1929, doi:10.1007/s00122-019-03348-7.
54. Yu, J.-K.; Chung, Y.-S. Plant Variety Protection: Current Practices and Insights. *Genes* **2021**, *12*, 1127.

6. Supplementary materials

Table S3. Multiple alignments results reporting read ID, *S. splendens* (GCA_004379255.2) accession number on NCBI database, Flavonoid/Terpenes product, KEGG ID assigned by KASS, amino acid substitution based on the polymorphic SNP in the 15 individuals of *Lavandula*

<i>Salvia splendens</i> V2 (GCA_004379255.2)				
FLAVONOIDS				
Read ID	S.sp CDS ID	product	KO-Ids from KAAS	SNP to AA Substitution
7930	KAG6403399.1	cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]	K00083	Leu -> Trp
13322	KAG6386750.1			Arg -> Cys
795	KAG6417397.1	peroxidase [EC:1.11.1.7]	K00430	Lys -> Arg
5487	KAG6392501.1			Thr -> Ile
	KAG6403428.1			
1515	KAG6394251.1	tyrosine aminotransferase [EC:2.6.1.5]	K00815	Asn -> Lys
1516	KAG6395555.1			Lys -> Arg
11728	KAG6395439.1	anthranilate synthase component I [EC:4.1.3.27]	K01657	Ser -> Pro
11729	KAG6393611.1			Asn -> Ser
11730	KAG6393611.1			Ser -> Ser
1270	KAG6383281.1	tryptophan synthase alpha chain [EC:4.2.1.20]	K01695	Pro -> Arg
1271	KAG6403024.1			Gly -> Glu
15032	KAG6420802.1	tryptophan synthase beta chain [EC:4.2.1.20]	K01696	Gly -> Ala
	KAG6423817.1			
6696	KAG6392655.1	3-dehydroquinate synthase [EC:4.2.3.4]	K01735	Asp -> Tyr
	KAG6394413.1			
1299	KAG6393529.1	phosphoribosylanthranilate isomerase [EC:5.3.1.24]	K01817	Pro -> Ala
	KAG6395367.1			
9011	KAG6435185.1	4-coumarate--CoA ligase [EC:6.2.1.12]	K01904	Gln -> Gln
9012	KAG6435185.1			Arg -> Gln
14333	KAG6430879.1	4-coumarate--CoA ligase [EC:6.2.1.12]	K01904	Gly -> Gly
14334	KAG6430879.1			Arg -> Arg
7488	KAG6436443.1	beta-glucosidase [EC:3.2.1.21]	K05349	Leu -> Ser
11456	KAG6436443.1			Ala -> Asp
7480	KAG6396700.1	arogenate/prephenate dehydratase [EC:4.2.1.91; 4.2.1.51]	K05359	Glu -> Val
	KAG6417935.1			
3421	KAG6386308.1	phenylalanine ammonia-lyase [EC:4.3.1.24]	K10775	Gly -> Val
7969	KAG6412059.1			Gln -> Arg
11587	KAG6434018.1			His -> Tyr
	KAG6427421.1			
10947	KAG6430482.1	anthocyanidin 3-O-glucosyltransferase [EC:2.4.1.115]	K12930	Arg -> Pro
	KAG6428875.1			
11569	KAG6382354.1	shikimate O-hydroxycinnamoyltransferase [EC:2.3.1.133]	K13065	Glu -> Glu

14519	KAG6421699.1 KAG6421697.1	isoflavone 4'-methoxyisoflavone 2'-hydroxylase [EC:1.14.14.90; 1.14.14.89]	K13260	Asn -> Asn
13190 13891	KAG6423171.1	3-dehydroquinate dehydratase shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]	K13832	Ala -> Ala Leu -> Leu
6706 10566	KAG6397466.1 KAG6399730.1	bifunctional aspartate aminotransferase and glutamate aspartate- prephenate aminotransferase [EC:2.6.1.1; 2.6.1.78; 2.6.1.79]	K15849	Val -> Ala Val -> Gly
9183 10093	KAG6430589.1 KAG6394378.1 KAG6385903.1	caffeoylshikimate esterase [EC:3.1.1.-]	K18368	Pro -> Ser Ala -> Ala

