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:
SCIENZE DELLE PRODUZIONI VEGETALI

CICLO: XXXII

Breeding new varieties of horticultural species by means of conventional genetics and biotechnological methods

Tesi redatta con il contributo finanziario di Blumen Group SpA (T&T Agricola)

Coordinatore: Ch.mo Prof. Sergio Casella

Supervisore: Ch.mo Prof. Gianni Barcaccia

Co-Supervisore: Ch.mo Dr Fabio Palumbo

Dottorando: Alice Patella

Index

Riassunto generale……1

General abstract……3

Chapter I……5

Molecular determination of hybridity and homozygosity estimates in breeding populations of lettuce (*Lactuca sativa* L.)

Chapter II……29

Genetic Structure of Cultivated Varieties of Radicchio (*Cichorium intybus* L.): A Comparison between F1 Hybrids and Synthetics

Chapter III……64

Assessing the genetic distinctiveness of endive (*Chicorium endivia*) experimental materials using SSR and SNP markers

Riassunto generale

La crescente domanda di nuove varietà sta spingendo le ditte sementiere a pianificare programmi di miglioramento genetico sempre più rapidi ed efficienti. Pertanto, i metodi convenzionali di costituzione di nuove varietà sono sempre più supportati da quelli biotecnologici. In particolare, la costituzione di nuove varietà assistita da marcatori (marker assisted breeding, MAB) riduce il tempo necessario per sviluppare nuove varietà grazie all'utilizzo di saggi molecolari. Nei seguenti tre lavori è discusso il ruolo cruciale svolto dai marcatori molecolari in questi programmi di miglioramento genetico, che sono stati condotti in diverse specie orticole appartenenti alla famiglia delle Asteraceae. Nel primo caso studio, i metodi convenzionali sono integrati dai metodi molecolari attraverso un tipico schema di miglioramento genetico in lattuga (*Lactuca sativa* L.). Sono stati impiegati marcatori SSR (*Simple Sequence Repeats*) o microsatelliti, 16 loci in totale, al fine di caratterizzare geneticamente 71 putative linee parentali e di pianificare 89 incroci, i quali sono stati progettati per massimizzare la resa delle impollinazioni manuali e il potenziale di ricombinazione genetica. Successivamente, la progenie risultante, costituita da 871 piante, è stata selezionata e i profili molecolari dei campioni sono stati confrontati con quelli dei loro putativi parentali. Il tasso di successo di questi incroci è risultato in media pari a 68 ± 33 % e il numero di ibridi F1 è risultato pari a 602. Infine, in una fase avanzata di questo programma genetico, sono state valutate 47 popolazioni sperimentali (generazione F3) in termini di omozigosi osservata e di similarità genetica. Tre di queste popolazioni sono risultate particolarmente idonee per accedere ai test pre-commerciali. In particolare questo materiale sperimentale presentava un elevato grado di omozigosi, superiore al 90 %, e gradi di similarità intra-popolazione superiori al 95 %. In conclusione, questo studio evidenzia l'effetto sinergico dei metodi convenzionali e biotecnologici in diverse fasi di un programma di miglioramento genetico. Oltre alla lattuga, sono state confrontate e caratterizzate molecolarmente diverse varietà di *Cichorium intybus* var. *foliosum* L., sia per il loro valore economico che per il grande interesse culturale di questa specie in Veneto. *C. intybus* var. *foliosum*, meglio conosciuto in Italia come radicchio, è un importante ortaggio a foglia coltivato localmente che si riproduce prevalentemente per allogamia. Le aziende sementiere si avvalgono ampiamente di marcatori molecolari per la costituzione di nuove varietà, soprattutto di ibridi F1 che si distinguono per la loro elevata uniformità e le rese colturali. In questo secondo caso studio abbiamo eseguito una caratterizzazione e un'indagine molecolare sulla struttura genetica di diverse popolazioni commerciali di Radicchio. Sono stati impiegati 29 marcatori microsatellite per la genotipizzazione di 504 campioni del biotipo rosso di Chioggia. In particolare, sono state inizialmente caratterizzate e confrontate in termini di similarità genetica, ricorrendo anche all'uso di statistiche sulla diversità genetica: due sintetiche, quattro ibridi F1 e due popolazioni F2 derivate. Utilizzando il medesimo saggio molecolare sono state valutate anche l'uniformità e la stabilità di tre lotti di tre anni di produzione di una varietà F1. I gradi medi di similarità risultanti hanno consentito la chiara

discriminazione molecolare delle sintetiche OP dalle varietà F1 e dalla loro progenie F2, oltre che la determinazione delle singole popolazioni. Inoltre, la struttura genetica degli ibridi F1 prodotti in 3 anni ha rivelato inaspettatamente due gruppi principali che discriminano i primi due anni dal terzo, principalmente a causa della presenza di alleli specifici non comuni e di diverse frequenze alleliche. Nel complesso, queste informazioni molecolari consentiranno ai costitutori di determinare la distinzione genetica, l'uniformità e la stabilità di popolazioni commerciali e sperimentali, nonché le loro relazioni genetiche. A differenza delle due precedenti specie più rilevanti e più studiate (lattuga e cicoria), la quantità di dati biologici e molecolari disponibili per l'indivia (*Cichorium endivia* Lam.) è incredibilmente minore. Questa specie è un'orticola a foglia verde con un sistema riproduttivo di tipo autogamo. L'indivia appartenente alla famiglia delle Asteraceae ed è caratterizzata da due tipi di cultivar: indivia riccia (*C. endivia* var. c*rispum* Lam.) e indivia scarola o liscia (*C. endivia* var. *latifolium* Lam.). Per le ditte sementiere la caratterizzazione genetica del loro materiale è fondamentale sia al momento della registrazione che in seguito per la protezione delle loro varietà. Nel nostro caso di studio, abbiamo verificato il carattere distintivo di 32 materiali sperimentali appartenenti ai due tipi di cultivar (indivia scarola e riccia). In un primo tentativo basato su marcatori SSR, solo 14 loci marcatori su 29 appartenenti a *C. intybus* sono stati trasferiti con successo in indivia e solo 8 di essi sono risultati polimorfici. A causa del livello limitato di discriminazione di questo saggio molecolare, è stato applicato un approccio alternativo basato su marcatori SNP (*Single Nucleotide Polymorphisms*) utilizzando il sequenziamento del DNA associato al sito di restrizione (RADseq). Complessivamente 4.621 marcatori SNP sono stati in grado di separare tutti gli individui di indivia riccia e scarola, in particolare, 50 di loro hanno discriminato i due tipi di cultivar. Inoltre, il dendrogramma e l'analisi PCoA, supportati dagli SNP, hanno diviso i due tipi cultivar di indivia in due sottogruppi distinti, discriminando in modo univoco tutti i materiali vegetali. Nel complesso, la nostra ricerca è stata in grado di valutare la distinguibilità di tutti i campioni analizzati, che rappresenta il primo requisito del test DUS. É stato valutato, inoltre, anche il grado di omozigosi, in modo da prevedere l'uniformità della progenie, ovvero il secondo requisito del DUS test, in quanto gli individui con un elevato grado di omozigosi sono noti per produrre popolazioni maggiormente uniformi. In conclusione, queste informazioni molecolari hanno permesso i costitutori di determinare la distinguibilità genetica dei 32 materiali d'élite e ogni possibile loro relazione genetica.

General abstract

The growing demand for new vegetable varieties is forcing seed companies to plan faster and more efficient breeding programs. Therefore, conventional breeding methods are increasingly supported by biotechnological methods. In particular, marker-assisted breeding (MAB) reduces the time needed to develop new varieties thanks to the use of molecular assays. The crucial role played by molecular markers in breeding new varieties is discussed in the three works carried out in different horticultural species belonging to Asteraceae family. In a first case study, conventional breeding methods are complemented by molecular methods throughout a typical breeding program in lettuce (*Lactuca sativa* L.). A total of 16 SSR (Simple Sequence Repeats) markers or microsatellites were efficiently used to genetically characterise 71 putative parental lines and to plan 89 crossings, designed to maximise the yield deriving from each manual pollination. After that, the resulting 871 progeny plants were screened, and their molecular profiles were compared with those expected considering their putative parents. The out-pollination success rate resulted in being, on average, 68 ± 33 % and 602 F1 hybrids were identified. Finally, in an advanced step of this breeding program, 47 different experimental populations (F3 generation) were evaluated in terms of observed homozygosity and genetic similarity within the population. Three of them resulted particularly suitable for pre-commercial trials due to observed median homozygosity above 90 % and an intra-genetic similarity value always higher than 95%. Hence, this study shows the synergetic effect of conventional and biotechnological methods in different steps of a breeding program. In addition to lettuce, different *Cichorium intybus* var. *foliosum* L. varieties were compared and molecularly characterised, both for the economic value and the great cultural interest of this species in Veneto region. *C. intybus* var. *foliosum*, better known in Italy as radicchio, is an important locally cultivated leafy vegetable that prevalently reproduces by allogamy. Considering that marker-assisted breeding is widely used by seed firms to develop new F1 hybrid varieties that are distinguished by high plant uniformity and crop yields, in this second case study we performed the molecular characterisation and the genetic structure investigation of different commercial populations of Radicchio. A total of 29 mapped microsatellite markers were used for genotyping 504 samples of the Red of Chioggia biotype. Two synthetics, four F1 hybrids and two derived F2 populations were initially characterised and compared in terms of genetic similarity and diversity statistics. Then, the uniformity and the stability of three years of production of an F1 variety were investigated by applying the same panel of molecular markers. As main finding, the mean similarity estimates enabled the clear molecular discrimination of OP synthetics from F1 varieties and their F2 progenies and the determination of individual plant memberships. Moreover, the genetic structure of F1 hybrids produced in 3 years unexpectedly revealed two main clusters that discriminate the first 2 years from the 3rd, mainly because of the presence of uncommon specific alleles and different allele frequencies. Overall, this molecular information will enable breeders to determine the genetic distinctness, uniformity

and stability of commercial and experimental populations, as well as their genetic relationships and relatedness. Unlike the previous two more relevant and more studied species (lettuce and chicory), the amount of biological and molecular data available for endive (*Cichorium endivia* Lam.) is incredibly scarce. This species is a self-pollinated leafy green vegetable, belonging to the Asteraceae family and it is characterized by two cultivar types: curly endive (*C. endivia* var. *crispum* Lam.) and escarole or smooth endive (*C. endivia* var. *latifolium* Lam.). For seed firms the genetic diversity characterisation of elite breeding material is crucial for the registration and protection of future varieties. In our case study, we verified the distinctiveness of 32 experimental materials belonging to the two cultivar types (escarole and curly endive). In a first SSR-based attempt, only 14 out of 29 microsatellite markers belonging to *C. intybus*, were successfully transferred to endive and only 8 of them resulted polymorphic. Due to the limited level of discrimination of this molecular assay, an alternative approach based on single nucleotide polymorphisms (SNP) was applied using the restriction site-associated DNA sequencing (RADseq). Overall, a set of 4,621 SNP markers was able to separate curly endive from escarole endive and, in particular 50 of them were able to discriminate the two cultivar types. Moreover, the resulting SNP-supported dendrogram and the PCoA analysis divided the two cultivar types of endive into two distinct clusters, discriminating univocally all the plant materials. Overall, our research was able to evaluate the distinctiveness requirement of the DUS testing. We also evaluated the observed homozygosity to predict the uniformity and stability of progenies, two additional requirements of the DUS testing: individuals with the highest homozygosity are known to produce more uniform and stable populations over generations. In conclusion, this molecular information enabled breeders to determine the genetic distinctness of the 32 elite breeding materials and to reconstruct their genetic relationships.

Chapter I

Molecular determination of hybridity and homozygosity estimates in breeding populations of lettuce (*Lactuca sativa* L.)

Alice Patella, Fabio Palumbo, Giulio Galla and Gianni Barcaccia

Published to *Genes*

Keywords: Pure lines, F1 hybrids, microsatellite markers, marker-assisted breeding, crop improvement, varieties.

Abstract:

The development of new varieties of horticultural crops benefits from the integration of conventional and molecular marker-assisted breeding schemes in order to combine phenotyping and genotyping information. In this study, a selected panel of 16 microsatellite markers were used in different steps of a breeding programme of lettuce (*Lactuca sativa* L., 2n=18). Molecular markers were first used to genotype 71 putative parental lines and to plan 89 controlled crosses designed to maximise recombination potentials. The resulting 871 progeny plants were then molecularly screened and their marker allele profiles compared with the profiles expected based on the parental lines. The average cross-pollination success rate was 68 ± 33 %, so 602 F1 hybrids were completely identified. Unexpected genotypes were detected in 5 % of cases, consistent with this species' spontaneous out-pollination rate. Finally, in a later step of the breeding programme, 47 different F3 progenies, selected by phenotyping for a number of morphological descriptors, were characterised in terms of observed homozygosity and within-population genetic uniformity and stability. Ten of these populations had a median homozygosity above 90% and a median genetic similarity above 95 %, and are therefore particularly suitable for pre-commercial trials. In conclusion, this study shows the synergistic effects and advantages of conventional and molecular methods of selection applied in different steps of a breeding programme aimed at developing new varieties of lettuce.

1 Introduction

Lettuce (*Lactuca sativa* L.) is a self-pollinating leafy vegetable species (2 n = 2 x = 18) of the Asteraceae family. It is cultivated on a large scale throughout the world for consumption as a fresh vegetable on its own or in combination with other ready-to-eat vegetable products [1]. Its growing economic importance has led seed companies to regularly develop new varieties with ever higher agronomic traits. However, breeding programmes are highly limited by the reproductive system of this species. The flower structure of lettuce determines a reproductive strategy known as cleistogamy, in which anther dehiscence and subsequent pollination take place before flower opening, resulting in a very high rate of self-pollination, very often equal to or close to 100% [2]. According to recent estimates, out-cross rates are limited to 1%–6% [3]. These reproductive barriers mean that in natural conditions the species spontaneously constitutes pure lines, characterised by phenotypic uniformity and genotypic stability, due to their very high homozygosity. In conventional breeding programmes, developing segregating and recombinant F2 populations traditionally requires crosses to be hand pollinated while self-pollination is prevented by emasculating the flowers. The most popular emasculation and hand-pollination technique is that described by Olivier [4]. Known as the "wash method", it involves hand-spraying the inflorescence with water during pistil emergence to remove the pollen attached to the female part of the flower. The inflorescence is then left to dry for a short period, after which it is rubbed with a ripe flower of the pollinating variety [5]. A slightly

different, but also widely used, technique is the "clip-and-wash method", which involves clipping the tip of the corolla before spraying with water. This guarantees more efficient pollen removal and cross rates close to 100% from the subsequent manual pollination [2]. However, these breeding techniques are time-consuming and technically highly demanding, and are only really effective if coupled with molecular analyses aimed at screening progeny plants and assessing their hybridity.

In recent years, many seed firms have begun using molecular markers to carry out assisted selection schemes and to speed up varietal development programmes [3]. Simple Sequence Repeat (SSR) markers are, so far, the most commonly used markers for these purposes [6–8] as they are codominant, have high reproducibility and multi-allelism, and can be detected at any stage of plant development, without being influenced by the environment [9]. There are a considerable number of SSR markers for lettuce in the literature [10]. Truco, et al. [11] produced an integrated genome map from 7 different linkage maps, which included 130 SSR loci organised in 9 linkage groups. Rauscher and Simko [10] augmented this genomic map with 54 genomic SSR and 52 EST-SSR (Expressed Sequence Tag) loci. Finally, with the publication of the *L. sativa* genome draft [12], tens of thousands of new SSR regions have become available for testing and use.

Given the availability of markers in lettuce, Marker-Assisted Selection (MAS) has started to be adopted in plant breeding programmes for various purposes, including identification of resistance genes [13,14] or Quantitative Trait Loci (QTLs) of phytopathogens [15,16], the study of QTLs controlling complex traits [17,18], and investigation of the genetic identity and purity of either experimental or commercial lines [19]. On the other hand, very few attempts have been made to prove the efficiency of molecular markers in Marker-Assisted Breeding (MAB) activities, where the genotypic background is molecularly investigated to complement traditional phenotypic selection [20].

In this work, SSR markers were used in three different steps of a conventional breeding scheme aimed at developing new varieties characterised by distinctiveness, uniformity, and stability (Figure 1).

We first examined the genetic background of a number of superior pure lines in order to plan experimental matings to produce F1 hybrids and then derived F2 progenies manifesting morphological variability as a result of genetic segregation and recombination (Figure 1). Each offspring in the F1 generation was analysed to distinguish the individuals resulting from planned out-crosses from those resulting from accidental selfing (Figure 1). After genotyping, the S1 individuals were discarded, and the F1 individuals were self-pollinated. In the F3 generations (Figure 1); experimental populations, previously selected according to their morphological traits, were also characterised by molecular markers due to the need to assess their stability and uniformity in order to run pre-commercial trials.

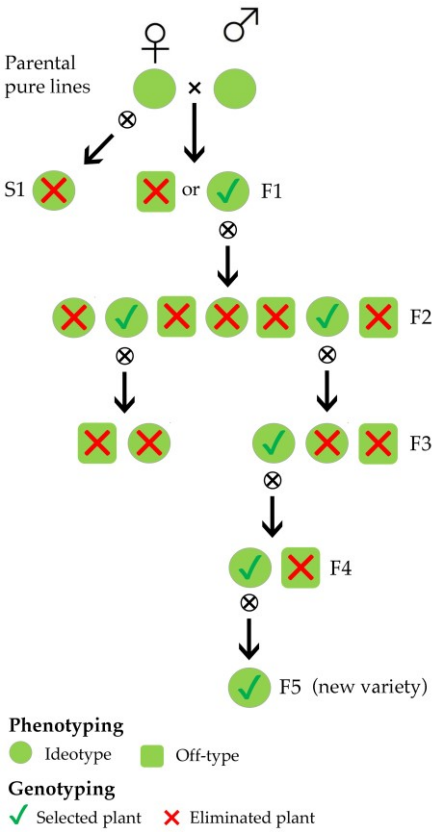


Figure 1. Simplified overview of a lettuce breeding scheme in which selection is based on both plant phenotyping and genotyping.

2 Materials and Methods

2.1 Plant materials and breeding techniques

Plant materials were developed and provided by Blumen Group SpA, Italy and belonged to five different lettuce cultivar types (Table S1). Seventy-one parental lines (germplasm composed of experimental, pre-commercial and commercial lines) were involved in 89 combinations of crosses, in which each progeny consisted of 6 – 12 individuals (871 progeny samples). Parental lines were grown in the spring of 2015, and the 89 programmed crosses were carried out in the summer using the clip-and-wash method [2]. This involved making an incision in the calyx and corolla and washing the anthers in the early morning before the pollen grains could settle naturally on the outermost stigmatic surface of pistils. The plants were then manually pollinated by rubbing anthers of the pollen donor on the stigma of the seed parent. For each planned cross, a bulk of 4/5 flowers from a pollinator parent was used to pollinate as many flowers of a seed plant. Seeds were collected from the seed plant and sown in early autumn for genotyping selection and agronomic evaluation (spring 2016).

Finally, to assess the uniformity of the 47 experimental lines, previously chosen for morpho-phenological traits and pathogen resistances, 940 samples belonging to the 47 F3 populations (labelled 1 to 47) were collected in the spring of 2018. Each experimental line comprised 20 individuals.

2.2 DNA isolation

A total of 100 mg of fresh leaves was collected from each of the 1882 lettuce samples (71 parents, 871 progeny and 940 F3) and ground to a fine powder using Tissue Lyser II (Qiagen, Valencia, CA, USA). Genomic DNA (gDNA) was extracted with the Dneasy® 96 Plant Kit (Qiagen), according to the manufacturer's protocols. After extraction, the integrity of the gDNA was assessed by electrophoresis on 1 % (*w*/*v*) agarose gel stained with SYBR Safe® 1 × DNA Gel Stain (Life Technologies, Carlsbad, CA, USA) in Tris-Acetate-EDTA (TAE) running buffer. Both the yield and purity of the extracted gDNA samples were evaluated using a NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific, Carlsbad, CA, USA). Following DNA quantification, all DNA samples were diluted to a final concentration of 20 ng/μL.

2.3 Primer design and testing of SSR marker amplification

Sixteen SSR marker loci were selected from those available in the scientific literature [10,21], according to i) chromosomal location; ii) polymorphism rate, expressed as PIC (Polymorphism Information Content); iii) allele size range; iv) annealing temperature of the locus-specific primers. Amplifications were performed according to the method previously described by Schuelke [22], with some minor modifications. Briefly, three primers were used to amplify each microsatellite locus: a pair of locus-specific primers, one with an oligonucleotide tail at the 5′ end (M13, PAN-1, PAN-2 or PAN-3, Table S2), and a third universal primer complementary to the tail and labelled with a fluorescent dye (6-FAM, VIC, NED, or PET). Primer pair efficiency was tested in silico using the PRaTo [23] web-tool and were organised in three multiplex reactions, as shown in Table 1.

Table 1. Microsatellite loci information. For each primer pair, the original simple sequence repeat (SSR) name, ID used in this work, linkage group [10,21], SSR motif, primer sequences (PAN1, PAN2, PAN3, or M13 tails at the 5' end are indicated in square brackets; for further details see Table S2), dye and the multiplex to which the SSR marker locus belongs is shown.

Marker name	ID	LG	Motif		Primer Sequence	Dye	Multiplex
LSSA27-2 [10]	Lsat1	1	$(AC)_7$	For	[M13]CACACTACCACCCAACACG	6-FAM	1
				Rev	ACCCTCTTCGCTTCTTCTT		
SML-045 [21]	Lsat2	2	$(AAG)_{9/12}$	For	ACAAAACCGTTTCACCCAAA	6-FAM	1
				Rev	[M13]AGCCCTGTCCTCTTCAGGAT		
LSSB54 [10]	Lsat3	8	$(GT)_{10}$	For	[PAN1]CTTGAGAGTGCTTGGAGAGGAT	VIC	1
				Rev	CACATACAACAAGACAAGTCCCA		
LSSA05 [10]	Lsat4	8	$(TC)_{18}$	For	AGAACAACGGTAGCTTGTTAAATTG	VIC	1
				Rev	[PAN1]ATCGTCGGTTAATCTTCGTCG		
LSSA04 [10]	Lsat5	4	$(TC)_{14}$	For	[PAN2]AAGGAAAGGAAGGGTTGACTTGT	NED	1
				Rev	TTGGTGAAGAAAAGAGAGAGTTT		
LSSA11 [10]	Lsat6	3	$(CT)_{20}$	For	[PAN2]ACTCCCACTATCCTCTTTGCAT	NED	1
				Rev	GCCCACATTCTTAATCTTGTCC		
LSSA14 [10]	Lsat7	9	$(AG)_{18}$	For	[PAN3]TGATGACTCCAGTCTTAGATACCA	PET	1
				Rev	AGTCCCCGACTATCAGTCTCA		
LSSB09 [10]	Lsat8	2	$(TG)_8$	For	AGAATGAGAAGGATGAAATGGCTG	6-FAM	2
				Rev	[M13]AAACACCTTTAGCATCAAAATACCC		
SML-029 [21]	Lsat9	9	$(GAG)_{7/8}$	For	[M13]AGCCCAGAAGAGCGTGATTA	6-FAM	2

Marker name	ID	LG	Motif		Primer Sequence	Dye	Multiplex
				Rev	TGCAGGGCTCCTTGATCTAC		
LSSB17-1 [10]	Lsat10	7	$(GT)_{11}$	For	ACTAGGGCTCTAATACAACTTGT	VIC	2
				Rev	[PAN1]TTGGCTTACAGTTATGGATTAAATG		
LSSA17 [21]	Lsat11	3	$(AG)_{21}$	For	[PAN1]AATGTGCGTGAGAGTTTCCTTT	VIC	2
				Rev	CAAGAAGGCAGTGATGAAGTTG		
LSSA12 [10]	Lsat12	5	$(GT)_{11}$	For	[PAN2]ACAAGGCCCAATCCTTTTCT	NED	2
				Rev	TCGAAAATTTGGAGAGAGTTTCTT		
LSSA15 [10]	Lsat13	1	$(AC)_{11}$	For	GCCCAACCCAAGAAGAGGAG	PET	2
				Rev	[PAN3]TGGAGAGGAGTGGAGAGTGTT		
LSSA28-1 [10]	Lsat14	4	$(GA)_{28}$	For	TTCATCTCTCTCCTCCTTCAGC	6-FAM	3
				Rev	[M13]ATCCCCATTGTCCTCCC		
LSSA21-1 [10]	Lsat15	8	$(TC)_{19}$	For	[PAN2]TTGTACCCAGTTGTCCAAACAG	NED	3
				Rev	CAGATTGTTGCAGATTTCTTCG		
LSSB68 [10]	Lsat16	na	$(CT)_{20}$	For	GTCTGTGTGGTTTTGGT	PET	3
				Rev	[PAN3]TGTGGTGGAGTGTGATTT		

The 16 primer pairs were first tested individually (singleplex reactions) using three randomly chosen lettuce gDNA to evaluate primer efficiency and to check the correspondence between expected and actual size of the bands; they were then evaluated in multiplex PCRs to assess possible primer interactions.

All amplification reactions (both singleplex and multiplex) were performed in a 10 μL reaction volume containing 1× Platinum® Multiplex PCR Master Mix (Thermo Scientific), 10% GC Enhancer (Thermo Scientific), 0.25 μM of non-tailed primer, 0.75 μM of tailed primer, 0.50 μM of fluorophore-labelled primer (universal primer) and 20 ng of genomic DNA. Thermal cycling conditions were as follows for multiplex 1 and 2: 94 °C for 5 minutes followed by 6 cycles of 94 °C for 30 seconds, 61 °C for 30 seconds, 72 °C for 45 seconds, with a 1 °C annealing temperature stepdown per cycle (from 61 °C to 56 °C). The annealing temperature for the following 35 cycles was set at 56 °C, with denaturation and extension phases as above and a final extension at 60 °C for 30 minutes. The multiplex 3 thermal cycling conditions were instead: 94 °C for 5 minutes followed by 6 cycles of 94 °C for 30 seconds, 56 °C for 30 seconds, 72 °C for 45 seconds, with a 1 °C annealing temperature stepdown per cycle (from 56 °C to 51 °C). The annealing temperature for the following 35 cycles was set at 51 °C with denaturation and extension phases as above and a final extension at 60 °C for 30 minutes. All amplifications were performed in a GeneAmp® PCR 9700 thermal cycler (Applied Biosystems, Carlsbad, CA, USA). PCR products were first checked on gel electrophoresis (2 % Ultrapure™ Agarose in TAE 1×, SYBR Safe® 1×, Life Technologies) then run on capillary electrophoresis with ABI 3730 DNA Analyzer (Applied Biosystem), using LIZ500 as the molecular weight standard. The size of each peak was determined with the Peak Scanner 1.0 software (Applied Biosystems).

2.4 Genotyping and data analysis

The 71 potential parents were genotyped at 16 SSR loci and statistical analyses were performed using NTSYS (Numerical Taxonomy and Multivariate Analysis System) version 2.2 (Exeter Software) [24]. Rohlf's

(or the simple matching) coefficient was used to calculate pairwise genetic similarity (GS) in all possible comparisons and to construct a genetic similarity matrix, according to the formula:

$$GS_{ij} = m / (m + n) \quad (1)$$

where "i" and "j" are two different individuals, while "m" and "n" represent the number of matching and non-matching attributes, respectively. An unweighted pair group method with an arithmetic mean (UPGMA) dendrogram and a Principal Coordinates Analysis (PCoA) of parental lines were carried out using the Jaccard coefficient in the PAST software v. 3.14 with 10,000 bootstrap repetitions [25]. The genetic structure of the lines was modelled using a Bayesian clustering algorithm implemented in STRUCTURE v. 2.2 [26]. Since no *a priori* knowledge of the origin of the populations under study was available, the admixture model and then the correlated allele frequencies model were used. Ten replicate simulations were conducted for each value of K, with the number of founding groups ranging from 1 to 8, using a burn-in of 200,000 and 1,000,000 iterations. The most likely K Estimates were determined using the method described by Evanno et al. [26]. Estimates of membership were plotted as a histogram in an Excel spreadsheet. Finally, observed homozygosity (Ho) was determined with the POPGENE software [27].

The 89 subsequent crosses were planned according to the following criteria: i) high genetic dissimilarity values among parents within the same lettuce cultivar type and between them, ii) availability of informative loci able to distinguish between the resulting offspring and individuals resulting from accidental self-pollitated. Only homozygous loci for different alleles were considered informative, whereas heterozygous loci were taken into account only if the origin of the parental alleles could be clearly discerned in the progenies. The resulting offspring (871 samples) were then screened, with the analysis restricted to those SSR loci which had previously proven to be informative for hybrid detection. This made it possible to determine whether individuals belonging to a given F1 population originated from cross-pollination or self-pollination. The successful crosses (S.C.) rate of 89 was calculated as follows:

$$S.C. = (No\ F1 \times 100) / (No\ Tot - No\ U.G.) \quad (2)$$

where "No F1" is the number of hybrid individuals, "No Tot" is the number of all individuals in a progeny population (No tot = No F1 + No U.G. + No SP) and "No U.G." is the number of unexpected genotypes deriving from unplanned crosses.

Finally, 940 samples from the 47 F3 populations were genotyped using the previously-described panel of SSR markers. The POPGENE software [27] was used to compute the mean values of observed homozygosity for each population (3), where n is the total number of samples). In addition, the median of genetic similarity between the 47 lines was calculated using Rohlf's coefficient, which was designed for codominant molecular markers [28,29]. Comparison of genetic similarity among ten selected populations

was instead calculated using the Jaccard coefficient, in accordance with the literature [30]. Genetic similarity matrices were generated in the NTSYS software [24].

$$\bar{\mathrm{H}}\mathrm{o} = \sum_n \mathrm{Ho}/\mathrm{n} \quad (3)$$

3 Results

3.1 Parental lines

The 16 SSR markers, organised in three multi-locus PCRs, were used firstly to amplify and score the 71 parental lines. Fourteen of the 16 SSR markers proved to be polymorphic among plant accessions. The similarity matrix constructed using Rohlf's coefficient revealed genetic similarity values ranging from 53 % to 100 % (Figure S1). The resulting unweighted pair group method with an arithmetic mean (UPGMA) dendrogram showed the samples clustering into two main sub-groups. Eighteen parental lines were not fully distinguishable, while the remaining 53 had unique genotypic profiles. The first principal coordinate from the PCoA accounted for 22 % of the total variation and divided the samples into two groups, analogous to the clustering in the tree. The second principal coordinate accounted for 12 % of the total variation. These results were confirmed by investigation of the genetic structure of the 71 parental lettuce lines based on allele frequencies; the best estimate of population size was $K = 2$ (Figure S2), such that the samples were grouped into two genetically distinct clusters (Figure 2). The lettuce cultivar types were reported in Table S1, but they did not correspond to different clusters in the UPGMA tree. The mean observed homozygosity was 82 %, with a minimum value of 69% and a maximum of 100 %. It is worth noting that 19 of the 71 parental lines (27 %) had observed homozygosity values greater than 90 %, while 30 of the 71 (42 %) had a medium-high observed homozygosity (Ho) between 81 % and 90 %. Fourteen of the 71 parental lines (20 %) had observed homozygosity ranging from 71 % to 80 %, and only 8 individuals had values lower than 70 % (Figure 3a and Figure S1).

