

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Comparison of selection methods for the establishment of a core collection using SSR markers for hazelnut (*Corylus avellana* L.) accessions from European germplasm repositories

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1837634> since 2022-02-01T14:52:38Z

Published version:

DOI:10.1007/s11295-021-01526-7

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

1 **Comparison of selection methods for the establishment of a core collection**
2 **using SSR markers for hazelnut (*Corylus avellana* L.) accessions from European**
3 **germplasm repositories**

4
5 Paolo Boccacci ¹⁾, Maria Aramini ²⁾, Matthew Ordidge ³⁾, Theo J.L. van Hintum ⁴⁾, Daniela Torello
6 Marinoni ⁵⁾, Nadia Valentini ⁵⁾, Jean-Paul Sarraquigne ⁶⁾, Anita Solar ⁷⁾, Mercè Rovira ⁸⁾, Loretta
7 Bacchetta ²⁾, Roberto Botta ⁵⁾

8
9 ¹⁾ Institute for Sustainable Plant Protection - National Research Council (IPSP-CNR). Strada delle Cacce, 73 -
10 10135 Torino, Italy

11 ²⁾ ENEA - Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile, Via
12 Anguillarese 301 - 00123 S.M. di Galeria (RM), Italy

13 ³⁾ School of Agriculture, Policy and Development, University of Reading, Whiteknights, RG6 6AR, Reading,
14 United Kingdom

15 ⁴⁾ Centre for Genetic Resources, the Netherlands, Wageningen Plant Research, P.O. Box 16, 6700 AA,
16 Wageningen, the Netherlands

17 ⁵⁾ Department of Agricultural, Forestry and Food Science - University of Torino, Largo Paolo Braccini, 2 - 10095
18 Grugliasco (TO), Italy

19 ⁶⁾ Association Nationale des Producteurs de Noisette (ANPN), 47290 Cancon, France

20 ⁷⁾ University of Ljubljana, Biotechnical Faculty, Department of Agronomy, Jamnikarjeva 101, 1000 Ljubljana,
21 Slovenia

22 ⁸⁾ Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre Mas Bové, Ctra. Reus-El Morell, km 3.8 -
23 43120 Constantí (Tarragona), Spain

24

25 **Corresponding author:** Paolo Boccacci, e-mail: paolo.boccacci@ipsp.cnr.it

26

27

28 **ORCID ID of the authors**

29 Paolo Boccacci: 0000-0001-8574-0478; Matthew Ordidge: 0000-0003-0115-5218; Theo van Hintum: 0000-0003-

30 4953-4700; Daniela Torello Marinoni: 0000-0002-3679-4813; Nadia Valentini: 0000-0002-8820-9006; Merce

31 Rovira: 0000-0002-0540-3752; Loretta Bacchetta: 0000-0001-8878-4054; Anita Solar: 0000-0001-9755-4998;

32 Roberto Botta: 0000-0002-1952-8775.

1 **Abstract**

2 Hazelnut (*Corylus avellana* L.) is one of the most important tree nut crops in Europe. Germplasm
3 accessions are conserved in *ex situ* repositories, located in countries where hazelnut production occurs.
4 In this work, we used ten simple sequence repeat (SSR) markers as the basis to establish a core collection
5 representative of the hazelnut genetic diversity conserved in different European collections. A total of
6 480 accessions were used: 430 from *ex situ* collections and 50 landraces maintained *on-farm*. SSR
7 analysis identified 181 genotypes, that represented our whole hazelnut germplasm collection (WHGC).
8 Four approaches (utilizing MSTRAT, Power Core, and Core Hunter's single- and multi-strategy) based
9 on the maximization (M) strategy were used to determine the best sampling method. Core Hunter's
10 multi-strategy, optimizing both allele coverage (Cv) and Cavalli-Sforza and Edwards (Dce) distance
11 with equal weight, outperformed the others and was selected as the best approach. The final core
12 collection (Cv-Dce30) comprised 30 entries (16.6% of genotypes). It recovered all SSR alleles and
13 preserved parameter variations when compared to WHGC. Entries represented all six gene pools
14 obtained from the population structure analysis of WHGC, further confirming the representativeness of
15 Cv-Dce30. Our findings contribute towards improving the conservation and management of European
16 hazelnut genetic resources and could be used to optimize future research by identifying a minimum
17 number of accessions on which to focus.

18

19

20 **Key words:** Filbert; Microsatellite; *Ex situ* and *in situ* conservation; Germplasm management; Plant
21 genetic resources

22

1 Introduction

2 The European hazelnut (*Corylus avellana* L.) is one of the most important tree nut crops in terms of
3 worldwide production (averaging 939,927 tons per annum in 2015-2019). The Black Sea countries
4 account for most of the annual world production (averaged from 2015-2019): Turkey (606,409 tons),
5 Azerbaijan (43,584 tons), and Georgia (25,440 tons). Other important producers are Italy (116,945 tons),
6 the USA (36,652 tons), Iran (15,583 tons), France (11,994 tons), and Spain (10,364 tons) followed by
7 Chile, Poland, Serbia, Kyrgyzstan, and Uzbekistan (FAOSTAT 2021). World production is based
8 entirely on cultivars selected over many centuries from local wild populations (Thompson et al. 1996)
9 and about 500 cultivars have been described in the literature and are available from one or more *ex situ*
10 germplasm repositories (Köksal 2000; Botta et al. 2019). However, only about 20 of these cultivars are
11 widely grown and another 30 are currently considered promising for breeding (Botta et al. 2019).
12 Collections consist primarily of cultivated forms of *C. avellana* and are mainly located in countries
13 where production occurs. A total of 510 hazelnut accessions, corresponding to 222 cultivars and 58
14 selections, are documented to be conserved in 12 European collection fields: four in Italy, three in
15 Portugal, two in Spain, and one each in Slovenia, France, and Greece (Bacchetta et al. 2015; Botta et al.
16 2019). Moreover, a relatively small collection of 48 cultivars, including accessions sourced from Europe
17 and the USA, is held in the United Kingdom (Köksal 2000). More than 700 *Corylus* accessions are
18 preserved in the major world hazelnut collection located in Oregon (USA) (Hummer 2001), while a
19 collection containing 20 registered cultivars and more than 400 accessions collected from the Black Sea
20 coast is held in Turkey (Öztürk et al. 2017). *In situ* conservation strategies have been applied only
21 recently, after *on-farm* explorations conducted in southern Europe (Ferreira et al. 2010; Boccacci et al.
22 2013).

