Genetic diversity investigation of the apple germplasm available at the Fondazione Edmund Mach

L. Poles[1], D. Micheletti[2], E. Banchi[2], L. Bianco[2], F. Costa[2], L. Lovatti[1] and R. Velasco[2]

[1]CIF, Fruit Innovation Consortium, via E. Mach 1, San Michele all'Adige, Trento, Italy; [2]CRI, Research and Innovation Centre, Foundation Edmund Mach, via E. Mach 1, San Michele all'Adige, Trento, Italy.

Abstract

Analysis of the genetic diversity harboured within a germplasm collection is a fundamental aspect for breeding, capable of providing information on the availability of an allelic pool potentially valuable for the creation of novel cultivars. The germplasm analysed in our study comprised 1639 *Malus* accessions available at the experimental orchards of Fondazione Edmund Mach (FEM), of which 1127 are unique accessions of *Malus* × *domestica*. This germplasm was collected to incorporate the maximum genetic variability of *M.* × *domestica* and to include representatives from other species of the genus *Malus*. In addition, the collection has been extended with the FEM's advanced selections and interspecific hybrids. Genetic identities and genetic diversity were investigated using a set of 17 simple sequence repeat (SSR) markers selected from the microsatellite inventory already available for apple, considering a representative for each apple chromosome. Markers were chosen on the base of important criteria, such as single locus targeting and high allelic variability. A total of 343 alleles were identified, with an average of 20 alleles for each SSR. Considering only *M.* × *domestica*, it was possible to identify all the cultivars unambiguously with just seven markers. The genetic stratification of the germplasm was investigated with principal component analysis (PCA) and model-based clustering, which allowed a clear separation of the domesticated and wild material. This analysis revealed a low level of structure and high admixture within the domesticated apple.

Keywords: fingerprint, SSR, microsatellite marker, population structure

INTRODUCTION

Breeding in apple can be traced back to the work of Thomas A. Knight (1759-1838), who produced the first cultivar of known parentage. Since then, many breeding programmes have been developed all over the world, but they seem to have been only moderately successful compared with other fruit crops. The main reasons for this are likely to be the relatively small number of elite cultivars employed as parental lines (Janick et al., 1996), the self-incompatibility property of this species, and its highly heterozygous genome.

Local and heritage accessions normally struggle to meet the standards of fruit quality and high production of modern cultivars (Liang et al., 2015); therefore, despite the number of cultivars that have been described in the literature (more than 7000), global apple production is still dominated by a few, closely related apple cultivars (Janick et al., 1996; Hokanson et al., 2001), leading to dramatic genetic erosion. To counterbalance this loss of genetic diversity, many apple collections have been established and preserved worldwide. The set-up of apple collections represents a valuable tool for breeders and researchers and, among other things, has facilitated the analysis of the allelic variability, since a considerable number of apple cultivars are planted in the same location, also enabling an easy exchange of accessions between different research institutions.

One of these apple collections is available at the Foundation Edmund Mach (FEM) in San Michele all'Adige, Italy, and includes 1639 accessions incorporating the maximum genetic variability of *Malus* × *domestica* (both heritage and local cultivars), as well as representatives from other species of the genus *Malus*.

Among the different types of molecular markers, short sequence repeats (SSR) are the ones of choice for genetic characterization of germplasm by virtue of their multi-allelism, co-dominance and reproducibility. Almost 300 apple-specific SSR markers are available nowadays, covering the entire genome (Guilford et al., 1997; Gianfranceschi et al., 1998; Liebhard et al., 2002). SSRs are routinely employed for several applications, such as fingerprinting (for cultivar identity and protection) and true-to-typeness checks (Liang et al., 2015; Hokanson et al., 2001). Thanks to high levels of allelic diversity within *Malus* species, a few, ad-hoc-chosen SSR markers are sufficient to efficiently distinguish apple cultivars thanks to their unique genetic fingerprints (Hokanson et al., 1998). Microsatellite markers have already been employed successfully to investigate genetic diversity and population structure among different apple collections from Belgium (Coart et al., 2003), Spain (Pereira-Lorenzo et al., 2007, 2008; Urrestarazu et al., 2012), the Netherlands (van Treuren et al., 2010) and France (Lassois et al., 2016).