TERPENES

Read ID	S.sp CDS ID	product	KO-Ids from KAAS	SNP to AA Substitution		
2113 13445 13821	KAG6400172.1 KAG6401460.1 KAG6401461.1 KAG6401462.1	hydroxymethylglutaryl-CoA reductase (NADPH) [EC:1.1.1.34]	K00021	Ala -> Pro Phe -> Val Tyr -> Ser		
13998 14741	KAG6401463.1 KAG6402379.1 KAG6403689.1 KAG6403690.1			Glu -> Val Arg -> Thr		
15739	KAG6403691.1 KAG6438234.1			Thr -> Ile		
9946	KAG6401668.1 KAG6386175.1 KAG6403927.1			1-deoxy-D-xylulose-5-phosphate reductoisomerase [EC:1.1.1.267]	K00099	Ala -> Thr
6406 11290 16152	KAG6426406.1 KAG6403942.1 KAG6412213.1 KAG6386168.1			aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	K00128	Gln -> Pro Lys -> Asn Val -> Leu
111	KAG6430119.1 KAG6427108.1			cytokinin dehydrogenase [EC:1.5.99.12]	K00279	Gly -> Gly
3603	KAG6400376.1 KAG6402590.1	squalene monooxygenase [EC:1.14.14.17]	K00511	Undetermined		
1515 1516	KAG6395555.1	tyrosine aminotransferase [EC:2.6.1.5]	K00815	Asn -> Lys Lys -> Arg		
5069 16201 16202	KAG6397147.1 KAG6436372.1 KAG6417367.1 KAG6414860.1	phosphomevalonate kinase [EC:2.7.4.2] diphosphomevalonate decarboxylase [EC:4.1.1.33]	K00938 K01597	X -> X Lys -> Glu Gly -> Val		
7456 7818 13637 13638	KAG6416305.1 KAG6418848.1 KAG6431799.1	hydroxymethylglutaryl-CoA synthase [EC:2.3.3.10]	K01641	Arg -> Leu Gly -> Ser Tyr -> His Pro -> Ser		
7267 8365 8976	KAG6435503.1 KAG6432198.1	1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]	K01662	Val -> Val Lys -> Arg Asp -> Asp		
4578 6515 13525	KAG6384495.1 KAG6404725.1 KAG6416140.1	isopentenyl-diphosphate Delta-isomerase [EC:5.3.3.2]	K01823	Undetermined		
9011 9012	KAG6435185.1	4-coumarate--CoA ligase [EC:6.2.1.12]	K01904	Gln -> Gln Arg -> Gln		

14333				Gly -> Gly
14334	KAG6430879.1			Arg -> Arg
11006	KAG6428791.1	15-cis-phytoene desaturase [EC:1.3.5.5]	K02293	Pro -> Ser
14152				
14154				
14685	KAG6404049.1	menaquinone-specific isochorismate synthase [EC:5.4.4.2]	K02552	Undetermined
14687				
15015				
7807	KAG6415409.1	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase		Val -> Phe
11373		[EC:1.17.7.1; 1.17.7.3]	K03526	Leu -> Val
11381	KAG6432822.1			Leu -> Ile
686	KAG6410751.1	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase		Lys -> Glu
10714	KAG6410735.1	[EC:1.17.7.4]	K03527	Leu -> Pro
13487				
13769	KAG6405342.1	ent-copalyl diphosphate synthase [EC:5.5.1.13]	K04120	Ala -> Val
				Arg -> Lys
6276	KAG6392464.1	ent-kaurene synthase [EC:4.2.3.19]	K04121	Pro -> Ala
7199	KAG6403159.1			
13444	KAG6400985.1	ent-kaurene oxidase [EC:1.14.14.86]	K04122	Val -> Met
				Thr -> Ala
5987	KAG6434513.1	ent-kaurenoic acid monooxygenase [EC:1.14.14.107]	K04123	Pro -> Thr
	KAG6437860.1			
	KAG6424438.1			
14513	KAG6421419.1	all-trans-nonaprenyl-diphosphate synthase [EC:2.5.1.84; 2.5.1.85]	K05356	Leu -> Phe
	KAG6385395.1			
	KAG6434439.1			
9534	KAG6437590.1	STE24 endopeptidase [EC:3.4.24.84]	K06013	Arg -> Gly
12560	KAG6437297.1	carotenoid epsilon hydroxylase [EC:1.14.14.158]	K09837	Asn -> Asp
	KAG6387826.1			
1426	KAG6409765.1	9-cis-epoxycarotenoid dioxygenase [EC:1.13.11.51]	K09840	Asn -> Ser
12154	KAG6408920.1			
12155	KAG6406577.1	(+)-abscisic acid 8'-hydroxylase [EC:1.14.14.137]	K09843	Tyr -> Asn
				Arg -> Gln
6830	KAG6413659.1			
6838	KAG6427098.1			
7041	KAG6427099.1	cis-zeatin O-glucosyltransferase [EC:2.4.1.215]	K13495	Undetermined
	KAG6427097.1			
3211	KAG6385247.1	o-succinylbenzoate---CoA ligase [EC:6.2.1.26]	K14760	Lys -> Arg
6208	KAG6434381.1			Lys -> Glu
8386	KAG6437655.1	beta-amyrin synthase [EC:5.4.99.39]	K15813	X -> Arg
10875				Pro -> Pro