23 Germplasm collections ensure the long-term conservation of genetic resources and provide easy
24 access to plant breeders, researchers, and other users. The management and use of large germplasm
25 collections requires significant economic costs for routine tasks, such as conservation, regeneration,
26 duplication, documentation, and evaluation. Moreover, collections invariably contain duplicate and
27 redundant accessions that may invalidate both the efficiency of the conservation and the effectiveness

1 of germplasm evaluation and use (van Hintum et al. 2000). Consequently, the long-term conservation
2 of collections can be endangered. Thus, a core collection concept was introduced in the 1980s to define
3 a subset of accessions from a collection that represents, with a minimum of repetitiveness, the genetic
4 diversity of a crop species and its wild relatives (van Hintum et al. 2000). Most core collections include
5 5-20% of the accessions and capture 70-90% of the diversity present in the collection (van Hintum et
6 al. 2000). Core collections do not replace the whole collections from which they are obtained, however
7 they can optimize the characterization and evaluation efforts by focusing on a subset of accessions (van
8 Hintum et al. 2000). Recognizing these objectives, core collections were recommended by the global
9 plan of action for the conservation and sustainable utilization of plant genetic resources for food and
10 agriculture as a necessary activity to encourage the use of genetic resources (FAO 1996).

11 The development of core collections was traditionally based on passport data, which are often
12 unreliable and incomplete, or phenotypic traits, which are influenced by environmental factors. DNA
13 markers, such as simple sequence repeats (SSRs) and single-nucleotide polymorphisms (SNPs), are now
14 often the tool of choice for the development of core collections. They can accurately represent the
15 genetic diversity of the whole collection and avoid the problems related to incomplete provenance and
16 environmental interaction, typically associated with passport and phenotypic data. SSRs generally show
17 a higher level of polymorphism than SNPs, resulting in the appearance of population-specific alleles
18 that are useful for revealing population structure. Nevertheless, SSRs are usually developed in small
19 numbers within a species, and they may reflect less genome-wide diversity in comparison to larger
20 numbers of SNPs (Bernard et al. 2020). The latter are much more frequent in the genome and many
21 SNPs can be identified using high-throughput genomics tools, allowing to develop panels of markers
22 useful for genetic diversity and fine mapping. However, in hazelnut, a total of 718 SSR markers have
23 been developed and more than 430 of them were used for the development of a reference linkage map
24 (Mehlenbacher 2018; Botta et al. 2019). They have been used to fingerprint cultivars, to identify
25 duplicated accessions and parents, to study genetic diversity in cultivated and wild populations, and in
26 association mapping studies (Mehlenbacher 2018; Botta et al. 2019). By comparison, SNPs have only
27 been used in hazelnut recently to develop two high-density genetic maps (Botta et al. 2019).

1 Different strategies and bioinformatic tools have been proposed to construct core collections
2 (Schoen and Brown 1993; Marita et al. 2000; Franco et al. 2005) and these have been compared in
3 annual (e.g., Franco et al. 2006) and perennial species (e.g., Escribano et al. 2008). Both studies
4 concluded that the maximization (M) strategy (Schoen and Brown 1993), which maximizes the number
5 of alleles, is highly suitable for constructing core collections. Several algorithms based on the M-strategy
6 have been developed and implemented in different software, such as MSTRAT (Gouesnard et al. 2001),
7 PowerCore (Kim et al. 2007), and Core Hunter (Thachuk et al. 2009; De Beukelaer et al. 2012).

8 Many studies concerning the construction of core collections have been performed in annual
9 species. Nevertheless, the benefits are perhaps most evident in woody perennial species, which are
10 usually maintained as clones in collection fields, due to higher management costs per accession than
11 annual seed germplasm (Escribano et al. 2008). Development of core collections using SSR markers has
12 been presented in several fruit tree species, including: apple (Liang et al. 2015; Lassois et al. 2016),
13 apricot (Wang et al. 2011), carob tree (Di Guardo et al. 2019), chestnut (Pereira-Lorenzo et al. 2017),
14 fig (Balas et al. 2014), grape (Le Cunff et al. 2008; Štajner et al. 2014), hazelnut (Öztürk et al. 2017),
15 olive (Belaj et al. 2012; Díez et al. 2012; El Bakkali et al. 2013), pear (Miranda et al. 2010; Liu et al.
16 2015), and walnut (Bernard et al. 2020).

17 The main objective of this work was to develop a core collection representative of the hazelnut genetic
18 diversity conserved in different *ex situ* and *in situ* European germplasm repositories. For that purpose,
19 first we used a range of different M-strategy approaches to build and select the respective best subsets
20 based on ten SSR markers. In a second step, the diversity parameters of subsets were compared to select
21 the final core collection. A quality evaluation of each sampling method was performed following the
22 model proposed by Odong et al. (2013). Finally, a population structure and relatedness among genotypes
23 were also investigated.

24

25

26

27

1 **Material and methods**

2 **Plant material and microsatellite genotyping**

3 A total of 410 hazelnut accessions were analysed from nine different *ex situ* germplasm repositories
4 located in six Countries: UK, Spain, France, Italy, Slovenia, and USA. Moreover, 6 landrace accessions
5 were also collected from an *on-farm* survey in the Nuoro province (Sardinia, Italy) (Online Resource 1).