In this study, we characterized 1639 accessions present in the FEM collection with 17 SSRs, in order to assess genetic diversity and to obtain useful insights about the genetic composition and structure of the FEM collection. We also investigated the smallest number of SSR markers required for optimal discrimination within *M.* × *domestica* accessions.

MATERIAL AND METHODS

Plant material

The germplasm collection analysed in the study comprised 1639 accessions. Plants are currently located in the experimental field of FEM, and included commercial and heritage cultivars together with accessions of wild *Malus*.

For each accession, young leaves were collected in spring, and each sample was labelled with a unique alphanumeric ID code and lyophilized. DNA was extracted following the CTAB protocol (Doyle and Doyle, 1991) optimized for 96-well plates. Reference cultivars were chosen and used as internal controls.

SSR analysis

Extracted DNA was quality checked and diluted to 10 ng μL^{-1}. Multiplex PCRs were performed using 17 pairs of fluorescently labelled primers, divided into four reactions (Table 1).

Table 1. Linkage group (LG), multiplex and dye of the 17 microsatellites used to fingerprint 1693 accessions.

LG	SSR	Multiplex	Dye	LG	SSR	Multiplex	Dye
1	CH05G08	M1	NED	10	CH02b03b	M1	FAM
2	CH05E03	M2	FAM	11	CH02D08	M3	NED
3	CH03G07	M1	HEX	12	CH01G12	M4	FAM
4	Hi23G02	M1	FAM	13	CH05H05	M3	HEX
5	CH04E03	M1	PET	14	CH01G05	M3	PET
6	CH03D12	M2	NED	15	NZ02B1	M4	NED
7	Hi03A10	M3	FAM	16	CH04F10	M4	HEX
8	CH01C06	M2	HEX	17	GD96	M4	PET
9	CH01F03b	M2	PET				

PCRs were performed in a final volume of 12.5 µL, with 0.2 µM of each primer and 0.64× of 2× Type-it Multiplex PCR Master Mix (Qiagen, Germany). The thermal cycle was as follows: initial denaturation step at 95°C for 5 min, followed by 30 cycles at 95°C for 30 s, 57°C for 90 s and 72°C for 30 s, and a final extension at 60°C for 30 min. PCR products were diluted to 1:50, and 1 µL of the dilution was mixed with 8.95 µL Hi-Di formamide (Thermo Fisher Scientific, USA) and 0.05 µL Genescan 500 LIZ size standard (Thermo Fisher Scientific,

USA).

Samples were denatured at 95°C for 15 min before being analysed with an ABI 3730 XL sequencing system (Thermo Fisher Scientific, USA). Electropherograms were analysed using the software GeneMapper4.0 (Thermo Fisher Scientific, USA) to size each fragment.

Data analysis

The genotypic profiles obtained with the 17 SSRs were checked for redundancy. An identity analysis was performed on the diploid accessions (two alleles at each locus maximum) with the CERVUS software (Marshall et al., 1998). Individuals with 15 or more loci in common were discarded, while the profiles of the putatively polyploid accessions were compared pairwise.

For each locus, the observed (A_o) and expected (A_e) number of alleles, and the observed (H_o) and expected (H_e) heterozygosity were calculated using CERVUS (Marshall et al., 1998). The null allele frequency (r) was determined as indicated by Brookfield (1996):

$$r = \frac{(H_e - H_o)}{(1 + H_e)}$$

The power of discrimination (PD) of the jth marker, defined as the ability of each locus to discriminate between two random cultivars, was calculated as follows:

$$PD = 1 - \sum p_i^2$$

where p_i is the frequency of the ith genotype (Kloosterman et al., 1993).

The methodology proposed by Tessier et al. (1999) to optimize the choice of molecular markers for varietal identification was followed. The first step was the calculation of the confusion probability (c_i), defined as the probability that two randomly chosen individuals from the subset have identical patterns or genotypes. For the ith pattern of the given j^{th} marker present at frequency p_i, the confusion probability (c_i) can be computed as follows:

$$c_i = p_i \frac{(Np_i - 1)}{N - 1}$$

where N represents the number of tested individuals. For all i genotypes generated by the jth marker, the global confusion probability (C_j) is equal to the sum of each c_i:

$$C_j = \sum_{i=1}^{I} c_i$$

The total number of non-differentiated pairs of cultivars for the jth marker (x_j) is given by:

$$x_j = \frac{N(N-1)}{2} C_j$$

which becomes:

$$X_k = \frac{N(N-1)}{2} \prod_{j=1}^{k} C_j$$

For a combination of *k* markers, X_k allows the calculation of the expected smallest possible number of SSRs able to discriminate the germplasm under analysis (Tessier et al., 1999).