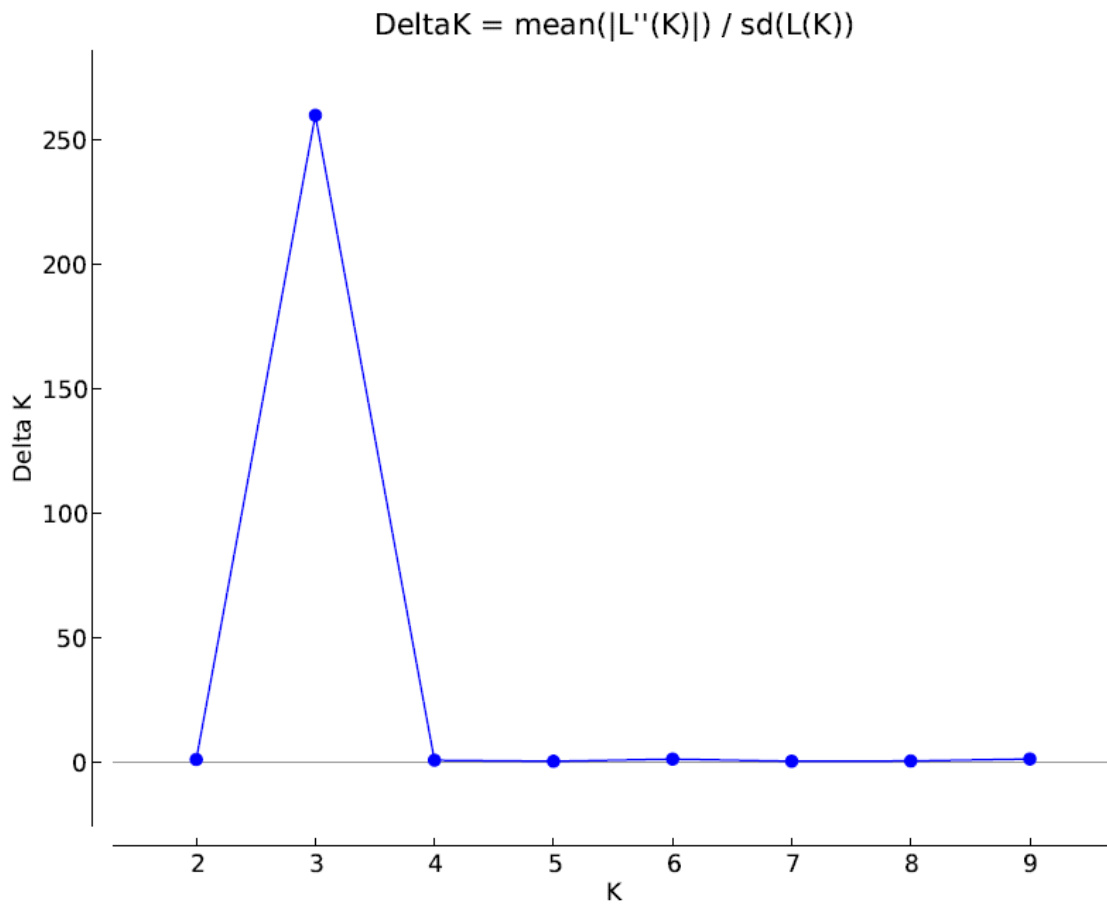


Figure S1. STRUCTURE Harvester software resulting ΔK chart

(a)

Obs. He	Genetic similarity (GS)																	
32.3%	Cluster A	A-1603	100.0%															
39.3%		B-2601	82.5%	100.0%														
25.4%		B-2604	79.1%	77.0%	100.0%													
22.2%	Cluster B	A-1605	81.0%	78.5%	74.7%	100.0%												
21.6%		1841	79.3%	77.3%	72.5%	86.8%	100.0%											
21.4%		1826	80.7%	78.0%	75.6%	87.1%	87.4%	100.0%										
13.9%		SD-041	77.6%	75.0%	71.7%	88.6%	84.5%	83.7%	100.0%									
10.2%	Cluster C	BPI	75.9%	74.0%	69.8%	83.8%	80.5%	84.4%	85.3%	100.0%								
14.8%		ST-913	76.2%	75.4%	71.6%	86.4%	81.8%	87.0%	84.8%	94.1%	100.0%							
16.4%		SD-332	77.0%	75.8%	71.9%	84.3%	80.3%	85.3%	86.4%	94.3%	93.9%	100.0%						
10.8%		1811	76.7%	74.0%	69.5%	86.5%	85.8%	86.0%	87.2%	90.4%	92.4%	90.0%	100.0%					
21.8%	Cluster D	ST-103	77.0%	74.2%	70.8%	82.1%	81.0%	78.7%	83.4%	83.7%	83.1%	83.1%	83.3%	100.0%				
23.0%		C-3601	74.3%	72.4%	69.7%	79.4%	78.1%	76.2%	80.8%	77.8%	78.5%	78.5%	81.5%	87.2%	100.0%			
13.1%	Cluster E	B-2603	66.1%	66.9%	68.2%	58.9%	62.3%	60.1%	58.1%	56.8%	57.1%	57.0%	55.5%	58.2%	57.3%	100.0%		
28.9%		B-2605	68.5%	69.8%	70.2%	64.1%	68.4%	65.1%	62.5%	60.7%	62.4%	62.1%	61.7%	65.8%	66.3%	74.2%	100.0%	
			A-1603	B-2601	B-2604	A-1605	SD-041	1841	1826	BPI	ST-913	SD-332	1811	ST-103	C-3601	B-2603	B-2605	
			Cluster A			Cluster B				Cluster C				Cluster D		Cluster E		

(b)