6 Total genomic DNA was extracted from 0.20 g of young leaves or immature catkins using the
7 modified procedure of Thomas et al. (1993). A total of 10 SSR loci selected by Boccacci and Botta
8 (2010) were analysed: CaT-B107, CaT-B501, CaT-B502, CaT-B503, CaTB504, CaT-B505, CaT-B507,
9 CaT-B508 (Boccacci et al. 2005), CaC-B020, and CaC-B028 (Bassil et al. 2005). PCR amplifications
10 were performed in a volume of 15 µl containing 40 ng DNA, 0.5 U Taq-DNA polymerase (Bioline,
11 Meridian Bioscience, OH, USA), 1.5 µl 10x PCR buffer (Bioline, Meridian Bioscience), 2.2 mM MgCl₂,
12 200 µM dNTPs, and 0.5 µM of each primer. The PCR conditions were: a first denaturation step at 95
13 °C for 9 min, followed by 26 cycles of denaturation (30 s at 95 °C), annealing (45 s at 55 °C and 50 °C
14 for CaT-B502), and extension (90 s at 72 °C). The final elongation step was carried out at 72 °C for 30
15 min.

16 Total genomic DNA was extracted from the UK samples using the Nucleospin[®] Plant II kit
17 (Macherey-Nagel GmbH & Co. KG, Deuren, Germany) according to manufacturer's instructions. SSR
18 loci were the same as above, but loci were amplified in two multiplex reactions: i) MP1: CaT B107,
19 CaT B501, CaT B502, CaT B504, and CaC B028; ii) MP2: CaT B503, CaT B505, CaT B507, CaT
20 B508, and CaC B020. PCR amplifications were performed in a volume of 11 µl containing 10 ng DNA
21 and using the Type-it Microsatellite PCR kit (QIAGEN, MD, USA) according to the manufacturer's
22 protocol. The PCR conditions were: a first denaturation step at 95 °C for 5 min, followed by 35 cycles
23 of denaturation (30 s at 95 °C), annealing (45 s at 55 °C decreasing by 0.5 °C per cycle for the first 10
24 cycles), and extension (60 s at 72 °C). The final elongation step was carried out at 72 °C for 15 min.

25 Amplification products were analysed using an ABI-PRISM 3130 Genetic Analyzer capillary
26 electrophoresis instrument; UK samples were analysed using an ABI 3730xl capillary electrophoresis
27 instrument (both, Applied Biosystems, Foster City, CA, USA). Results were processed with

1 GeneMapper software (Applied Biosystems), and alleles were designated by their size in base pairs
2 using a GeneScan-500 LIZ standard (Applied Biosystems). UK data were aligned to the main dataset
3 by applying a simple conversion based on a series of overlapping cultivars between the two datasets.

4 Data obtained at the same SSR loci reported by Boccacci et al. (2013) for 17 reference cultivars
5 from the Hazelnut Research Institute (HRI) at Giresun (Turkey), 3 reference cultivars from the National
6 Agricultural Research Foundation - Pomology Institute (NAGREF-PI) at Naoussa (Greece), and 44
7 landraces surveyed *on-farm* in southern Europe were also added. Thus, microsatellite data from a total
8 of 480 samples (430 from *ex situ* collections and 50 landraces maintained *on-farm*) were processed using
9 the software Identity 4.0 (Wagner and Sefc 1999) to calculate the total probability of identity (Paetkau
10 et al. 1995) and to identify samples with identical SSR genotypes. When two or more samples had
11 identical SSR genotype, only one was retained for further analysis.

12

13 **Construction of the core collections by different M-strategies**

14 Three different approaches based on the maximization (M) strategy were used to build core collections:

- 15 • The standard M-strategy described by Schoen and Brown (1993) was employed as implemented in
16 MSTRAT (Gouesnard et al. 2001). Nei's diversity index (Nei 1987) was used as diversity criterion.
17 A total of five subsets containing 10, 20, 30, 40 and 50 entries, respectively, were developed, together
18 with the optimized subset selected by the software algorithm. For each sampling size, 100
19 independent replicates and 200 iterations were generated. The replicates that maintained the highest
20 number of alleles and genetic diversity scores were selected;
- 21 • The advanced M-strategy proposed by Kim et al. (2007) was carried out as implemented in
22 PowerCore v. 1.0;
- 23 • The advanced stochastic local search (SLS) algorithm, replica exchange Monte Carlo, developed by
24 Thachuk et al. (2009) and implemented in Core Hunter II (De Beukelaer et al. 2012) was used for
25 the third approach. The software can select core subsets using different allocation strategies by
26 optimizing one genetic parameter or many parameters simultaneously. By maximizing only genetic
27 distance parameters the software selects the most genetically distant accessions, whereas by

1 optimizing diversity index accessions are selected with the highest allelic variability. In this study,
2 six allocation methods were used: i) optimizing each of the following measures independently:
3 average Cavalli-Sforza and Edwards (Dce) and Modified Rogers (Mr) as genetic distance indices,
4 expected proportion of heterozygous loci (He) and Shannon's diversity index (Sh) as allelic diversity
5 indices, and allele coverage (Cv); ii) optimizing simultaneously both Cv and Dce (Cv-Dce) with
6 equal weight assigned to each parameter. Indeed, when a weight of 50% was assigned to Cv and 50%
7 to Dce, all observed alleles were captured in the sampled subset (Online Resource 2, Fig. S1). For
8 each strategy, five subsets containing 10, 20, 30, 40, and 50 entries, respectively, were developed.