Genetic diversity analysis

A principal component analysis (PCA) was performed using the Python library scikit-sklearn (Pedregosa et al., 2012) after conversion of the allele to a presence (1)/absence (0) state in all the accessions. The population stratification of the *M.* × *domestica* cultivars was inferred with Structure version 2 (Pritchard et al., 2000) software. This program uses a clustering method that identifies *K* subgroups of individuals with distinctive allele frequencies. Individuals can be members of multiple subpopulations with a different coefficient, provided that their sum equals 1. The program was run under the admixture model assumption. Each run comprised 1,000,000 iterations of the Monte Carlo Markov chain (MCMC) after a burn-in length of 100,000. Five independent runs were performed, each with a *K* (number of clusters) ranging from 1 to 15. The most probable number of subpopulations was estimated using the method proposed by Evanno et al. (2005).

RESULT AND DISCUSSION

SSRs analysis

One SSR per chromosome was chosen from the list of markers available at the Hidras website (http://users.unimi.it/hidras/). A set of 204 individuals (from the total of 1639) were discarded because of their low amplification rate, as they showed more than 20% missing data. The identity analysis carried out on the entire dataset showed that the initial dataset contained 21.3% redundancy. In particular, 121 samples were discarded because an profile identical to other samples was shown, and 186 individuals were removed because their profile matched with other samples for at least 15 SSRs. Among the final set of 1127 genotypes, 960 were diploids and 167 were considered as putative polyploids (showing more than two distinct alleles for more than at least three SSR loci). Statistical analyses were performed on two different datasets: one considering all diploid accessions from all species (not shown) and a second considering only the accessions of *M.* × *domestica* (Table 2).

The total number of alleles detected in the overall dataset including different apple species was 555, with a mean of 32.6 per marker. It was reduced to 343 for *M.* × *domestica* accessions (mean 20.4, ranging from a minimum of 15 in CH02B03B to a maximum of 27 in CH03D12; Table 2) and finally to 215 total alleles (mean 12.6; not shown) when the selections of the FEM breeding program (198 plants) were considered. The genetic diversity also decreased from 0.864 (entire dataset) to 0.827 (*M.* × *domestica*), and to 0.746 (FEM selections), which confirms the reduction of variability.

For the *M.* × *domestica* accessions, about 2.4% of the genotypes were missing for the 17 SSR loci (Table 2), with a higher incidence for CH04F10 (14.6%) and HI23G02 (7.2%). In addition, these two markers showed an estimated frequency of null allele $r>0.1$, and were therefore discarded and not considered for further analysis. The mean of the expected number of alleles per locus was 6.3 (varying from 3.2 for CH05G08 to 9.3 for CH04F10). Gene diversity for individual markers (Nei, 1978) ranged from 0.892 for CH01G12 to 0.686 for CH05G08 with a mean of 0.827, similar to what was reported previously by Lassois et al. (2016) and Liang et al. (2015). The SSR that showed the largest number of rare alleles (frequency less than 0.02) was CH05G08, with 77% of rare alleles, followed by HI03A10 (74%) and HI23G02 (71%). The marker with a lowest percentage (33%) of rare alleles was CH02b03b.

Table 2. Genetic diversity for all 17 microsatellites. LG, Linkage group; A_o, observed number of alleles; A_e, expected number of alleles; H_o, observed heterozygosity; H_e, expected heterozygosity; $F_{(null)}$, null allele frequency; rare alleles, alleles with frequencies less than 0.02; PD, power of discrimination; C_j, confusion probability.