Obs. Ho	Sample	Avg. Genetic similarity											
67.7% ± 4.0%	Cluster A	79.5% ± 1.6%											
80.2% ± 2.0%	Cluster B	76.8% ± 0.9%		86.4% ± 0.8%									
86.9% ± 1.5%	Cluster C	74.0% ± 0.8%		84.7% ± 0.5%		92.5% ± 0.8%							
77.6% ± 0.4%	Cluster D	73.1% ± 1.1%		80.0% ± 0.8%		81.2% ± 0.9%		87.2% ± N/A					
79.0% ± 7.9%	Cluster E	68.3% ± 0.7%		62.5% ± 1.2%		59.2% ± 1.0%		61.9% ± 2.4%	74.2% ± N/A				
79.0% ± 2.3%	A+B+C+D	62.7% ± 0.9%						80.7% ± 0.7%					
		Cluster A		Cluster B		Cluster C		Cluster D		Cluster E		A+B+C+D	

Figure S2: (a) Genetic Similarity matrix of 15 *Lavandula* individuals based the BLASTN analysis against *S. splendens* exome, and relative observed homozygosity (Obs. Ho) and heterozygosity (Obs. He). (b) Average genetic similarity of Clusters identified through the construction of the UPGMA dendro-gram, and average observed homozygosity (Avg. Obs. Ho)

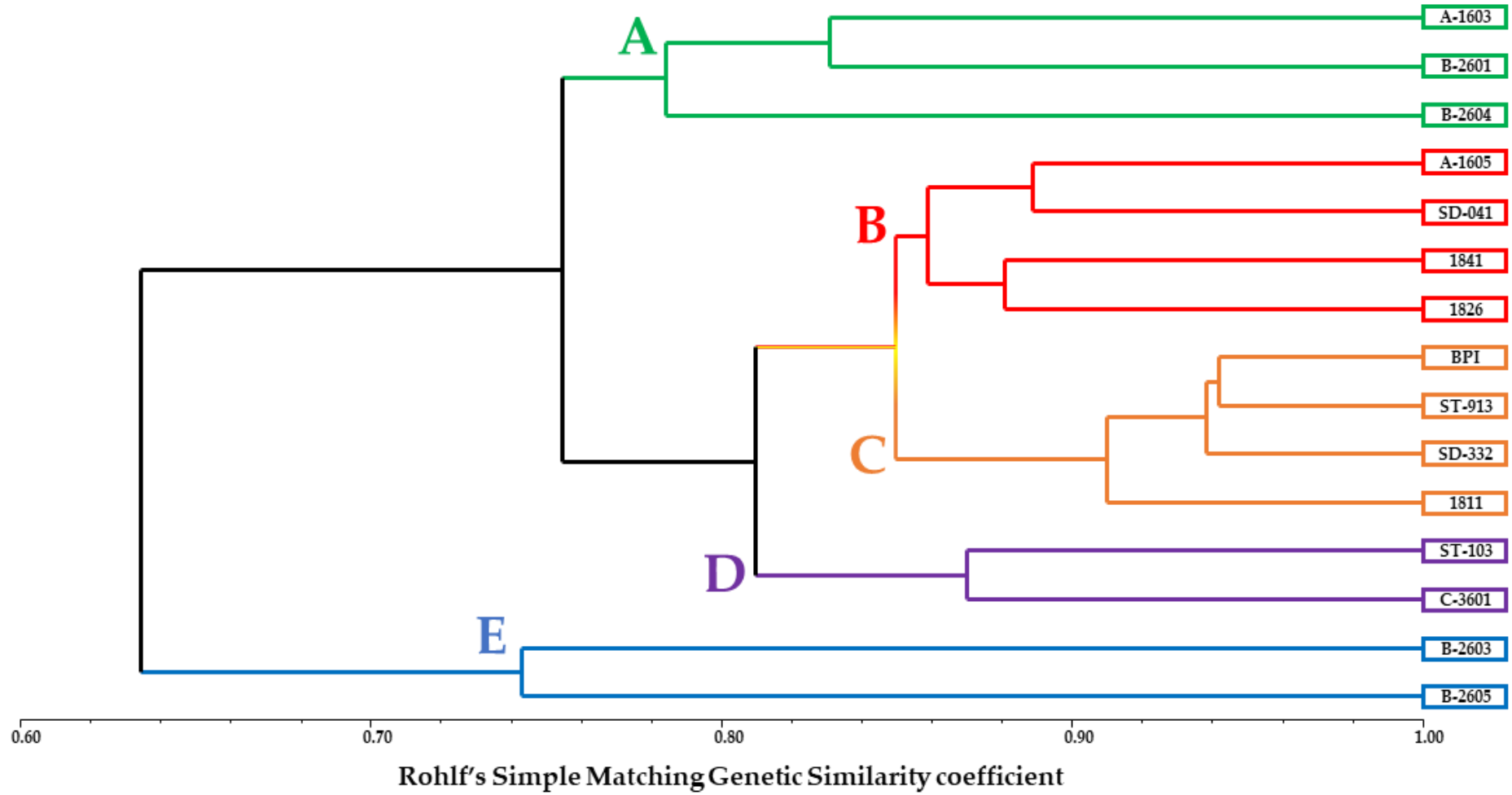


Figure S3. UPGMA dendrogram of the genetic similarity calculated on the *Lavandula* reads matching the *S. indicum* exome