10 **Characterization and validation of the representativeness of the core collections**

11 In order to evaluate the ability of each sampling strategy to capture the diversity represented in the whole
12 germplasm collection, different parameters were considered: number of alleles (A), genetic diversity
13 (GD), observed heterozygosity (Ho), polymorphism information content (PIC), and allele frequencies
14 (Fr), that were calculated using PowerMarker v.3.25 (Liu and Muse 2005). Significant differences in A,
15 GD, Ho, and PIC values between the cores and the whole collection were examined using a *post hoc*
16 Dunnett's test ($P < 0.05$), computed by the ANOVA analysis with the SPSS software (IBM, Armonk,
17 NY, USA). Significant differences in Fr at each locus were analysed independently, comparing the
18 frequency of each allele between the whole collection and each core subset by the chi-squared test ($P <$
19 0.05).

20 After selecting the best subset for each sampling strategy, based on the above parameters, the
21 representativeness of the subsets was validated against the criteria proposed by Escribano et al. (2008),
22 which expected a representative core collection to: i) all SSR alleles present in the original collection
23 captured ; ii) no significant differences in frequency distribution of alleles in at least 95% of the loci
24 from that of the whole collection; iii) no significant differences in diversity indices, GD and Ho, between
25 the core and the whole collection.

1 **Comparison and quality of sampling strategies**

2 Once the most representative core subsets were determined for each strategy, the effectiveness of each
3 sampling method was evaluated following the criteria reported by Thachuk et al. (2009) which expected
4 the best subset to have the: i) highest average genetic distance between accessions; ii) highest allele
5 richness; iii) lowest proportion of non-informative alleles and the highest allele coverage. In order to
6 assess this, Modified Rogers (MR) and Cavalli-Sforza and Edwards (CE) genetic distances, Shannon's
7 diversity index (SH), expected proportion of heterozygous loci (HE), number of effective alleles (NE),
8 proportion of non-informative alleles (PN), and allele coverage (CV) were calculated for each core
9 collection and compared with respect to the whole collection. Each parameter was optimized
10 independently by performing 20 runs using Core Hunter II (De Beukelaer et al. 2012).

11 The quality of each sampling method was also determined against two criteria proposed by Odong
12 et al. (2013):

- 13 • Average distance between each accession in the whole collection and the nearest entry in the core
14 collection (A-NE), a criterion to indicate the representativeness of a core collection. If the A-NE
15 realized value is low, there is always an entry close to each accession;
- 16 • Average distance between each entry in the core collection and the nearest neighbouring entry in the
17 core collection (E-NE), a criterion to indicate to what extent the entries are spaced in the diversity
18 space, represented by the whole collection. If the E-NE realized value is high, the entries cover the
19 entire space, and each entry should be as different as possible from each other.

20 In order to create a baseline to evaluate these criteria, 1,000 random subsets were generated of the
21 equivalent sizes of each selected subset. Values for each criterion (rA-NE and rE-NE) were determined
22 for each random set and the standard deviation of the results were calculated for each size. To indicate
23 the potential value of the criteria for each strategy (pA-NE and pE-NE), an optimisation was done for
24 both criteria for all optimal core sizes. All calculations were performed following the genetic distance
25 optimisation (GDOpt) procedure described by Odong et al. (2011).

26

27

1 **Genetic structure analysis**

2 The genetic structure analysis and the relatedness among genotypes were performed within those
3 obtained from the Identity analysis, that constituted our whole hazelnut germplasm collection (WHGC).
4 The population structure was explored using STRUCTURE v. 2.3.4 (Pritchard et al. 2000), a model-
5 based Bayesian clustering method, assigning individuals to subpopulations with no *a priori* grouping
6 assumptions. The admixture model was applied, and allele frequencies were assumed to be correlated.
7 A burn-in period of 1,000,000 generations and 2,000,000 Markov chain Monte Carlo replications were
8 used. STRUCTURE was run 10 independent times for each K value ranging from 1 to 20. The most
9 likely K value was determined using the ΔK method (Evanno et al. 2005), as implemented in
10 CLUMPAK (Kopelman et al. 2015). The resulting matrices of estimated group membership coefficients
11 (Q) were permuted using the *Greedy* algorithm implemented in CLUMPP (Jakobsson and Rosenberg
12 2007) and bar plots were drawn using STRUCTURE PLOT v 2.0 (Ramasamy et al. 2014). Genotypes
13 with probability of membership $\geq 80\%$ ($Q \geq 0.8$) were assigned to the same group, while those with
14 intermediate admixture coefficients ($Q < 0.8$) in any group were classified as “admixed” and were
15 clustered in a separate mosaic group (M). The genetic relationships among genotypes were also
16 investigated using the weighted Neighbor-Joining (NJ) dendrogram and the principal coordinate
17 analysis (PCoA) implemented in DARwin v. 6.0 (Perrier and Jacquemoud-Collet 2006). The NJ tree
18 and the two-dimensional PCoA scatterplot were both constructed based on Dice dissimilarity scores
19 (10,000 bootstraps).

20

21 **Results**

22 **Identification of matching genotypes**

23 The genetic profiles of 480 samples across 10 SSR loci were analysed using the Identity software to
24 identify samples with identical genotypes. Among them, 430 accessions are conserved in 11
25 international *ex situ* germplasm repositories, while 50 accessions are landraces maintained *on-farm* in
26 five southern European countries (Portugal, Spain, Italy, Slovenia, and Greece). Prior to the analysis,

1 only 106 of the accessions were identified with a unique cultivar name, while the remaining 374 was
2 comprised of groups of two or more accessions labelled with the same name.