LG	SSR	Missing data (%)	A_o	A_e	H_o	H_e	$F_{(null)}$	Rare alleles (%)	PD	C_j
1	CH05G08	0.8	22	3.2	0.710	0.686	-0.014	77	0.857	0.141
2	CH05E03	2.4	23	9.2	0.635	0.891	0.135	52	0.970	0.028
3	CH03G07	1	20	4.9	0.804	0.797	-0.004	70	0.932	0.066
4	Hi23G02	7.2	17	4.5	0.478	0.778	0.169	71	0.905	0.093
5	CH04E03	0.6	22	5.8	0.807	0.827	0.011	55	0.950	0.048
6	CH03D12	1.6	27	4.6	0.748	0.780	0.018	67	0.928	0.071
7	Hi03A10	2.2	23	6.3	0.798	0.842	0.024	74	0.945	0.053
8	CH01C06	0.2	18	5.2	0.794	0.807	0.007	61	0.937	0.062
9	CH01F03b	2	18	5.9	0.802	0.832	0.016	44	0.940	0.058
10	CH02b03b	0.6	15	7.4	0.885	0.865	-0.011	33	0.961	0.037
11	CH02D08	1.8	20	6.5	0.837	0.845	0.004	65	0.957	0.041
12	CH01G12	0.8	22	9.2	0.901	0.892	-0.005	50	0.975	0.023
13	CH05H05	0.6	16	4.9	0.811	0.795	-0.009	63	0.921	0.077
14	CH01G05	0.2	17	6.4	0.814	0.844	0.016	47	0.953	0.046
15	NZ02B1	2	20	5.8	0.816	0.827	0.006	55	0.943	0.055
16	CH04F10	14.6	23	9.3	0.684	0.893	0.110	52	0.971	0.027
17	GD96	2.6	20	7.5	0.881	0.866	-0.008	50	0.964	0.034
Mean		2.4	20.2	6.3	0.777	0.827	0.027	59	0.942	0.056
Total			**343**	**106.6**						

Genetic identity

Identity analysis showed 121 samples with an identical profile, including several clones of 'Golden Delicious', 'Fuji' and 'Gala'. Some cultivars were entered in multiple copies with different names (for example, 'Enterprise'/Coop 30, 'Priscilla'/Coop 4, 'Goldrush'/Coop 38, 'Mela di cera'/'Cerine' and 'Belfiore di Trento'/'Gelber Bellefleur') or wrong or unknown names, indicated only by a numeric code and field coordinates. Genetic identity based on the SSR profile excluded mistakes and resolved these problems of identification.

A total of 186 samples showed matching profiles for 16 or 15 SSRs, which were compared pairwise to confirm possible base shifts or genotyping errors that could have hampered the scoring of the electropherograms. Among these samples, we also identified somatic mutants ('Stark Delicious', 'Topred Delicious', 'Auvil Spur', 'Morspur Red', 'Top Crop', etc.) that were further removed from the analysis. In addition, 35 unlabelled plants were characterized thanks to an identical profile with another cultivar(s) within the germplasm.

Structure identification

The PCA performed on 763 accessions from 47 species highlighted a clear separation between domesticated and wild accessions (Figure 1). The total variance explained by the first two principal components represents only the 4.4% of the variance present in the dataset.

The model based on Bayesian clustering adopted here to group the 500 accessions of *M.* × *domestica* showed that the most probable number of subpopulations is 4, and the Δk is 124, with a smaller peak Δk=24 identified at K=8. At K=4, the number of individuals assigned with a Q value higher than 0.7 was 189, and the largest population included 64 individuals. One of those subpopulations included all the heritage cultivars (green in Figure 2), while the major apple cultivars were spread over the other three groups: 'Cox's Orange Pippin' and 'McIntosh' are grouped together (red in Figure 2), 'Jonathan' and 'Golden Delicious' are depicted in black, while 'Fuji', 'Delicious' and 'Braeburn' showed higher

similarity and all belong to the same group (blue in Figure 2). The high proportion of individuals with high admixture indicates a weak population structure in *M. × domestica.* This could be due to the high level of gene flow between different accessions and to the relatively small number of generations since the domestication that took place in the Tian Shan region about 4000 years ago (Cornille et al., 2012).