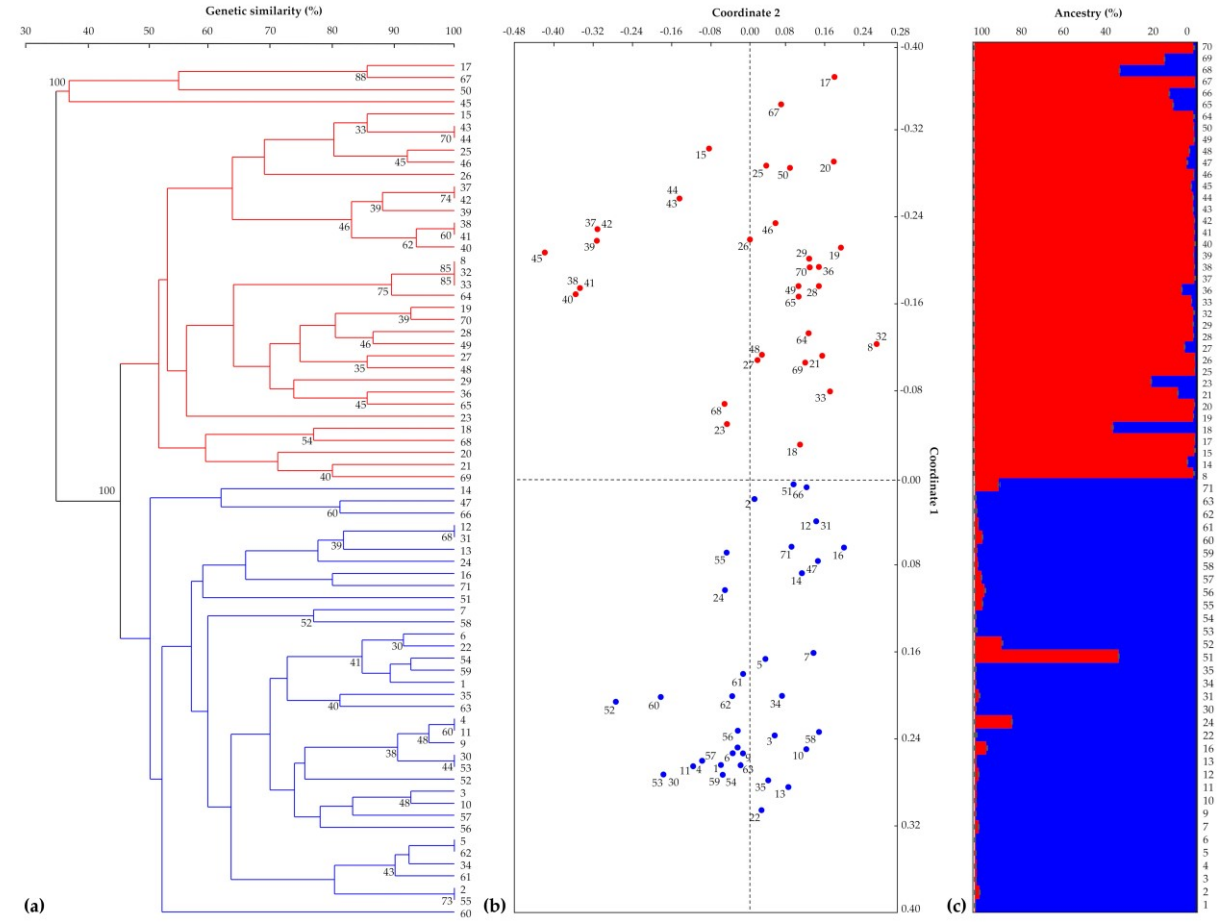


Figure 2. (a) Genetic similarity-based unweighted pair group method with an arithmetic mean (UPGMA) dendrogram of 71 parental lines calculated using the Jaccard coefficient. Bootstrap estimates ≥30% are reported next to the nodes (red and blue branches indicate the two clusters identified). **(b)** Principal coordinate analysis (PCoA). The 71 samples are shown in red or blue according to the clustering shown in the UPGMA tree. **(c)** The population genetic structure of the 71 lines as estimated by STRUCTURE. Each sample is represented by a vertical histogram partitioned into $K = 2$ coloured segments (red or blue, in accordance with (**a**) and (**b**)) representing the estimated membership. The proportion of subgroup membership (%) is reported on the ordinate axis, and the identification number of each accession is reported below each histogram.

3.2 Determination of hybridity

Using a combination of genotypic and phenotypic data, 89 cross combinations were planned (Table S3). Before proceeding, we also checked the availability of informative loci able to distinguish between offspring resulting from out-cross and those obtained by accidental self-pollination. Screening identified 1 discriminant locus in 16 % of cases, 2 informative loci in 36 % of cases, and 3 to 7 informative molecular markers in 48 % of the crosses (Figure 3b). The three most informative loci were Lsat3, Lsat7, and Lsat6, while Lsat4 and Lsat13 were monomorphic in almost all parental groups. It is worth noting that the Lsat8 marker was in a heterozygous state in all but four parental genotypes (7, 45, 58, and 60).

We were able to take advantage of these informative loci to calculate the success rate of each cross. In 30% of manual pollinations (27 out of 89), a success rate of 100 % was achieved (i.e., all the offspring were hybrids), whereas in 18 % of crosses (16 out of 89), the S.C. ranged from 71 % to 90 %. A hybridity rate fluctuating between 51 % and 70 % was reported in 15 % of cases (13 out of 89), while 26 % of the crosses

produced fewer than 50 % hybrids each. Finally, in only 7 % of crosses (6 out of 89) were all the offspring the result of self-pollination (Figure 3c and Table S3). Overall, the mean hybridization rate (the average number of hybrids per crosses) was 68 ± 33 %, and out of a total of 871 individuals, 602 (69 %) were hybrids, and 556 were derived from programmed crosses. The remaining 46 individuals (5 % of the total) had a unexpected genotypes (U.G.) compared with their putative parents (Table S3).

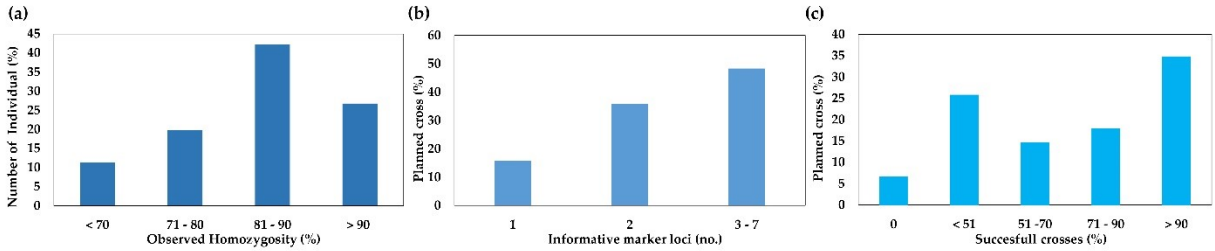


Figure 3. (a) Observed homozygosity of 71 lettuce parental individuals belonging to as many pure lines. **(b)** Histogram of discriminating loci in 89 cross combinations (in percentages). **(c)** Histogram of the percentages of pollination success in 89 programmed lettuce crosses.

3.3 Lettuce breeding populations

The 47 F3 experimental lines were genotyped using the same set of 16 SSR loci as for the previous analyses. The homozygosity estimates of all samples (940) ranged from 67 % to 93 % (Figure 4a). Ten experimental populations had a median observed homozygosity ≥90 %. Outliers—with homozygosity values consistently deviating from the median—were present in only three experimental populations (11, 14, and 32).

The median genetic similarity observed within each line was always greater than 90% (Figure 4b), and 37 experimental populations had a median genetic similarity ≥95 %. Outliers were present in 21 of the 47 lines (Figure 4b).

After assembling the data, we found 10 breeding populations, belonging to butterhead type (Table S4), to have Ho values ≥ 90 %, and a median genetic similarity ≥ 95 %; the box-plots of these populations are labelled in red in Figure 5a,b. Finally, in the genetic similarity matrix calculated from all pairwise comparisons of these ten populations, the Jaccard's index ranged from 44 % ± 3 % (between populations 3 and 18) to 96 ± 5 % (between populations 45 and 47, Figure S3). Moreover, the populations called 45 and 47 were constituted starting from the same parents (2 × 6, Table S4).

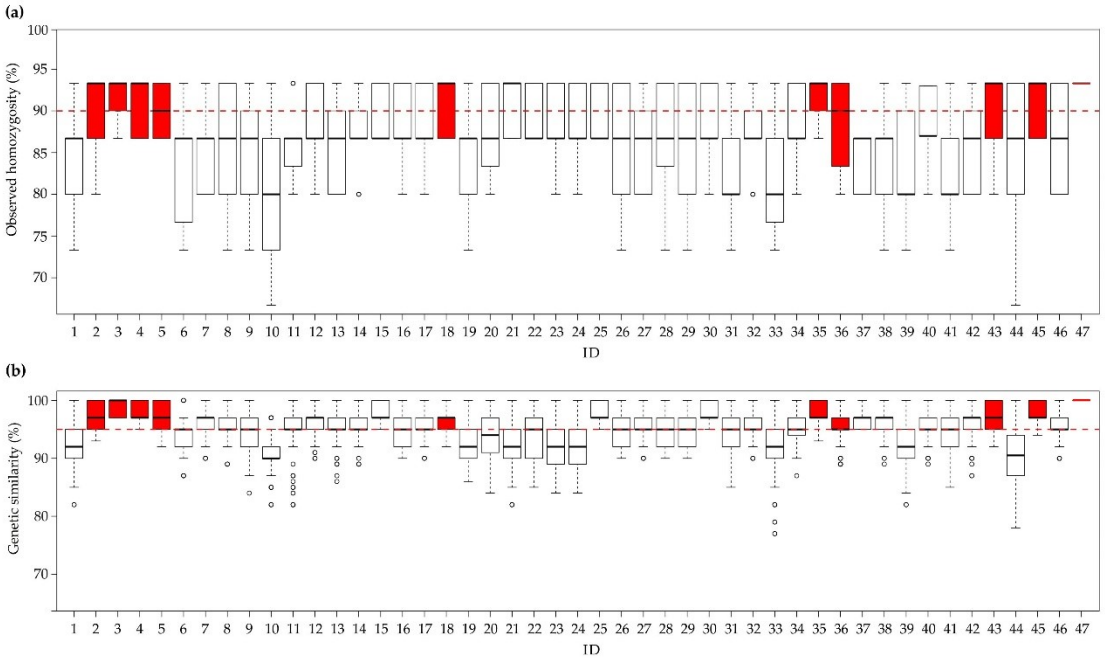


Figure 4. Statistics relating to the observed homozygosity and genetic similarity among lines. (**a**) Box-plot of the median observed homozygosity (in percentages) in each of the 47 populations. The red dotted line represents the homozygosity threshold set at 90 %. (**b**) Box plot of the median genetic similarity in each experimental population (in percentages). The red dotted line represents the genetic similarity threshold set at 95 %. The red box-plots represent the ten best experimental populations (observed homozygosity ≥ 90 % and genetic similarity values ≥ 95 %). The second and third quartiles are marked inside the square and are divided by a bold bar (median). Dots show outlier samples.

4 Discussion

The last decade has seen major advances in the acquisition of knowledge concerning the genetics of lettuce and, in particular, the development of molecular markers [1,11,21]. This has facilitated marker-assisted selection programmes, especially those aimed at countering the onset of new diseases. For example, several studies have dealt with identifying the QTLs associated with biotic and abiotic stress resistance [17,31,32]. Molecular markers have also been extensively used to assess genetic variation and relationships in lettuce germplasm [19] and to identify possible duplicate varieties [33]. However, although the benefits derived from exploitation of these molecular tools have also been discussed in marker-assisted breeding programmes [34] and demonstrated in several species [35,7], there are only a few studies on this type in lettuce [20]. The aim of our work, therefore, was to integrate conventional and biotechnological methods in three different steps of a breeding programme to show that this strategy is also effective in *L. sativa* (Figure 1). This is of pivotal importance if we consider the economic impact of lettuce (the world production of lettuce and chicory in 2017 was 26.8 million tons [36]) and the need to regularly develop new varieties.

Commercial lettuce varieties are usually characterised by pure lines due to the autogamous nature of this species. In order to introduce variability, manual pollination is usually carried out to cross genetically stable parent lines with agronomic traits of interest. Progeny selection is a crucial step, but despite the efficiency of some emasculation and hand pollination methods developed over the years [2], a major

problem—distinguishing unequivocally and rapidly F1 individuals from self-pollinated progeny—still remains. The use of molecular assays to quickly and accurately screen progeny makes it possible to overcome most of the conventional breeding limits in this species.

In this context, our SSR-based analysis has i) facilitated selection of the best parents to cross in order to maximise the variability of the progeny both within the same cultivar type and among them, ii) allowed accurate evaluation of the resulting offspring, and iii) sped up the screening of experimental F3 lines for their stability and uniformity.

The first part of our work focused on pre-screening 71 parental lines for crossing with the aim of maximising the gains obtained from each out-pollination within cultivar type and, in some cases, among them. As expected, the similarity matrix and the unweighted pair group method with an arithmetic mean (UPGMA) dendrogram showed varying levels of similarity among the different parental genotypes. Parental germplasm appeared to divide into two different groups, as revealed by the Principal Coordinates Analysis (PCoA) results and particularly by the genetic structure analysis. However, samples did not separate in UPGMA tree and PCoA according to the cultivar type, but we may assume that increasing the number of markers it could be possible to clarify this clustering. Although 53 parental lines were found to be fully distinguishable, with similarity values ranging from 53 % to 98 % and characterised by a unique genotypic profile, it was impossible to identify unequivocally the remaining 18. This is not surprising if we consider that some of the parental lines were closely related. We may speculate that increasing the number of SSRs would allow us to address these remaining issues. Given the aim of this study, these data were useful to avoid crosses between parents with 100 % similarity. To introduce variability according to the phenotype and the lowest similarity values, we carried out 89 crossing combinations. Another aspect that needs to be considered when planning crosses is the stability of the parental line in terms of homozygosity. In our study, the median observed homozygosity of the parental lines was lower than expected (82 %), especially in light of the strictly autogamous reproductive system of lettuce [37]. Overall, the fact that only one individual in four had homozygosity values greater than 90% showed that some of these lines were not entirely stable. However, it must be borne in mind that, although the observed homozygosity was not optimal, some of these lines, experimental lines, were chosen to produce F1 partly because they displayed resistance to multiple pathogens and had interesting phenotypic traits.

Before proceeding with hand pollination, in order to distinguish between F1 individuals resulting from cross-pollination and those resulting from self-pollination, we first examined the informative loci among the parental lines used in the crosses. Only homozygous loci for different alleles in parental lines were considered informative. Our analysis showed at least 2 informative loci in 84% of the programmed crosses. It is worth pointing out that restricting the analysis to the informative loci brought us considerable savings in terms of time and costs.

Overall, the molecular determination of hybridity was successful: F1 individuals represented at least 51 % of the offspring in 67 % of the manual crosses, and 100 % of the offspring in 30 % of the crosses (100 % success rate), in agreement with the estimates originally reported by the developer of the pollination technique [2]. Unexpected genotypes (U.G.) were identified in 5 % of the individual progeny. In these cases, the progeny genotypes appeared to diverge from what would be expected given the parents. This percentage is consistent with the spontaneous or undesired occurrences of cross-pollination (1 % – 6 %) reported in the literature for this species [3], mainly due to pollinator insects. However, we cannot exclude human error during manual pollination or seed collection.

Finally, at an advanced step of the breeding programme, we genetically assessed 47 different experimental F3 populations (940 samples), previously selected for their morpho-phenotypic traits and resistances of interest (Figure 1). Interestingly, the findings in terms of both homozygosity and intra-line similarity were very good. This would suggest that in strictly autogamous species, such as lettuce, three cycles of self-pollination may already be sufficient to reach desired outcomes in terms of genetic uniformity and homozygosity. This also confirms that the use of molecular markers could speed up the process by making it possible to select the best individuals on the basis of their genotype, thereby reducing the number of generations needed to develop new varieties. The ten experimental populations with the highest homozygosity estimates (≥ 90 %) and the highest intra-genetic similarity values (≥ 95 %) were considered suitable for pre-commercial trials (red box plot, Figure 4). However, a pairwise comparison of two of them (identified as 45 and 47) showed them to be genetically too similar (96 ± 5 % genetic similarity, Figure S3), in agreement with phenotypic data and their common origin (Table S4), to be registered and marketed as distinct varieties. According to the most recent guidelines concerning the protection of new plant varieties, the similarity threshold to define two lettuce varieties as distinct is set at 96 % [30]. The next step will be to integrate molecular data and morphological observations in order to the select the best genotypes (positive selection) for evaluation as pre-commercial varieties. In particular, the eligible genotypes will be self-pollinated to multiply the seed so that their agronomic performance can be compared in different locations and periods of the year, and with the best commercial varieties already on the market.

For the remaining experimental populations (white box blot, Figure 4), an attempt could be made to increase their genetic uniformity through negative selection to remove the most genetically divergent individuals (i.e., outlier samples). Moreover, if necessary, the remaining genotypes can undergo a further selfing cycle aimed at reaching optimum values of homozygosity.

5 Conclusions

The results of this study demonstrate the advantages of mutual integration of traditional and biotechnological methods and show the added value that molecular markers can give to breeding programmes. We used microsatellite markers in three different steps of a conventional lettuce breeding program (see Figure 1) and demonstrated, firstly, the efficiency of SSR markers not only in selecting the best parental plants for crossing based on their observed homozygosity and dissimilarity values, but also in screening the resulting F1 progeny to distinguish between the offspring resulting from cross-pollination and those resulting from self-pollination. Furthermore, using the same SSR panel, we were able to act downstream of the breeding scheme to assess the uniformity of some pre-commercial cultivars. Our molecular assay could therefore also be used by seed firms to assess newly developed varieties for distinctiveness, uniformity and stability (DUS test), three major requirements for registering plant materials [6]. Finally, molecular characterisation of a new variety could also be used to register it in national or international varietal catalogues. In fact, the genotype or molecular profile of a registered variety can be crucial in solving cases of fraudulent practices, and in curbing plagiarism and unfair free-riding on the original plant breeder's time and investment [30].

Supplementary Materials

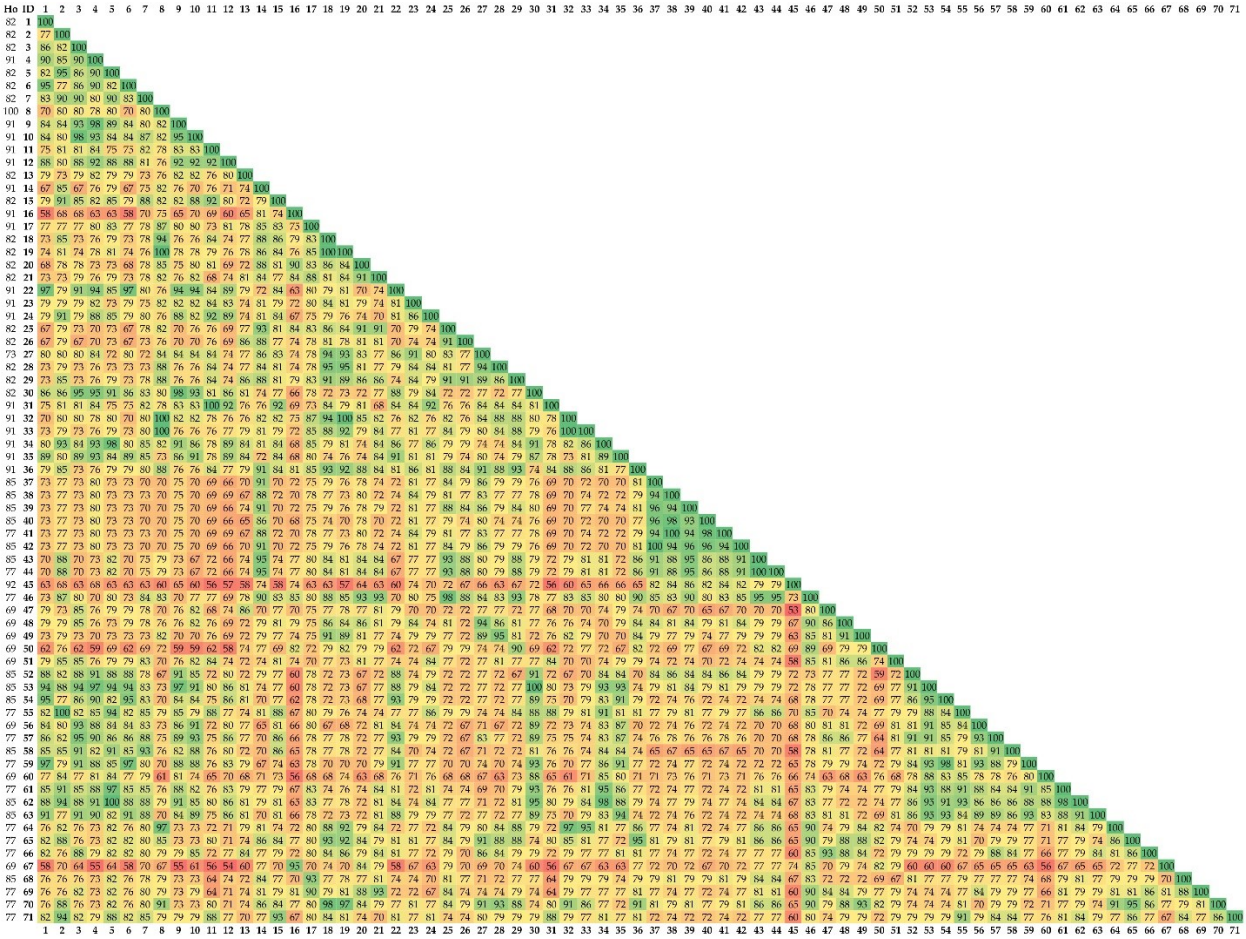


Figure S1. Pairwise genetic similarity matrix of the 71 individuals analysed (in percentages) based on Rohlf's genetic similarity coefficient. High genetic similarity values are labelled in green, the low values in red, and intermediate values are coloured on a scale from green to red. The observed homozygosity values of the 71 putative parental lines are reported to the left of each ID name.

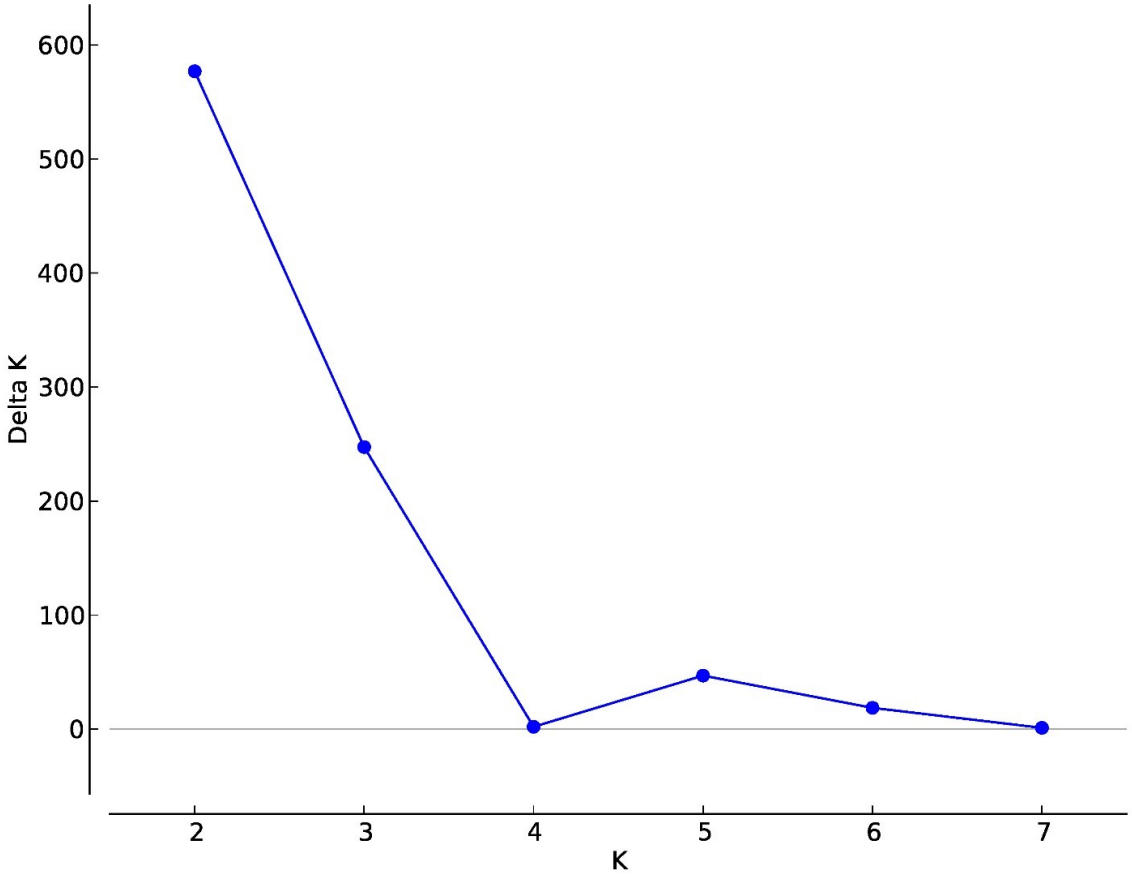


Figure S2. Definition of the subgroup number of parental lines based on the SSR marker dataset. Mean ΔK is calculated as |L″ (K)|/(SD(L(K)), following Evanno et al. [23]. The blue line represents the ΔK values.

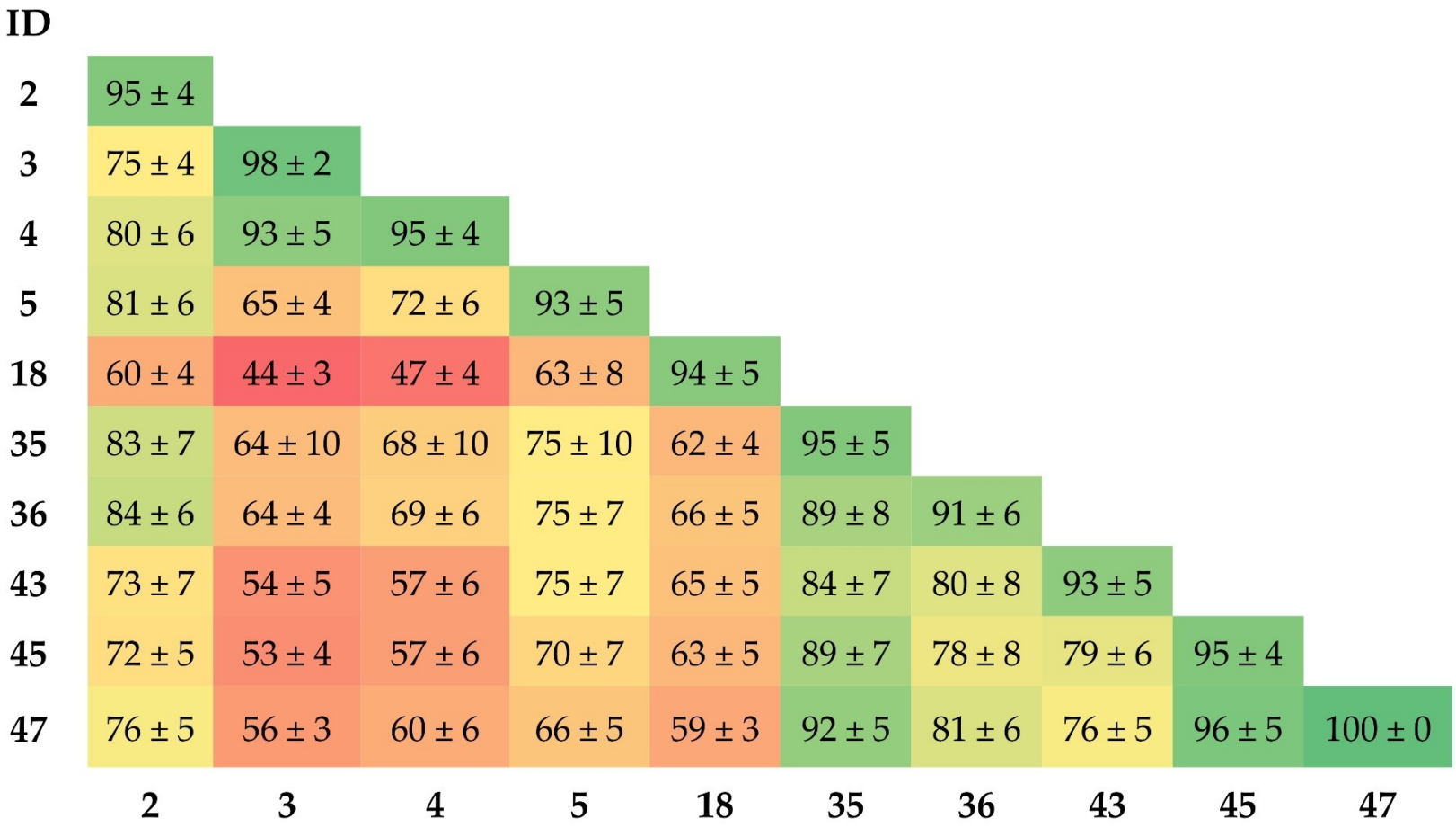


Figure S3. Pairwise genetic similarity matrix of ten selected populations (in percentages) based on the Jaccard coefficient. The high genetic similarity values are labelled in green, the low values in red, and intermediate values are coloured on a scale from green to red.

Table S1. Lettuce parental lines information, including ID of accessions and cultivar type of materials.

Parental ID	Cultivar type
1	butterhead
2	butterhead
3	butterhead
4	butterhead
5	butterhead
6	butterhead
7	butterhead
8	batavia
9	butterhead
10	butterhead
11	butterhead
12	butterhead
13	butterhead
14	butterhead
15	romaine
16	batavia
17	romaine
18	romaine
19	romaine
20	romaine
21	iceberg
22	iceberg
23	iceberg
24	iceberg
25	butterhead
26	butterhead
27	butterhead
28	butterhead
29	butterhead
30	butterhead
31	butterhead
32	batavia
33	batavia
34	butterhead
35	butterhead
36	leaf
37	leaf
38	leaf
39	leaf
40	leaf
41	leaf
42	leaf
43	leaf
44	leaf
45	leaf
46	butterhead
47	butterhead
48	leaf
49	leaf
50	leaf
51	butterhead
52	butterhead
53	butterhead

Parental ID	Cultivar type
54	butterhead
55	butterhead
56	butterhead
57	butterhead
58	butterhead
59	butterhead
60	butterhead
61	butterhead
62	butterhead
63	butterhead
64	butterhead
65	butterhead
66	romaine
67	romaine
68	romaine
69	romaine
70	romaine
71	romaine

Table S2. SSR primer tails and dyes. List of the primer tails used with their sequences and corresponding dyes.

Universal primer	Sequence 5'-3'	Dye
M13	TTGTAAAACGACGGCCAGT	6-FAM
PAN1	GAGGTAGTTATTGTGGAGGAC	VIC
PAN2	GGAATTAACCGCTCACTAAAG	NED
PAN3	TGTAGAAAGACGAAGGGAAGG	PET

Table S3. Lettuce plant material information, including ID of accessions used in the crosses, total number of plants per programmed cross, number of informative marker loci, hybrid plants, selfed plants and unexpected genotypes, and the mean hybridisation values (in percentages) for all the programmed crosses.