3 SSR analysis identified a total of 181 individual genotypes showing a unique profile, with a total
4 probability of identity of 1.85×10^{-12} (Online Resource 1). By comparison, 252 accessions (52.5% of
5 the total) were deemed to be duplicates and 47 accessions (9.7% of the total) were classified as probable
6 planting or labeling mistakes. Among the accessions deemed to be duplicates, a total of 18 known or
7 likely synonym groups were identified (Online Resource 1). Each group contained accessions with
8 similar nut and husk morphology and most of them were already reported in the literature (Bocchacci et
9 al. 2006, 2008, 2013; Gökirmak et al. 2009; Gürcan et al. 2010; Bacchetta et al. 2015). Nevertheless,
10 some potential new synonyms were also identified: i) the German accessions ‘Kurzhuellige Zellernuss’,
11 ‘Minna's Zellernuss’, ‘Volle Zellernuss’, and ‘Gunslebenert Zellernuss’ showed the same SSR profile
12 of the cultivars ‘Barr's Zellernuss’, ‘Gustav's Zellernuss’, ‘Merveille de Bollwiller’ (syn. ‘Hall’s Giant’),
13 and ‘Gunslebert’, respectively; ii) the English accession ‘Inghilterra’ was genetically identical to
14 ‘Bandnuss’ (syn. ‘Apolda’); iii) the Spanish accessions ‘Closca molla’ and ‘Punxenc’ revealed the same
15 genetic profile as ‘Comun Alava’ and ‘Pere Mas’, respectively; iv) the local cultivars ‘Negret primerenc’
16 and ‘Negret primerenc cort’ showed the same microsatellite profile as ‘Negret’.

17

18 **Development of core collections and comparison to the whole collection**

19 The 181 individual genotypes, representing our whole hazelnut germplasm collection (WHGC), were
20 used to construct core collections by means of three different approaches based on the M-strategy. The
21 performance of each sampling strategy for assembling core collections was evaluated over a range of
22 putative core subset sizes. Thus, a total of five subsets with 10 (5.5%), 20 (11.0%), 30 (16.6%), 40
23 (22.1%), and 50 (27.6%) entries, respectively, were developed, except for the PC strategy where only
24 one subset can be obtained.

25 The results of the variability parameters (A, GD, Ho, and PIC) obtained from a total of 37 subsets
26 compared with the initial collection are reported in Table 1. No SSR loci with significantly different
27 allele frequencies (Fr) were observed and thus, the criterion of having no significant differences ($P <$

1 0.05) in at least 95% of the loci was met in all subsets. The characterization of the WHGC resulted in
2 118 amplification fragments (A) with a mean GD, Ho and PIC of 0.79, 0.80 and 0.76, respectively.
3 Among the subsets obtained by the MSTRAT (MS) strategy, only MS10 showed a significant difference
4 ($P < 0.05$) in the number of alleles (A). MS50 was deemed the best subset capturing all the alleles present
5 in the WHGC, while the optimized MS19 subset given by the algorithm captured a total of 99 alleles
6 (83.9%). The core collection obtained by the Power Core (PC) strategy, representing a full coverage of
7 all the alleles existing in the WHGC, comprised 53 entries (29.3%) and no significant differences were
8 observed with the whole collection. Among the subsets obtained with the Core Hunter (CH) single-
9 strategy, optimizing the Dce, Mr, Cv, He, and Sh indices independently, significantly different values
10 ($P < 0.05$) were detected for the number of alleles (A) at Dce10, Dce20, Mr10, Mr20, Mr30, Cv10,
11 He10, and Sh10 subsets, for Ho in all Mr subsets, and for GD and PIC in all Dce, He and Sh subsets.
12 Thus, only by optimizing the Cv index was it possible to build a core collection that respected the criteria
13 proposed by Escribano et al. (2008) and the best subset was recovered with a minimum of 30 entries
14 (Cv30). For the CH multi-strategy, where both Cv and Dce were optimized simultaneously with equal
15 weight (50%), only Cv-Dce30 respected the criteria proposed by Escribano et al. (2008) and this was
16 selected as the best subset.

17

18 **Selection of the final core collection**

19 A total of four core collections were selected as best representatives of each sampling method: MS50
20 from MSTRAT (MS), PC53 from Power Core (PC), Cv30 and Cv-Dce30 from Core Hunter (CH) single-
21 and multi-strategy, respectively.

22 In Table 2 the best representatives are compared with the whole collection (WHGC) using the
23 mean values of the independent runs for each of the following parameters: the genetic distances MR
24 and CE, the genetic diversity indices SH, HE, and NE, and of the auxiliary values PN and CV (Thachuk
25 et al. 2009). All core subsets showed higher average genetic distance between entries and higher allelic
26 richness than the WHGC. Moreover, all sampling strategies were optimal in minimizing PN and

1 maximizing CV (0.0 and 100.0, respectively). Among them, the CH multi-strategy (Cv-Dce30) showed
2 slightly higher values at CE, SH, and HE and the highest NE value.

3 In Table 3 the best representatives are compared with the WHGC using the A-NE and E-NE
4 quality parameters for each strategy: the realized values (A-NE and E-NE), the potential optimal value
5 (pA-NE and pE-NE), the average value from 1,000 random sets and the corresponding standard
6 deviation (rA-NE and rE-NE). All four realized values for the A-NE parameter were higher than their
7 respective potential optimal values but were not different from those of the random sets. Therefore, no
8 strategies performed better against the A-NE criterion when compared to the random sets. On the
9 contrary, all four E-NE values were considerably higher than random values and considerably lower
10 than potential values. In proportion, none of them reached more than 80% (PC53 and Cv-Dce30) of the
11 potential optimum, while the random sets reached 67-69% of the potential optimum. Thus, all four
12 strategies performed better against the E-NE criterion, as compared to a random set, and the CH multi-
13 strategy (Cv-Dce30) outperformed the others.

14 The subset judged best to form the final core collection of our whole collection was obtained by
15 the CH multi-strategy (Cv-Dce30) and was composed by 30 entries (16.6%); the relationship among
16 frequencies of alleles between this subset and the WHGC was very highly correlated ($R^2=0.93$) (Online
17 Resource 2, Fig. S2).