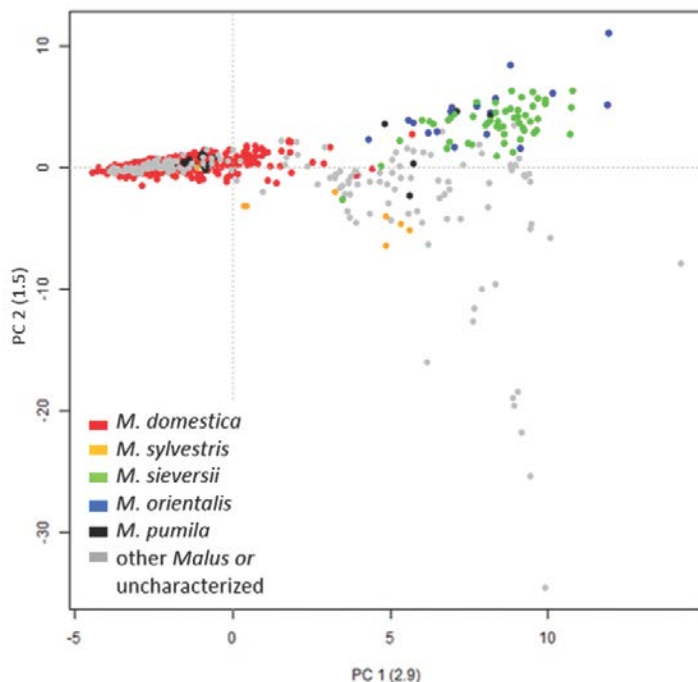

Figure 1. Principal component analysis of the entire dataset. Accessions of *M. × domestica* and closely related species are highlighted with different colours.

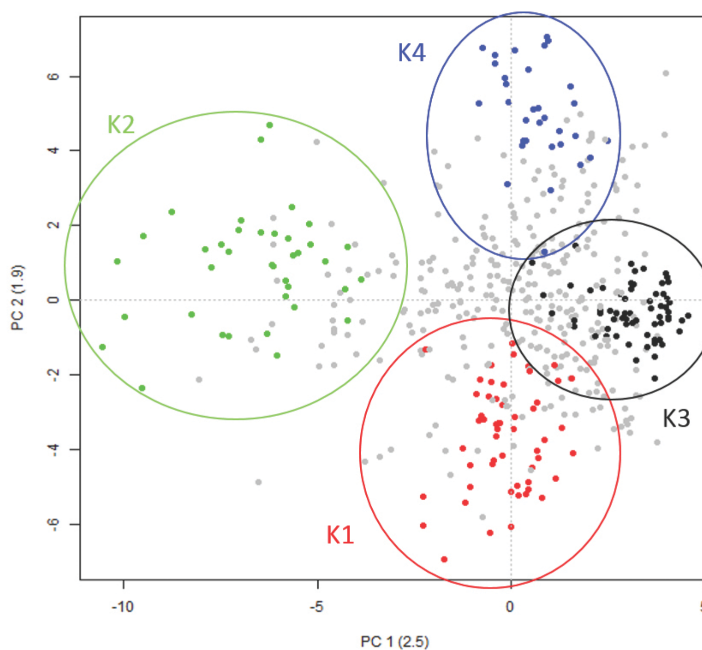

Figure 2. PCA of the *M. × domestica* subset integrated with Structure results. Samples in grey present high admixture, while the coloured ones were assigned with a Q value greater than 0.7 to one of the four subpopulations.

Optimization of the choice of markers

Most of the SSRs employed in this survey were used previously for apple fingerprinting: CH03G07 and CH01C06 were included by Pereira-Lorenzo et al. (2007, 2008); CH05E03 and CH02D08 by Gasi et al. (2010); and CH05G08, CH02D08 and CH04F10 by Urrestarazu et al. (2012). CH01F03B and CH02D08 belonged to the 17 SSRs recommended by the European Collaborative Programme for Crop Genetic Resources (ECPGR) *Malus*/*Pyrus* working group (Lateur et al., 2013). CH02D08 and GD96 overlap with the set of markers used to fingerprint the USDA-ARS National Plant Germplasm System (NPGS) by the US Department of Agriculture (Gross and Volk, 2012), while CH01F03B, CH02D08, CH03D12 and CH01G05 were included in the set used by Lassois et al. (2016). Despite the existence of reference cultivars and SSRs shared among the different sets used worldwide, which allow harmonization and comparison of the results, a unique recognized set of markers used for apple varietal identification is still lacking.