ID Cross	No plants analysed	Informative marker loci	No hybrid plants	No selfed plants	No unexpected genotypes	Hybridisation (%)
1 × 6	10	2	0	10	0	0
1 × 34	10	2	10	0	0	100
1 × 35	8	1	7	1	0	88
1 × 5	10	2	0	10	0	0
6 × 30	11	1	8	0	3	100
6 × 34	12	2	8	0	4	100
30 × 34	8	1	6	2	0	75
7 × 6	10	4	4	6	0	40
7 × 34	10	3	4	6	0	40
7 × 35	10	2	10	0	0	100
7 × 5	9	2	3	0	6	100
35 × 6	9	1	8	1	0	89
35 × 30	10	1	3	5	2	38
35 × 34	11	2	4	7	0	36

ID Cross	No plants analysed	Informative marker loci	No hybrid plants	No selfed plants	No unexpected genotypes	Hybridisation (%)
35 × 5	9	2	5	4	0	56
10 × 6	11	2	7	4	0	64
10 × 34	11	2	7	4	0	64
10 × 5	9	2	3	6	0	33
36 × 8	12	3	8	4	0	67
11 × 12	8	2	7	1	0	88
22 × 34	10	3	6	4	0	60
12 × 31	10	1	5	5	0	50
2 × 6	9	3	3	6	0	33
2 × 35	10	3	4	6	0	40
2 × 5	9	2	9	0	0	100
3 × 6	8	3	0	8	0	0
3 × 35	10	2	7	3	0	70
3 × 5	10	1	1	9	0	10
33 × 8	9	1	8	1	0	89
15 × 18	11	2	3	3	5	50
15 × 19	11	1	3	7	1	30
16 × 18	10	4	3	5	2	38
17 × 16	10	4	0	10	0	0
17 × 18	12	4	5	0	7	100
17 × 19	8	2	6	0	2	100
20 × 18	11	2	6	5	0	55
20 × 19	8	2	8	0	0	100
21 × 16	12	4	1	11	0	8
21 × 18	9	3	1	5	3	17
21 × 19	11	2	5	6	0	45
4 × 6	10	2	8	2	0	80
4 × 34	11	1	10	1	0	91
4 × 35	8	1	8	0	0	100
4 × 5	10	1	7	3	0	70
5 × 6	12	2	11	1	0	92
5 × 30	8	1	6	2	0	75
26 × 13	12	4	4	2	6	67
26 × 14	11	3	8	0	3	100
26 × 25	11	3	11	0	0	100
24 × 23	10	3	9	1	0	90
23 × 24	10	3	10	0	0	100
27 × 28	12	2	11	1	0	92
29 × 27	12	2	0	10	2	0
45 × 41	10	5	10	0	0	100
45 × 39	7	3	2	5	0	29
45 × 37	10	3	9	1	0	90
45 × 40	10	4	10	0	0	100
45 × 42	10	4	10	0	0	100

ID Cross	No plants analysed	Informative marker loci	No hybrid plants	No selfed plants	No unexpected genotypes	Hybridisation (%)
45×38	10	4	3	7	0	30
45×43	10	3	10	0	0	100
59 × 61	10	3	3	7	0	30
57 × 58	10	1	7	3	0	70
57 × 63	9	2	6	3	0	67
56 × 61	10	2	10	0	0	100
56 × 62	10	2	9	1	0	90
54 × 62	9	2	8	1	0	89
54 × 60	10	2	10	0	0	100
54 × 57	10	4	10	0	0	100
54 × 53	9	2	8	1	0	89
54 × 55	10	3	0	10	0	0
54 × 61	10	2	8	2	0	80
54 × 58	10	3	8	2	0	80
54 × 56	10	4	8	2	0	80
54 × 63	10	2	10	0	0	100
71 × 67	10	5	4	6	0	40
67 × 66	10	4	10	0	0	100
68 × 67	10	5	10	0	0	100
70 × 66	10	3	10	0	0	100
69 × 67	10	3	3	7	0	30
51 × 50	9	5	4	5	0	44
51 × 49	6	4	2	4	0	33
50 × 48	10	3	9	1	0	90
50 × 49	8	5	8	0	0	100
49 × 48	8	4	7	1	0	88
52 × 50	8	7	8	0	0	100
52 × 49	8	7	8	0	0	100
47 × 46	8	3	5	3	0	63
47 × 44	9	6	8	1	0	89
65 × 64	10	2	2	8	0	20
Total	871	242	556	269	46	
Mean						68

References

1. Wang, S.; Wang, B.; Liu, J.; Ren, J.; Huang, X.; Zhou, G.; Wang, A. Novel polymorphic EST-based microsatellite markers characterized in lettuce (*Lactuca sativa*). *Biologia* **2017**, *72*, 1300–1305.
2. Nagata, R. Clip-and-wash Method of Emasculation for Lettuce. *HortScience* **1992**, *27*, 907–908.
3. Barcaccia, G.; Falcinelli, M. *Genetica e Genomica*; Liguori Editore, S.r.l.: Naples, Italy, 2006; Volume 3.
4. Oliver, G.W. New methods of plant breeding. *J. Hered.* **1910**, *1*, 21–30.
5. Prohens-Tomás, J.; Nuez, F. *Vegetables I: Asteraceae, Brassicaceae, Chenopodicaceae, and Cucurbitaceae*; Springer Science & Business Media: Berlin, Germany, 2007; Volume 1.
6. 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* **2019**, *8*, 213.
7. 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. Breed.* **2018**, *38*, 122.
8. Jha, N.; Jacob, S.; Nepolean, T.; Jain, S.; Kumar, M. SSR markers based DNA fingerprinting and it's utility in testing purity of eggplant hybrid seeds. *Qual. Assur. Saf. Crop. Foods* **2016**, *8*, 333–338.
9. Singh, D.; Singh, C.K.; Tomar, R.S.S.; Taunk, J.; Singh, R.; Maurya, S.; Chaturvedi, A.K.; Pal, M.; Singh, R.; Dubey, S.K. Molecular Assortment of Lens Species with Different Adaptations to Drought Conditions Using SSR Markers. *PLoS ONE* **2016**, *11*, e0147213.
10. Rauscher, G.; Simko, I. Development of genomic SSR markers for fingerprinting lettuce (*Lactuca sativa* L.) cultivars and mapping genes. *BMC Plant Biol.* **2013**, *13*, 11.
11. Truco, M.J.; Antonise, R.; Lavelle, D.; Ochoa, O.; Kozik, A.; Witsenboer, H.; Fort, S.B.; Jeuken, M.J.W.; Kesseli, R.V.; Lindhout, P.; et al. A high-density, integrated genetic linkage map of lettuce (*Lactuca* spp.). *Theor. Appl. Genet.* **2007**, *115*, 735–746.
12. Reyes-Chin-Wo, S.; Wang, Z.; Yang, X.; Kozik, A.; Arikit, S.; Song, C.; Xia, L.; Froenicke, L.; Lavelle, D.O.; Truco, M.-J.; et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* **2017**, *8*, 14953.
13. Bull, C.T.; Goldman, P.H.; Hayes, R.; Madden, L.V.; Koike, S.T.; Ryder, E. Genetic Diversity of Lettuce for Resistance to Bacterial Leaf Spot Caused by *Xanthomonas campestris* pv. *vitians*. *Plant Health Prog.* **2007**, *8*, 11.
14. Giesbers, A.K.J.; Pelgrom, A.J.E.; Visser, R.G.F.; Niks, R.E.; Ackerveken, G.V.D.; Jeuken, M.J.W. Effector-mediated discovery of a novel resistance gene against *Bremia lactucae* in a nonhost lettuce species. *New Phytol.* **2017**, *216*, 915–926.
15. Macias-González, M.; Truco, M.J.; Bertier, L.D.; Jenni, S.; Simko, I.; Hayes, R.J.; Michelmore, R.W. Genetic architecture of tipburn resistance in lettuce. *Theor. Appl. Genet.* **2019**, *132*, 2209–2222.

16. Mamo, B.E.; Hayes, R.J.; Truco, M.J.; Puri, K.D.; Michelmore, R.W.; Subbarao, K.V.; Simko, I. The genetics of resistance to lettuce drop (*Sclerotinia* spp.) in lettuce in a recombinant inbred line population from Reine des Glaces × Eruption. *Theor. Appl. Genet.* **2019**, *132*, 2439–2460.
17. Hartman, Y.; Hooftman, D.A.P.; Uwimana, B.; Schranz, M.E.; Van De Wiel, C.C.M.; Smulders, M.J.M.; Visser, R.G.F.; Michelmore, R.W.; Van Tienderen, P.H. Abiotic stress QTL in lettuce crop–wild hybrids: Comparing greenhouse and field experiments. *Ecol. Evol.* **2014**, *4*, 2395–2409.
18. Su, W.; Tao, R.; Liu, W.; Yu, C.; Yue, Z.; He, S.; Lavelle, D.; Zhang, W.; Zhang, L.; An, G.; et al. Characterization of four polymorphic genes controlling red leaf colour in lettuce that have undergone disruptive selection since domestication. *Plant Biotechnol. J.* **2019**, doi:10.1111/pbi.13213.
19. El-Esawi, M.A. Molecular Genetic Markers for Assessing the Genetic Variation and Relationships in Lactuca Germplasm. *Annu. Res. Rev. Biol.* **2015**, *8*, 1–13.
20. Hayes, R.; Simko, I. Breeding lettuce for improved fresh-cut processing. *Acta Hortic.* **2016**, *1141*, 65–76.
21. Simko, I. Development of EST-SSR markers for the study of population structure in lettuce (*Lactuca sativa* L.). *J. Hered.* **2009**, *100*, 256–262, doi:10.1093/jhered/esn072.
22. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **2000**, *18*, 233–234.
23. Nonis, A.; Scortegagna, M.; Nonis, A.; Ruperti, B. PRaTo: A web-tool to select optimal primer pairs for qPCR. *Biochem. Biophys. Res. Commun.* **2011**, *415*, 707–708.
24. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
25. Hammer, Ø.; Harper, D.A.; Ryan, P.D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* **2001**, *4*, 1–9.
26. 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.
27. Yeh, F.C.; Yang, R.-C.; Boyle, T. *POPGENE Version 1.31*; University of Alberta: Edmonton, AB, Canada, 1999.
28. 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* **2017**, *9*, 32.
29. Feng, J.Y.; Li, M.; Zhao, S.; Zhang, C.; Yang, S.T.; Qiao, S.; Tan, W.F.; Qu, H.J.; Wang, D.Y.; Pu, Z.G. Analysis of evolution and genetic diversity of sweetpotato and its related different polyploidy wild species I-trifida using RAD-seq. *BMC Plant Biol.* **2018**, *18*, 181.
30. Lawson, C. *Plant Breeder's Rights and Essentially Derived Varieties: Still Searching for Workable Solutions*; European Intellectual Property Review 499: Griffith University Law School Research, Australia, 2016; 16–17.

31. Hand, P.; Kift, N.; McClement, S.; Lynn, J.; Grube, R.; Schut, J.; Van der Arend, A.; Pink, D. Progress towards mapping QTLs for pest and disease resistance in lettuce. In Proceedings of the Eucarpia Leafy Vegetables Conference, Centre for Genetic Resources, Wageningen, The Netherlands, 22–24 May 2003; pp. 31–35.
32. Šimko, I.; Pechenick, D.; McHale, L.; Truco, M.; Ochoa, O.; Michelmore, R.; Scheffler, B. Development of Molecular Markers for Marker-Assisted Selection of Dieback Disease Resistance in Lettuce (*Lactuca sativa*). *Acta Hortic.* **2010**, *859*, 401–408.
33. Sochor, M.; Jemelková, M.; Doležalová, I. Phenotyping and SSR markers as a tool for identification of duplicates in lettuce germplasm. *Czech J. Genet. Plant Breed.* **2019**, *55*, 110–119.
34. Bhat, J.A.; Shivaraj, S.M.; Ali, S.; Mir, Z.A.; Islam, A.; Deshmukh, R. Genomic Resources and Omics-Assisted Breeding Approaches for Pulse Crop Improvement. In *Pulse Improvement*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 13–55.
35. Phing Lau, W.C.; Latif, M.A.; Rafii, M.Y.; Ismail, M.R.; Puteh, A. Advances to improve the eating and cooking qualities of rice by marker-assisted breeding. *Crit. Rev. Biotechnol.* **2016**, *36*, 87–98, doi:10.3109/07388551.2014.923987.
36. FAO Site. Availabe online: http://www.fao.org/faostat/en/#data/QC (accessed on 11 September 2019).
37. Uwimana, B.; Smulders, M.J.M.; Hooftman, D.A.P.; Hartman, Y.; Van Tienderen, P.H.; Jansen, J.; McHale, L.K.; Michelmore, R.W.; Van De Wiel, C.C.M.; Visser, R.G.F. Hybridization between crops and wild relatives: The contribution of cultivated lettuce to the vigour of crop-wild hybrids under drought, salinity and nutrient deficiency conditions. *Theor. Appl. Genet.* **2012**, *125*, 1097–1111.

Chapter II

Genetic Structure of Cultivated Varieties of Radicchio (*Cichorium intybus* L.): A Comparison between F1 Hybrids and Synthetics

---

Alice Patella, Francesco Scariolo, Fabio Palumbo and Gianni Barcaccia

Published to *Plants*

Keywords: SSR markers; red-chicory; genotyping; microsatellite; DUS test; plant breeders' rights

Abstract:

Cichorium intybus L., well known in Italy with the common name "Radicchio", is an important leafy vegetable that is prevalently reproduced by allogamy due to very efficient barriers of self-incompatibility. Marker-assisted breeding is widely used by seed firms to develop new hybrid varieties that manifest genetic distinctiveness, uniformity and stability. A total of 29 mapped microsatellite markers were used for genotyping 504 samples of the Red of Chioggia biotype: First, two synthetics, four F1 hybrids and two derived F2 populations were compared to assess the distinctiveness of their gene pool and structure; then, the uniformity and stability of 3 years of production of a commercial F1 variety were also investigated. Genetic similarity and diversity statistics as well as the genetic structure of populations were analysed, including allele and genotype frequencies. The mean estimates and ranges of genetic similarity enabled the molecular discrimination of OP synthetics from F1 varieties and their F2 progenies and the determination of individual plant memberships. Moreover, the genetic structure of F1 hybrids produced in 3 years unexpectedly revealed two main clusters that discriminate the first 2 years from the 3rd, mainly because of the presence of uncommon specific alleles and different allele frequencies. Overall, this molecular information will enable breeders to determine the genetic distinctness, uniformity and stability of commercial and experimental varieties, as well as their genetic relationships and relatedness. Hence, this work provides a useful tool for achieving the molecular characterisation and genetic identification of different radicchio populations.

1 Introduction

Radicchio (*Cichorium intybus* subsp. *intybus* var. *foliosum* L., $2n = 2x = 18$) is the Italian name of an important locally-cultivated leaf chicory belonging to the Asteraceae, one of the largest families among flowering plants. From a reproductive perspective, radicchio is prevalently allogamous due to an efficient sporophytic self-incompatibility system and presents entomophilous pollination [1]. Moreover, outcrossing is promoted by a floral morpho-phenology that creates a physical barrier to self-pollination in the absence of pollen donors and favourable competition of allo-pollen grains and tubes [2].

Among the different biotypes available in radicchio, Red of Chioggia is one of the most commercially relevant. Historically, commercial varieties were developed by recurrent mass selection, but in recent years, synthetics have been constituted by breeders through inter-crossing or poly-crossing a number of mother individuals or clonal lines selected on the basis of their morpho-phenological and agronomic traits and, eventually, by performing progeny tests to assess their general combining ability [3]. Currently, owing to the economic benefits, newly released varieties are mainly F1 hybrids developed by Italian or European seed firms through large-scale single crosses between inbred lines selected according to their specific combining ability and exploiting molecular marker-assisted breeding (MAB) strategies. Thus, radicchio breeding programmes have improved significantly in recent years due to more efficient biotechnological tools [4].

In this regard, several linkage maps saturated with DNA markers and spanning the entire genome size (approximately 2.6 Gb) are available for leaf chicory [5–9]. These maps are particularly relevant considering that biotechnology and molecular genetics are largely utilised in programmes for breeding radicchio [10], as well as the vast majority of crop plant species [11]. In this context, the linkage map developed by Cadalen et al. [5] in chicory (*C. intybus*) is of particular interest. This genetic map is based on 431 SSR and 41 STS markers and includes nine linkage groups obtained after the integration and organisation of molecular marker data derived from one witloof chicory and two industrial chicory progenies [12]. Among codominant molecular markers, SNPs are advantageous for their abundance and high frequency in the genome and for their efficiency, but most SNPs are limited by their biallelic nature. In contrast, SSR markers are characterised by multiallelism, a mostly single-locus inheritance with a relatively lower cost. Moreover, a robust and reliable genotyping method based on SSR markers is already available for radicchio [10]. In this study, we present the implementation of the research of Ghedina et al. [10], who identified an efficient method for assessing a multi-locus genotype of plant individuals and lineages aimed at the selection of new varieties and the certification of local firm products. Our research includes the implementation of a DNA genotyping method useful for assessing the genetic distinctness and population structure of various commercial open pollinated (OP) synthetics, F1 hybrids and their F2 progenies belonging to the biotype Red of Chioggia. In addition, our research addresses the molecular characterisation and comparison of an F1 hybrid variety produced in 3 years (2014, 2015 and 2016) and obtained by open field crossings. The aim consists of evaluating the genetic uniformity and stability of these plants in different commercial lots, exploiting the same set of markers used for genotyping the other samples.

2 Results

2.1 Genetic Structure of Commercial Open Pollinated (OP) Varieties, F1 Hybrids and their F2 Progenies

A total of 216 samples belonging to two synthetics lines (OP-1 and OP-2), four F1 hybrids (F1-A, F1-B, F1-C, F1-D) and two F2 progenies (F2-C and F2-D produced from F1-C and F1-D, respectively) were investigated with 29 SSR markers. Samples were also chosen among short and medium–long term development cycle materials to equally represent the Chioggia biotypes currently available on the market (Table 1).

Table 1. Plant materials information, including accession IDs, number of individuals per population, population type and varietal cycles in days (d), are reported.

Accession ID	No. of Individuals	Population Type	Varietal Cycles (d)
OP-1	30	OP	70
OP-2	30	OP	110
F1-A	30	F1	70
F1-B	30	F1	110
F1-C	18	F1	100
F1-D	18	F1	70
F2-C	30	F2	100
F2-D	30	F2	70

All SSR markers were polymorphic, and the mean polymorphic information content (PIC) value was 0.61, with a maximum equal to 0.84 for the M2.6 SSR locus. Of note, 24 of 29 SSR markers were highly informative [13], with PIC values higher than 0.5. Two other loci were considered informative (M5.15 and M7.21, 0.25 < PIC < 0.5), while M3.7, M8.22 and M5.14 were less informative (PIC < 0.20, Supplementary Table S1). The total number of scored alleles was 220, with 2.7 observed alleles (Na) per locus and 2.0 expected alleles (Ne) per locus (Supplementary Table S1). The mean Na for a single locus was higher in synthetics (4.5 alleles/locus) than in F1 hybrids (2.3 alleles/locus) and F2 progenies (2.8 alleles/locus).

Additionally, the mean Ne was higher in the OP populations (2.6 alleles/locus) than in the F1 hybrids and F2 progenies (1.8 and 1.9 alleles/locus, respectively, Table 2). Private alleles were observed in 25 of 29 SSR loci, but in only 11 of these SSR loci was the allele frequency higher than 15 %. Specifically, the F1-B hybrid had private alleles in four loci with frequencies >15.0 % (M2.4 50.0 %, M1.1 40.0 %, M4.12 and M6.17 15.0 %); instead, the F1-C hybrid had one locus showing a private allele with a 25.0 % frequency (M6.17). In the case of OP-1, two private alleles were detected at two different loci with frequencies >30.0 % (M6.17 at 50.0 % and M6.18 at 31.0 %) and five other loci with frequencies >15.0 % (Supplementary Table S2). The locus with the highest number of private alleles across the population was M6.17, although it did not represent the locus with the higher PIC value, being equal to 0.7 (Supplementary Table S1).

Table 2. Number of alleles found across populations for each F1 hybrid, F2 population and OP synthetic. In particular, statistics refer to the mean number of alleles (Na) and number of expected alleles (Ne).

ID	OPs		F1s		F2s	
	Na	Ne	Na	Ne	Na	Ne
Mean	4.5	2.6	2.3	1.8	2.8	1.9
St. Dev.	2.4	1.2	0.9	0.6	0.9	0.5

Allele frequencies were calculated per locus within each population, permitting the identification of the most common genotype. In contrast to synthetics, within the hybrid pool, a fixed genotype was found across several loci (Supplementary Tables S3 and S4). Specifically, we found a mean of 8.8 fixed genotypes for hybrids and only 3.0 for synthetics.

The observed heterozygosity (Ho) of F1 hybrids, considered as a whole, was 50.0 % on average (ranging from 37.9 % to 65.5 %, Figure 1a), and in particular, F1-C and F1-D exhibited the median highest values (51.7 % and 65.5 %, respectively). The two OP synthetics showed a median heterozygosity as low as 40.5 % (from 39.7 % to 41.4 %). Moreover, the median percentage of heterozygosity within the two F2 progenies (F2-D and F2-C) was 41.4 % (from 37.9 % to 44.8 %), considerably lower than the one calculated for the two F1 hybrids used as parents (F1-C and F1-D) (Figure 1a). The median observed heterozygosity of F1-A, and F1-B resulted in 37.9 % and 44.8 %, respectively (Figure 1a). Additionally, the expected heterozygosity (He) was lower than the observed heterozygosity in the F1 populations and F2-C (Supplementary Table S5).

According to the similarity analysis conducted using NTSYS software, the median estimate of genetic similarity within each population was higher for hybrid varieties (95.2 % on average, ranging from 94.0 % to 97.1 %) than synthetics (84.3 %). In the case of F2 progenies, the genetic similarity within each population was 89.1 % and 90.7 % for F2-C and F2-D, respectively (Figure 1b). The mean genetic similarity (MGS) calculated among F1 hybrids ranged from 76.4 ± 1.5 % (F1-B vs. F1-D) to 87.2 ± 1.6 % (F1-A vs. F1-B), while those calculated between F1-C vs. F2-C and F1-D vs. F2-D were both slightly over 90 % (90.2 ± 2.5 %). Moreover, F1-A and F1-B were highly similar to each other (87.2 ± 1.6 %) (Supplementary Figure S1).

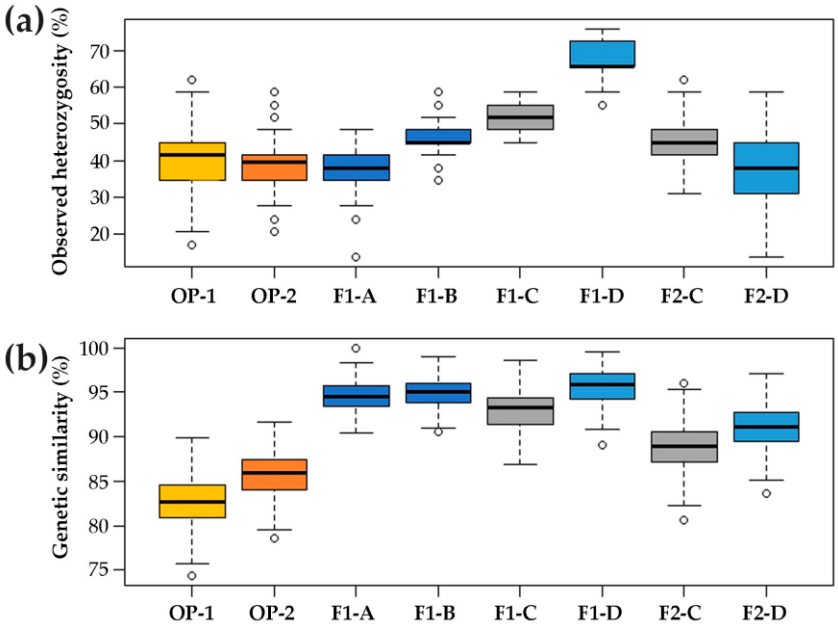


Figure 1. Statistics of genetic diversity and similarity for OP synthetics, F1 hybrids and F2 progenies. **(a)** Box plot of the observed heterozygosity within each population. **(b)** Box plot of the median genetic similarity (MGS) within each ID (in percentage). The second and third quartiles are marked inside the square and are divided by a bold bar (median). Dots show outlier samples.

An unweighted pair group method with arithmetic mean (UPGMA) dendrogram was also constructed on the basis of the genetic similarity matrix, whose coefficients were computed in all possible pair-wise combinations of the 216 samples. The dendrogram clustered the entire collection in six main subgroups. Each cluster, with few exceptions, included the whole pool of samples used to represent each of the six commercial lines. Additionally, as shown by the genetic similarity matrix, OP-1 resulted in the most dissimilar group compared to the rest of the samples, and the corresponding branch of the dendrogram was clearly separated from the rest of the tree. Notably, the UPGMA did not enable the full discrimination of the F2 progenies that grouped together with their respective F1 parents (Supplementary Figure S1).

According to the principal coordinate analysis (PCoA), similar to what was highlighted by the UPGMA tree, the two synthetic populations clustered independently from the rest of the groups, although a partial overlap was observed between OP-2 and F1-A. In contrast, F1-C and F1-D formed unique clusters with their respective F2 progenies. The first principal coordinate accounted for 8 % of the total variation and clearly separated a group including F1-A, F1-B and OP-2 from the group of F1-C, F2-C, F1-D, and F2-D; the second principal coordinate accounted for 5 % of the total variation and separated F1-C and F2-C from F1-D and F2-D. Finally, the hybrid varieties F1-A and F1-B were highly similar genetically, forming a cluster divided into two subgroups (Figure 2).

Regarding the investigation of the genetic structure of the radicchio core collection, the best estimate of population size was $K = 5$ such that the 216 samples were grouped into five genetically distinct clusters. Each genotype was plotted as a vertical histogram divided into $K = 5$ coloured segments representing the estimated membership in each hypothesised ancestral genotype. A total of 212 of 216 individuals showed strong ancestry association (>90.0 %), and two samples scored a slightly lower level of association (equal to 87 % and 88 %, respectively); only two additional samples had a very low degree of ancestry association (75.0 % and 70.0 %, respectively), and they were hence considered admixed. Moreover, the two F2 progenies were not distinguished from their original hybrids, and F1-A and F1-B hybrids were grouped as if they had only one common ancestor (Figure 2), analogous to what was observed in the PCoA.

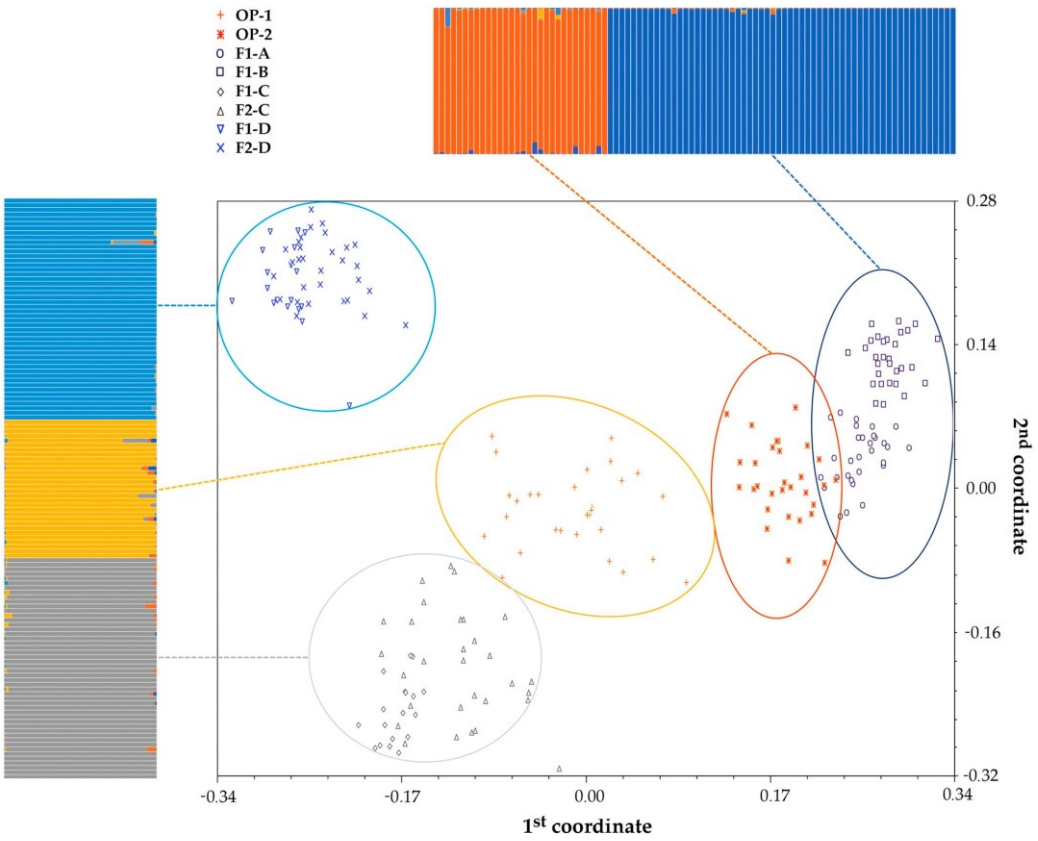


Figure 2. Principal coordinate analysis (PCoA). The centroids of all radicchio samples (n = 216) deriving from the mean genetic similarity coefficients plotted according to the first two main coordinates. For each population, the PCoA output is coupled with the results of the population genetic structure study, estimated by STRUCTURE software using the same SSR marker data set. Each sample is represented by a vertical histogram partitioned into K = 5 coloured segments that report the estimated membership.

2.2 Genomic Comparison among Three Years Production of An F1 Hybrid Variety

Regarding the genomic comparison of the 3 production years of commercial lots of an F1 hybrid (2014, 2015 and 2016), the level of genetic differentiation between samples was investigated by calculating the genetic similarity in all possible pair-wise comparisons among all 288 individual plants (96 samples per year).

The Rohlf's coefficient of genetic similarity ranged from 65.0 % to 100 %, with an estimated average equal to 91 %. Additionally, a comparison within and between the 3-year populations was performed to assess the stability of the F1 hybrid variety. Regarding the year 2014, the median genetic similarity observed was 91.8 %, ranging from 66.4 % to 100 %; 2015 went from 77.0 % to 100 % with a median genetic similarity of 95.4 %, while 2016 showed a median genetic similarity equal to 89.7 %, ranging within the population from 80.0 % to 100 %. Moreover, comparisons among the different populations were made. The median values were 92.0 %, 89.0 % and 90.8 % for the 2014 vs. 2015, 2014 vs. 2016 and 2015 vs. 2016 comparisons, respectively. For the minimum genetic similarity data, the lowest value was observed in 2014 vs. 2016 (65.4 %), followed by 2014 vs. 2015 (67.0 %), and the highest value was in 2015 vs. 2016 (75.2 %).

Figure 3 shows the genetic similarity statistics for the 3 years of production of the analysed F1 hybrid variety, calculated within and among individuals of years 2014, 2015 and 2016.

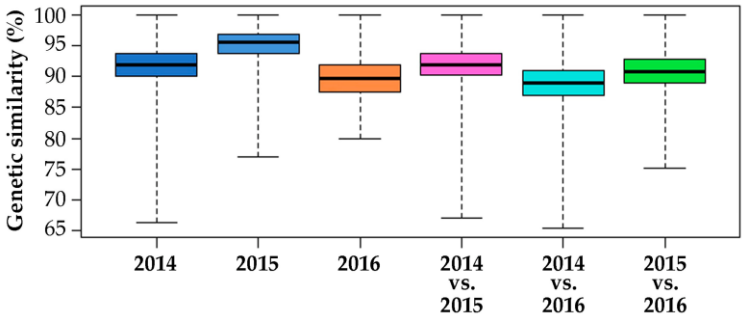


Figure 3. Box plot of the genetic similarity of 3 years of production of an F1 hybrid variety, calculated within and among individuals of years 2014, 2015 and 2016. The second and third quartiles are marked inside the square and divided by a bold bar (median).

With the data obtained from the genetic similarity analysis, a UPGMA dendrogram was also computed, highlighting the clustering in two main subgroups. The first group included individuals belonging to both years 2014 and 2015, while the second group, strongly divided from the previous years, prevalently clustered 2016 plants (Figure 4a).