18

19 **Genetic population structure**

20 The estimation of ΔK (Online Resource 2, Fig. S3) from the analysis of 181 individual genotypes
21 revealed the highest value for $K = 3$ ($\Delta K = 246.82$), but high values were also obtained for $K = 2$ (ΔK
22 $= 205.22$) and $K = 5$ ($\Delta K = 184.24$). In $K = 2$, genotypes were grouped in two gene pools (Fig. 1): one
23 composed mainly by cultivars from Central Europe (CEU) and the British Islands (BI), and another
24 composed by cultivars from the Iberian Peninsula (IbeP), the Italian Peninsula (ItaP), and the Balkans-
25 Black Sea (BBS). A total of 63 genotypes were not clearly placed in these groups ($Q < 0.8$) and were
26 classified as admixed. In $K = 3$ genotypes were grouped in three gene pools composed mainly by
27 cultivars from CEU and BI, ItaP, and BBS, respectively (Fig. 1). Cultivars from the Iberian Peninsula

1 (IbeP) were widespread within all three groups, while 62 genotypes were classified as admixed. In $K =$
2 5 genotypes were classified into five groups (Fig. 1). Cultivars from CEU and BI were placed in two
3 separate groups: Q1 was composed by 9 cultivars from CEU and 2 accessions of unknown origin
4 ('Mogulnuss' and 'Pallagrossa'), while Q2 included 8 cultivars from BI, 2 from CEU, and 3 accessions
5 of unknown origin ('Apolda', 'Bearn', and 'Sodlinger'). 18 cultivars from IbeP showed the tendency to
6 constitute a separate group (Q3), together with 4 cultivars and 3 landraces from ItaP, 1 landrace from
7 BBS, and 1 accession of unknown origin ('Comen'). Q4 grouped 13 cultivars and 10 landraces from
8 ItaP, 4 cultivars and 1 landrace from IbeP, and 1 cultivar from CEU. Q5 clustered 20 cultivars from
9 BBS, 5 landraces from ItaP, and 2 accessions of unknown origin ('Fructo rubro' and 'Jann's'). A total
10 of 74 genotypes were classified as admixed ($Q < 0.8$) and were deemed "mosaics" (M group).

11 The NJ dendrogram and PCoA scatterplot showed a clustering of the 181 genotypes similar to
12 that obtained from STRUCTURE analysis. In the NJ dendrogram (Fig. 2), genotypes were grouped in
13 three main clusters (I, II, and III), corresponding to $K = 3$, that showed a substructure similar to that
14 observed in $K = 5$: CEU (Q1) and BI (Q2) largely constituted two distinct subgroups in cluster I; IbeP
15 (Q3) and ItaP (Q4) were largely separated in several subgroups within cluster II; BBS (Q5) corresponded
16 to the cluster III. Admixed genotypes (M) were distributed in all three main clusters. In the PCoA
17 scatterplot (Fig. 3), the projection of the genotypes on a two-dimensional plane defined by the first two
18 PCs (15.85 % of the cumulative variation) showed: i) a separation between groups CEU and BI (right
19 half of the graph) and groups IbeP, ItaP, and BBS (left half of the graph), as in $K = 2$; ii) a separation
20 between group ItaP (top left), group BBS (lower left), and groups CEU and BI (right half), as in $K = 3$;
21 iii) a general tendency to separate each Q group obtained with the $K = 5$ stratification. CEU (Q1) was
22 placed in the upper right of the graph, while BI (Q2) was positioned in the centre of the right half. ItaP
23 (Q4) was placed in the upper left, while IbeP (Q3) and BBS (Q5) were located separately in the lower
24 left. M genotypes were scattered in all four parts of the graph.

25 Considering the WHGC population structure obtained, the genotypes included in the Cv-Dce30 core
26 collection covered all six groups: 1 from Q1 (9.1%), 2 from Q2 (15.4%), 4 from Q3 (15%), 7 from Q4
27 (24.1%), 5 from Q5 (18.5%), and 11 from M (15%).

1 **Discussion**

2 **Identification of matching genotypes**

3 Mislabeling and duplication are important challenges for germplasm conservation. Duplication
4 through the existence of synonyms is a characteristic challenge in cultivars of vegetatively propagated
5 woody perennial species. Thus, SSR markers have become very valuable tools in the management of *ex*
6 *situ* and *in situ* hazelnut collections. Genotypes that showed the same SSR profile were considered as
7 duplicates, and were 52.5 % of the total number of accessions analysed. Among them were two types of
8 duplication: i) accessions labelled with the same name and collected from different collection fields (i.e.
9 true duplicates). In this first case the duplication allowed us to define the true SSR genotype of many
10 cultivars by comparing profiles obtained from several accessions and to identify some mislabeling
11 among them; ii) accessions labelled with differing names, conserved either in different collection fields
12 or in the same collection field. In this second case, the duplication allowed us to identify some probable
13 mislabeling (9.7% of the total number of accessions analysed) due to planting or labeling mistakes, as
14 happened to Bassil et al. (2009) during a backup of the USDA collection. Moreover, it was also possible
15 to confirm several synonyms reported in the literature (Boccalci et al. 2006, 2008, 2013; Gökirmak et
16 al. 2009; Gürcan et al. 2010; Bacchetta et al. 2015) and identify potential new ones. Among them, the
17 local cultivars ‘Negret primerenc’ and ‘Negret primerenc cort’ were found to have the same genetic
18 profile as ‘Negret’, although they are known to be more productive and their fruits mature earlier (Rovira
19 et al. 2017), and they represent a possible case of clonal mutation. A similar result was observed between
20 ‘Tonda di Biglini’ and ‘Tonda Gentile delle Langhe’ by Valentini et al. (2014). Consequently, to
21 construct our core collection it was important to identify mislabeling, duplicates, and synonyms from
22 the whole hazelnut germplasm collection (WHGC), to delete a significant source of redundancy and
23 build the core collection only from true-to-type genotypes.