Our set of SSRs showed a PD ranging from 0.857 to 0.975, with an average of 0.942 (Table 2), which is comparable to that observed by Urrestarazu et al. (2012) and Lassois et al. (2016), with CH01G12 and CH05E03 having the greatest ability to discriminate different cultivars. PD can guide the selection of the most discriminative SSRs. In our dataset, just seven SSRs can fully differentiate the investigated *M.* × *domestica* germplasm (Table 3): the first five were ranked according to their PD value, while HI03A10 and NZ02B1 were added to discriminate the 12 remaining pairs of individuals. CH01G05 and CH04E03 have higher PD but, as the last primers of the set, they were not able to conclude the discrimination of the dataset (observed non-distinguishable pairs: 10 for CH01G05, 8 for CH01G05+CH04E03). According to our study, these seven SSRs are therefore the most strongly indicated for routine analysis of varietal identification (Table 3).

Table 3. The combination of primers that most reduced the number of undifferentiated pairs, comparing the real and theoretical efficiency according to Tessier et al. (1999).

Marker combination	No. undifferentiated pairs	
	Observed	Expected
CH01G12	464	2794.6855
CH01G12+CH05E03	152	78.8584
CH01G12+CH05E03+GD96	33	2.6974
CH01G12+CH05E03+GD96+CH02b03b	20	0.0989
CH01G12+CH05E03+GD96+CH02b03b+CH02D08	12	0.0041
CH01G12+CH05E03+GD96+CH02b03b+CH02D08+Hi03A10	8	0.0002
CH01G12+CH05E03+GD96+CH02b03b+CH02D08+Hi03A10+NZ02B1	0	0.0000

CONCLUSIONS

The FEM germplasm collection has been characterized in this study by using 17 SSRs, and their genetic diversity has been assessed. SSRs was shown to be an effective and reliable type of marker for such a purpose. The high stratification within our germplasm will represent a valuable tool for breeders and researchers to counterbalance the genetic erosion resulting from intensive breeding programmes. At the same time, we demonstrated that an accurate choice of SSRs based on their genetic descriptive statistics could reduce the time and cost of discriminating genotypes.

Literature cited

Brookfield, J.F. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. Mol. Ecol. *5* (*3*), 453–455 https://doi.org/10.1111/j.1365-294X.1996.tb00336.x. PubMed

Coart, E., van Bockstaele, E., Roldán-Ruiz, I., Wagner, I., van Huylenbroeck, J., Smulders, M.J.M., and Vekemans, X. (2003). Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. Mol. Ecol. *12* (*4*), 845–857

https://doi.org/10.1046/j.1365-294X.2003.01778.x. PubMed

Cornille, A., Gladieux, P., Smulders, M.J.M., Roldán-Ruiz, I., Laurens, F., Le Cam, B., Nersesyan, A., Clavel, J., Olonova, M., Feugey, L., et al. (2012). New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet. *8* (*5*), e1002703 https://doi.org/10.1371/journal.pgen.1002703. PubMed

Doyle, J., and Doyle, J. (1991). DNA isolation from a small amount of plant tissue. Phytochem. Bull. *57*, 13–15.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14* (*8*), 2611–2620 https://doi.org/10.1111/j.1365-294X.2005.02553.x. PubMed

Gasi, F., Simon, S., Pojskic, N., Kurtovic, M., and Pejic, I. (2010). Genetic assessment of apple germplasm in Bosnia and Herzegovina using microsatellite and morphologic markers. Sci. Hortic. (Amsterdam) *126* (*2*), 164–171 https://doi.org/10.1016/j.scienta.2010.07.002.

Gianfranceschi, L., Seglias, N., Tarchini, R., Komjanc, M., and Gessler, C. (1998). Simple sequence repeats for the genetic analysis of apple. Theor. Appl. Genet. *96* (*8*), 1069–1076 https://doi.org/10.1007/s001220050841.

Gross, B., and Volk, G. (2012). Identification of "duplicate" accessions within the USDA-ARS National Plant Germplasm System *Malus* collection. J. Am. Soc. Hortic. Sci. *137*, 333–342.

Guilford, P., Prakash, S., Zhu, J.M., Rikkerink, E., Gardiner, S., Bassett, H., and Forster, R. (1997). Microsatellites in *Malus* × *domestica* (apple): abundance, polymorphism and cultivar identification. Theor. Appl. Genet. *94* (*2*), 249–254 https://doi.org/10.1007/s001220050407.