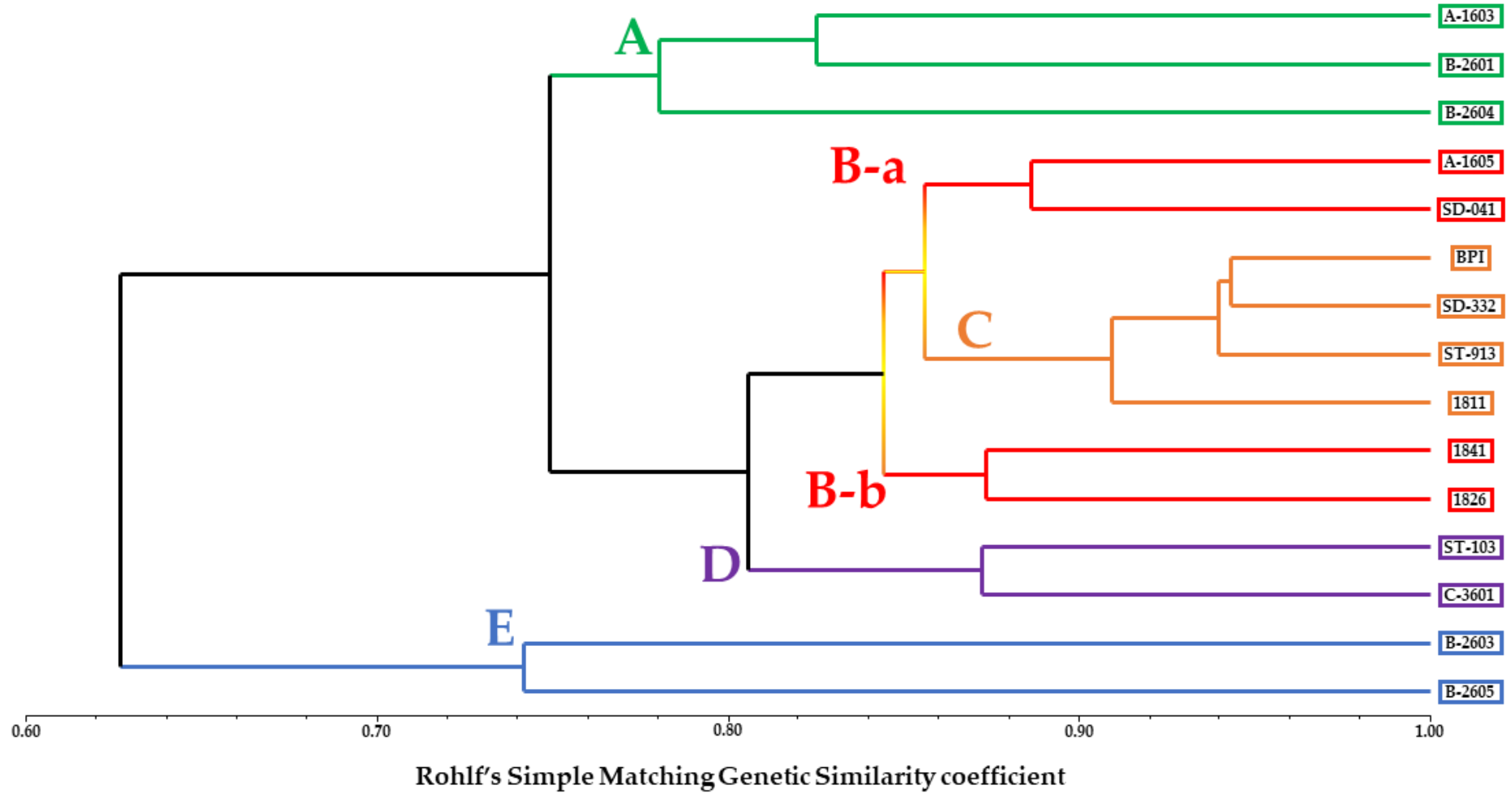


Figure S4. UPGMA dendrogram of the genetic similarity calculated on the *Lavandula* reads matching the *S. splendens* exome