24

25 **Building the core collection**

26 The main strategies used to construct core collections from molecular marker data can be
27 classified into two groups. The first methods are based on genetic distance, with or without stratified

1 sampling techniques, that cluster the accessions and then select entries from each cluster using different
2 allocation approaches (van Hintum et al. 2000; De Beukelaer et al. 2012). The second methods are based
3 on the M-strategy that construct cores with high allelic richness, maximizing the number of observed
4 alleles at each marker locus (Schoen and Brown 1993). M-methods are the only approaches that recover
5 all the alleles of the whole collection, including rare alleles, and retain the original allele frequencies at
6 each locus, favouring smaller subsets, reducing redundancy, and capturing most of the genetic diversity
7 (Marita et al. 2000; Gouesnard et al. 2001). In fruit and nut tree genera, such as *Annona*, *Ficus* and
8 *Castanea*, the M-strategy was the most efficient method to develop core collections, outperforming other
9 strategies (Escribano et al. 2008; Balas et al. 2014; Pereira-Lorenzo et al. 2017); and for this reason was
10 largely used by many other authors (Le Cunff et al. 2008; Miranda et al. 2010; Belaj et al. 2012; Díez
11 et al. 2012; El Bakkali et al. 2013; Štajner et al. 2014; Liang et al. 2015; Liu et al. 2015; Öztürk et al.
12 2017; Bernard et al. 2020). Nevertheless, the choice of the most appropriate evaluation measures
13 depends on the purpose of the core collection. Genetic distance methods based on the allele
14 representativeness tend to be preferred by plant breeders, while methods based on the allele richness,
15 including rare and localized alleles, interest taxonomists and geneticists (Marita et al. 2000).

16 The M-strategy, based on MSTRAT (Gouesnard et al. 2001), PowerCore (Kim et al. 2007), and
17 Core Hunter (Thachuk et al. 2009, De Beukelaer et al. 2012) algorithms, was used to develop our core
18 collections (Table 1). The best subsets obtained from each sampling strategy (MS50, PC53, Cv30, and
19 Cv-Dce30) captured all the alleles within the minimum number of accessions, without significant
20 differences in allele frequencies, and preserved the parameter variations when compared to the WHGC
21 (Table 2). The core subsets obtained from the Core Hunter (CH) simple- and multi-strategies contained
22 a smaller number of entries (30 accessions) compared to those obtained from MSTRAT (50 accessions,
23 MS50) and PowerCore (53 accessions, PC53). As reported by Thachuk et al. (2009), our results
24 confirmed that the CH strategy was able to select significantly smaller core subsets that retain all unique
25 alleles within a whole collection and a similar result was also observed in olive by Díez et al. (2012).

26 The main advantage of the Core Hunter software is its ability to build core collections using
27 different allocation strategies by optimizing one parameter or many parameters simultaneously.

1 Generally, core subsets optimized using multiple criteria perform worse than those obtained using
2 individual measures (Thachuk et al. 2009; Díez et al. 2012). Nevertheless, our core subset obtained from
3 CH multi-strategy (Cv-Dce30) showed slightly higher values at MR, CE, SH, and HE and the highest
4 NE value, compared to that obtained from CH single-strategy (Cv30) (Table 2). Thus, the CH multi-
5 strategy approach, optimizing Cv and Dce simultaneously with equal weight, was selected as the best
6 strategy to build our final core collection. It satisfied both the breeders' and geneticists'/taxonomists'
7 perspectives despite including only 16.6 % of the WHCG genotypes. This value sits within the 5–20 %
8 proposed by van Hintum et al. (2000) and is lower than the 19 % obtained from the Turkish national
9 hazelnut collection (Öztürk et al. 2017). Other studies on fruit tree crops reported a minimum
10 requirement of 4 % inclusion in grape (Le Cunff et al. 2008; Štajner et al. 2014), a common range from
11 13 % in fig (Balas et al. 2014) to 15-19 % in olive (Belaj et al. 2012; Díez et al. 2012; El Bakkali et al.
12 2013), and a maximum of 28.6 % in apple (Liang et al. 2015) and 29.8 % in chestnut (Pereira-Lorenzo
13 et al. 2017).

14 According to Ondong et al. (2013), the criterion of choice for evaluating the quality of core
15 collections should be determined by the objectives or type of the collection itself. Thus, they proposed
16 two genetic distance-based criteria, A–NE and E–NE, for evaluating the quality of two important types
17 of core collections, respectively: i) a core collection (CC-I) where each entry represents one (itself) or
18 more accessions of the whole collection, in order to maximize the representativeness of genetic diversity
19 (where lower values for A-NE indicate increased representation); ii) a core collection (CC-X) where the
20 diversity of the traits of the entries is maximized, in order to represent the total genetic diversity (where
21 higher values for E-NE indicate lower redundancy and a better coverage of the diversity space). Using
22 these criteria, our results indicated that all the best subsets obtained from each sampling strategy (MS50,
23 PC53, Cv30, and Cv-Dce30) aimed at covering the range of the genetic diversity, rather than
24 representing the accessions from WHGC. All values for the A-NE parameter were not significantly
25 different from that of a random set. On the contrary, all four strategies resulted in higher E-NE values,
26 as compared to a random set, and the CH multi-strategy (Cv-Dce30) outperformed the others (Table 3).
27 Since the objective of our final core collection was the maximisation of the allelic richness, including

1 rare and localized alleles, the E-NE parameter was the most appropriate to evaluate the quality of the
2 four core collections obtained. Nevertheless, the fact that the size of the core collections varied made E-
3 NE comparison difficult (Table 3). In the case of the Cv30 and Cv-Dce30 subsets (30 entries), the CH
4 multi-strategy outperformed the CH single-strategy maximizing E-NE, but not by a large factor (80%
5 vs 77% of the maximum achievable). Comparing these two subsets with the larger core collections
6 MS50 (50 entries) and PC53 (53 entries), all four were in a similar range. Nevertheless, PC53 did best
7 in terms of standard deviation from the random E-NE, whereas Cv-Dce30 did best in terms of
8 approaching the potential maximum.