Hokanson, S.C., Szewc-McFadden, A.K., Lamboy, W.F., and McFerson, J.R. (1998). Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a *Malus*×*domestica* Borkh. core subset collection. Theor. Appl. Genet. *97* (*5-6*), 671–683 https://doi.org/10.1007/s001220050943.

Hokanson, S.C., Lamboy, W.F., Szewc-McFadden, A.K., and McFerson, J.R. (2001). Microsatellite (SSR) variation in a collection of *Malus* (apple) species and hybrids. Euphytica *118* (*3*), 281–294 https://doi.org/10.1023/A:1017591202215.

Janick, J., Cummins, J.N., Brown, S.K., and Hemmat, M. (1996). Apples. In Fruit Breeding, Tree and Tropical Fruits, Vol.1, J. Janick, and J.N. Moore, eds. (New York, NY, USA: John Wiley & Sons), p.1–78.

Kloosterman, A.D., Budowle, B., and Daselaar, P. (1993). PCR-amplification and detection of the human D1S80 VNTR locus. Amplification conditions, population genetics and application in forensic analysis. Int. J. Legal Med. *105* (*5*), 257–264 https://doi.org/10.1007/BF01370382. PubMed

Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrand-Saint-Oyant, L., Poncet, C., Lasserre-Zuber, P., Feugey, L., and Durel, C.E. (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. Plant Mol. Biol. Report. *34* (*4*), 827–844 https://doi.org/10.1007/s11105-015-0966-7.

Lateur, M., Ordidge, M., Engels, J., and Lipman, E. (2013). Report of a Working Group on *Malus*/*Pyrus*. Fourth Meeting, 7-9 March 2012, Weggis, Switzerland (Rome, Italy: Bioversity International), p.26–27.

Liang, W., Dondini, L., De Franceschi, P., Paris, R., Sansavini, S., and Tartarini, S. (2015). Genetic diversity, population structure and construction of a core collection of apple cultivars from Italian germplasm. Plant Mol. Biol. Report. *33* (*3*), 458–473 https://doi.org/10.1007/s11105-014-0754-9.

Liebhard, R., Gianfranceschi, L., Koller, B., Ryder, C.D., Tarchini, R., van de Weg, E., and Gessler, C. (2002). Development and characterisation of 140 new microsatellites in apple (*Malus* × *domestica* Borkh.). Mol. Breed. *10* (*4*), 217–241 https://doi.org/10.1023/A:1020525906332.

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

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89* (*3*), 583–590. PubMed

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2012). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Pereira-Lorenzo, S., Ramos-Cabrer, A.M., and Diaz-Hernandez, M.B. (2007). Evaluation of genetic identity and variation of local apple cultivars (*Malus* × *domestica* Borkh.) from Spain using microsatellite markers. Genet. Resour. Crop Evol. *54* (*2*), 405–420 https://doi.org/10.1007/s10722-006-0003-7.

Pereira-Lorenzo, S., Ramos-Cabrer, A.M., González-Díaz, A.J., and Díaz-Hernández, M.B. (2008). Genetic assessment of local apple cultivars from La Palma, Spain, using simple sequence repeats (SSRs). Sci. Hortic.

(Amsterdam) *117* (*2*), 160–166 https://doi.org/10.1016/j.scienta.2008.03.033.

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

Tessier, C., David, J., This, P., Boursiquot, J.M., and Charrier, A. (1999). Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera* L. Theor. Appl. Genet. *98* (*1*), 171–177 https://doi.org/10.1007/s001220051054.

Urrestarazu, J., Miranda, C., Santesteban, L.G., and Royo, J.B. (2012). Genetic diversity and structure of local apple cultivars from Northeastern Spain assessed by microsatellite markers. Tree Genet. Genomes *8* (*6*), 1163–1180 https://doi.org/10.1007/s11295-012-0502-y.

van Treuren, R., Kemp, H., Ernsting, G., Jongejans, B., Houtman, H., and Visser, L. (2010). Microsatellite genotyping of apple (*Malus* × *domestica* Borkh.) genetic resources in the Netherlands: application in collection management and variety identification. Genet. Resour. Crop Evol. *57* (*6*), 853–865 https://doi.org/10.1007/s10722-009-9525-0.