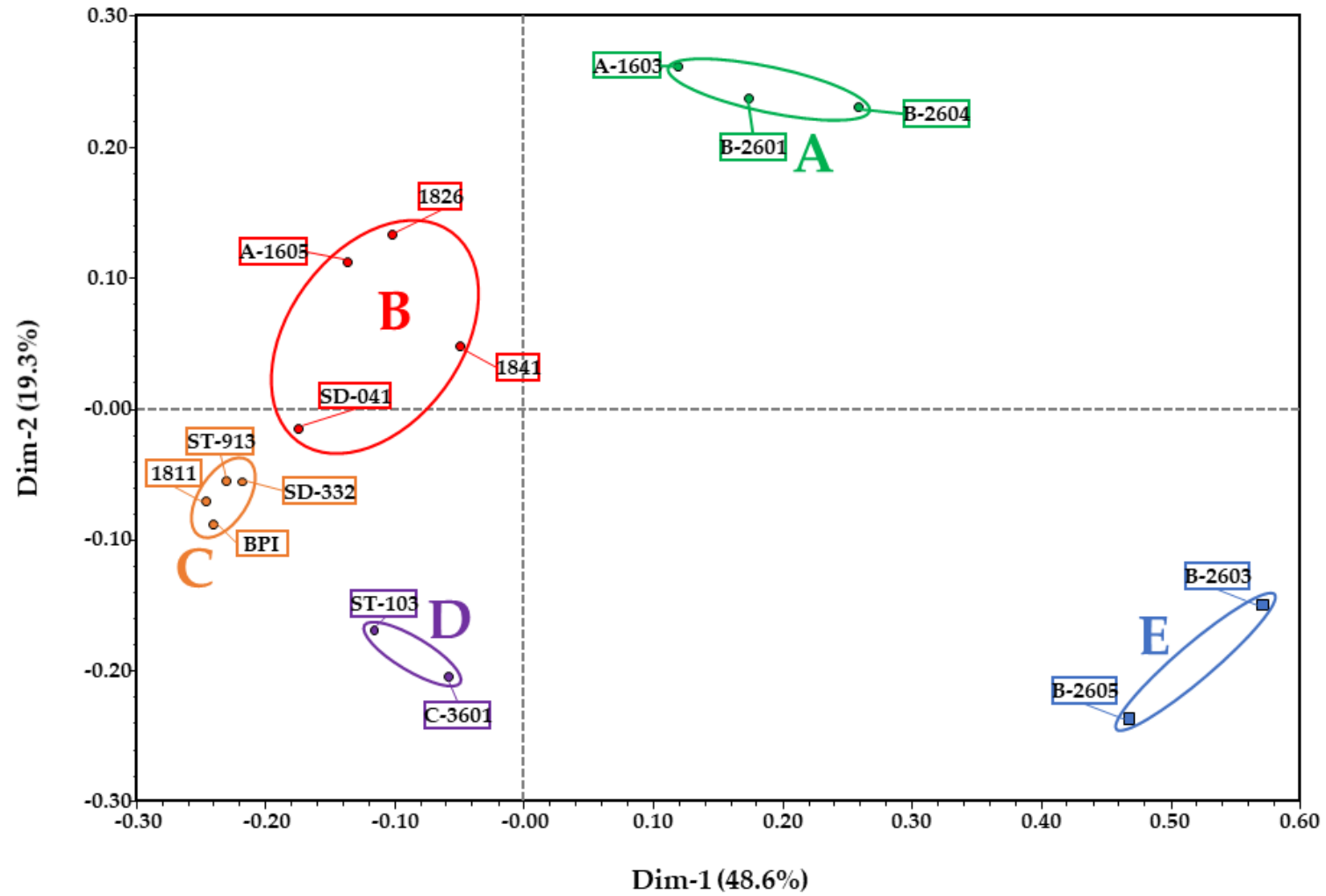


Figure S5. Principal Coordinate Analysis (PCoA) of the genetic similarity calculated on the *Lavandula* reads matching the *S. indicum* exome