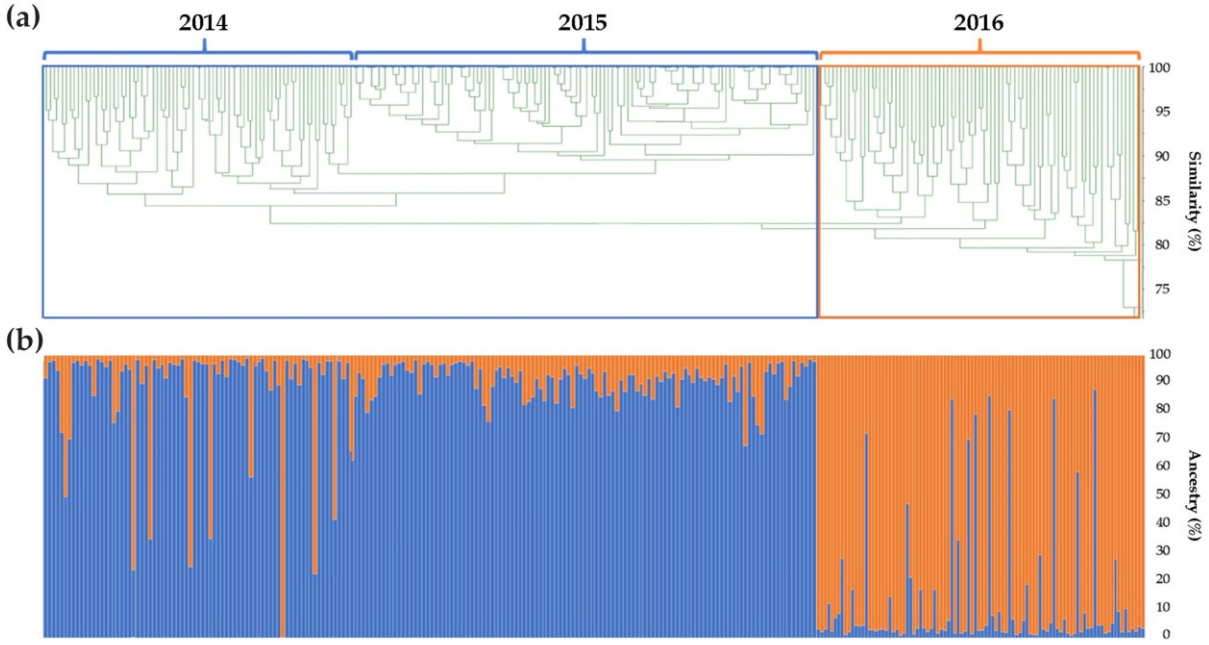


Figure 4. Analysis of the genetic structure of the 3 years of seed production of the F1 hybrid variety analyzed. **(a)** Unweighted pair group method with arithmetic mean (UPGMA) dendrogram, based on the calculated genetic similarity among individuals of years 2014, 2015 and 2016. Blue and orange squares highlight the two main identified clusters. **(b)** STRUCTURE software results. Data are disposed of in a vertical histogram, labelled for $K = 2$ colours concerning cluster membership of each individual sample belonging to years 2014, 2015 or 2016.

From the analysis of the genetic structure of the 3 years of seed production, the most likely value of K was 2 for the population as a whole. Each genotype of the analysed F1 hybrid variety was plotted in a vertical

histogram divided into $K = 2$ coloured segments representing the estimated membership in each hypothesised ancestral (Figure 4b). From this analysis, reflecting the results obtained from the UPGMA dendrogram, almost all of the 2014 and 2015 samples grouped in cluster 1, while 2016 individuals grouped in cluster 2. Clustering revealed that 246 of 288 samples (85.4 % of samples) showed a strong ancestry association (>85.0 %). Twenty-three admixed samples were part of the first 2 years of production (2014 and 2015, cluster 1), while the remaining 19 admixed genotypes belonged to the 2016 production (cluster 2). It is relevant to note that few samples scored a membership perfectly fitting the second cluster with matching values over 99 %. At the same time, relative to the year 2016, which was mainly grouped in cluster 2, some admixed samples showed very low membership values.

From the analysis of the allele frequencies, it emerged that for 2014 and 2015, 51.7 % of loci (15 out of 29 loci) matched the hypothetical profile by which both parental genotypes used for crossing are homozygous for the same allele, while for 2016, 48.7 % of loci (14 out of 29 loci) matched. Regarding the second putative profile ($A_1A_1 \times A_1A_2$), 13.8 %, 17.2 % and 20.7 % of loci (4, 5 and 6 out of 29 loci, respectively) matched for the years 2014, 2015 and 2016, respectively. The year that best represented the third case ($A_1A_1 \times A_2A_2 / A_1A_2 \times A_1A_2$) was 2015 (27.6 % of loci, i.e., 8 out of 29 loci) followed by 2014 (24.1 %, i.e., 7 out of 29 loci) and 2016 (10.3 %, i.e., 3 out of 29 loci). In contrast, 2016 matched the fourth hypothesis ($A_1A_1 \times A_2A_3 / A_1A_1 \times A_1A_3$) with a percentage equal to 17.2 %, which is representatively higher than for 2014 and 2015, and fits 6.9 % (2 out of 29 loci) and 3.4 % (1 out of 29 loci), respectively. Regarding the last cluster of the bar chart ($A_1A_2 \times A_3A_4$), where both parental genotypes are heterozygous for different alleles, 3.4 % of the loci showed correspondence for 2014 and 2016, while there were none for 2015 (Figure 5).

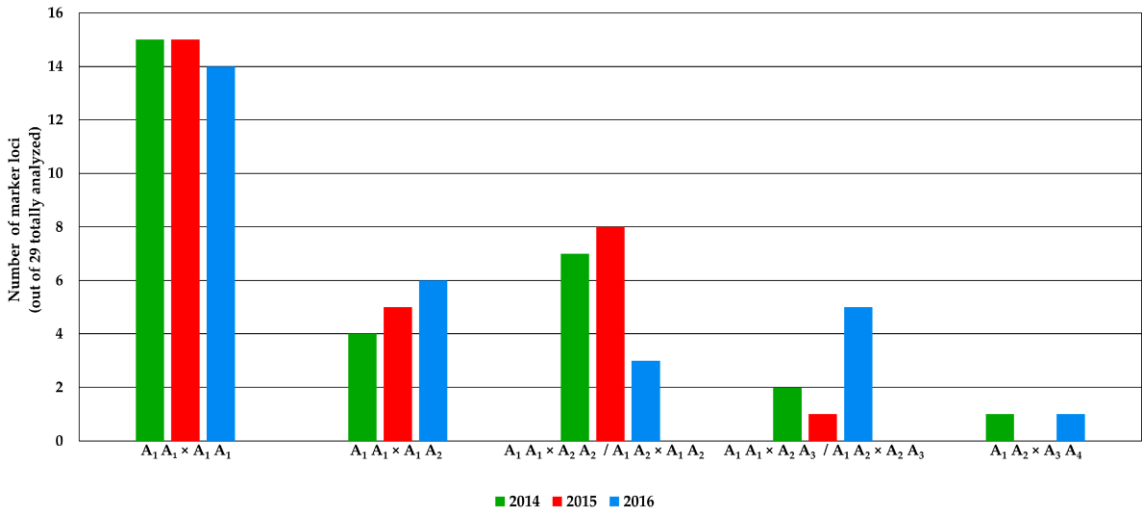


Figure 5. Number of loci corresponding to the five allelic frequency profiles derived from hypothetical crosses between representative genotypes in the 3 different analysed years (2014, 2015 and 2016).

Considering the allele frequencies, it is noteworthy that five loci (M1.2, M1.3, M2.6, M8.24 and M9.25) showed new alleles in 2016, with rates extending from 8.6 % (M1.3) to 15.5 % (M1.2) (Table S6a). Moreover, in

2014, at locus M9.26, allele A_1 was present, but it disappeared in 2015, and in 2016, it was replaced by a new allelic variant (allele A_3). In contrast, two polymorphisms at the M8.23 (allele A_7) and M9.27 (allele A_6) loci were detected in 2014 and 2016 but not in 2015 (Table S6b). Another result emerging from the allele frequencies analysis is that there were polymorphisms at many loci (82.8 %) among all 3 years, with percentages under 3.0 % being observed. These results were not considered during the analysis because they are probably derived from pollen contamination.

3 Discussion

Traditional methods have recently been integrated with biotechnological methods to accelerate breeding programmes. Marker-assisted breeding (MAB), in fact, is widely utilised for the development of improved lines by firms and research institutes, allowing breeding activity based not only on the evaluation of phenotypes but also on plant genotypes. Moreover, molecular assays have become useful tools for verifying the distinctness, uniformity and stability of varieties (DUS test), three major requirements for the registration of new plant materials. Microsatellite markers can identify essentially-derived varieties (EDV) in the context of variety registration; therefore, these markers represent useful tools for cultivar protection against plagiarism. [14]. In particular, molecular assays based on SSR markers overcome the mostly subjective system of morpho-phenological characterisation. Our genotyping investigation of the radicchio cultivated lines of the Chioggia biotype fits well with this scenario. Specifically, the present study enabled the comparison of different commercial varieties of high breeding value by improving the method of Ghedina et al. [10] based on SSR markers. Two loci were added to the original number (27), and the multiplex PCRs were reduced from 4 to 3. The 29 SSR molecular markers used in this work were chosen because they were equally distributed throughout the whole genome and dispersed over the nine LGs (minimum of three markers for each LG), making the molecular assay efficient. Moreover, this method was shown to be an inexpensive and fast tool for genotyping analyses. This panel of SSR markers was first used to evaluate the genetic variation within and among a core collection of commercial OP synthetics, F1 hybrids and their F2 progenies. Moreover, the same set of microsatellites was also used to investigate the genetic structure and genetic similarity of 3 production years of an F1 hybrid variety. Additionally, this study provides useful tools for protecting registered varieties against plagiarism.

3.1 Genetic Structure of a Core Collection of OP, F1 and F2 Populations

Marker tools differ in their information content, depending on their polymorphism degree. The PIC calculated for the 29 SSR marker loci showed an average value of 0.61; therefore, the microsatellite markers used in this study were determined to be generally highly informative. According to Botstein et al. [13], 24 of the selected SSR markers could be considered highly informative (PIC > 0.5), while only five SSR markers were

considered less informative. The highest Na and Ne values (Na = 4.5, Ne = 2.6), observed within the two OP synthetics, directly correlate with the population size and with a high genetic diversity, which was later found within them both. In fact, a high Na should produce many genetically possible genotypes and thus low genetic similarity within the population. In contrast, this property was not found for the F1 progenies. Considering the F1 generation as a whole (i.e., 96 samples), the population size was larger than that of the OP synthetics, but both Na and Ne were significantly lower (2.3 and 1.8, respectively) (Table 2). However, this finding is consistent with the progenies' breeding history and with the high similarity highlighted within the F1 populations.

A high number of private alleles were detected within the synthetic populations, especially in OP-1, where private allelic variants were found at seven SSR loci; notably, in two cases, the frequency was > 30 %. Interestingly, private alleles with high frequencies were also identified in F1-B and F1-C, making them attractive tools for protecting the rights of plant breeders [15]. Additionally, a fixed genotype across several loci of the hybrid varieties was found. Specifically, we observed a mean of 8.8 fixed genotypes for hybrids and a mean of 3.0 for synthetics. Thus, it is reasonable to think that a combination of those SSR loci exhibiting private alleles could be profitably exploited to protect and trace registered varieties, as well as their derivative food products. High uniformity and the ability to trace hybrid products explains the high exploitation of F1 seeds. Indeed, since these varieties have the same genotypes, farmers adopted hybrids for combining such qualities as maturation contemporaneity and productivity traits.

The significant variability, in terms of heterosis, shown by the four F1 hybrid populations, relies on the genetic distance between the relative parents. In fact, information derived from genotyping data are exploited more and more for planning crosses and predicting plant vigour traits (i.e., heterosis) of experimental F1 hybrids on the basis of the overall genetic distance and allelic divergence between parental inbred lines, as an estimate of their specific combining ability [10,16]. As occurs for most open-pollinated species, detectable heterotic effects are well-known also in Radicchio [16,17]. In this species, it has been demonstrated that hybridization between selected genotypes provides uniform and heterotic populations due to increased heterozygosity: Field trials performed using commercial F1 hybrids showed that the genetic diversity between paternal and maternal lines is positively correlated not only with the observed degree of heterozygosity of their hybrid progenies, but also with the realized crop yield potentials of individual plants. In particular, F1 hybrids of Radicchio can manifest an increase of leaf yields per single plant equal to 25–30 %, on average, if compared to OP synthetics of the same varietal cycle length (unpublished data, Blumen Group SpA). This is why recently private breeders and seed firms have implemented methods for the development of F1 hybrids [8]. Therefore, we can speculate that those F1 offspring that showed lower heterozygosity (i.e., F1-A and F1-B with 37.9 % and 44.8 %, respectively) are the results of crosses between parents that are homozygous for the same alleles at the considered loci. As expected, both the OP synthetics and the F2 progenies showed a lower heterozygosity compared to the F1 hybrids, F1-A excluded. However, the reduction

in heterozygosity observed in the F2 populations is the direct consequence of segregation events, while the low levels found in the OP synthetics are the result of a long breeding process aiming to constitute highly uniform populations. Notably, the genetic similarity calculated for F1-C and F1-D is shown to be higher than those reported in their direct F2 progenies. The reduction in the genetic similarity values of the F2 populations can still be attributed to the segregation events that cause progenies to be less similar within them but with an increment of their genetic similarity range. The median estimate of genetic similarity within populations was higher in hybrid (95.2 %) than in synthetic (84.3 %) varieties, which was in full agreement with the breeding strategies exploited for their development. This result explains and gives rise to the greater stability that is usually appreciated in the hybrid cultivated varieties compared to the non-hybrids. Moreover, the genetic stability of a population also facilitates the employment of molecular data for breeders' rights protection. On the one hand, it is feasible to identify and exploit private alleles for the unequivocal identification of commercial hybrids; on the other hand, the higher genetic variability found within the OP synthetics highlights the difficulty of protecting them from frauds.

Other information regarding the genetic similarity and the genetic structure of the analysed populations were provided by both the UPGMA dendrogram and the PCoA-based centroid, revealing well-separated sub-populations corresponding to the cultivated varieties object of this study. It is worth noting that the high similarity shown by the two hybrids F1-A and F1-B (median of genetic similarity equal to 87.2 %) is also graphically evident in both analyses and is clearly the direct consequence of a common genetic root. The shared origin of these two hybrids is also visible from the STRUCTURE analysis, which confirms a common ancestor. The findings reported for F1-A and F1-B are also transposable to the two F2s (F2-C and F2-D) and their related F1 parents (F1-C and F1-D). The graphical overlapping in the PCoA and the high genetic similarity among the two generations are the expressions of a direct lineage, and their common progenitor further confirmed this finding. The only two samples considered admixed from the structure analysis can be assumed as off-types of these populations; thus, they will be removed from the core collection. Considering that a new variety, before the release on the market, needs to be distinct from all other commercially available cultivars, the results of this study indicated that SSR markers represent an efficient tool to evaluate the distinctness of commercial varieties or to investigate their common origin, as in the case of EDV.

3.2 Genomic Comparison among Three Different Production Years of a Commercial F1 Hybrid

From the genetic similarity analysis of the 3 years of hybrid seed production, samples belonging to 2015 showed the highest intra-similarity scores, with a median value of 95.4 %, reflecting a generally higher uniformity of this population. In contrast, 2016 exhibited the lowest median similarity within the population, even being the year with the most contained variability among its individuals. At the same time, when considering the pairwise comparison between the different years, 2014 vs. 2015 appeared to be most similar, with a median value of 92.0 %, followed by 2015 vs. 2016 (90.8 %) and 2014 vs. 2016 (89.0 %). Moreover, the

2015 vs. 2016 comparison showed the lowest variability, as the genetic similarity values ranged from 75.2 % to 100 % (Figure 3).

4 Materials and Methods

4.1 Plant Materials and DNA Isolation

The plant material used in this study, including OP and F1 varieties, represents part of a high-breeding value collection of commercial lots belonging to the "Red of Chioggia" biotype of radicchio. In addition, F2 populations obtained ad hoc for the purpose of this investigation were also used.

In this study, 30 samples were used to represent each of the first two hybrids (F1-A and F1-B) and each of the two synthetic populations (OP-1 and OP-2); 36 samples were collected for F1-C (18) and F1-D (18) hybrids, and eventually, the two F2 progenies (F2-C and F2-D) were composed of 30 plants each and obtained by selfing single F1 individuals (respectively, F1-C and F1-D; Table 1). The F2s were specifically produced to evaluate their genetic structure compared to the synthetic genetic structure and to investigate the segregating pattern of some relevant F1 hybrids. The plant material was also chosen considering the varietal cycle: Two hybrids and a synthetic were characterised by a short cycle, while the remaining two hybrids and the second synthetic were distinguished for medium-long cycles (Table 1).

The second set of samples consisted of 288 plants belonging to three different populations obtained in 3 different years of production (2014, 2015 and 2016) from an F1 hybrid variety obtained in an open-field system without physical barriers.

On the whole, 504 samples were collected, and approximately 100 mg of fresh leaves were ground to fine powder using a TissueLyser II mill (Qiagen, Valencia, CA, USA). Genomic DNA was extracted with a DNeasy® 96 Plant Kit (Qiagen) following the manufacturer's protocol. The quality, purity and quantity of gDNA were assessed by gel electrophoresis in 1 % agarose/1× TAE gels containing 1× SYBR Safe DNA Gel Stain (Life Technologies, Carlsbad, CA, USA) and a NanoDrop spectrophotometer (Thermo Scientific, Pittsburgh, PA, USA). DNA samples were diluted to 20–30 ng/µl to be used as a template in a multilocus PCR.

4.2 Genotyping by SSR Markers

The composition of the PCR multiplex reactions was designed to improve a previous panel of 27 SSR markers [10]. A modification of the dye-labels system [18] permitted the analysis of two additional loci developed thanks to the genome draft from radicchio [9], for a total of 29 SSR markers by using only three different PCR multiplex reactions (Table 3). The primers were combined into three different multiplex groups based on their annealing temperatures and their attitude to specifically and efficiently amplify the target microsatellite in multiplex reactions. Briefly, the amplification procedure was applied to 504 samples based on a three-primer system. This method consists of using SSR-targeting specific primer pairs, one of which is

anchored in 5' with a tail. This added sequence is complementary to one of the four universal (M13, PAN-1, PAN-2 and PAN-3) and fluorophore-labelled (6-FAM, VIC, NED and PET, respectively) primers used to discriminate the different loci during capillary electrophoresis (Table S7).

The genetic similarity data, also used for the construction of a UPGMA dendrogram, highlights a division in two main clusters. The first cluster consists of the years 2014 and 2015, while the other cluster consists of the year 2016. From the computed analysis, the dendrogram confirmed that two main clusters divide the year 2016 from the other two, 2014 and 2015, grouped together. It was also possible to observe that few samples were admixed (<85.0 % of membership with their main group) among both clusters. Information was also confirmed by the UPGMA analysis, in which the highly different samples outstood the primary roots of the tree. To corroborate the clustering analyses, a comparison between observed allele frequencies at single loci per year and those obtainable in progenies after crossing different hypothetical parental genotypes was performed. Thus, five main patterns were supposed to categorise the different frequency profiles (e.g., $A_1A_1 \times A_1A_1$ gives $p(A_1) = 100$ %; $A_1A_1 \times A_1A_2$ gives $p(A_1) = 75$ % and $q(A_2) = 25$ %) among the 29 analysed loci over the 3 seed production years. After this step, all profiles across all loci were matched with the hypothetical profiles and then counted for each year. As described in Figure 5, most loci within each year (51.7 %) showed having ancestors with a monomorphic genotype, but in 2016, this value decreased to 48.3 %. In contrast, the second pattern ($A_1A_1 \times A_1A_2$) showed an average increase of 3.5 % per year from 2014 to 2016, rising from 13.8 % to 20.7 %. This finding was considered to be an increase in the number of heterozygous loci in the parental lines used for the crossing. Additionally, the number of loci fitting the third and the fourth hypothetical profiles scored a substantial decrease in 2016 and increase in 2016, respectively (see Figure 5). This result is a further demonstration of the low genetic stability of the parental lines over the 3 years of production of the F1 hybrid. Another relevant result emerging from these data is that regarding the pattern obtained with crosses as $A_1A_1 \times A_2A_2$ and $A_1A_1 \times A_2A_3$ by which the resulting hybrid genotypes were observed in an overall frequency much lower than the one expected for F1 hybrid varieties. Notably, the fifth profile, characteristic of progenies obtained by crossing divergent heterozygous parental genotypes, was matched in years 2014 and 2016 for only 3.5 % but not recorded in 2015. The investigation of unique alleles determined that in 2016, five different loci manifested having private alleles that were not found in the other two populations. Moreover, two alleles at different loci (M8.23 and M9.27) were present in 2014 and not 2015 and reappeared in 2016 with the same frequencies as in the first year, meaning that a correlation between the parental plants used for the constitution of this hybrid population is certain but with strong differences shown by the presence of specific alleles with elevated frequencies in the 2016 population. Interestingly, these results indicated the presence of a unique allele for marker M9.26 in 2014, substituted by an already existing polymorphism in 2015 and then replaced in 2016 by an entirely new polymorphism. These data enabled us to speculate that in 2014, parental genotypes were $A_2A_4 \times A_1A_2$; in 2015, they were $A_2A_4 \times A_2A_2$; and in 2016, they were $A_2A_4 \times A_3A_3$ (Supplementary Table S6).

Overall, these findings, from the genetic similarity comparison to the investigation of unique alleles, enabled a major consideration to explain the clustering and the poor genetic uniformity displayed by the 3 years' productions. In particular, the fact that year 2016 clustered separately from the previous two can be interpreted as a strong contamination in the genetic pool of probably one parental line used in the crossing programme. Considering that both the parents and the F1 hybrid object of this study are grown in an open-field system without physical barriers; the high variability in terms of genetic similarity observed both within and among the 3 years of production could be consistent with some possible pollen contaminations, mainly from wild-type radicchio. This hypothesis is further corroborated by the presence of unique alleles (with frequencies under 3 %), especially in 2016. Moreover, this second case study demonstrated the efficiency of microsatellite markers in assessing uniformity and stability through the generation of a commercial variety, two crucial requisites for its release and survival on the market.

Table 3. Sequences of the primer pairs used to amplify the SSR molecular markers. For each primer pair, ID, SSR linkage group (LG), motif, multiplex to which the SSR marker locus belongs, and tailed primers used (PAN1, PAN2, PAN3 or M13) are reported. All the microsatellite used in this study derive from Ghedina et al. [10], except for the two underlined SSR loci, which were newly introduced.

ID	Motif		Primer Sequence and Tail	Multiplex
1.1	$(GA)_{40}$	F	[PAN3]CCAACGGATACCAAGGTGTT	1
		R	AACCGCACGGGTTCTATG	
2.4	$(GA)_{25}$	F	[M13]CCGCTCTCTCATCACTCCTC	1
		R	GCTCGAAAATCGGCTACAAC	
2.5	$(CT)_5CC(CT)_{13}TT(CT)_5$	F	[PAN1]GTGCCGGTCTTCAGGTTACA	1
		R	CGCCTACCGATTACGATTGA	
3.7	$(CT)_{22}$	F	TTCGAGTCTTGCCTTAATTGTT	1
		R	[PAN1]CAGACGACCTTACGGCAACT	
<u>4.10a</u>	$(CT)_{22}$	F	[PAN2]CATCACCTTCACGAAAAGCA	1
		R	CGAAGACCATCCATCACCA	
4.10b	$(CT)_{21}CATA(CA)_5CT(CA)_5$	F	[M13] CCATTATTGGGCAGCA	1
		R	CACCAACGAACTCCTTACAAAG	
<u>4.11a</u>	$(CT)_{12}N_5(CA)_{11}$	F	[PAN3]GAAGGAACCTATGAACCAACCACTCA	1
		R	GTTTTGAGCCTGAGCCAGA	
5.15	$(CT)_{11}N_7(CAA)_5$	F	AGCACGACTCTGCTGTCTTTT	1
		R	[PAN1]CGAGCCATGTTAGGGTTTGT	
8.22	$(CA)_5AA(CA)_9$	F	[PAN1]TCGTCATCAGAAACAAAGCAA	1
		R	CAAAGAAGGCACTCTTGTCG	
9.26	$(GATA)_3N_{19}(GA)_9$	F	[PAN2]CCTACACTCGGCCACCTACT	1
		R	TCGACGGTATAACAACACCTG	
3.8	$(CT)_{16}$	F	[PAN1]AGGAAGCGGTGTCATCTGT	2
		R	CGCCCACATATTCATTCTCA	
6.16	$(CT)_{12}TT(CT)_{15}TT(CT)_2TT(CT)_4$	F	[PAN1]TATTGCATTGTTGTTCCTTG	2
		R	TATTTAGAAGAGGGAAATAGATG	
	$(CT)_{18}$	F	ATGTCGGAGCAAAATCGTTC	2

ID	Motif		Primer Sequence and Tail	Multiplex
7.19		R	[PAN1]CATGTTCCCGCTCATGAATA	
1.2	$(CT)_{19}$	F	CCGGCAGAATTTTTAGGG	2
		R	[PAN3]CAGGTCATAGGTCCATGTGAAA	
1.3	$(CT)_{17}$	F	[PAN3]TGGAGAAAAATGAAGCAC	2
		R	GAATGAGTGAGAGAATGATAGGG	
5.13	$(CT)_{23}$	F	[M13]AGGCATAAAGAGGTGTGG	2
		R	TCAAACATGAAAACCGCTC	
6.17	$(CA)_8(CT)_{18}$	F	CGTGTCCAAACGCAAACATTAT	2
		R	[PAN2]GCACAATTTTCCTACCACTTATCC	
5.14	$(TC)_{11}$	F	[M13]AAAGTCACACATCGCATTTCCT	2
		R	GTAGCAGCAGCAGCCATCTT	
4.11b	$(TG)_5CG(TG)_7$	F	[M13]GCCATTCCTTTCAAGAGCAG	2
		R	AACCCAAAACCGCAACAATA	
4.12	$(CT)_8TT(CT)_5CC(CT)_3TT(CT)_7$	F	GGCATCGGGATAGAAAAACA	2
		R	[PAN2]TCAATGCCTCAACAGAAATCC	
3.9	$(CA)_{12}$	F	CTGCTATGGACAGTTCCAGT	3
		R	[PAN3]CAATTCAGTTGTGATAGACGC	
7.20	$(CT)_{31}$	F	[PAN2]ACACTCACTCACACTCCGTAA	3
		R	GTCATGATGGCGTAAAAGTC	
8.23	$(CA)_{11}(CT)_9$	F	TGTAGACACACAAAATGCACA	3
		R	[M13]ACCGGTTGAAAACATGAAAT	
8.24	$(TC)_{16}(CA)_{13}$	F	[PAN2]GGTCCGTAGACTGCAGACTTTT	3
		R	CACCGTCCCACTTTTTAGG	
9.25	$(CA)_{11}$	F	[M13]GTGTGGGTGTTTGAAGAGC	3
		R	TCAAGAACATCAACGCGTAA	
7.21	$(CT)_{13}$	F	GGACACCGAGCTGGAGAA	3
		R	[PAN1]TTCCACTTTCGGGAGTTACC	
9.27	$(GA)_{10}TAAA(GA)_5$	F	GCTAAAAGAAGTGCAAGGAGA	3
		R	[PAN1]TGTTCTTTCAAGTGCCAA	
6.18	$(CT)_{16}$	F	[PAN3]CTCAACGAATGCTTTGGACA	3
		R	CCTCGCGGTAGCTTATTGTT	
2.6	$(CT)_{26}$	F	GGAGCAGGTAGAGTCCCATC	3
		R	[PAN1]CGTTTGAAAATTTATACCAAAATG	

Every multilocus PCR was performed in a total volume of 20 μL containing 2X Platinum ®Multiplex PCR Master Mix (Thermo Scientific), 10 % GC Enhancer (Thermo Scientific), 0.25 μM non-tailed primer, 0.75 μM tailed primer, 0.50 μM fluorophore-labelled primer, 20–30 ng of genomic DNA and distilled water up to volume. All amplifications were performed in a GeneAmp® PCR 9700 thermal cycler (Applied Biosystems, Carlsbad, CA, USA). The following thermal conditions were adopted for reactions of multiplex 1: 5 min at 95 ◦C, followed by five cycles at 95 ◦C for 30 s and at 60 ◦C for 30 s, which decreased by 1 ◦C with each cycle, and at 72 ◦C for 30 s; and then 35 cycles at 95 ◦C for 30 s, at 56 ◦C for 30 s, and at 72 ◦C for 30 s. For multiplex 2 and 3, the annealing temperature was modified. Three cycles were undertaken at 95 ◦C for 30 s and at 56 ◦C for 30

s, which decreased by 1 ◦C with each cycle followed by 35 cycles at 95 ◦C for 30 s, at 54 ◦C for 30 s, and at 72 ◦C for 30 s. All reactions were terminated with a final extension of 30 min at 72 ◦C.

Finally, the quality of the PCR amplicons was checked by electrophoresis on 2 % agarose/1× TAE gels containing 1× SYBR Safe DNA Gel Stain (Life Technologies). PCR products were dried at 65 ◦C. Capillary electrophoresis was performed in an ABI 3730 DNA Analyzer (Applied Biosystems). The SSR alleles were scored using PeakScanner 1.0 software (Applied Biosystems).

4.3 Genetic Structure of Populations

All of the analyses described below were performed for both the case study objects of this work with several exceptions that will be specified. Statistical analyses were performed with the GenAlEx6.5 [19] and POPGENE software 1.32 [20]. Specifically, the mean values of observed heterozygosity (Ho = H/n) were computed and compared to the expected heterozygosity ($He = 1 - p_i2$). Moreover, the PIC for each of the 29 SSR loci, along with the Na, Ne and number of private alleles, was also computed. Among private alleles, we considered only those with a frequency higher than 15 % in specific hybrids or populations and absent in others. Finally, marker allele and genotype frequencies for each locus were also determined.

For the comparison of 3 different years of seed production of an F1 hybrid variety, allele frequencies at each locus were also used to hypothesise the putative parental genotypes at the relative loci (e.g., $A_1A_1 \times A_1A_1$, $A_1A_1 \times A_1A_2$, $A_1A_1 \times A_2A_2/A_1A_2 \times A_1A_2$, $A_1A_1 \times A_2A_3/A_1A_2 \times A_1A_3$ and $A_1A_2 \times A_3A_4$). Moreover, the observed allele frequencies for each locus and year were matched to the most representative hypothetical profiles, and a bar chart showing the locus number of correspondences (in percentage) was drawn. Frequencies under 3 % were considered deriving from external contaminations.

A UPGMA dendrogram was constructed on the base of the genetic similarity matrix, whose mean coefficients were computed between all possible pairwise combinations of the individuals belonging to the core collection. Centroids were plotted to graphically represent the genetic similarity calculated with Rohlf's coefficient, which was calculated with the following formula GSij = m/(m + n). All genetic similarity analyses were conducted using the NTSYS software package v.2.21c [21].

The population structure of the sample collection was assessed using the clustering algorithm of STRUCTURE software [22]. All simulations were obtained by setting an admixture model without preliminary information on the population. We run a Markov Chain Monte Carlo (MCMC) model with 1,000,000 iterations and a burn-in of 200,000 samples under the assumption that the allele frequencies in the populations were correlated. Ten iterations were conducted for each value of the number of populations (K), with K ranging from 1 to 8. The method described by Evanno et al. [23] was used to evaluate the most likely estimation of K. The best value of K was calculated according to Evanno et al. [23]. Individuals showing a membership coefficient <85.0 % were considered admixed.

5 Conclusions

The results of this study indicated that our mapped SSR marker-based method developed for genotyping analyses is reliable and informative when applied to the characterisation of the genetic structure of different cultivated populations of radicchio. The comparison among eight different subgroups of plant materials, including two OPs, four F1s and two F2s, allowed us to discover fixed marker alleles and genotypes, which is potentially useful information for assessing the genetic identity of varietal seed lots and for protecting the legal rights of breeders. The comparative analysis of different representative commercial populations also verified that OP synthetics are characterised by lower similarity and heterozygosity than F1 hybrids (which can widely vary for these two parameters) and that F2 progenies usually show intermediate mean estimates of genetic uniformity and diversity. Moreover, from the analyses of 3 production years of an F1 hybrid variety, the same set of SSR markers highlighted significant differences among the commercial seed productions. This finding was observed due to the multi-allelic nature of the derived SSR genotypes and the high PIC values found for this set of marker loci. Furthermore, our panel of SSR markers was shown to be highly informative when exploited to test the stability of an F1 hybrid variety over different production years. Overall, this study provides a cost-efficient method to genotype the Red of Chioggia biotype at different levels of hybrid varieties development, ranging from the pre-selection of parental plants to the post-production verification of the obtained seeds, as well as for the identification of seed and plant lots to prevent potential frauds. In particular, this work supports the exploitation of highly discriminant SSR markers to test whether a new variety possesses uniformity and distinctiveness traits and so can legally gain access to registration and commercialisation. In fact, to be registered and released as new plant variety, some rigorous and specific requirements concerning the distinctness (D), uniformity (U), and stability (S) of the newly bred cultivar need to be satisfied. Although most of the existing methods exploited for DUS testing are expensive and time consuming, in this work, we demonstrated, through two different case studies, that SSR markers can be adopted as a complementary tool for morphological descriptors in DUS tests.