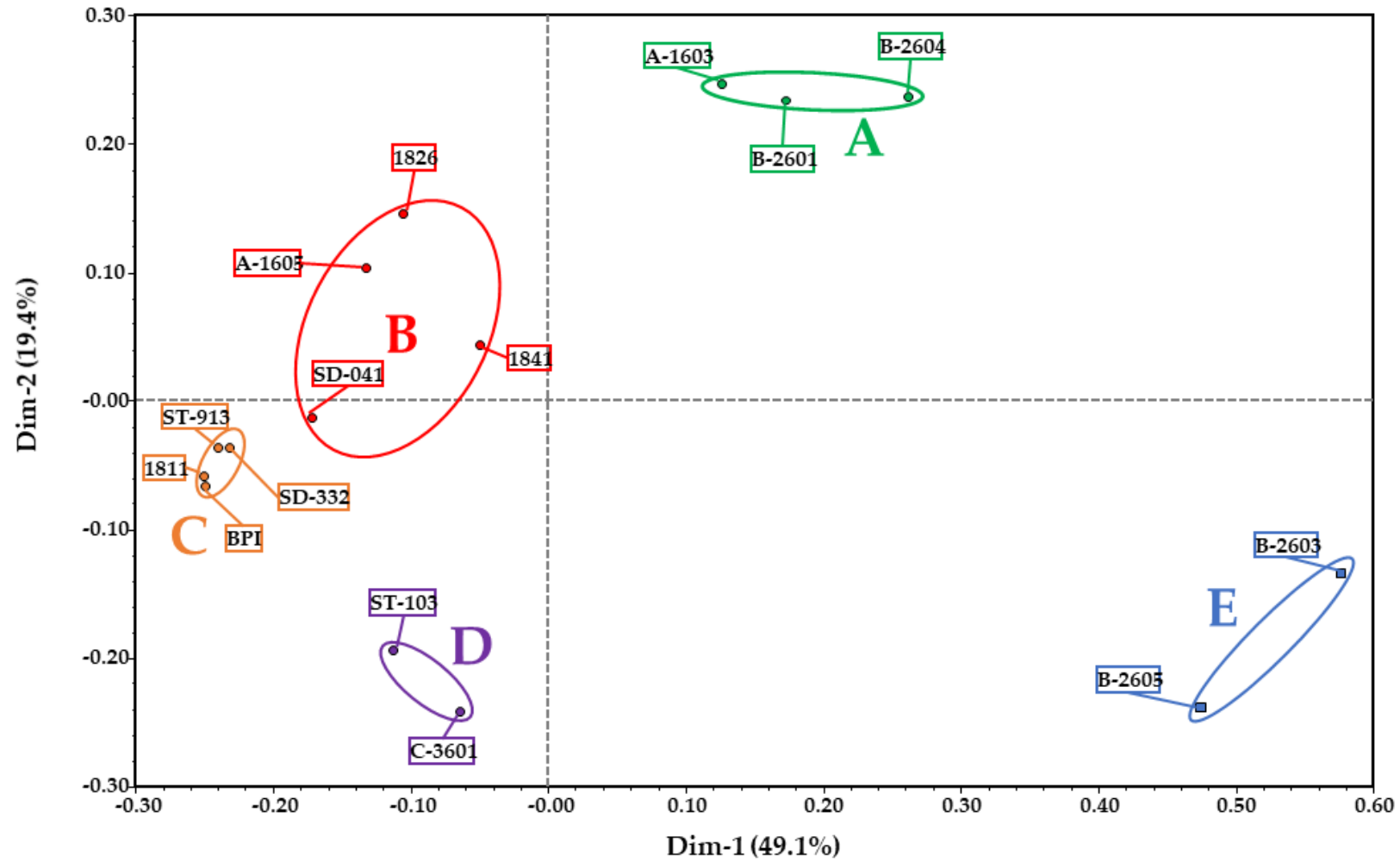


Figure S6. Principal Coordinate Analysis (PCoA) of the genetic similarity calculated on the *Lavandula* reads matching the *S. splendens* exome

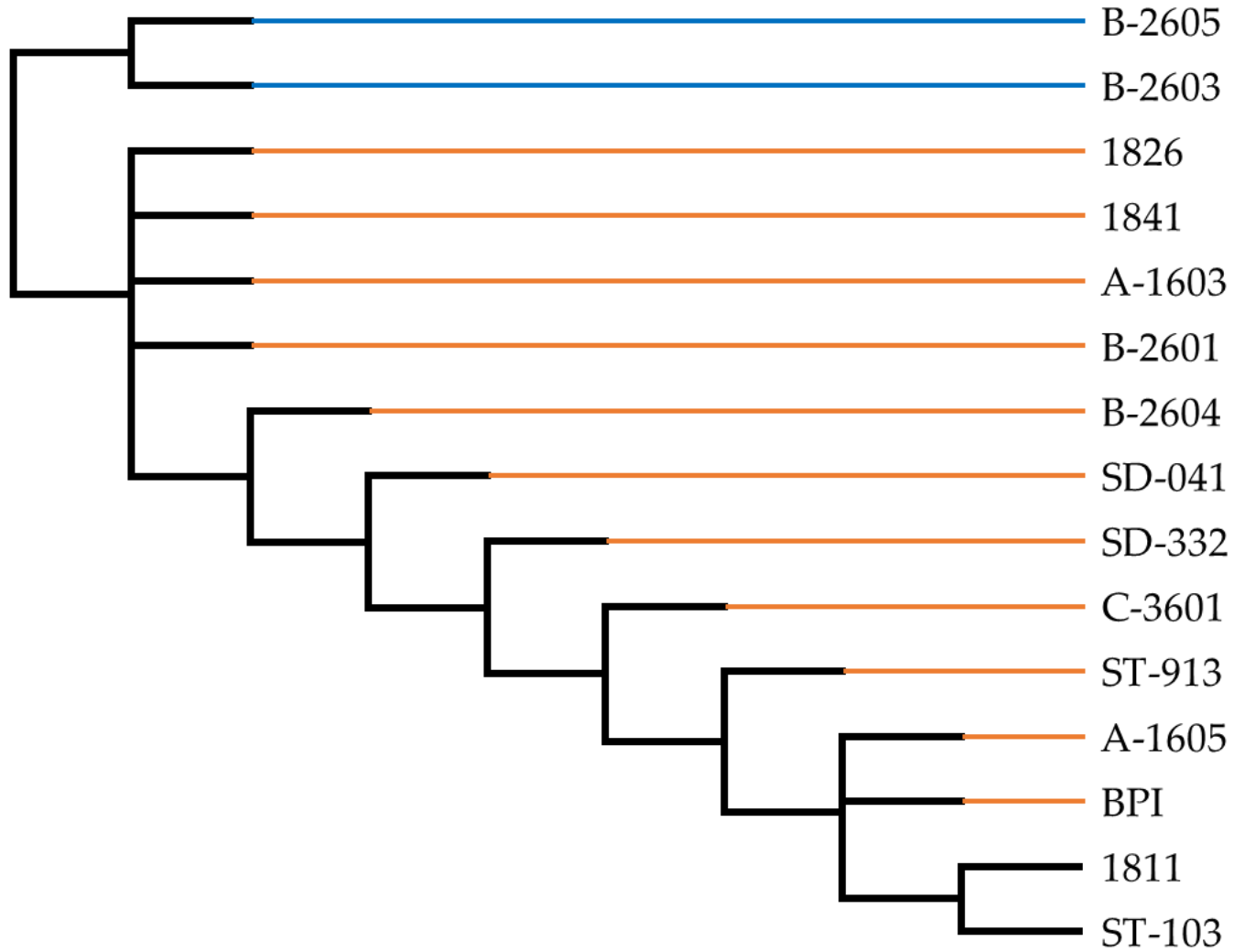


Figure S7. Neighbour Joining tree based on Gencious® software plug-in “multiple alignment” of ITS nuclear region of the 15 *Lavandula* individuals