Supplementary materials

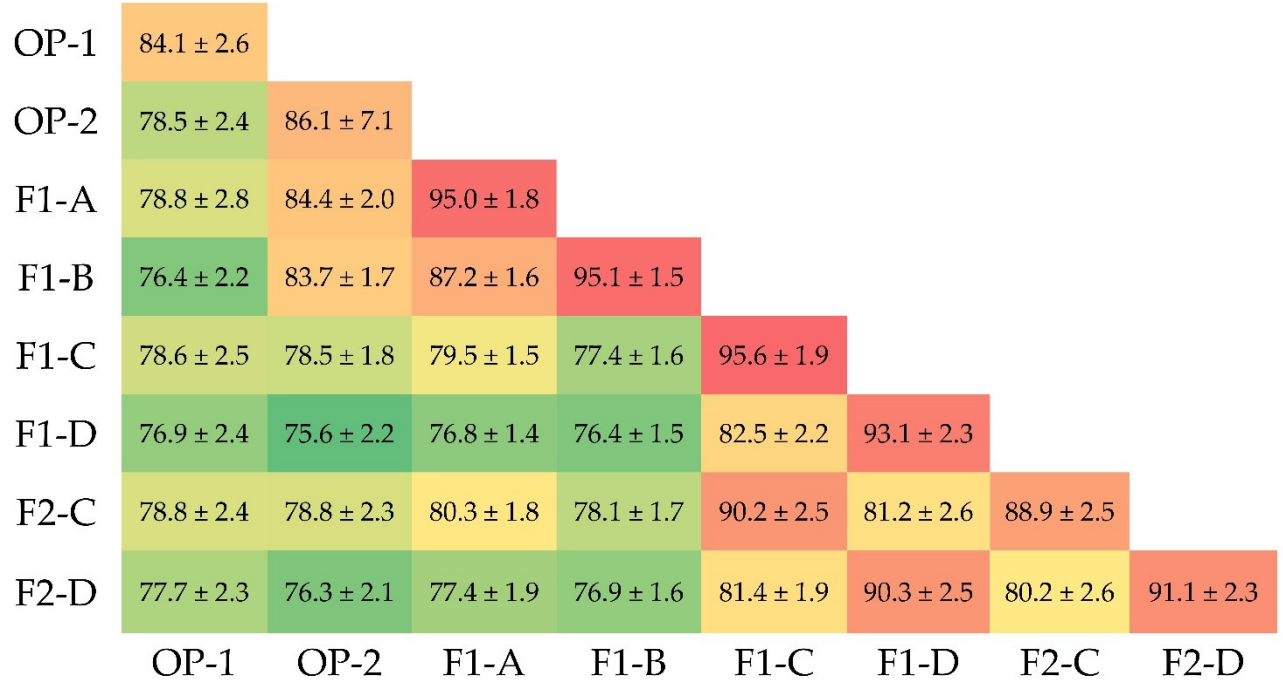


Figure S1. Pairwise genetic similarity matrix of the eight analysed populations (in percentage) based on Rohlf's genetic similarity coefficient. The high genetic similarity values are labelled in red, and the low values are labelled in green. Intermediate values are coloured in scale.

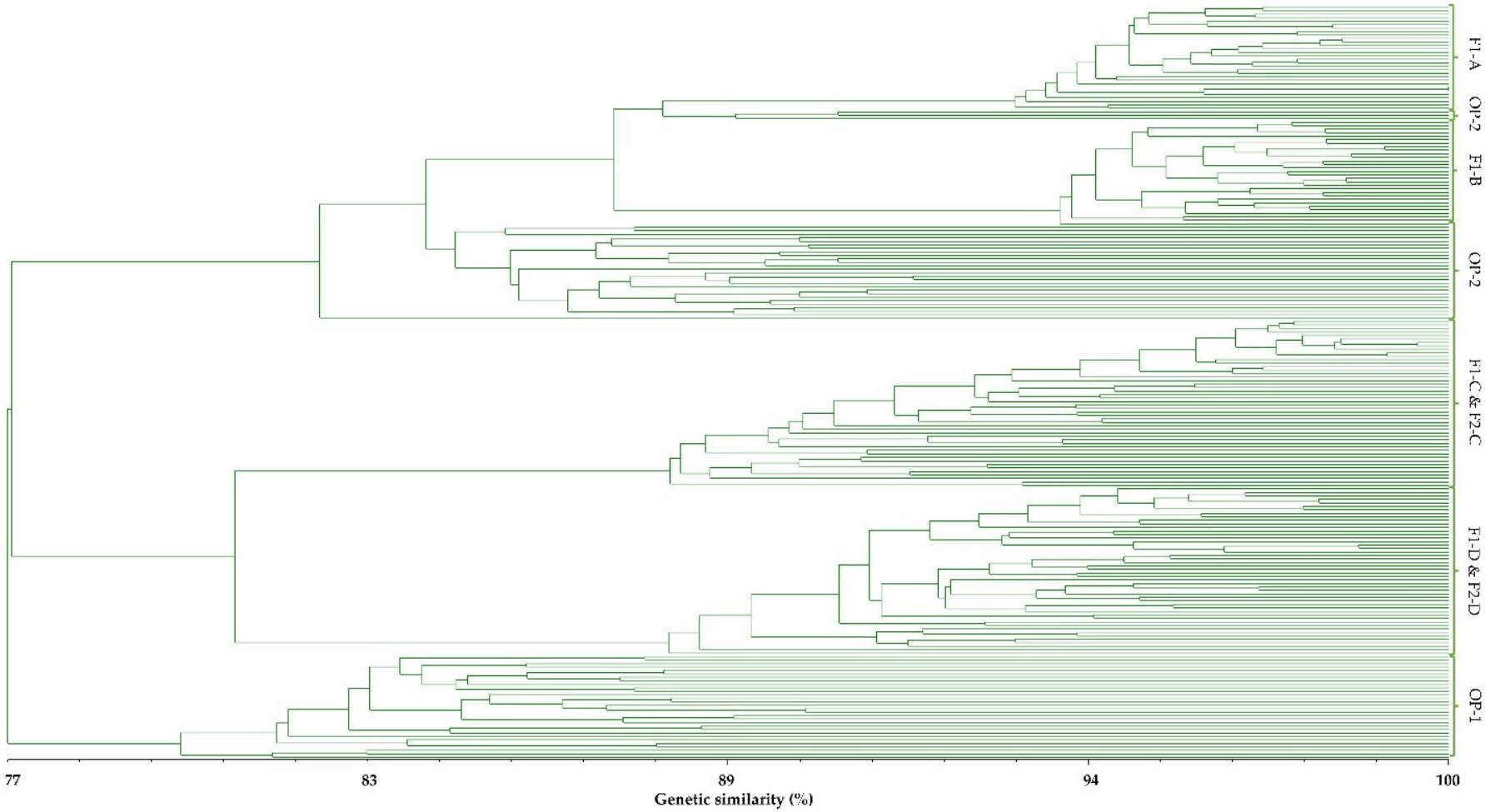


Figure S2. UPGMA tree of 216 samples analysed. The dendrogram was computed using the genetic similarity matrix of all pair-wise comparisons among samples.

Table S1. Number of alleles and PIC found across populations and loci for each F1 hybrid, F2 progeny and OP variety. In particular, statistics refer to the mean number of alleles (Na) and number of effective alleles (Ne) for each locus, population, and type of population and for the whole population. Moreover, PIC values for each locus of each population are presented.

ID	Mean	Mean	PIC	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
Locus	Na/Locus	Ne/Locus		Na	Ne	Na	Ne	Na	Ne	Na	Ne	Na	Ne	Na	Ne	Na	Ne	Na	Ne
M 2.4	3.1	2.2	0.7	5.0	3.1	8.0	5.1	2.0	1.0	2.0	2.0	2.0	1.1	2.0	2.0	4.0	1.7	2.0	2.0
M 4.10b	2.4	1.8	0.6	5.0	3.2	3.0	1.3	1.0	1.0	1.0	1.0	3.0	1.6	4.0	2.5	3.0	2.5	2.0	1.6
M 3.7	1.1	1.1	0.1	1.0	1.0	1.0	1.0	2.0	1.1	1.0	1.0	3.0	1.4	1.0	1.0	2.0	1.3	1.0	1.0
M 8.22	1.0	1.1	0.0	2.0	1.4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
M 2.5	2.6	2.1	0.8	4.0	2.9	5.0	3.0	2.0	2.0	2.0	1.9	1.0	1.0	2.0	2.0	4.0	2.1	2.0	2.0
M 5.15	2.4	1.5	0.3	5.0	3.0	3.0	1.2	1.0	1.0	1.0	1.0	1.0	1.0	3.0	1.7	5.0	2.3	1.0	1.0
M 4.11a	3.1	2.1	0.7	4.0	1.8	6.0	1.7	2.0	1.7	2.0	1.7	3.0	2.6	3.0	2.3	4.0	2.0	4.0	2.6
M 1.1	3.9	2.3	0.8	9.0	3.4	7.0	3.1	2.0	1.8	1.0	1.0	2.0	2.0	3.0	2.1	4.0	2.8	3.0	2.1
M 9.26	3.0	2.0	0.6	6.0	2.3	3.0	1.9	2.0	1.0	2.0	1.9	3.0	2.1	4.0	3.2	4.0	2.3	3.0	1.5
M 4.10a	1.6	1.7	0.5	2.0	1.7	2.0	1.4	1.0	1.0	2.0	1.9	2.0	2.0	2.0	1.9	2.0	2.0	2.0	1.9
M 4.11b	2.0	2.1	0.5	3.0	2.6	3.0	2.1	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
M 5.14	1.1	1.1	0.1	2.0	1.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0
M 5.13	2.1	2.1	0.7	4.0	3.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	4.0	2.9	2.0	2.0
M 6.16	2.9	1.8	0.6	6.0	2.6	4.0	2.4	2.0	1.2	3.0	1.7	2.0	2.0	2.0	1.2	4.0	2.2	2.0	1.4
M 7.19	2.4	2.0	0.6	2.0	1.9	5.0	2.5	3.0	1.7	1.0	1.0	3.0	2.5	2.0	2.0	4.0	2.2	2.0	2.0
M 3.8	3.5	2.1	0.7	4.0	1.6	8.0	4.3	2.0	1.1	2.0	1.0	1.0	1.0	4.0	3.3	4.0	1.3	4.0	3.3
M 4.12	3.9	2.5	0.8	7.0	4.1	9.0	6.1	3.0	1.3	4.0	2.7	1.0	1.0	2.0	1.3	2.0	1.8	4.0	1.4
M 6.17	2.9	2.3	0.7	2.0	2.0	8.0	4.0	2.0	1.1	4.0	3.0	4.0	3.3	2.0	1.1	2.0	1.7	3.0	2.5
M 1.2	3.1	2.1	0.8	6.0	2.5	6.0	3.1	3.0	1.6	2.0	1.2	3.0	2.1	2.0	1.9	4.0	2.6	2.0	1.5
M 1.3	3.1	2.2	0.6	4.0	1.7	7.0	4.1	4.0	2.1	2.0	2.0	2.0	2.0	2.0	2.0	4.0	2.2	2.0	2.0
M 9.25	2.5	2.1	0.7	4.0	2.4	3.0	2.1	4.0	2.7	3.0	1.7	3.0	2.1	2.0	1.5	2.0	2.0	2.0	2.0
M 8.23	2.6	1.8	0.7	3.0	2.7	3.0	1.5	4.0	1.6	2.0	2.0	2.0	1.1	3.0	2.1	3.0	1.3	3.0	1.9
M 2.6	3.9	3.0	0.8	10.0	5.2	7.0	4.8	3.0	2.4	2.0	2.0	3.0	2.3	3.0	2.1	4.0	2.9	2.0	2.0
M 7.21	1.9	1.7	0.5	1.0	1.0	2.0	2.0	2.0	1.8	2.0	2.0	2.0	1.7	2.0	2.0	3.0	1.2	3.0	2.1
M 9.27	3.4	2.4	0.8	7.0	4.2	5.0	2.6	4.0	3.9	3.0	2.3	3.0	1.5	2.0	1.5	4.0	1.3	2.0	2.0
M 7.2	2.5	2.2	0.7	6.0	4.3	1.0	1.0	2.0	1.9	3.0	2.1	2.0	2.0	3.0	2.1	3.0	2.1	2.0	1.9
M 8.24	3.3	2.5	0.8	6.0	3.9	4.0	2.0	2.0	1.3	3.0	2.6	4.0	2.2	4.0	3.1	3.0	2.4	4.0	2.1
M 6.18	3.1	2.2	0.7	6.0	4.6	4.0	1.4	1.0	1.0	3.0	2.6	2.0	1.1	4.0	3.1	4.0	1.5	3.0	2.1
M 3.9	3.5	2.6	0.8	8.0	4.0	7.0	3.2	2.0	2.0	2.0	2.0	3.0	2.5	3.0	2.5	3.0	2.7	3.0	2.1
Mean/ Pop	2.7	2.0	0.6	4.6	2.8	4.4	2.5	2.2	1.6	2.1	1.7	2.3	1.8	2.5	2.0	3.2	2.0	2.4	1.9

Table S2. Frequency of private alleles and alleles size for each locus of the populations with frequencies ≥15 % and population specific.

ID	Locus	Allele size (bp)	Freq
OP-1	M8.22	204	16.1 %
	M4.11b	196	16.7 %
	M6.17	254	50.0 %
	M1.3	256	16.7 %
	M9.27	278	28.3 %
	M 7.2	184	22.9 %
	M 6.18	173	19.0 %
		175	31.0 %
OP-2	M2.4	172	15.0 %
	M3.8	171	20.0 %
F1-B	M2.4	139	50.0 %
	M1.1	247	40.0 %
	M4.12	240	15.0 %
	M6.17	274	15.0 %
F1-C	M6.17	298	25.0 %

Table S3. The sizes and the frequencies of the most frequent alleles for each locus of all the varieties and F2 progenies are reported.

ID	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
LOCUS	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**
M2.4	166	22.0 %	160 - 164 - 174	< 5.0 %	170	98.0 %	139	50.0 %	170	97.0 %	166	47.0 %	170	75.0 %	166	58.0 %
	168	43.0 %	162 - 170 - 172	< 16.0 %							170	53.0 %	164	12.0 %	170	42.0 %
	170	30.0 %	166	20.0 %	168	2.0 %	166	50.0 %	166	3.0 %			168	10.0 %		
	158 - 164	< 3.0 %	168	32.0 %									166	3.0 %		
M4.10b	253 - 257	> 30.0 %							271	78.0 %	255	50.0 %	271	48.0 %		
	255	12.0 %	255	86.5 %	255	100.0 %	255	100.0 %	255	19.0 %	271	38.0 %	255	40.0 %	255	75.0 %
	259	4.0 %	257	9.6 %					253	3.0 %	252	6.0 %	252	13.0 %	271	25.0 %
	271	8.0 %	253	3.8 %							259	6.0 %				
M3.7	162	100.0 %	162	100.0 %	162	93.0 %	162	100.0 %	162	83.0 %	162	100.0 %	162	87.0 %	162	100.0 %
					164	7.0 %			164	11.0 %			164	13.0 %		
									157	6.0 %						
M8.22	201	84.0 %	201	100.0 %	201	100.0 %	201	100.0 %	201	100.0 %	201	100.0 %	201	100.0 %	201	100.0 %
	204	16.0 %														
M2.5	215	40.0 %	215	47.0 %	215	57.0 %	215	62.0 %	225	100.0 %	219	47.0 %	225	67.0 %	219	47.0 %
	219	42.0 %	219 - 225	10.0 %	221	43.0 %	221	38.0 %			221	53.0 %	221	15.0 %	221	53.0 %
	225	10.0 %	221	30.0 %									215	12.0 %	215	2.0 %
	221	8.0 %	223	3.0 %									219	7.0 %		
M5.15	257 - 271	40.0 %	271	91.7 %	271	100.0 %	271	100.0 %	271	100.0 %	255	22.0 %	271	60.0 %	271	100.0 %
	272	6.7 %	272	6.7 %							271	72.0 %	265	24.0 %		
	259	10.0 %	269	1.7 %							257	6.0 %	255	12.0 %		
	265	3.3 %											261	2.0 %		
M4.11a	187	58.3 %	187	66.7 %	185	98.0 %	185	60.0 %	191	61.0 %	185	39.0 %	191	57.0 %	185	79.0 %
	185	30.0 %	185	28.3 %					185	33.0 %	191	36.0 %	185	32.0 %	183	16.0 %

ID	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
LOCUS	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**
	172 - 183	3.3 %	183	5.0 %	187	2.0 %	187	40.0 %	187	6.0 %	187	17.0 %	187	7.0 %	187	5.0 %
	177	1.7 %									183	8.0 %	183	5.0 %		
M1.1	253	70.0 %	253	16.0 %	255	100.0 %	247	40.0 %	253	50.0 %	253	41.0 %	253	57.0 %	253	41.0 %
	255	30.0 %	255	84.0 %			255	60.0 %	255	50.0 %	255	59.0 %	255	43.0 %	255	59.0 %
M9.26	130	73.0 %	130	75.0 %	140	70.0 %	140	72.0 %	140	50.0 %	130	33.0 %	140	60.0 %	140	55.0 %
	132	17.0 %	162	13.0 %	162	30.0 %	162	28.0 %	130	31.0 %	140	56.0 %	130	37.0 %	130	24.0 %
	128 - 134	< 7.0 %	134 - 140 - 150 - 156	< 5.0 %					128	19.0 %	132	11.0 %	128 - 134	2.0 %	132 - 143	10.0 %
M4.10a	286	46.7 %	247	51.7 %	247	69.0 %	247	100.0 %	249	50.0 %	261	47.0 %	249 - 265	37.0 %	261	50.0 %
	247 - 249 - 265 - 272	7.0 % - 5.0 %	265	15.0 %	265	31.0 %			265	50.0 %	265	50.0 %	286	12.0 %	265	48.0 %
	261 - 267 - 273 - 284	< 5.0 %	267 - 273	10.0 % - 11.0 %							267	3.0 %	247	8.0 %	270	2.0 %
			249 - 255 - 269	< 0.08												
M4.11b	187	58.0 %	187	67.0 %	199	48.0 %	199	50.0 %	199	50.0 %	199	50.0 %	199	50.0 %	199	50.0 %
	185	30.0 %	185	28.0 %			203	50.0 %								
	172 - 183	3.0 %	183	5.0 %	203	52.0 %			203	50.0 %	203	50.0 %	203	50.0 %	203	50.0 %
	177	17.0 %														
M5.14	217	30.0 %	219	100.0 %	219	100.0 %	219	100.0 %	219	100.0 %	219	100.0 %	219	98.0 %	219	100.0 %
	219	70.0 %											217	2.0 %		
M5.13	243	40.0 %			243	55.0 %	243	100.0 %	245	50.0 %	245	53.0 %	245	43.0 %	245	53.0 %
	245	40.0 %	245	57.0 %	245	45.0 %			247	50.0 %	247	47.0 %	247	31.0 %	247	47.0 %
	247	20.0 %	243	43.0 %									243	24.0 %	241	2.0 %
													241	2.0 %		
M6.16	238	60.0 %	268	60.0 %	268	75.0 %	268	75.0 %	238	53.0 %	238	91.0 %	238	57.0 %	238	81.0 %

ID	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
LOCUS	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**
	246	30.0 %	266	30.0 %	270	18.0 %	265	20.0 %	268	47.0 %	268	9.0 %	268	31.0 %	268	19.0 %
	268 - 270	10.0 %	238 - 270	10.0 %	265	7.0 %	244	5.0 %					228 - 248	6.0 %		
M7.19	195	60.0 %	195	47.0 %	195	71.0 %	195	100.0 %	195	53.0 %	195	47.0 %	195	50.0 %	195	52.0 %
	203	40.0 %	199	7.0 %	203	28.0 %			203	33.0 %	199	53.0 %	203	45.0 %	199	47.0 %
			203	42.0 %	193	2.0 %			201	14.0 %			201	3.0 %	197	2.0 %
													197	2.0 %		
M3.8	169	20.0 %	178	40.0 %			169	98.0 %			181	36.0 %	181	88.0 %	181	34.0 %
	181	80.0 %	171	20.0 %	169	93.0 %	189	2.0 %	181	100.0 %	185	36.0 %	187	5.0 %	185	34.0 %
			169 - 181 - 183 - 189	< 20.0 %	189	7.0 %					183	19.0 %	183	5.0 %	178	24.0 %
											178	8.0 %			183	7.0 %
M4.12	224	40.0 %	226	30.0 %	226	87.0 %	234	52.0 %	228	100.0 %	224	86.0 %	228	68.0 %	224	84.0 %
	226	30.0 %	232 - 234	> 10.0 %	228	12.0 %	226	27.0 %			216	14.0 %	226	32.0 %	216	9.0 %
	220	20.0 %	220 - 224 - 228 -230	< 10.0 %	224	2.0 %	238	15.0 %							219 - 225	4.0 %
	214 - 222	10.0 %					222	7.0 %								
M6.17			264	38.0 %			266	50.0 %	298	25.0 %	264	94.0 %			264	44.0 %
	254	50.0 %	266	12.0 %	264	98.0 %	288	18.0 %	304	44.0 %			264	69.0 %	304	44.0 %
	264	50.0 %	284	8.0 %	284	3.0 %	274	15.0 %	264	19.0 %	304	6.0 %	304	31.0 %	284	13.0 %
							284	18.0 %	296	13.0 %						
M1.2	201	50.0 %	201	7.0 %	210	75.0 %	210	91.0 %	201	56.0 %	195	36.0 %	213	50.0 %	201	80.0 %
	206	10.0 %	206	21.0 %	206	23.0 %			213	42.0 %	201	64.0 %	201	35.0 %	195	20.0 %
	213	40.0 %	211	48.0 %	215	2.0 %			210	3.0 %			206	10.0 %		
			213	19.0 %									199	5.0 %		
M1.3	244	75.0 %	250	31.0 %	244	63.0 %	244	55.0 %	244	50.0 %	242	53.0 %	244	58.0 %	242	52.0 %
	250	7.0 %	232	22.0 %	232	25.0 %	250	45.0 %	250	50.0 %	244	47.0 %	250	35.0 %	244	48.0 %

ID	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
LOCUS	Allele size	Freq	Allele size	Freq	Allele size	Freq	Allele size	Freq	Allele size	Freq	Allele size	Freq	Allele size	Freq	Allele size	Freq
	252	2.0 %	244	29.0 %	227	4.0 %							248	2.0 %		
	256	17.0 %	246	10.0 %									242	5.0 %		
			242 - 248 - 252	< 3.0 %												
M9.25	177	57.0 %	175	3.0 %	177	34.0 %	177	73.0 %	169	47.0 %	175	22.0 %	169	50.0 %	175	53.0 %
	195	27.0 %	177	48.0 %	195	47.0 %	183	25.0 %	183	50.0 %	177	78.0 %	183	50.0 %	177	47.0 %
	199	13.0 %	183	48.0 %	183	17.0 %	193	2.0 %	177	3.0 %						
	183	3.0 %			187	2.0 %										
M8.23	231	50.0 %	221	81.0 %	221	78.0 %	221	50.0 %	229	97.0 %	229	53.0 %	229	87.0 %	229	62.0 %
	229	27.0 %	231	14.0 %	223	14.0 %	229	50.0 %	231	3.0 %	247	44.0 %	231	13.0 %	247	36.0 %
	221	23.0 %	229	5.0 %	229	5.0 %					243	3.0 %			243	2.0 %
					243	3.0 %										
M2.6	187	35.0 %	193	33.0 %	193	50.0 %	193	50.0 %	193	44.0 %	197	53.0 %	207	42.0 %	201	57.0 %
	211	13.0 %	197 - 205	17.0 % - 18.0 %	187	40.0 %	205	50.0 %	197	47.0 %	201	44.0 %	197	33.0 %	197	43.0 %
	183 - 185 - 193 - 197 - 199	5.0 %- 1.0 %	195 - 213	12.0 %- 13.0 %	205	10.0 %			207	8.0 %	203	3.0 %	193	20.0 %		
	175 - 205 - 213	< 7.0 %	187 - 199	< 5.0 %									199	5.0 %		
M7.21	247	100.0 %	237	47.0 %	237	31.0 %	237	50.0 %	247	72.0 %	237	47.0 %	247	90.0 %	237	54.0 %
			247	53.0 %	247	69.0 %	247	50.0 %	266	28.0 %	247	53.0 %	237 - 266	5.0 %	247	46.0 %
M9.27	264	59.0 %	306	32.0 %	266	36.0 %	264	59.0 %	301	81.0 %	301	81.0 %	301	87.0 %	301	57.0 %
	303	22.0 %	266 -278	21.0 % - 28.0 %	301	25.0 %	303	22.0 %	299	17.0 %	303	19.0 %	303	7.0 %	303	43.0 %
	301	19.0 %	262 - 301 - 303 - 308	< 8.0 %	264	21.0 %	301	19.0 %	303	3.0 %			297	5.0 %		

ID	OP-1		OP-2		F1-A		F1-B		F1-C		F1-D		F2-C		F2-D	
LOCUS	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**	**Allele size**	**Freq**
					305	14.0 %							299	2.0 %		
M7.2	175	33.0 %	167	100.0 %	132	38.0 %	167	60.0 %	132	58.0 %	132	50.0 %	132	60.0 %	132	63.0 %
	167 - 184	23.0 %			167	62.0 %	132	35.0 %	175	42.0 %	175	47.0 %	175	33.0 %	175	37.0 %
	163 - 165 - 173	< 8.0 %					169	5.0 %			162	3.0 %	167	7.0 %		
M8.24	242	35.0 %	242	67.0 %	242	86.0 %	250	50.0 %	242	47.0 %	267	39.0 %	242	52.0 %	267	50.0 %
	250 - 268	20.0 % - 28.0 %	252	20.0 %			267	30.0 %	267	47.0 %	273	33.0 %	267	38.0 %	273	47.0 %
	270	11.0 %	268	5.0 %			242	20.0 %	244	3.0 %	269	25.0 %	251	10.0 %	242	2.0 %
	232 - 252	4.0 % - 2.0 %	270	8.0 %					251	3.0 %	271	3.0 %			265	2.0 %
M6.18	175	31.0 %	187	83.0 %	187	100.0 %	187	50.0 %	191	97.0 %	191	41.0 %	191	81.0 %	187	54.0 %
	187	21.0 %	181 - 191	7.0 %			191	30.0 %	193	3.0 %	187	31.0 %	187 - 189	9.0 %	189	45.0 %
	173 - 179	19.0 %	179	2.0 %			189	20.0 %			183	25.0 %	193	2.0 %	181	2.0 %
	183 - 191	< 7.0 %									193	3.0 %				
M3.9	223 - 239	32.0 % - 35.0 %	223	40.0 %	223	52.0 %	225	50.0 %	223	36.0 %	221	53.0 %	223	48.0 %	233	52.0 %
	221 - 233	13.0 % - 8.0 %	239	37.0 %	225	48.0 %	239	50.0 %	239	50.0 %	239	28.0 %	239	28.0 %	221	47.0 %
	217	5.0 %	225	12.0 %					221	14.0 %	233	19.0 %	221	23.0 %	239	2.0 %
	227 - 229 - 231	< 3.0 %	221 - 227 - 229 - 231	< 5.0 %												

Table S4. The sizes of the most frequent genotype, frequency (Freq) and observed heterozygosity (Ho) for each locus of all the varieties and F2 progenies are reported.

ID	OP-1			OP-2			F1-A			F1-B			F1-C			F1-D			F2-C			F2-D		
Locus	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**
M2.4	168 - 168 166 - 168 168 - 170	26.7 % 26.7 % 23.3 %	66.7 %	163 - 168 170 - 170 166 - 170	20.0 % 13.3 % 13.3 %	80.0 %	170-170	96.7 %	3.3 %	139-166	100.0 %	100.0 %	170-170	94.44 %	5.6 %	166-170	94.4 %	94.4 %	170-170 170-164 170-168	50.0 % 23.3 % 20.0 %	50.0 %	166 - 170 166 - 166 170 - 170	50.0 % 33.3 % 16.7 %	50.0 %
M4.10b	257 - 257 253 - 253 255 - 255	30.0 % 23.3 % 6.7 %	26.9 %	255 - 255 255 - 257 253 - 253	76.9 % 19.2 % 3.8 %	19.2 %	255-255	100.0 %	0.0 %	255-255	100 %	0.0 %	271-271	61.11 %	33.3 %	255-255	44.4 %	62.5 %	271-255 271-271 255-255	23.3 % 20.0 % 20.0 %	45.8 %	271 - 271 251 - 271 251 - 251	60.0 % 20.0 % 13.3 %	21.4 %
M3.7	162-162	100.0 %	0.0 %	162 - 162	100.0 %	0.0 %	162-162	93.3 %	0.0 %	162-162	100 %	0.0 %	162-162	100 %	0.0 %	162-162	100.0 %	0.0 %	162-162 164-164	86.7 % 13.3 %	0.0 %	162-162	100.0 %	0.0 %
M8.22	201 - 201 201 - 204	71.4 % 25.0 %	25.0 %	201 - 201	100.0 %	0.0 %	201-201	100.0 %	0.0 %	201-201	100 %	0.0 %	201-201	100 %	0.0 %	201-201	100.0 %	0.0 %	201-201	100.0 %	0.0 %	201-201	100.0 %	0.0 %
M2.5	215 - 221 219 - 219	32.1 % 25.0 %	56.7 %	215 - 215 221 - 221	32.1 % 25.0 %	33.3 %	215-215 221-221	46.7 % 33.3 %	17.2 %	215-215 215-221	0.5 % 20.0 %	20.7 %	225-225	100 %	0.0 %	219-221	94.4 %	94.4 %	225-225 215-225	56.7 % 16.7 %	36.7 %	219 - 221 219 - 219	56.7 % 23.3 %	64.3 %
M5.15	257 - 257 271 - 271 259 - 271	26.7 % 26.7 % 6.7 %	40.0 %	225 - 215 271 - 271 273 - 273	21.4 % 90.0 % 6.7 %	3.3 %	271-271	100.0 %	0.0 %	271-271	100 %	0.0 %	271-271	100 %	0.0 %	255-271 271-271	33.3 % 50.0 %	44.4 %	271-271 265-265 255-271	53.3 % 20.0 % 10.0 %	17.2 %	271-271	100.0 %	0.0 %
M4.11a	187 - 187 185 - 185	46.7 % 16.7 %	23.3 %	187 - 187 185 - 185	53.3 % 16.7 %	40.0 %	185-185	96.7 %	3.3 %	185-187	73.3 %	73.3 %	185-185 191-191	61.11 % 33.33 %	0.0 %	185-185 187-191	27.8 % 33.3 %	72.2 %	191 - 191 185 - 185	40.0 % 26.7 %	33.3 %	185 - 185 183 - 183	76.7 % 13.3 %	3.5 %
M1.1	253 - 253 255 - 255	63.3 % 23.3 %	53.3 %	255 - 255 253 -253	82.8 % 13.8 %	60.0 %	255-255	100.0 %	0.0 %	247-247 247-255	20.0 % 20.0 %	40.0 %	253-255	100 %	100.0 %	253-255	77.8 %	82.4 %	253 - 253 253 - 255	40.0 % 33.3 %	33.3 %	253 - 255 253 - 253	40.0 % 36.7 %	42.9 %
M9.26	130 - 130 132 - 132	66.7 % 10.0 %	33.3 %	130 - 130 130 - 162	63.0 % 18.5 %	26.7 %	140-162	60.0 %	60.0 %	140-140 140-162	43.3 % 56.7 %	56.7 %	130-140 128-140	61.11 % 38.89 %	100.0 %	130-140	66.7 %	88.9 %	130 - 140 140 - 140	66.7 % 23.3 %	80.0 %	130 - 140 132 - 144	46.7 % 20.0 %	69.0 %
M4.10a	273 - 286 286 - 286 273 - 273	26.7 % 26.7 % 6.7 %	13.3 %	247 - 247 247 -265 247 - 267	33.3 % 13.3 % 13.3 %	3.5 %	247-265	60.0 %	62.1 %	247-247	100.0 %	0.0 %	249-265	88.89 %	88.9 %	261-265	88.9 %	94.4 %	249 - 265 265 - 265 249 - 249	20.0 % 13.3 % 10.0 %	76.7 %	261 - 265 261 - 261	40.0 % 26.7 %	44.8 %
	199 - 203	60.0 %		199 - 203	86.7 %	90.0 %	199-203	96.7 %	96.7 %	199-203			199-203	100 %		199-203			199-203			199-203	100.0 %	

ID	OP-1			OP-2			F1-A			F1-B			F1-C			F1-D			F2-C			F2-D		
Locus	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**	**Genotype**	**Freq**	**Ho Locus**
M4.11b	196 - 199	33.3 %	93.3 %	199 - 199	6.7 %						100.0 %	100.0 %			100.0 %		100.0 %	100.0 %		100.0 %	100.0 %			100.0 %
M5.14	219 - 219 217 - 219	56.7 % 20.0 %	20.7 %	219 - 219	100.0 %	0.0 %	219-219	100.0 %	0.0 %	219-219	100.0 %	0.0 %	219-219	100 %	0.0 %	219-219	100.0 %	0.0 %	219-219	96.7 %	3.3 %	219-219	100.0 %	0.0 %
M5.13	243 - 247 243 - 245 243 - 243	28.0 % 28.0 % 20.0 %	63.3 %	243 - 245 245 -245 243 -243	60.0 % 26.7 % 13.3 %	60.0 %	243-245	90.0 %	90.0 %	243-243	100.0 %	0.0 %	245-247	100.00 %	100.0 %	245-247	94.4 %	94.4 %	243 - 245 245 - 247 245 - 245	30.0 % 26.7 % 20.0 %	79.3 %	247 - 247 249 - 249 245 - 245	53.3 % 26.7 % 16.7 %	56.7 %
M6.16	238 - 238 246 - 246 238 - 246	43.3 % 20.0 % 6.7 %	27.6 %	268 - 268 238 - 268 266 - 266	33.3 % 16.7 % 16.7 %	50.0 %	268-268	30.0 %	21.4 %	268-268	43.3 %	31.8 %	238-268	94.44 %	94.4 %	238-238	77.8 %	17.7 %	238 - 238 268 - 268 238 - 268	46.7 % 20.0 % 16.7 %	20.7 %	238-238	63.3 %	31.0 %
M7.19	195 - 203 195 - 195	73.3 % 23.3 %	73.3 %	195 - 203 195 - 195	73.3 % 6.7 %	90.0 %	195-203	53.3 %	58.6 %	195	100.0 %	0.0 %	195-203	66.67 %	94.4 %	195-199	94.4 %	94.4 %	193 - 203	86.7 %	96.7 %	195 -199 195 - 195	43.3 % 30.0 %	46.7 %
M3.8	181 - 181 169 - 181 169 - 187	90.5 % 4.8 % 28.6 %	36.7 %	178 - 178 171 - 171 171 - 181	30.0 % 10.0 % 10.0 %	53.3 %	169-169	86.7 %	13.3 %	169-169	93.3 %	3.5 %	181-181	100.00 %	0.0 %	181-185	61.1 %	88.9 %	181 - 181	76.7 %	17.2 %	181 - 185 181 - 181 185 - 185	30.0 % 16.7 % 16.7 %	44.8 %
M4.12	224 - 224 226 - 226	36.7 % 23.3 %	10.0 %	226 - 234 220 - 220	13.3 % 6.7 %	60.0 %	226-226	73.3 %	26.7 %	226-234 234-234	30.0 % 33.3 %	56.7 %	228-228	100 %	0.0 %	224-224	77.8 %	16.7 %	228 - 228 226 - 228	46.7 % 16.7 %	24.0 %	224 - 224 216 - 224	78.6 % 10.7 %	10.7 %
M6.17	254 - 254 264 - 264 254 - 264	42.3 % 42.3 % 15.4 %	15.4 %	264 - 264 286 - 286 266 - 266	32.0 % 24.0 % 12.0 %	12.0 %	264-264	63.3 %	5.0 %	266-266 266-288	20.0 % 20.0 %	60.0 %	298-304 264-304	22.22 % 16.67 %	87.5 %	264-264	38.9 %	12.5 %	264 - 264 304 - 304 264 - 304	36.7 % 13.3 % 10.0 %	16.7 %	264 - 264 304 - 304 264 - 304	37.5 % 37.5 % 12.5 %	12.5 %
M1.2	201 - 201 201 - 213 213 - 213	32.1 % 32.1 % 21.4 %	44.8 %	211 - 211 206 - 206 213 - 213	44.8 % 20.7 % 17.2 %	6.9 %	206-210 210-210	43.3 % 46.7 %	50.0 %	210-210	76.7 %	17.9 %	201-213	83.33 %	88.9 %	195-201	72.2 %	72.2 %	201 - 213 213 - 213 201 - 201	40.0 % 23.3 % 13.3 %	56.7 %	201 - 201 195 - 201	60.0 % 40.0 %	40.0 %
M1.3	244 - 244 244 - 256 244 - 250	53.3 % 30.0 % 10.0 %	46.7 %	244 - 244 232 - 250 250 - 250	20.7 % 20.7 % 17.2 %	44.8 %	232-244 244-244	33.3 % 36.7 %	60.7 %	244-244 244-250	33.3 % 43.3 %	43.3 %	244-250	100 %	100.0 %	242-244	94.4 %	94.4 %	244 - 250 244 - 244	40.0 % 33.3 %	53.3 %	242 - 242 244 - 244 242 - 244	36.7 % 33.3 % 30.0 %	30.0 %
M9.25	177 - 195	40.0 %		177 - 183	93.3 %		177-195	56.7 %	89.7 %	177-177	50.0 %	46.7 %	169-183	94.4 %		175-177	66.7 %	22.2 %	169 - 183			175 - 177	53.3 %	53.3 %

ID	OP-1			OP-2			F1-A			F1-B			F1-C			F1-D			F2-C			F2-D		
Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus	Genotype	Freq	Ho Locus
	177 - 177	30.0 %	66.7 %	175 - 183	3.3 %	100.0 %				177-183	43.3 %				100.0 %					100.0 %	100.0 %	175 - 175	26.7 %	
	177 - 199	10.0 %		175 - 177	3.3 %																	177 - 177	20.0 %	
M8.23	231 - 231	37.5 %	29.2 %	169 - 169	65.5 %	34.5 %	221-221	66.7 %	17.2 %	221-229	100 %	100.0 %	229-229	94.4 %	5.6 %	229-247	88.9 %	94.4 %	229 - 229	73.3 %	26.7 %	229 - 247	72.4 %	75.9 %
	221 - 221	16.7 %		169 - 183	24.1 %														221 - 229	23.3 %		229 - 229	24.1 %	
	187 - 187	20.0 %		193 - 197	13.3 %														193 - 207	40.0 %		197 - 201	53.3 %	
M2.6	187 - 213	10.0 %	63.3 %	193 - 205	13.3 %	73.3 %	187-193	76.7 %	100.0 %	193-205	100.0 %	100.0 %	193-197	77.8 %	94.4 %	197-201	88.9 %	94.4 %	197 - 207	23.3 %	73.3 %	201 - 201	30.0 %	53.3 %
	187 - 211	10.0 %		193 - 213	13.3 %														197 - 197	20.0 %		197 - 197	16.7 %	
				237 - 247	37.9 %														247 - 247	80.0 %		237 - 247	40.0 %	
M7.21	247 - 247	100.0 %	0.0 %	247 - 247	34.5 %	37.9 %	237-247	60.0 %	62.1 %	237-247	100.0 %	100.0 %	247-266	55.6 %	55.6 %	237-247	94.4 %	94.4 %	247 - 266	10.0 %	20.0 %	237 - 237	33.3 %	43.3 %
				237 - 237	27.6 %																	247 - 247	23.3 %	
	266 - 278	16.7 %		264 - 266	31.0 %					264-303	33.3 %					301-301	61.1 %					301 - 303	43.3 %	
M9.27	266 - 306	10.0 %	80.0 %	264 - 264	27.6 %	58.6 %	266-301	30.0 %	92.9 %	264-301	26.7 %	72.4 %	301-301	72.2 %	16.7 %	301-303	38.9 %	38.9 %	301 - 301	73.3 %	26.7 %	301 - 301	33.3 %	44.8 %
	278 - 306	10.0 %		266 - 266	13.8 %																	303 - 303	20.0 %	
M7.2	174 - 184	26.1 %	37.5 %	167 - 167	100.0 %	0.0 %	132-167	70.0 %	70.0 %	132-167	66.7 %	70.0 %	132-175	83.3 %	83.3 %	132-175	94.4 %	100.0 %	132 - 175	66.7 %	80.0 %	132 - 175	7333.3 %	73.3 %
	167 - 167	21.7 %																	132 - 132	20.0 %		132 - 132	2666.7 %	
	242 - 268	22.2 %		242 - 242	60.0 %											269-273	38.9 %		242 - 267	40.0 %		269 - 273	4000.0 %	
M8.24	250 - 268	18.5 %	70.4 %	252 - 270	13.3 %	26.7 %	242-242	70.0 %	27.6 %	250-267	60.0 %	100.0 %	242-267	94.4 %	100.0 %	267-273	27.8 %	83.3 %	242 - 242	26.7 %	60.0 %	269 - 269	2666.7 %	48.3 %
	242 - 270	11.1 %		252 - 252	10.0 %																	273 - 273	2333.3 %	
	173 - 173	10.7 %		187 - 187	73.3 %											183-187	44.4 %		191 - 191	63.3 %		187-187	4000.0 %	
M6.18	175 -175	10.7 %	51.7 %	181 - 181	3.3 %	3.7 %	187-187	100.0 %	0.0 %	187-191	60.0 %	100.0 %	191-191	94.4 %	5.6 %	191-191	33.3 %	56.3 %	187 - 191	16.7 %	34.5 %	183-183	3333.0 %	21.4 %
	223 -239	20.0 %		223 - 239	73.3 %											221-239	55.6 %		223 - 223	40.0 %		221-233	5333.0 %	
M3.9	239 - 239	16.7 %	70.0 %	225 - 225	6.7 %	86.7 %	223-225	96.7 %	96.7 %	225-239	100.0 %	100.0 %	223-239	72.2 %	100.0 %	233-239	38.9 %	94.4 %	221 - 239	40.0 %	56.7 %	233-233	2333.0 %	56.7 %

Table S5. The mean degree of expected and observed heterozygosity (He ± Std.Dev. and Ho ± Std.Dev.) for each ID: F1 hybrids, F2 progenies and OPs, singularly and by category (*).

ID	He (%)	Ho (%)
OP-1	57 ± 0.04	43 ± 9.73
OP-2	49 ± 0.05	40 ± 8.29
F1-A	29 ± 0.40	39 ± 6.59
F1-B	35 ± 0.05	49 ± 5.23
F1-C	36 ± 0.05	53 ± 4.2
F1-D	45 ± 0.04	67 ± 11.9
F2-C	45 ± 0.03	46 ± 8.48
F2-D	42 ± 0.04	40 ± 10.19
F1*	36 ± 6.42	52 ± 11.56
F2*	44 ± 2.18	43 ± 4.35
OP*	53 ± 5.91	42 ± 1.75

Table S6. **(a)** Marker loci showing allele frequencies of the private alleles only present in year 2016 (italic-bold), compared to the 2 previous production years of the hybrid variety analysed. **(b)** Marker loci that show alleles with meaningful differential frequencies (italic-bold) among the 3 years.

(a)

Locus		**M1.2**			**M1.3**			**M2.6**			**M8.24**			**M9.25**	
Allele\year	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**
A_1	48.9 %	50.0 %	34.5 %	26.6 %	19.8 %	22.8 %	0.6 %			99.5 %	98.9 %	87.4 %	99.5 %	100.0 %	85.2 %
A_2	51.1 %	50.0 %	50.0 %			2.5 %	30.8 %	49.5 %	39.0 %			0.5 %	0.5 %		
A_3			***15.5 %***	72.9 %	79.3 %	66.1 %	0.6 %				1.1 %				***14.8 %***
A_4				0.5 %	0.9 %		67.0 %	50.5 %	50.0 %	0.5 %					
A_5						***8.6 %***	1.1 %					***12.1 %***			
A_6									***11.1 %***						

(b)

Locus		**M8.23**			**M9.26**			**M9.27**	
Allele\year	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**	**2014**	**2015**	**2016**
A_1	1.1 %			***25.8 %***	0.5 %		30.8 %	24.2 %	25.0 %
A_2	21.7 %	50.0 %	34.6 %	48.4 %	75.8 %	30.3 %	27.3 %	49.5 %	35.8 %
A_3	0.5 %			0.6 %		***46.1 %***		0.5 %	
A_4	50.0 %	50.0 %	48.9 %	24.7 %	23.7 %	23.6 %	20.4 %	25.8 %	24.3 %
A_5			1.1 %	0.6 %			0.6 %		1.4 %
A_6	1.6 %						***19.8 %***		***13.5 %***
A_7	***24.5 %***		***15.4 %***				0.6 %		
A_8	0.5 %							0.6 %	

Table S7. SSR primer tail and dye. List of the primer tails used with their sequence and corresponding dye.

Universal primer	Sequence 5'-3'	Dye
M13	TTGTAAAACGACGGCCAGT	6-FAM
PAN1	GAGGTAGTTATTGTGGAGGAC	VIC
PAN2	GGAATTAACCGCTCACTAAAG	NED
PAN3	TGTAGAAAGACGAAGGGAAGG	PET

References

1. Funk, V.A.; Susanna, A.; Stuessy, T.F.; Bayer, R.J. *Systematics, Evolution, and Biogeography of Compositae*; International Association for Plant Taxonomy, Institute of Botany, University of Vienna: Vienna, Austria, 2009; ISBN 3950175431.
2. Bellamy, A.; Mathieu, C.; Vedel, F.; Bannerot, H. Cytoplasmic DNAs and nuclear rDNA restriction fragment length polymorphisms in commercial witloof chicories. *Theor. Appl. Genet.* **1995**, *91*, 505–509.
3. 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 Breed.* **2003**, *122*, 178–183.
4. Charcosset, A.; Moreau, L. Use of molecular markers for the development of new cultivars and the evaluation of genetic diversity. *Euphytica* **2004**, *137*, 81–94.
5. Cadalen, T.; Mörchen, M.; Blassiau, C.; Clabaut, A.; Scheer, I.; Hilbert, J.L.; Hendriks, T.; Quillet, M.C. Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Mol. Breed.* **2010**, *25*, 699–722.
6. Gonthier, L.; Blassiau, C.; Mörchen, 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.
7. Muys, C.; Thienpont, C.N.; Dauchot, N.; Maudoux, O.; Draye, X.; Cutsem, P.V. Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var. sativum). *Plant Breed.* **2014**, *133*, 130–137.
8. Palumbo, F.; Qi, P.; Batista Pinto, V.; 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. *Front. Plant Sci.* **2019**, *10*, 276.
9. 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.). *Plant Genom.* **2016**, *1*, 65–87.
10. 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 Genetics. *BMC Res. Notes* **2015**, *8*, 1–12.
11. Singh, B.D.; Singh, A.K. *Marker-Assisted Plant Breeding: Principles and Practices*; Springer: New Delhi, India, 2015; ISBN 9788132223160.

12. Van Stallen, N.; Vandenbussche, B.; Verdoodt, V.; De Proft, M. Construction of a genetic linkage map for witloof (*Cichorium intybus* L. var. foliosum Hegi). *Plant Breed.* **2003**, *122*, 521–526.
13. Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **1980**, *32*, 314–331.
14. Brumlop, S.; Finckh, M.R. *Applications and Potentials of Marker Assisted Selection (MAS) in Plant Breeding*; Bundesamt für Naturschutz: Bonn, Germany, 2011; Volume 298, ISBN 9783896240330.
15. Palumbo, F.; Galla, G.; Martínez-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* **2017**, *9*, 32.
16. Barcaccia, G.; Ghedina, A.; Lucchin, M. Current Advances in Genomics and Breeding of Leaf Chicory (*Cichorium intybus* L.). *Agriculture* **2016**, *6*, 50.
17. Barcaccia, G.; Tiozzo, C.S. New Male Sterile *Chicorium* spp. Mutant, Parts or Derivatives, Where Male Sterility is Due to a Nuclear Recessive Mutation Linked to a Polymorphic Genetic Marker, Useful for Producing Mutant F1 Hybrids of *Chicorium* spp. EU Patent No. WO2012163389-A1, 6 December 2012.
18. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **2000**,*18*, 233.
19. Peakall, R.; Smouse, P.E. GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **2012**, *28*, 2537–2539.
20. Yeh, F.; Boyle, T. Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian J. Bot.* **1997**, *130*, 129–157.
21. Rohlf, F.J. *NTSYS-pc: Microcomputer Programs for Numerical Taxonomy and Multivariate Analysis*; American Statistician: New York, NY, USA, 1987; Volume 41, ISBN 0925031313.
22. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **2003**, *164*, 1567–1587.
23. 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.

Chapter III

Assessing the genetic distinctiveness of endive (*Chicorium endivia*) experimental materials using SSR and SNP markers

---

Keywords: RADseq, genetic marker discovery, microsatellite markers, genetic diversity, plants breeder rights

Abstract

The characterisation of genetic diversity in elite breeding material is crucial for registration and protection of future varieties. Moreover, population structure and information about genetic distances of firms' material is essential for crop breeding programs. The purpose of our research was to analyse the genetic diversity of plants belonging to 32 endive (*Cichorium endivia* L.) breeding plants using both heterologous chicory-derived microsatellite markers and single nucleotide polymorphism (SNP) markers. Only 14 out of 29 heterologous SSR markers retrieved from *Cichorium intybus* were successfully transferred and 6 of them resulted monomorphic. In order to overcome the limits deriving from the use of a low number of informative microsatellite loci, a second SNP-based approach was attempted. A set of 4,621 SNPs, produced by means of a Radseq approach, was able to discriminate the 32 endive materials and, in particular, 50 loci separated the curly endive group from the escarole endive one. Also, the SNP-based dendrogram and the PCoA analysis support the clear separation of these two cultivars types and the unambiguous discrimination of plant materials. Finally, our work was able to evaluate the DUS test requirements. Firstly, we evaluated distinctiveness among phenotypically similar breeding plants; secondly, we calculated observed homozygosity in order to predict the uniformity and stability of progenies. Overall, our study represents the first genotypic analysis of endive breeding materials in which thousands of discriminant SNP markers were identified at the genomic level.

1 Introduction

Endive (*Cichorium endivia* L., 2n = 2x = 18) is a leafy green vegetable, belonging to the Asteraceae family [1,2]. It can be further classified into two cultivar types: curly endive (*C. endivia* var. *crispum Lam*.) and escarole or smooth endive (*C. endivia* var. *latifolium* Lam.) [3]. From a reproduction point of view, endive is an autogamous species with a rate of outcrossing around 1%. The natural populations are composed by a mixture of highly homozygous lines [4,5]. This species is utilised for the preparation of salads along with lettuce and chicory, which are consumed in increasing amounts due to their healthy properties. There are several biological activities and properties attributed to this vegetable, such as anti-inflammatory, antioxidant, and hepatoprotective effects [6,7]. Consequently, an increased interest by consumers for ready-to-eat food with health benefits are leading companies to invest in breeding programs and varieties protection also in minor species such as endive. Typically, commercialised cultivars mostly consist of pure lines [8]; thus the risk of elite genotype plagiarism phenomena in this species is very high. Therefore, the protection of a registered variety is important for identification of any essentially-derived variety (EDV) as well as legal protection of germplasm stocks [9]. As a matter of fact, the International Seed Federation (ISF) requires the owner to provide the proofs to resolve EDV disputes about different crops. Depending on the species, different coefficients are exploited and a similarity threshold is set to genetically

differentiate the varieties (i.e. in lettuce the coefficient is Jaccard and the threshold is 96.00 %) [10]. Companies verify distinctiveness, uniformity and stability (DUS test) of plant materials, that are three major requirements for the registration of varieties. Currently, molecular markers, among other advantages, are useful methods for adding specific information at the registration step and for legal protection of varieties. The most used biotechnological methods are based on the use of Simple Sequence Repeats (SSR) and Single Nucleotide Polymorphism (SNP) markers for genotyping plant materials due to their codominant nature, high frequency in all genomes and high reproducibility among laboratories. Moreover, these DNA markers are easy to use, cheap, flexible, quick and multiallelic [9].

The most studied species among the genus *Cichorium* is *C. intybus* , because of the availability of its draft genome [11] and relevant molecular assays pertaining to breeding [12-15]. In particular, for *C. intybus* an informative panel of SSR markers [13,14] demonstrated to be very effective for genetic characterisation both of hybrids parental materials and synthetic varieties. Another research study provides a genetic consensus map for the *Cichorium spp*. [16] that includes markers from a *C. intybus* × *C. endivia* cross [17]. To date, except for this molecular information, the only source of molecular data available for *C. endivia* derives from a transcriptomic analysis performed by Testone et al., [5] where the differences among curly and smooth leafed endive accessions were investigated analysing allelic and gene transcriptional variation.

Our research evaluated 32 elite endive lines (F5) using heterologous chicory-derived microsatellites and single nucleotide polymorphisms for their applicability in the DUS test. In particular, we assessed the genetic distinctiveness (D) of materials and we calculated the observed homozygosity in order to predict the uniformity (U) of progenies.

2 Materials and Methods

2.1 Plant materials

Thirty-two experimental lines of endive, belonging to Blumen Group SpA, Italy, were used in this study. Specifically, 18 samples (numbered from 1 to 18) belonging to *C. endivia* var. *latifolium* (escarole or smooth endive) and 14 individuals (numbered from 19 to 32) of *C. endivia* var. *crispum* (curly endive) were analysed. The genomic DNA was isolated from 100 mg of fresh leaf using the DNAeasy Plant Mini Kit (Qiagen, Valencia, CA, USA), following the procedure provided by the suppliers. Both quality and quantity of genomic DNA samples were assessed by agarose gel electrophoresis (1 % agarose/1× TAE gel containing 1× Sybr Safe DNA stain, Life Technologies, Carlsbad, CA, USA) and NanoDrop 2000c UV-Vis spectrophotometer (Thermo Scientific, Pittsburgh, PA, USA), respectively.

2.2 Heterologous chicory-derived microsatellites

Three randomly samples, diluted to 15 ng/µl, were primarily amplified using 29 heterologous SSR primer couples selected from *C. intybus* [14], in order to evaluate their transferability within the genus. Firstly, the microsatellite loci were tested individually (singleplex reactions), then markers were arranged into three multiplex reactions (Table 1). PCR reactions were performed following the method described by Schuelke et al., [18], with some variations. Briefly, three primers were used for amplifying each heterologous microsatellite locus: a couple of locus-specific primers, one of which had a oligonucleotide tail at the 5' end (PAN-1, PAN-2, PAN-3 and M13, Table S1), and a third common primer complementary to the tail and labelled with a fluorescent dye (VIC, NED, PET and 6-FAM, respectively).

Table 1. Sequences of the primer pairs that produced amplicons in endive species. For each primer pair, original ID, SSR linkage group (LG), motif, tailed primers used (PAN1, PAN2, PAN3 or M13), and multiplex to which the SSR marker locus belongs are reported. All the microsatellite used in this study derives from Patella et al., [14].

ID	LG	Motif		Primer Sequence and Tail	Multiplex
M2.4	2	$(GA)_{25}$	F	[PAN3]CCAACGGATACCAAGGTGTT	1
			R	AACCGCACGGGTTCTATG	
M2.5	2	$(CT)_5CC(CT)_{13}TT(CT)_5$	F	[PAN1]GTGCCGGTCTTCAGGTTACA	1
			R	CGCCTACCGATTACGATTGA	
M3.7	3	$(CT)_{22}$	F	TTCGAGTCTTGCCTTAATTGTT	1
			R	[PAN1]CAGACGACCTTACGGCAACT	
M4.10a	4	$(CT)_{22}$	F	[PAN2]CATCACCTTCACGAAAAGCA	1
			R	CGAAGACCATCCATCACCA	
M4.11a	4	$(CT)_{12}N_5(CA)_{11}$	F	[PAN3]GAAGGAACCTATGAACCAACCACTCA	1
			R	GTTTTGAGCCTGAGCCAGA	
M1.3	1	$(CT)_{17}$	F	[PAN3]TGGAGAAAAATGAAGCAC	2
			R	GAATGAGTGAGAGAATGATAGGG	
M5.13	5	$(CT)_{23}$	F	[M13]AGGCATAAAGAGGTGTGG	2
			R	TCAAACATGAAAACCGCTC	
M6.17	6	$(CA)_8(CT)_{18}$	F	CGTGTCCAAACGCAAACATTAT	2
			R	[PAN2]GCACAATTTTCCTACCACTTATCC	
M5.14	5	$(TC)_{11}$	F	[M13]AAAGTCACACATCGCATTTCCT	2
			R	GTAGCAGCAGCAGCCATCTT	
M4.11b	4	$(TG)_5CG(TG)_7$	F	[M13]GCCATTCCTTTCAAGAGCAG	2
			R	AACCCAAAACCGCAACAATA	
M3.9	3	$(CA)_{12}$	F	CTGCTATGGACAGTTCCAGT	3
			R	[PAN3]CAATTCAGTTGTGATAGACGC	
M7.20	7	$(CT)_{31}$	F	[PAN2]ACACTCACTCACACTCCGTAA	3
			R	GTCATGATGGCGTAAAAGTC	
M6.18	6	$(CT)_{16}$	F	[PAN3]CTCAACGAATGCTTTGGACA	3
			R	CCTCGCGGTAGCTTATTGTT	
M2.6	2	$(CT)_{26}$	F	GGAGCAGGTAGAGTCCCATC	3
			R	[PAN1]CGTTTGAAAATTTATACCAAAATG	

All PCR reactions (both the singleplex and multiplex reactions) were set up in a 10 μl reaction volume, containing 1× Platinum® Multiplex PCR Master Mix (Thermo Scientific), 10% GC Enhancer (Thermo Scientific), 0.25 μM of non-tailed primer, 0.75 μM of tailed primer, 0.50 μM of fluorophore-labelled primer (universal primer) and 15 ng of genomic DNA. Thermal cycling conditions were as follows for all multiplex: 94°C for 5 minutes followed by 8 cycles of 94°C, 30 seconds, 61°C for 30 seconds, 72°C for 30 seconds; annealing temperature stepdown every cycle of 1°C (from 61°C to 54°C). The annealing temperature for the following 37 cycles was set to 54°C, with denaturation and extension phases as above and a final extension hold at 60°C for 30 minutes. PCR products were performed in a GeneAmp® PCR 9700 thermal cycler (Applied Biosystems, Carlsbad, CA, USA). The amplifications were first tested on gel electrophoresis (2 % Ultrapure™ Agarose in TAE 1×, SYBR Safe® 1×, Life Technologies) and then run on capillary electrophoresis with ABI 3730 DNA Analyzer (Applied Biosystem), adopting LIZ500 as molecular weight standard. The size of each peak was determined using Peak Scanner software 1.0 (Applied Biosystems).

For those markers that produced amplicons, the polymorphic information content (PIC) was calculated using POPGENE software [19]. The UPGMA dendrogram and principal coordinate analysis (PCoA) centroids were constructed applying the Jaccard coefficient and plotted, using PAST software v. 3.14 with 1,000 bootstrap repetitions [20]. Finally, a Bayesian clustering algorithm implemented in STRUCTURE v.2.2 [21] was used to model the genetic structure of the endive core collection. The number of founding groups ranged from 1 to 15, and 10 replicate simulations were conducted for each value of K, setting a burn-in of 200,000 and a final run of 1,000,000 Markov chain Monte Carlo (MCMC) steps. STRUCTURE HARVESTER [22] 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 RADseq analyses

The set of SNPs were identified using restriction-site associated DNA (RAD) sequencing of 32 individuals of endive. 1 μg of DNA was digested with the restriction enzymes *PstI* and *MspI* (New England Biolabs). For library preparation, DNAs were diluted at concentrations of 3 ng/μL. Library preparation, sequencing run and bioinformatics analyses were carried out according to the protocol described by Stevanato et al., [23]. RAD sequencing was carried out according to the protocol by Stevanato et. al., [23] using Ion S5 sequencer (Thermo Scientific). Reads were trimmed according to the enzyme recognition sequence, cleaned after a quality check, removing all the artefacts and the reads with Ns. Variants were called using Stacks v2.41 software [24] and SNPs were filtered according to the following criteria: (1) SNPs with more than 10% of missing data, (2) SNPs with a sequence depth ≤ 3, (3) tri- and tetra-allelic SNPs, (4) SNPs with allele frequencies across all samples ≤ 5 % and ≥ 95 % were all removed.

SNP-containing reads were annotated using a set of 62,656 CDS available for lettuce (*Lactuca sativa*), retrieved from Phytozome [25] and deriving from Chin-Wo S et al., [26]. A local BLASTn-based approach (E-value ≤1e-07, BLAST+ v.2.3.0) was used. For a further enrichment analysis [27] in terms of Gene Ontology (GO) [28] STRING [29] was exploited.

RADseq data were used for calculating similarity analysis (Jaccard coefficient) and similarity matrix, which was produced by NTSYS (Numerical Taxonomy and Multivariate Analysis System) version 2.2 (Exeter Software) [30]. Moreover, UPGMA dendrogram and PCoA centroids were constructed and plotted with Jaccard coefficient, as previously described for the SSR analysis and population genetic structure of the 32 endive samples as estimated by STRUCTURE v.2.2 [21]. The number of founding groups ranged from 2 to 5, and other parameters used were as previously described. Finally, SNPs markers were used to estimated observed homozygosity value with POPGENE software [19].

3 Results

3.1 Heterologous chicory-derived microsatellites

In a preliminary analysis aimed to investigate the transferability of SSR markers among species of the Cichorium genus, 14 out of 29 heterologous SSR primer couples (48 %) retrieved from *C. intybus* [14] produced amplicons also in three randomly chosen samples of *C. endivia.* The 14 SSR markers were then organised in three multi-locus PCRs (Table 2) and used to genotype the whole set of samples (32 individuals). The number of polymorphic microsatellites resulted to be 8 (57 % of the total): 4 scored PIC values highly informative ≥ 0.50 (from 0.58 to 0.70) and others 4 showed reasonably informative (0.42 < PIC < 0.50) (Table S2). The UPGMA dendrogram constructed with Jaccard coefficient divided the samples into three main clusters, with bootstrap supports always lower than 50 %, except for some nodes (Figure 1, panel a). From the PCoA, the first principal coordinate accounted for 35 % of the total variation and separated samples in two groups while the second coordinate accounted for 16 % of the total variation (Figure 1, panel b). The dendrogram and PCoA did not distinguish individuals and the differences between the two cultivar types of endive (Figure 1, panel b). From the genetic structure analysis, following the procedure of Evanno et al, [21] a clear maximum for ΔK value at K = 2 was found (ΔK = 240, Figure S1). Although the population size K = 2 also corresponds to the number of varieties used in this study (var. *crispum* and var. *latifolium*), the estimated membership of each sample to the ancestral genotypes, did not reflect the subspecies classification, as already shown by the UPGMA dendrogram and the PCoA, (Figure 1, panel c).