7. Informatic material

Informatic material can be accessed at the following link:

https://drive.google.com/drive/folders/11KNrlGp_IKjSfpETooIL_3k0WGqbPZza?usp=sharing

Table I1. BLASTN result of the RAD-seq obtained reads of Lavandula against the *S. indicum* exome

Table I2. BLASTN result of the RAD-seq obtained reads of Lavandula against the *S. splendens* exome

Table I3. BLASTN results for Lavandula reads matching genes involved in the Flavonoids and Terpenes biosynthetic pathways of *S. indicum*

Table I4. BLASTN results for Lavandula reads matching genes involved in the Flavonoids and Terpenes biosynthetic pathways of *S. splendens*

Publications :

1. Patella, A., **Scariolo, F.**, Palumbo, F., & Barcaccia, G. (2019). Genetic structure of cultivated varieties of radicchio (*Cichorium intybus* L.): A comparison between fl hybrids and synthetics. *Plants*, 8(7), 213.
2. Palumbo, F., **Scariolo, F.**, Vannozzi, A., & Barcaccia, G. (2020). NGS-based barcoding with mini-COI gene target is useful for pet food market surveys aimed at mislabelling detection. *Scientific reports*, 10(1), 1-8.
3. Barcaccia, G., Palumbo, F., **Scariolo, F.**, Vannozzi, A., Borin, M., & Bona, S. (2020). Potentials and challenges of genomics for breeding cannabis cultivars. *Frontiers in plant science*, 11, 1472.
4. Sica, P., Galvao, A., **Scariolo, F.**, Maucieri, C., Nicoletto, C., Pilon, C., ... & Franklin, D. (2021). Effects of drought on yield and nutraceutical properties of beans (*Phaseolus* spp.) traditionally cultivated in Veneto, Italy. *Horticulturae*, 7(2), 17.
5. Sica, P., **Scariolo, F.**, Galvao, A., Battaglia, D., Nicoletto, C., Maucieri, C., ... & Barcaccia, G. (2021). Molecular Hallmarks, Agronomic Performances and Seed Nutraceutical Properties to Exploit Neglected Genetic Resources of Common Beans Grown by Organic Farming in Two Contrasting Environments. *Frontiers in plant science*, 12.
6. **Scariolo, F.**, Palumbo F., Vannozzi A., Sacilotto B., Gazzola M. and Barcaccia G. Genotyping Analysis by RAD-Seq Reads is Useful to Assess the Genetic Identity and Relationships of Breeding Lines in Lavender Species Aimed at Managing Plant Variety Protection. *Genes MDPI* (on going revision)
7. Borin M, Palumbo F, Vannozzi A, **Scariolo F**, Gazzola M, Sacilotto B, and Barcaccia G. Developing and Testing Molecular Markers in Cannabis sativa (Hemp) for their Use in Variety and Dioecy Assessments. *Plants MDPI* (on going revision)
8. Basso A., **Scariolo F.**, Negrisolo E. and Barcaccia G. Molecular relationships and phylogenies of Venetian Radicchio (leaf chicory, *Cichorium intybus* subsp. *intybus* var. *foliosum*, $2n=2x=18$) varietal groups. *Diversity MDPI* (on going revision)
9. **Scariolo, F.**, Fabio Palumbo, Alessandro Vannozzi, Gio B. Sacilotto, Marco Gazzola, and Gianni Barcaccia. 2021. "Genotyping Analysis by RAD-Seq Reads Is Useful to Assess the Genetic Identity and Relationships of Breeding Lines in Lavender Species Aimed at Managing Plant Variety Protection" *Genes* 12, no. 11: 1656. <https://doi.org/10.3390/genes12111656>

Congresses :

1. **Scariolo F.**, Palumbo F., Barcaccia G. “Molecular identification of species in pet-food products by mini-barcoding combined with Next-Generation Sequencing”, poster presentation at SIGA annual congress (September 2019)
2. Patella A., **Scariolo F.**, Palumbo F., Barcaccia G. “Molecular marker-based DUS testing for radicchio F1 and OP varieties”, poster presentation at SIGA annual congress (September 2019)
3. Oral speech at the international congress ExoFlowMetry, 13th-15th November 2019 - Rome, titled “Flow Cytometry Applied to Plant Reproductive Biology” as part of the “PLANT AND ANIMAL CYTOMETRY AND BIOTECH” session.
(https://www.enea.it/it/seguici/events/exoflowmetry_13-15nov2019/exoflowmetry-2019)
4. **Scariolo F.**, Basso A., Negrisolo E., Barcaccia G. Venetian Radicchio biotypes, who came first? Poster presentation at SIGA LXIV congress (September 2021)