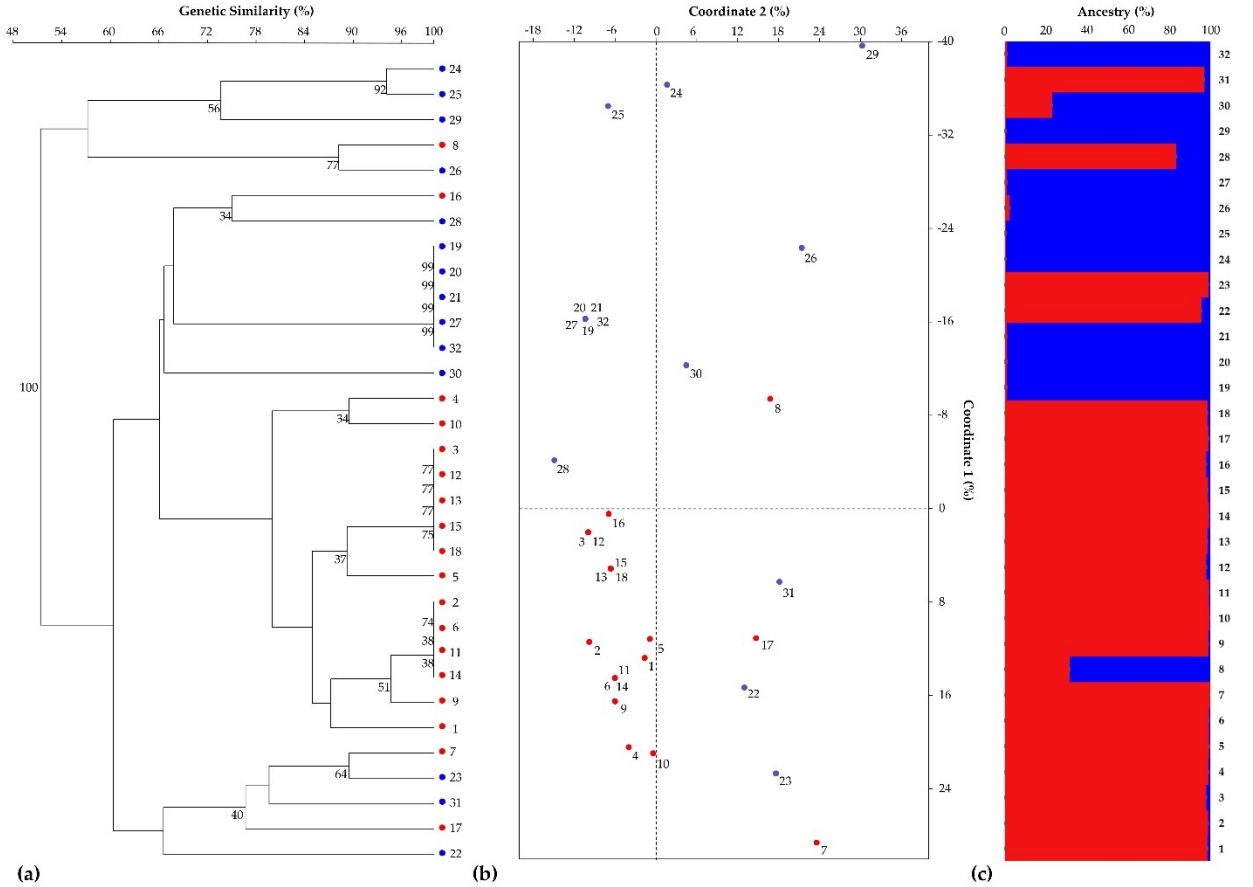


Figure 1. Grouping analysis of 32 samples of endive based on 14 SSR markers. **(a)** UPGMA dendrogram of genetic similarity estimates computed among pairwise comparisons of individual samples using the Jaccard coefficient. Bootstrap estimates ≥ 30 % are reported next to the nodes (red and blue dots highlight the two endive cultivar types, *C. endivia* var. *latifolium* and *C. endivia* var. *crispum*). **(b)** PCoA centroids deriving from the analysis of genetic similarity estimated with Jaccard coefficient (red and blue dots correspond to the individual samples of the two cultivar types, *C. endivia* var. *latifolium* and *C. endivia* var. *crispum*). **(c)** Population genetic structure of the 32 endive samples as estimated by STRUCTURE. Each sample is represented by a vertical histogram partitioned into K = 2 coloured segments (red or blue, in accordance with (a) and (b)) representing the estimated membership. The proportion of ancestry (%) is reported on the ordinate axis, and the identification number of each accession is reported below each histogram.

3.2 RADseq analyses

RADseq approach was applied to the same 32 samples already tested using 14 heterologous SSR primer couples. The Ion S5 sequencer produced of 81,200,000 raw reads, on average 2,030,891 ± 150 per samples. After quality and adapter trimming, we obtained 72,670,760 reads that were used to create a catalog of 18,806 consensus loci, used as reference for the variant calling. A raw pool of 6,242 SNPs was firstly identified, and after the filtering step, 4,621 SNPs distributed in 4,482 RAD sequence tags were retained.

From BLASTn analysis (E-value ≤1e-07) performed aligning the 4,482 SNP-carrying sequences against the lettuce genome [26] 578 (12.90%) showed at least one significant match. Among these, the average of similarity was equal to 95.5 % and 61.8 % of matches exhibited similarity scores higher than 95

% (Figure S2, panel b). 42.7 % of E-values ranged from 10e-15 to 10e-19 (Figure S2, panel b), and the average of total E-values was 1.40e-9. Based on the BLASTn analysis, our sequences have matched with uniformly distributed lettuce loci on all the LGs of the genome. The analysis performed using STRING revealed 578 sequence tags involved in multiple processes. 4,418 GO terms were resulted assigned to biological process, 1,975, were related to a molecular function, and 3,915 were associated with specific cellular component (Tables S3 to S5). From the enrichment analysis, the most important category resulted to be the biological process. Specifically, three processes were significantly represented: negative regulation of apoptotic process, glycogen catabolic process and negative regulation of autophagy (Table S3). Generally speaking, 46 out of 578 reads annotated were in common with Testone [5] results. Among the 4,621 SNPs, 50 markers fully discriminated curly endive from escarole. Only 12 out of the 50 SNP-carrying reads resulted annotated from the BLASTn alignment, and 4 SNPs were previously shown also in a study by Testone et al.,[5] (Table 2). Additionally, 3 sequence tag containing SNPs were annotated in *Arabidopsis thaliana* as genes involved in the stage of development of leaves and among 12 annotated SNPs, only 6 showed non-synonymous mutation at protein level (Table 2).

Table 2. Information on the 12 annotate sequence tags that distinguished escarole to curly endive – each constituted by polymorphism, the best *L. sativa* match, similarity value, E-value, Arabidopsis match, predicted function, synonym mutation or not and common SNP with Testone [5] are reported.

Reads	Polymorphism	Best *L. sativa* match	Similarity	E-value	*Arabidopsis thaliana* match	Predicted function	Synonymous / non synoumous	Testone et al., [5]
82	16:T>A	Lsat_1_v5_gn_8_14760.1	98.438	1.67e-025	AT3G55610.2	delta 1-pyrroline-5-carboxylate synthase 2	synonymous	n.a.
238	50:T>C	Lsat_1_v5_gn_8_19481.2	94.118	6.08e-015	AT4G27290.1	S-locus lectin protein kinase family protein	synonymous	Up-regolated; SNP
414	36:A>C	Lsat_1_v5_gn_1_64300.1	87.5	7.87e-014	AT2G37170.1	Plasma membrane intrinsic protein 2	non-synonymous	n.a.
857	60:G>A	Lsat_1_v5_gn_8_40821.2	98	1.01e-017	AT4G38630.1	Regulatory particle non-ATPase 10	non-synonymous	n.a.
894	63:T>C	Lsat_1_v5_gn_2_24380.3	95.313	3.61e-022	AT5G05010.1	Clathrin adaptor complexes medium subunit family protein	synonymous	SNP
1310	60:G>C	Lsat_1_v5_gn_8_45601.2	95.313	3.61e-022	AT2G44140.1	Peptidase family C54 protein	non-synonymous	Up-regolated
1611	13:C>T	Lsat_1_v5_gn_9_57180.3	96.875	7.76e-024	AT4G39420.2	Spatacsine carboxy terminus protein	synonymous	SNP
3058	7:A>G	Lsat_1_v5_gn_8_15760.1	90	6.08e-015	AT4G26055.1	transmembrane protein	synonymous	n.a.
3451	7:T>C	Lsat_1_v5_gn_1_75900.1	93.22	1.01e-017	AT1G08730.1	Myosin family protein with Dil domain	synonymous	n.a.
3508	52:G>A	Lsat_1_v5_gn_8_16481.1	96.875	7.76e-024	AT5G13680.1	IKI3 family protein	non-synonymous	n.a.
3614	12:T>A	Lsat_1_v5_gn_8_18921.2	97.222	3.65e-010	AT2G28370.1	Uncharacterised protein family (UPF0497)	non-synonymous	n.a.
4087	46:A>T	Lsat_1_v5_gn_2_115321.1	97.222	5.40e-010	AT2G39290.1	Phosphatidylglycerolphosphate synthase 1	non-synonymous	n.a.

The average similarity values calculated in all possible pair-wise comparisons between all samples are reported in Figure S3. When calculated within two cultivar types, these similarity values varied from 63.9% to 99.7 % (escarole endive) and from 52.0% to 99.5 % (curly endive) with an average of 83.9 % and 74.9 %, respectively. In the pair-wise comparisons between escarole and curly types accessions, the average similarity values (67.7 %) ranged from a minimum of 57.1 % to a maximum of 78.4 %. Two samples (2 and 3) presented the highest genetic similarity (99.7 %) and resulted discriminable only for 23 SNP loci. The extent of genetic variation and relationships among all samples, along with comparisons with their relative cultivar types, was measured and visualized using the whole SNP data set.

The PCoA enabled the definition of centroids of the endive accessions and provided, along with UPGMA dendrogram, useful information for plants discrimination (Figure 2). The ordination of the centroids, while revealing a relatively tight aggregation among samples belonging to curly or escarole types, highlighted clear discrimination of the most genetically differentiated samples. It is worth mentioning that the first two components were able to explain together the 53.8 % of the molecular variation, accounting for 36.7 % and 17.0 %, respectively (Figure 2, panel b). Moreover, STRUCTURE analysis was used to investigate the genetic structure of the endive core collection and following the procedure of Evanno et al., [21] a maximum for ΔK value at $K = 3$ was found. Therefore, ancestral analysis showed three different roots of endive plants. All samples marked in red were attributed to escarole endive, while plants in blue belonging to curly endive; expect for two samples, namely 16 and 28, seemed to completely derive from a third ancestor (Figure 2, panel c).

In addition, the observed homozygosity computed across all individual DNA samples ranged from 74.2 % to 97.6 %, with an average estimate of 95.9 ± 4.1 % (Figure S3).

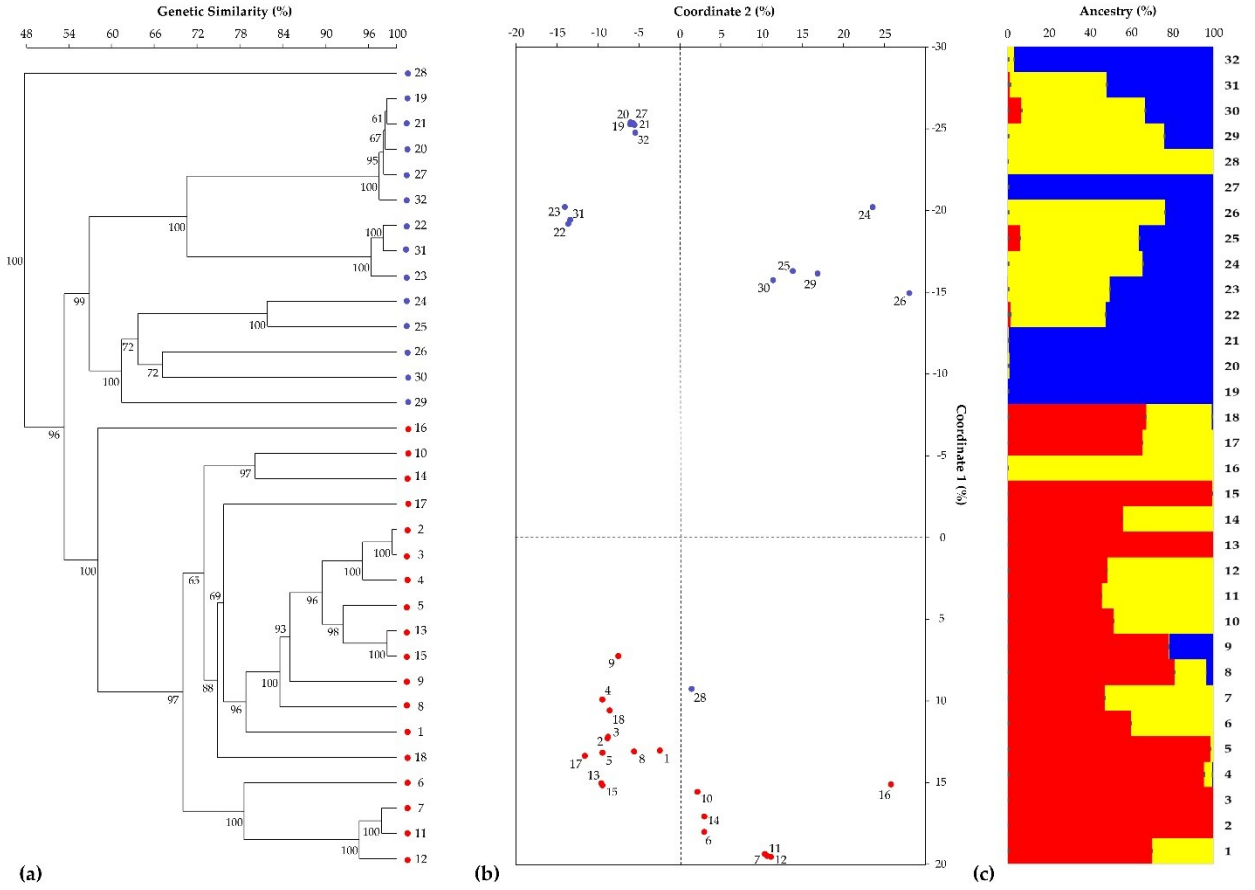


Figure 2. Grouping analysis of 32 samples of endive based on 4,621 SNP markers. **(a)** UPGMA dendrogram of genetic similarity estimates computed among pairwise comparisons of individual samples using the Jaccard coefficient. Bootstrap estimates ≥ 30 % are reported next to the nodes (red and blue dots highlight the two endive cultivar types, *C. endivia* var. *latifolium* and *C. endivia* var. *crispum*). **(b)** PCoA centroids deriving from the analysis of genetic similarity estimated with Jaccard coefficient (red and blue dots correspond to the individual samples of the two cultivar types, *C. endivia* var. *latifolium* and *C. endivia* var. *crispum*). **(c)** Population genetic structure of the 32 endive samples as estimated by STRUCTURE. Each sample is represented by a vertical histogram partitioned into K = 3 coloured segments (red, blue, in accordance with (a) and (b), or yellow) representing the estimated membership. The proportion of ancestry (%) is reported on the ordinate axis, and the identification number of each accession is indicated below each histogram.

4 Discussion

Lettuce, chicory and endive are leafy vegetables popular for the preparation of ready-to-use salads, which attract consumers' interest. Their fresh-like nature and convenience along with their pro-health properties, are chief for their continual use as vegetables in salads [31,32]. The economic importance of endive species is leading companies to invest first in breeding programs and then in varieties protections. Noteworthy, cultivated varieties of endive are usually pure lines [8], so deposition of variety genotype is crucial for its legal protection. Currently, molecular markers are useful in many species not only to assess the overall genetic diversity among varieties but also for their registration and the protection of plant breeders' rights. In particular, for some crops of great commercial interest, the ISF already set species specific thresholds to define essentially derived varieties, and in general, to protect intellectual property

rights. Although in species like lettuce this threshold is well established and widely exploited to solve legal disputes (0.96, calculated using the Jaccard coefficient [10]), in endive, the situation is far from being defined. In fact, the total lack of an informative and robust panel of markers, makes any genotyping analysis impossible. For this reason, the first aim of this work was to develop a molecular assay in this species able to assess the genetic distance existing among elite breeding materials (F5 generations) of *C. endivia*. Eventually, the newly developed markers could also be used for the registration of new cultivars and for their legal protection.

In the first section of the work, an approach based on the exploitation of heterologous SSR primers retrieved from *C. intybus* [14] was evaluated. Despite the inter fertility between *C. endivia* and *C. intybus*, and the availability of molecular linkage maps referred to such hybrids [16], only 48 % of microsatellites were successfully transferred. Generally, the use of transferable cross species/genera microsatellite markers is considered a cost-effective approach to ensure the ubiquitous applicability of markers in genomic resources [33,34]. For example, recently, Bombonato et al., [35] tested cross-genera microsatellite loci in the family Cactaceae using 20 heterologous markers previously developed for the genera Ariocarpus, Echinocactus, Polaskia and Pilosocereus, in four taxa of the genus Cereus. Nine SSR loci were amplified in Cereus resulting in 35.2 % of success in transferability. Harijan et al., [36] demonstrated that this approach can be a very efficient also for cotton species. 46.6 % of primer pairs deriving from safflower and pulses were found transferable from one species to another but only 15.9 % were shown polymorphic [36]. In the scientific literature the transferability of molecular markers has been extensively studied and our work partially reflected these results. Despite the transferability of microsatellite in endive, the limited number of microsatellites and low number of the polymorphic loci (8, 57.1 %) were not enough to discriminate the two cultivar types and morphologically different plant materials. This lack of informativeness was highlighted by the UPGMA dendrogram and the PCoA centroids: samples were clustered in three main sub-groups deviating from the expected cultivar types (see Figure 1, panels a and b) and, in general, from the morphological observations. Also, STRUCTURE analysis showed the limits of SSR panel, because recognised two different ancestors without discriminating completely escarole from curly endive (see Figure 1, panel c). To overcome some of the limits deriving from the use of a low number of informative SSR markers, a second SNP-based approach was attempted. On the contrary, the next-generation RAD-based sequencing adopted is well established as a powerful method for recovering thousands of polymorphic loci across the genomes of many crop species [37-39]. We have identified 4,621 biallelic SNP loci and thus 9,242 possible alleles discriminating the smooth and curly varieties of *C. endivia*. The genetic characterization based on SNP markers increase a lot our existing knowledge of the genetic diversity of endive. RAD-seq is a technology routinely exploited to generate thousands of SNPs and able to provide accurate estimates of genetic relatedness. Looking at the scientific literature, usually this method allows to yield, on average, from 12,000 to 40,000 SNPs depending on the species, data processing filters and the

inherent genetic diversity in plant material [23,39-41]. It is also worth noting here that more large number of shared polymorphic sites is not necessarily associated with higher genetic differentiation. Our analyses yielded SNPs in the order of a few thousand. This could be specifically related to the nature of this crop, the small size of the genome, the low levels of genetic diversity in the plant material and the autogamous nature of reproduction in this crop. Nevertheless, the benefits of this technology are well pronounced even in small populations with low genetic diversity, as shown in the present study. Moreover, the use of this technique represents an important step towards improving the discovery of molecular markers linked to specific traits of interest. The discriminant SNP markers identified in our work represent a valuable tool that could be used by breeders to discriminate endive genetic groups in germplasms of high breeding value. We believe that the use of this method has broader application for genetic analysis of many other non-model organisms including population structure analysis, genotype-phenotype correlations and evolutionary analyses.

The 4,621 SNP markers were distributed in 4,482 RAD sequence tags and the BLASTn analysis allowed annotation of 12.90 % of them. Consequently, considering that the genic regions in a genome are usually represented by 2-3 % of whole genome, [42] RADseq analysis once again proved to be an extremely versatile method. In particular, the selection of the enzymes used for the preparation of the sequencing libraries, allows to choose whether focusing the analysis more on coding regions or on intergenic regions. Furthermore, the BLASTn analysis support two more considerations. First, the fact that the overall genetic similarity calculated by aligning the endive RAD sequence tags against the lettuce CDS regions was as high as 95.5 %, confirmed the close phylogenetic relationship between these two species. Second, the distribution of the RADseq markers in all linkage groups of lettuce, reflects a good representativeness of the entire endive genome.

Interestingly, among the 50 SNP markers able to fully discriminate curly and escarole endive, 4 resulted particularly worthy since they were reported earlier in the study by Testone et al., [5]. Of these, three resulted also annotated as S-locus lectin protein kinase family protein, clathrin adaptor complexes medium subunit family protein and peptidase family C54 protein. To the best of our knowledge, these genes may not be directly responsible for leaves morphology, but further analysis is needed to draw conclusions about any possible involvement. This is particularly true for the peptidase protein belonging to the C54 family, since the SNP resulted as non-synonymous. In this regard, 6 SNPs out of the 50 above mentioned, resulted as non-synonymous mutations, but, according to their putative function, none of them seems to be directly involved in the definition of floral morphology, except for SNP 3508. In fact, the read containing this non-synonymous SNP, aligned with Arabidopsis gene AT5G13680.1 and corresponds to ELONGATA 2 a sub-unit of Elongator complex [43]. In An C. [44] work, Arabidopsis mutant *ien2* (mutant for ELONGATA 3) was characterized and showed serrated and curly leaves when compared to the wild. Therefore, further analyses will be required to elucidate any possible association between this gene and the

leaf morphology in endive. Moreover, 3 out the 6 genes carrying SNPs with non-synonymous mutations (marker 3508, AT5G13680.1; marker 3614, AT2G28370.1 and marker 4087, AT2G39290.1, Table 2) resulted particularly expressed in some stages of leaf development [45].

RADseq data were also used to conduct genetic similarity analysis among the 32 samples. Combining the genetic similarity estimates in all possible pair-wise comparisons using the UPGMA dendrogram and PCoA analyses of the core collection, it was possible to obtain extremely useful information for breeding purposes. High levels of genetic similarity scored within escarole materials (83.9 %) confirmed the relatively low genetic variability among these materials already, as observed at phenotypic level. This was contrasting from what was concluded by means of the SSR-based analysis, all samples were univocally discriminated, even if, only 23 SNPs discriminated sample 2 and 3. This latest finding is in full agreement with the results produced using microsatellite markers in the first section of the work. In this case, to avoid the registration and the release on the market of two almost-identical (and thus essentially-derived) varieties, a choice will be made by breeders according to phenotypic data and to pre-commercial trials.

Interestingly, STRUCTURE analysis based on 4,621 SNP loci showed three different ancestral of groups for the endive materials, highlighting the importance of analysed plant material as the result of crosses between different germplasms. Accordingly to the phenotype information, in STRUCTURE analysis samples were attributed as escarole or curly endive, expect for two samples, namely 16 and 28, which were derived from another ancestor. These findings confirmed UPGMA dendrogram results, in which the two samples appeared as out-groups.

The overall observed homozygosity was 95.9 ± 4.1 % on average, consistently with the autogamous reproductive system of this species. Moreover, the high homozygosity values of lines are in agreement with the five self-pollinated cycles that these materials have undergone. Therefore, the low heterozygosity values guaranteed progenies with the desired genetic stability and, consequently, phenotypic uniformity[8]. Therefore, individuals with high homozygosity could be selected to produce pre-commercial varieties characterised by genetic stability and thus uniformity.

5 Conclusions

In conclusion, the results of the study carried out in endive have demonstrated the advantages of using a molecular, genome-wide approach to distinguish phenotypically similar breeding stocks. We firstly documented the inefficiency of heterologous SSR markers derived from *C. intybus*, due to both a limited transferability from one species to another and the low number. Following this, we were able to discriminate the 32 endive lines using 4,621 SNP markers and to predict the two main different cultivar types of endive based on subset of 50 SNPs. Overall, our research was able to evaluate the distinctiveness requirement of DUS test. This is a key aspect considering that the genotype and the molecular profile of a

registered variety can be crucial to solve cases of fraudulent practices. Currently, no specific protocol to assess the distinctiveness among varieties is available for this species. Finally, we evaluated the observed homozygosity to predict the uniformity and stability of progenies which are additional requirements of DUS test: individuals with the highest homozygosity are known to produce more uniform and stable populations over generations. As a future objective, lab-scale SNP validation will be achieved by an HRM technique performed on randomly selected events. In conclusion, our study represents the first genotypic analysis of endive breeding materials in which thousands of discriminant SNP markers were identified at genomic level.

Supplementary materials

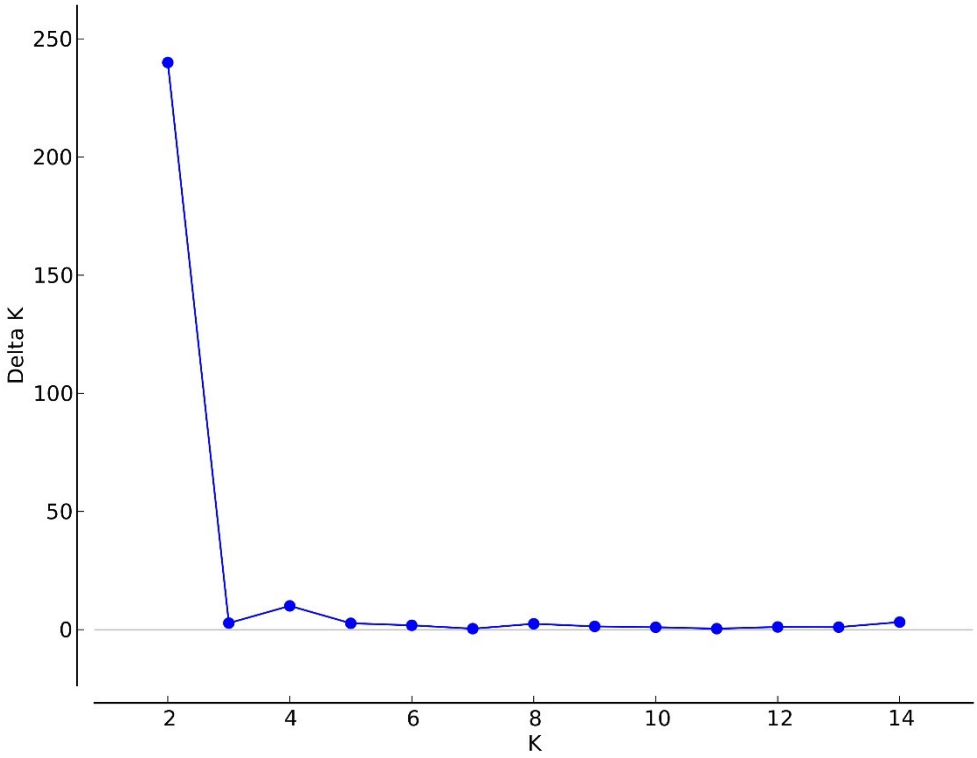


Figure S1. Definition of the number of ancestral parental lines based on the SSR marker dataset. Mean ΔK is calculated as |L″ (K)|/(SD(L(K)), following Evanno et al. [21]. The blue line represents the ΔK values.

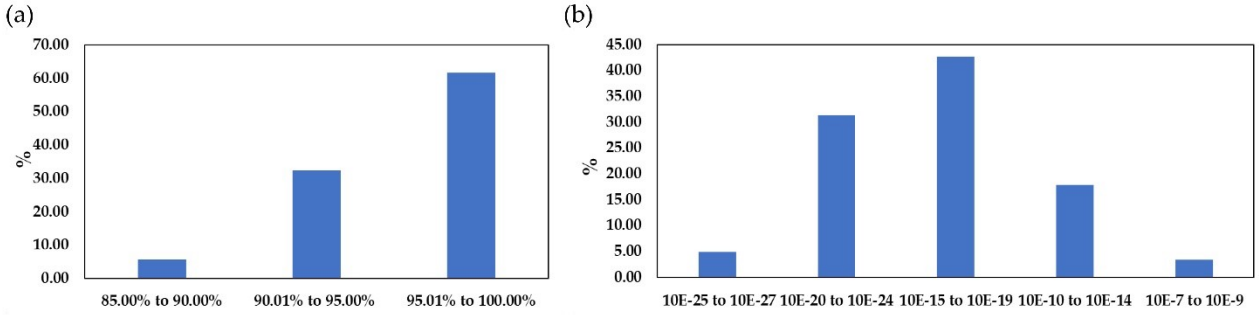
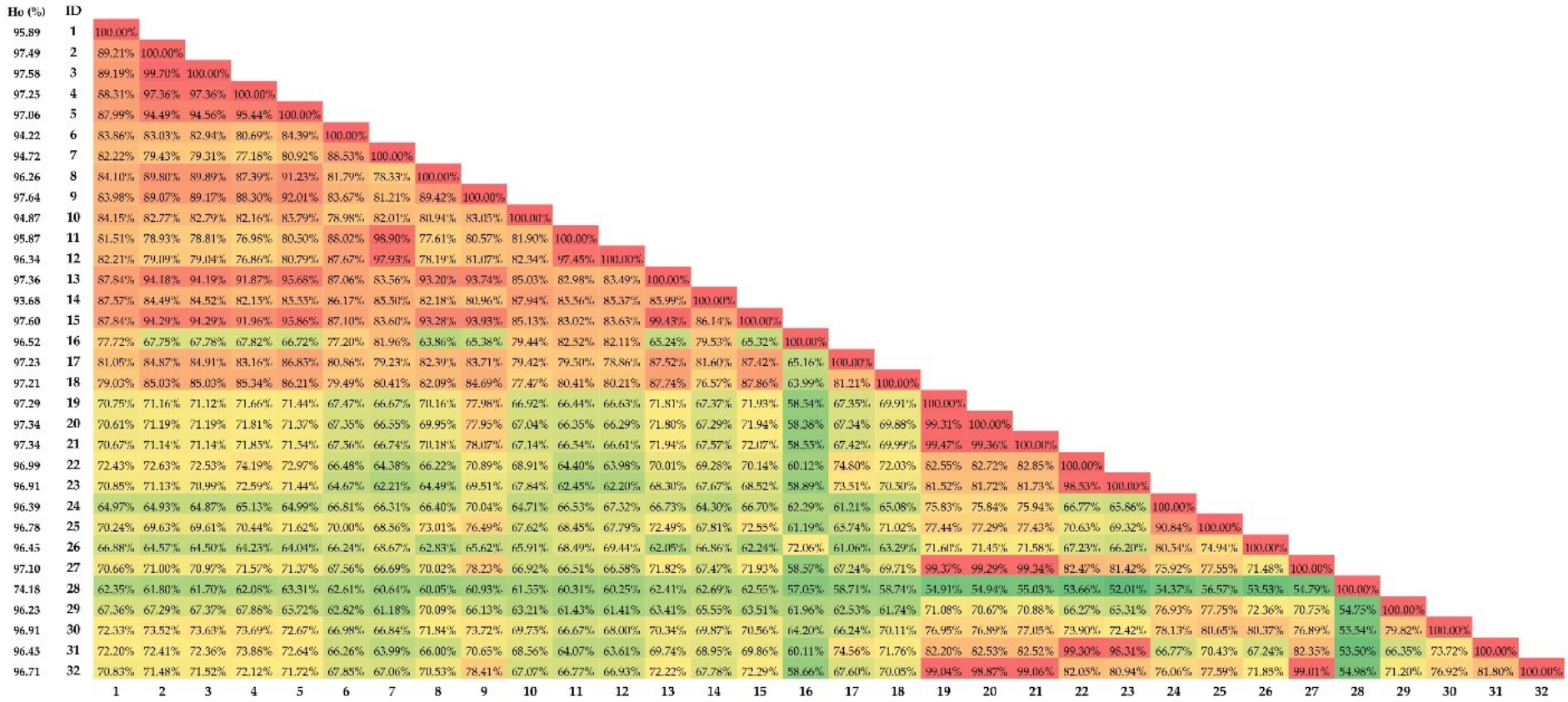


Figure S2. (a) Distribution of similarity values. **(b)** E-value distribution.



Ho (%)	ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
95.89	1	100.00%																															
97.49	2	89.21%	100.00%																														
97.58	3	89.19%	99.70%	100.00%																													
97.25	4	88.31%	97.36%	97.36%	100.00%																												
97.06	5	87.99%	94.49%	94.56%	95.44%	100.00%																											
94.22	6	83.86%	83.03%	82.94%	80.69%	84.39%	100.00%																										
94.72	7	82.22%	79.43%	79.31%	77.18%	80.92%	88.53%	100.00%																									
96.26	8	84.10%	89.80%	89.89%	87.39%	91.23%	81.79%	78.33%	100.00%																								
97.64	9	83.98%	89.07%	89.17%	88.30%	92.01%	83.67%	81.21%	89.42%	100.00%																							
94.87	10	84.15%	82.77%	82.79%	82.16%	85.79%	78.98%	82.01%	80.94%	83.05%	100.00%																						
95.87	11	81.51%	78.93%	78.81%	76.98%	80.50%	88.02%	98.90%	77.61%	80.57%	81.90%	100.00%																					
96.34	12	82.21%	79.09%	79.04%	76.86%	80.79%	87.67%	97.93%	78.19%	81.07%	82.34%	97.45%	100.00%																				
97.36	13	87.84%	94.18%	94.19%	91.87%	95.68%	87.06%	83.56%	93.20%	93.74%	85.03%	82.98%	83.49%	100.00%																			
93.68	14	87.57%	84.49%	84.52%	82.15%	85.35%	86.17%	85.50%	82.18%	80.96%	87.94%	83.56%	85.37%	85.99%	100.00%																		
97.60	15	87.84%	94.29%	94.29%	91.96%	95.86%	87.10%	83.60%	93.28%	93.93%	85.13%	83.02%	83.63%	99.43%	86.14%	100.00%																	
96.52	16	77.72%	67.75%	67.78%	67.82%	66.72%	77.20%	81.96%	63.86%	65.38%	79.44%	82.52%	82.11%	65.24%	79.53%	65.32%	100.00%																
97.23	17	81.05%	84.87%	84.91%	83.16%	86.83%	80.86%	79.23%	82.30%	83.71%	79.42%	79.50%	78.86%	87.52%	81.60%	87.42%	65.16%	100.00%															
97.21	18	79.03%	85.03%	85.03%	85.34%	86.21%	79.49%	80.41%	82.09%	84.69%	77.47%	80.41%	80.21%	87.74%	76.57%	87.86%	63.99%	81.21%	100.00%														
97.29	19	70.75%	71.16%	71.12%	71.66%	71.44%	67.47%	66.67%	70.16%	77.98%	66.92%	66.44%	66.63%	71.81%	67.37%	71.93%	58.54%	67.35%	69.91%	100.00%													
97.34	20	70.61%	71.19%	71.19%	71.81%	71.37%	67.35%	66.55%	69.95%	77.95%	67.04%	66.35%	66.29%	71.80%	67.29%	71.94%	58.38%	67.34%	69.88%	99.31%	100.00%												
97.34	21	70.67%	71.14%	71.14%	71.85%	71.54%	67.56%	66.74%	70.18%	78.07%	67.14%	66.54%	66.61%	71.94%	67.57%	72.07%	58.55%	67.42%	69.99%	99.47%	99.36%	100.00%											
96.99	22	72.43%	72.63%	72.53%	74.19%	72.97%	66.48%	64.38%	66.22%	70.89%	68.91%	64.40%	63.98%	70.01%	69.28%	70.14%	60.12%	74.80%	72.03%	82.55%	82.72%	82.85%	100.00%										
96.91	23	70.85%	71.13%	70.99%	72.59%	71.44%	64.67%	62.21%	64.49%	69.51%	67.84%	62.45%	62.20%	68.30%	67.67%	68.52%	58.89%	73.51%	70.50%	81.52%	81.72%	81.73%	98.53%	100.00%									
96.39	24	64.97%	64.93%	64.87%	65.13%	64.99%	66.81%	66.31%	66.40%	70.04%	64.71%	66.53%	67.32%	66.73%	64.30%	66.70%	62.29%	61.21%	65.08%	75.83%	75.84%	75.94%	66.77%	65.86%	100.00%								
96.78	25	70.24%	69.63%	69.61%	70.44%	71.62%	70.00%	68.56%	73.01%	76.49%	67.62%	68.45%	67.79%	72.49%	67.81%	72.55%	61.19%	65.74%	71.02%	77.44%	77.29%	77.43%	70.63%	69.32%	90.84%	100.00%							
96.45	26	66.88%	64.57%	64.50%	64.23%	64.04%	66.24%	68.67%	62.83%	65.62%	65.91%	68.49%	69.44%	62.05%	66.86%	62.24%	72.06%	61.06%	63.29%	71.61%	71.45%	71.58%	67.23%	66.20%	80.54%	74.94%	100.00%						
97.10	27	70.66%	71.00%	70.97%	71.57%	71.37%	67.56%	66.69%	70.02%	78.23%	66.92%	66.51%	66.58%	71.82%	67.47%	71.93%	58.57%	67.24%	69.71%	99.37%	99.29%	99.34%	82.47%	81.42%	75.92%	77.55%	71.48%	100.00%					
74.18	28	62.35%	61.80%	61.70%	62.08%	63.31%	62.61%	60.64%	60.05%	60.93%	61.55%	60.31%	60.25%	62.41%	62.69%	62.55%	57.05%	58.71%	58.74%	54.91%	54.94%	55.03%	53.66%	52.01%	54.37%	56.57%	53.53%	54.79%	100.00%				
96.23	29	67.36%	67.29%	67.37%	67.88%	65.72%	62.82%	61.18%	70.09%	66.13%	63.21%	61.43%	61.41%	63.41%	65.55%	63.51%	61.96%	62.53%	61.74%	71.08%	70.67%	70.88%	66.27%	65.31%	76.93%	77.75%	72.36%	70.75%	54.75%	100.00%			
96.91	30	72.33%	73.52%	73.63%	73.69%	72.67%	66.98%	66.84%	71.84%	73.72%	69.75%	66.67%	68.00%	70.34%	69.87%	70.56%	64.20%	66.24%	70.11%	76.95%	76.89%	77.05%	73.90%	72.42%	78.13%	80.65%	80.37%	76.89%	55.54%	79.82%	100.00%		
96.45	31	72.20%	72.41%	72.36%	73.88%	72.64%	66.26%	63.99%	66.00%	70.65%	68.56%	64.07%	63.61%	69.74%	68.95%	69.86%	60.11%	74.56%	71.76%	82.20%	82.53%	82.52%	99.30%	98.31%	66.77%	70.43%	67.24%	82.35%	53.50%	66.35%	73.72%	100.00%	
96.71	32	70.83%	71.48%	71.52%	72.12%	71.72%	67.85%	67.06%	70.53%	78.41%	67.07%	66.77%	66.93%	72.22%	67.78%	72.29%	58.66%	67.60%	70.05%	99.04%	98.87%	99.06%	82.05%	80.94%	76.06%	77.59%	71.85%	99.01%	54.98%	71.20%	76.92%	81.80%	100.00%
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

Figure S3. Pairwise genetic similarity matrix of 32 endive lines (in percentages) based on the Jaccard coefficient. The high genetic similarity values are labelled in red, the low values in green, and intermediate values are coloured on a scale from red to green. Moreover, observed homozygosity in percentage (Ho) of 32 plants are reported.

Table S1. Microsatellite primer tails and dyes. List of the primer tails used with their sequences and corresponding dyes.

Universal primer	Sequence 5'-3'	Dye
M13	TTGTAAAACGACGGCCAGT	6-FAM
PAN1	GAGGTAGTTATTGTGGAGGAC	VIC
PAN2	GGAATTAACCGCTCACTAAAG	NED
PAN3	TGTAGAAAGACGAAGGGAAGG	PET

Table S2. PIC values for each locus found across 32 elite materials.

ID	PIC
M2.4	0.61
M2.5	0.7
M3.7	0.49
M4.10a	0
M4.11a	0.65
M1.3	0
M5.13	0.47
M6.17	0.44
M5.14	0
M4.11b	0.58
M3.9	0
M7.20	0
M6.18	0
M2.6	0.42

Table S3. Information of biological process with GO-term, functional description, background gene count and false discovery rate

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0009987	cellular process	226	10581	1.52E-06
GO:0051179	localization	68	2244	0.00011
GO:0051234	establishment of localization	65	2170	0.00021
GO:0044237	cellular metabolic process	179	8432	0.00023
GO:0006091	generation of precursor metabolites and energy	20	360	0.00035
GO:0006793	phosphorus metabolic process	53	1677	0.00035
GO:0006810	transport	63	2140	0.00035
GO:0007049	cell cycle	23	448	0.00035
GO:0008152	metabolic process	197	9671	0.00035
GO:0044238	primary metabolic process	172	8114	0.00035
GO:0071704	organic substance metabolic process	181	8632	0.00035
GO:0006796	phosphate-containing compound metabolic process	51	1636	0.00038
GO:1901564	organonitrogen compound metabolic process	99	4116	0.00068
GO:0019637	organophosphate metabolic process	24	547	0.0012
GO:0016043	cellular component organization	62	2271	0.0015
GO:0050896	response to stimulus	114	5064	0.0018
GO:0006807	nitrogen compound metabolic process	150	7152	0.0019
GO:0022402	cell cycle process	15	253	0.0019
GO:0040007	growth	17	325	0.0021
GO:0050793	regulation of developmental process	25	622	0.0021
GO:0071840	cellular component organization or biogenesis	65	2467	0.0021
GO:0090407	organophosphate biosynthetic process	18	367	0.0025
GO:0009741	response to brassinosteroid	9	96	0.0027
GO:0006950	response to stress	73	2932	0.003
GO:0044281	small molecule metabolic process	44	1503	0.004
GO:0051301	cell division	16	315	0.004
GO:0016192	vesicle-mediated transport	17	358	0.0049
GO:0016049	cell growth	14	265	0.0068
GO:0098657	import into cell	8	87	0.0068
GO:0009628	response to abiotic stimulus	47	1699	0.0069
GO:0048589	developmental growth	14	271	0.008

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0006996	organelle organization	38	1283	0.0082
GO:0042221	response to chemical	65	2654	0.0098
GO:0019693	ribose phosphate metabolic process	13	246	0.0101
GO:1901700	response to oxygen-containing compound	40	1398	0.0101
GO:0015979	photosynthesis	12	215	0.0104
GO:0051641	cellular localization	21	553	0.011
GO:0006090	pyruvate metabolic process	7	74	0.0113
GO:0015980	energy derivation by oxidation of organic compounds	9	127	0.0113
GO:0048585	negative regulation of response to stimulus	11	186	0.0113
GO:0065007	biological regulation	111	5235	0.0113
GO:0009259	ribonucleotide metabolic process	12	227	0.0143
GO:0009826	unidimensional cell growth	11	196	0.0152
GO:0060560	developmental growth involved in morphogenesis	12	231	0.0158
GO:0033554	cellular response to stress	26	800	0.0178
GO:0016310	phosphorylation	32	1077	0.0181
GO:0002683	negative regulation of immune system process	4	21	0.0202
GO:0048583	regulation of response to stimulus	23	681	0.0207
GO:0051640	organelle localization	8	113	0.0207
GO:0051704	multi-organism process	40	1475	0.0207
GO:0009150	purine ribonucleotide metabolic process	11	210	0.0217
GO:0019538	protein metabolic process	71	3107	0.0217
GO:2000026	regulation of multicellular organismal development	17	433	0.0217
GO:0002682	regulation of immune system process	8	117	0.0218
GO:0050789	regulation of biological process	98	4623	0.0218
GO:0071702	organic substance transport	32	1101	0.0218
GO:0098662	inorganic cation transmembrane transport	15	358	0.0218
GO:0022414	reproductive process	37	1348	0.022
GO:0009117	nucleotide metabolic process	14	323	0.0222
GO:0051716	cellular response to stimulus	58	2428	0.0223

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0051239	regulation of multicellular organismal process	18	485	0.0226
GO:0009408	response to heat	10	184	0.023
GO:1901135	carbohydrate derivative metabolic process	23	701	0.023
GO:0006897	endocytosis	6	68	0.0249
GO:0008654	phospholipid biosynthetic process	8	123	0.0249
GO:0014070	response to organic cyclic compound	14	331	0.0249
GO:1903047	mitotic cell cycle process	8	124	0.0252
GO:0043170	macromolecule metabolic process	129	6502	0.0262
GO:0055086	nucleobase-containing small molecule metabolic process	16	414	0.0262
GO:0009260	ribonucleotide biosynthetic process	9	159	0.0285
GO:0055085	transmembrane transport	32	1142	0.0285
GO:0000902	cell morphogenesis	12	265	0.0286
GO:0048229	gametophyte development	14	341	0.0286
GO:0071407	cellular response to organic cyclic compound	9	160	0.0286
GO:0019752	carboxylic acid metabolic process	26	863	0.0294
GO:0033036	macromolecule localization	25	818	0.0296
GO:0006732	coenzyme metabolic process	12	271	0.0302
GO:0009856	pollination	11	232	0.0302
GO:0032502	developmental process	58	2492	0.0302
GO:0043066	negative regulation of apoptotic process	2	2	0.0302
GO:0044262	cellular carbohydrate metabolic process	15	386	0.0302
GO:0007017	microtubule-based process	9	166	0.0311
GO:0043436	oxoacid metabolic process	28	973	0.0327
GO:0048868	pollen tube development	8	135	0.0327
GO:0044249	cellular biosynthetic process	85	4013	0.0328
GO:0003006	developmental process involved in reproduction	31	1125	0.0344
GO:0006644	phospholipid metabolic process	10	205	0.0344
GO:0034220	ion transmembrane transport	21	656	0.0344
GO:0050776	regulation of immune response	7	107	0.0344
GO:1901566	organonitrogen compound biosynthetic process	33	1229	0.0352
GO:0016052	carbohydrate catabolic process	11	244	0.0355

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0044272	sulfur compound biosynthetic process	8	140	0.0355
GO:0051649	establishment of localization in cell	16	443	0.0355
GO:0070838	divalent metal ion transport	7	110	0.0375
GO:0050794	regulation of cellular process	87	4167	0.0379
GO:0005980	glycogen catabolic process	2	3	0.0387
GO:0009152	purine ribonucleotide biosynthetic process	8	143	0.0387
GO:0010182	sugar mediated signaling pathway	4	33	0.0387
GO:0010507	negative regulation of autophagy	2	3	0.0387
GO:0033169	histone H3-K9 demethylation	2	3	0.0387
GO:1901576	organic substance biosynthetic process	85	4083	0.0426
GO:0006816	calcium ion transport	5	58	0.0429
GO:0006790	sulfur compound metabolic process	13	335	0.0442
GO:0009240	isopentenyl diphosphate biosynthetic process	3	16	0.0442
GO:0030004	cellular monovalent inorganic cation homeostasis	4	35	0.0442
GO:0046490	isopentenyl diphosphate metabolic process	3	16	0.0442
GO:0048856	anatomical structure development	54	2361	0.0442
GO:0048519	negative regulation of biological process	25	872	0.0443
GO:0006399	tRNA metabolic process	7	118	0.0449
GO:0015833	peptide transport	20	642	0.0449
GO:1901701	cellular response to oxygen-containing compound	19	596	0.0449
GO:0006403	RNA localization	6	89	0.047
GO:0006637	acyl-CoA metabolic process	4	37	0.047
GO:0007143	female meiotic nuclear division	2	4	0.047
GO:0009617	response to bacterium	14	385	0.047
GO:0010235	guard mother cell cytokinesis	2	4	0.047
GO:0010501	RNA secondary structure unwinding	5	61	0.047
GO:0031347	regulation of defense response	10	224	0.047
GO:0070988	demethylation	3	17	0.047
GO:0044267	cellular protein metabolic process	62	2826	0.0476
GO:0044255	cellular lipid metabolic process	19	606	0.0483
GO:0065003	protein-containing complex assembly	14	387	0.0483
GO:0071705	nitrogen compound transport	25	888	0.0483

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0008104	protein localization	20	654	0.0489
GO:0009058	biosynthetic process	87	4258	0.0489
GO:0009743	response to carbohydrate	7	123	0.0489
GO:0010035	response to inorganic substance	23	795	0.0489
GO:0022607	cellular component assembly	18	564	0.0489
GO:0071456	cellular response to hypoxia	3	18	0.0489
GO:0080134	regulation of response to stress	12	307	0.0489
GO:0006468	protein phosphorylation	23	798	0.0493

Table S4. Information of molecular function with GO-term, functional description, background gene count and false discovery rate

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0005524	ATP binding	74	1939	5.65E-10
GO:0008144	drug binding	77	2074	5.65E-10
GO:0032555	purine ribonucleotide binding	79	2179	5.65E-10
GO:0032559	adenyl ribonucleotide binding	75	1971	5.65E-10
GO:0035639	purine ribonucleoside triphosphate binding	78	2147	5.65E-10
GO:0043168	anion binding	88	2629	5.65E-10
GO:0097367	carbohydrate derivative binding	80	2233	5.65E-10
GO:0000166	nucleotide binding	83	2461	1.11E-09
GO:0036094	small molecule binding	85	2633	4.39E-09
GO:0043167	ion binding	129	5070	1.15E-07
GO:0003824	catalytic activity	166	7239	3.57E-07
GO:0005488	binding	186	8611	2.22E-06
GO:0016462	pyrophosphatase activity	33	761	7.44E-06
GO:0017111	nucleoside-triphosphatase activity	32	722	7.44E-06
GO:1901363	heterocyclic compound binding	135	5835	9.68E-06
GO:0097159	organic cyclic compound binding	135	5841	9.77E-06
GO:0016772	transferase activity, transferring phosphorus-containing groups	39	1112	7.55E-05
GO:0016740	transferase activity	72	2847	0.001
GO:0016887	ATPase activity	21	498	0.001
GO:0140098	catalytic activity, acting on RNA	14	300	0.008
GO:0004386	helicase activity	10	167	0.008
GO:0046873	metal ion transmembrane transporter activity	11	200	0.008
GO:0016301	kinase activity	30	986	0.009
GO:0015318	inorganic molecular entity transmembrane transporter activity	21	597	0.011

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0022890	inorganic cation transmembrane transporter activity	15	362	0.013
GO:0046872	metal ion binding	67	2940	0.017
GO:0015386	potassium:proton antiporter activity	3	11	0.019
GO:0016773	phosphotransferase activity, alcohol group as acceptor	27	910	0.020
GO:0000146	microfilament motor activity	2	2	0.021
GO:0008184	glycogen phosphorylase activity	2	2	0.021
GO:0016780	phosphotransferase activity, for other substituted phosphate groups	3	12	0.021
GO:0102250	linear malto-oligosaccharide phosphorylase activity	2	2	0.021
GO:0102499	SHG alpha-glucan phosphorylase activity	2	2	0.021
GO:0042623	ATPase activity, coupled	15	391	0.021
GO:0005215	transporter activity	31	1138	0.026
GO:0022857	transmembrane transporter activity	29	1047	0.028
GO:0005388	calcium-transporting ATPase activity	3	16	0.034
GO:0047334	diphosphate-fructose-6-phosphate 1-phosphotransferase activity	2	4	0.039
GO:0015075	ion transmembrane transporter activity	19	609	0.043

Table S5. Information of cellular component with GO-term, functional description, background gene count and false discovery rate

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0005737	cytoplasm	197	7481	2.01E-14
GO:0044446	intracellular organelle part	138	4389	2.01E-14
GO:0044444	cytoplasmic part	172	6244	3.90E-14
GO:0005622	intracellular	235	10570	1.69E-10
GO:0005623	cell	259	12120	1.69E-10
GO:0043226	organelle	215	9369	1.69E-10
GO:0044424	intracellular part	233	10448	1.69E-10
GO:0043229	intracellular organelle	214	9362	2.17E-10
GO:0044464	cell part	258	12106	2.17E-10
GO:0043227	membrane-bounded organelle	206	9036	9.72E-10
GO:0043231	intracellular membrane-bounded organelle	202	8914	3.14E-09
GO:0009536	plastid	69	2064	5.89E-08
GO:0009507	chloroplast	68	2026	6.24E-08
GO:0016020	membrane	136	5592	3.27E-07
GO:0031090	organelle membrane	65	1985	3.27E-07
GO:0044434	chloroplast part	47	1205	3.27E-07

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0031967	organelle envelope	41	1001	6.54E-07
GO:0032991	protein-containing complex	65	2105	2.07E-06
GO:0098805	whole membrane	35	835	3.55E-06
GO:0005829	cytosol	52	1663	2.85E-05
GO:0044425	membrane part	96	3934	5.91E-05
GO:0009526	plastid envelope	26	598	6.32E-05
GO:0098588	bounding membrane of organelle	40	1178	6.78E-05
GO:0009570	chloroplast stroma	27	649	8.31E-05
GO:0031982	vesicle	23	517	1.40E-04
GO:0098796	membrane protein complex	21	469	2.90E-04
GO:0031410	cytoplasmic vesicle	21	502	7.20E-04
GO:0009941	chloroplast envelope	23	584	7.40E-04
GO:0005739	mitochondrion	36	1163	9.90E-04
GO:0030660	Golgi-associated vesicle membrane	6	43	1.00E-03
GO:0043232	intracellular non-membrane-bounded organelle	40	1369	1.30E-03
GO:0031969	chloroplast membrane	13	250	2.10E-03
GO:0005774	vacuolar membrane	20	536	3.30E-03
GO:0044431	Golgi apparatus part	20	534	3.30E-03
GO:0031984	organelle subcompartment	37	1306	3.40E-03
GO:0044433	cytoplasmic vesicle part	12	239	4.10E-03
GO:0005773	vacuole	27	869	5.00E-03
GO:0005794	Golgi apparatus	27	868	5.00E-03
GO:0030135	coated vesicle	7	88	5.00E-03
GO:0012510	trans-Golgi network transport vesicle membrane	3	9	5.10E-03
GO:0030120	vesicle coat	4	23	5.10E-03
GO:0005768	endosome	15	367	6.00E-03
GO:0009706	chloroplast inner membrane	6	68	6.50E-03
GO:0044428	nuclear part	30	1030	6.50E-03
GO:0016021	integral component of membrane	76	3460	7.70E-03
GO:0012505	endomembrane system	44	1753	8.60E-03
GO:0030662	coated vesicle membrane	5	50	9.20E-03
GO:0031224	intrinsic component of membrane	78	3602	9.20E-03
GO:0098791	Golgi subcompartment	18	520	1.01E-02
GO:0070013	intracellular organelle lumen	28	984	1.12E-02
GO:0098797	plasma membrane protein complex	4	31	1.12E-02
GO:0009506	plasmodesma	22	709	1.15E-02
GO:0005802	trans-Golgi network	10	215	1.30E-02
GO:0030136	clathrin-coated vesicle	5	57	1.30E-02
GO:0031981	nuclear lumen	24	813	1.30E-02
GO:0005886	plasma membrane	55	2406	1.33E-02
GO:0010319	stromule	4	35	1.42E-02

GO-term	Description	Observed gene count	Background gene count	False discovery rate
GO:0044459	plasma membrane part	13	339	1.53E-02
GO:0016459	myosin complex	3	17	1.59E-02
GO:0019866	organelle inner membrane	12	305	1.73E-02
GO:0019898	extrinsic component of membrane	5	64	1.90E-02
GO:0010287	plastoglobule	5	76	3.46E-02
GO:0030130	clathrin coat of trans-Golgi network vesicle	2	7	3.46E-02
GO:0009579	thylakoid	15	483	4.20E-02

References

1. Raulier, P.; Maudoux, O.; Notté, C.; Draye, X.; Bertin, P. Exploration of genetic diversity within Cichorium endivia and Cichorium intybus with focus on the gene pool of industrial chicory. *Genetic Resources and Crop Evolution* **2015**, *63*, 243-259, doi:10.1007/s10722-015-0244-4.
2. Kowalczyk, K.; Gajc-Wolska, J.; Marcinkowska, M.; Jabrucka-Pióro, E. Assessment of quality attributes of endive (Cichorium endivia L.) depending on a cultivar and growing conditions. *Acta Sci. Pol. Hortorum Cultus* **2015**, *14*, 13-16.
3. D'Antuono, L.F.; Ferioli, F.; Manco, M.A. The impact of sesquiterpene lactones and phenolics on sensory attributes: an investigation of a curly endive and escarole germplasm collection. *Food chemistry* **2016**, *199*, 238-245.
4. Lucchin, M.; Varotto, S.; Barcaccia, G.; Parrini, P. Chicory and endive. In *Vegetables I*, Springer: 2008; pp. 3-48.
5. Testone, G.; Mele, G.; di Giacomo, E.; Tenore, G.C.; Gonnella, M.; Nicolodi, C.; Frugis, G.; Iannelli, M.A.; Arnesi, G.; Schiappa, A., et al. Transcriptome driven characterization of curly- and smooth-leafed endives reveals molecular differences in the sesquiterpenoid pathway. *Hortic Res* **2019**, *6*, 1, doi:10.1038/s41438-018-0066-6.
6. Wang, F.-X.; Deng, A.-J.; Li, M.; Wei, J.-F.; Qin, H.-L.; Wang, A.-P. (3S)-1, 2, 3, 4-Tetrahydro-β-carboline-3-carboxylic Acid from Cichorium endivia. L Induces Apoptosis of Human Colorectal Cancer HCT-8 Cells. *Molecules* **2012**, *18*, 418-429.
7. Wang, S.Z.; Wang, B.C.; Liu, J.; Ren, J.; Huang, X.X.; Zhou, G.L.; Wang, A.H. Novel polymorphic EST-based microsatellite markers characterized in lettuce (Lactuca sativa). *Biologia* **2017**, *72*, 1300-1305, doi:10.1515/biolog-2017-0154.
8. Ryder, E. Physiology of germination, growth and development. In *Lettuce, endive chicory. Crop production science in horticulture*, Publishing, C., Ed. Nueva York, 1998; pp 54-78.
9. Van Inghelandt, D.; Melchinger, A.E.; Lebreton, C.; Stich, B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor Appl Genet* **2010**, *120*, 1289-1299, doi:10.1007/s00122-009-1256-2.
10. Lawson, C. Plant Breeder's Rights and Essentially Derived Varieties: Still Searching for Workable Solutions. *European Intellectual Property Review 499, Griffith University Law School Research Paper No. 16-17* **2016**.
11. 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*, 2016; 10.5772/61747pp. 67-87.

12. 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* **2003**, *122*, 178-183, doi:10.1046/j.1439-0523.2003.00786.x.
13. 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*, 831-843, doi:10.1186/s13104-015-1819-z.
14. 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* **2019**, *8*, 1-16, doi:10.3390/plants8070213.
15. 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:ARTN 27610.3389/fpls.2019.00276.
16. Cadalen, T.; Morchen, M.; Blassiau, C.; Clabaut, A.; Scheer, I.; Hilbert, J.L.; Hendriks, T.; Quillet, M.C. Development of SSR markers and construction of a consensus genetic map for chicory (Cichorium intybus L.). *Molecular Breeding* **2010**, *25*, 699-722, doi:10.1007/s11032-009-9369-5.
17. De Simone, M.; Morgante, M.; Lucchin, M.; Parrini, P.; Marocco, A.J.M.B. A first linkage map of Cichorium intybus L. using a one-way pseudo-testcross and PCR-derived markers. **1997**, *3*, 415-425.
18. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* **2000**, *18*, 233-234, doi:10.1038/72708.
19. Yeh, F.C.; Yang, R.-C.; Boyle, T.J.M.w.-b.f.f.p.g.a.U.o.A., Edmonton, Canada. POPGENE version 1.31. **1999**.
20. Hammer, Ø.; Harper, D.A.; Ryan, P.D. PAST: paleontological statistics software package for education and data analysis. *Palaeontologia electronica* **2001**, *4*, 1 - 9.
21. 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.
22. Earl, D.A.; Vonholdt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **2012**, *4*, 359-361, doi:10.1007/s12686-011-9548-7.
23. 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* **2014**, *32*, 691-696, doi:10.1007/s11105-013-0685-x.
24. 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**, 10.1111/mec.15253, 615385, doi:10.1111/mec.15253.

25. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N., et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **2012**, *40*, D1178-1186, doi:10.1093/nar/gkr944.
26. Reyes-Chin-Wo, S.; Wang, Z.; Yang, X.; Kozik, A.; Arikit, S.; Song, C.; Xia, L.; Froenicke, L.; Lavelle, D.O.; Truco, M.J., et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat Commun* **2017**, *8*, 14953, doi:10.1038/ncomms14953.
27. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.J.N.a.r. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **2018**, *47*, D607-D613.
28. Gene Ontology Consortium Internet. Availabe online: http://geneontology.org/ (accessed on 23rd September 2019).
29. STRING Availabe online: https://string-db.org (accessed on 26th September 2019).
30. Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **1987**, *4*, 406-425, doi:10.1093/oxfordjournals.molbev.a040454.
31. Chisari, M.; Todaro, A.; Barbagallo, R.N.; Spagna, G. Salinity effects on enzymatic browning and antioxidant capacity of fresh-cut baby Romaine lettuce (Lactuca sativa L. cv. Duende). *Food Chemistry* **2010**, *119*, 1502-1506, doi:10.1016/j.foodchem.2009.09.033.
32. Wulfkuehler, S.; Kurfiss, L.; Kammerer, D.R.; Weiss, A.; Schmidt, H.; Carle, R. Impact of different washing procedures on quality of fresh-cut iceberg lettuce (Lactuca sativa var. capitata L.) and endive (Cichorium endivia L.). *Eur Food Res Technol* **2013**, *236*, 229-241, doi:10.1007/s00217-012-1878-5.
33. Hoshino, A.A.; Bravo, J.P.; Angelici, C.M.L.C.D.; Barbosa, A.V.G.; Lopes, C.R.; Gimenes, M.A. Heterologous microsatellite primer pairs informative for the whole genus Arachis. *Genetics and Molecular Biology* **2006**, *29*, 665-675, doi:Doi 10.1590/S1415-47572006000400016.
34. Wang, M.; Gillaspie, A.; Newman, M.; Dean, R.; Pittman, R.; Morris, J.; Pederson, G. Transfer of simple sequence repeat (SSR) markers across the legume family for germplasm characterization and evaluation. *Plant Genetic Resources: Characterization Utilization* **2004**, *2*, 107-119.
35. Bombonato, J.R.; Bonatelli, I.A.S.; Silva, G.A.R.; Moraes, E.M.; Zappi, D.C.; Taylor, N.P.; Franco, F.F. Cross-genera SSR transferability in cacti revealed by a case study using Cereus (Cereeae, Cactaceae). *Genetics molecular biology* **2019**, *42*, 87-94.
36. Harijan, Y.; Nishanth, G.; Katageri, I.; Khadi, B. Research Note Transferability of heterologous SSR Markers to cotton genotypes. *Electronic Journal of Plant Breeding* **2017**, *8*, 379-384.
37. Kim, C.; Guo, H.; Kong, W.; Chandnani, R.; Shuang, L.S.; Paterson, A.H. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci* **2016**, *242*, 14-22, doi:10.1016/j.plantsci.2015.04.016.

38. Li, Z.; Zhang, Z.; Yan, P.; Huang, S.; Fei, Z.; Lin, K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* **2011**, *12*, 540, doi:10.1186/1471-2164-12-540.
39. Feng, J.Y.; Li, M.; Zhao, S.; Zhang, C.; Yang, S.T.; Qiao, S.; Tan, W.F.; Qu, H.J.; Wang, D.Y.; Pu, Z.G. Analysis of evolution and genetic diversity of sweetpotato and its related different polyploidy wild species I-trifida using RAD-seq. *Bmc Plant Biol* **2018**, *18*, 181, doi:ARTN 18110.1186/s12870-018-1399-x.
40. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **2016**, *17*, 81-92, doi:10.1038/nrg.2015.28.
41. Díaz-Arce, N.; Rodríguez-Ezpeleta, N. Selecting RAD-seq data analysis parameters for population genetics: the more the better? *Frontiers in genetics* **2019**, *10*, 533, doi: 10.3389/fgene.2019.00533.
42. Michael, T.P. Plant genome size variation: bloating and purging DNA. *Brief Funct Genomics* **2014**, *13*, 308-317, doi:10.1093/bfgp/elu005.
43. Nelissen, H.; De Groeve, S.; Fleury, D.; Neyt, P.; Bruno, L.; Bitonti, M.B.; Vandenbussche, F.; Van Der Straeten, D.; Yamaguchi, T.; Tsukaya, H. Plant Elongator regulates auxin-related genes during RNA polymerase II transcription elongation. *Proceedings of the National Academy of Sciences* **2010**, *107*, 1678-1683, doi:10.1073/pnas.0913559107.
44. An, C.; Ding, Y.; Zhang, X.; Wang, C.; Mou, Z. Elongator Plays a Positive Role in Exogenous NAD-Induced Defense Responses in Arabidopsis. *Mol Plant Microbe Interact* **2016**, *29*, 396-404, doi:10.1094/MPMI-01-16-0005-R.
45. TAIR. Availabe online: http://www.arabidopsis.org (accessed on 26th September 2019).