9

10 **Characteristics of the WHGC and core collection**

11 The population structure and relatedness among the 181 independent genotypes from the WHGC,
12 indicated the existence of three levels of genetic structure (Fig 1, Fig. 2, and Fig. 3). In the first level (K
13 =2) we observed a geographic pattern with one gene pool dominating western and central Europe (BI
14 and CEU) and another gene pool dominating southern Europe (IbeP, ItaP, and BBS). In the second level
15 ($K = 3$), there appeared a third gene pool (BBS) most frequent in the Balkans and the Anatolian
16 Peninsula. Finally, in the third level ($K = 5$) genotypes from northern Europe were further subdivided
17 into the BI and CEU gene pools and those from southern Europe into the BBS, IbeP, and ItaP gene
18 pools.

19 The high level of genetic similarity between cultivars grown in the Iberian and Italian Peninsulas,
20 observed in $K = 2$ and $K = 3$ (Fig. 1), was already reported by other authors (Bocacci et al. 2006;
21 Gökirmak et al. 2009; Gürcan et al. 2010) and was appears to be a consequence of a high gene flow
22 between western and central Mediterranean basin (Bocacci and Botta 2010). Nevertheless, in $K = 5$
23 (Fig. 1) cultivars from the Iberian Peninsula were separated from Italian ones and a significant genetic
24 differentiation between the Spanish and Italian gene pools was reported in subsequent studies (Bocacci
25 and Botta 2010; Bocacci et al. 2013), indicating that northern Spain and southern Italy were two
26 independent hazelnut domestication areas (Bocacci et al. 2013). On the contrary, the genetic similarity
27 between cultivars from the British Islands and the Central Europe obtained with $K = 2$ and $K = 3$ (Fig.

1) was not observed by Gökirmak et al. (2009). Indeed, the authors reported that these cultivars clustered into separate groups as observed in our K = 5 stratification (Fig. 1). Thus, considering the data reported in the literature, the most likely genetic structure of WHGC was composed by five Q groups (CEU, Q1; BI, Q2; IbeP, Q3; ItaP, Q4; and BBS, Q5), and a more complex group of mosaics (M). According to several authors (Gökirmak et al. 2009; Boccacci and Botta 2010; Boccacci et al. 2013) these gene pools would be the result of an independent domestication of *C. avellana* that occurred in different geographical areas: Central Europe, the British Islands, Spain, Italy, and Black Sea. In contrast, the mosaic genotypes, which are found throughout our sampling range, represent a heterogeneous group that may be indicative of recent admixture between distinct groups of cultivars.

The Bayesian clustering and admixture analysis can be considered a standard method to identify the ancestral populations from which cultivars originated and quantify genetic relationships with probabilities and proportions. Thus, it was helpful for suggesting the unknown origin of some cultivars. ‘Mogulnuss’ (syn. ‘Riekchen’s Zellernuss’) and ‘Pallagrossa’ were placed into the CEU group, ‘Comen’ in the IbeP group and ‘White Filbert’ (syn. ‘Fructo rubro’) in the BI group, confirming the results obtained by Gökirmak et al. (2009). On the contrary, ‘Jann’s/ Jean’s’ was placed in the group mainly from the Italian Peninsula rather than from the Black Sea and ‘The Shah’ was placed in the admixed group instead of the English group 2, as would have been expected from the findings of Gökirmak et al. (2009). In our analysis ‘Sodlinger’ clustered into the BI group, although it was placed in a Spanish–Italian group by Muehlbauer et al. (2014).

The 30 entries (Online Resource 1) included in our final core collection (Cv-Dce30) were from different countries: Italy (17 entries, 56.7%), Spain (4 entries, 13.3%), Germany (4 entries, 13.3%), Turkey (3 entries, 10%), and Slovenia (2 entries, 6.7%). They covered all six genetic groups obtained, further confirming that Cv-Dce30 core collection was representative of the WHGC. Interestingly, half of the entries were cultivars from *ex situ* collections, while the other half were landraces from *in situ* collections. The high number of landraces included in the Cv-Dce30 core collection indicated that the hazelnut *on-farm* exploration conducted in southern Europe (Boccacci et al. 2013) has genuinely contributed to expanding the existent hazelnut biodiversity in our collections. No reference accessions

1 were included as “kernel” in our core collections (van Hintum et al. 2000), but the most popular hazelnut
2 cultivar ‘Tonda Gentile delle Langhe’ (TGL), particularly appreciated by the industry for the
3 morphological, organoleptic, and nutritional characteristics of its nuts and kernels, was included in all
4 MS50, PC53, Cv30, and Cv-Dce30 subsets. Different reference cultivars could be added on a case-by-
5 case basis in different places where this core collection could be studied, such as: ‘Negret’ and ‘Casina’
6 in Spain, ‘Barcelona’ (syn. ‘Fertile de Coutard’) in France, ‘Tonda Gentile Romana’ and ‘Tonda di
7 Giffoni’ in Italy, and ‘Tombul’ in Turkey.

8

9 **Conclusions**

10 The M-strategies employed in this work to build our core collections may be considered useful tools for
11 the conservation and characterization of hazelnut genetic resources. Among them, the CH multi-
12 strategy, optimizing Cv and Dce simultaneously with equal weight, was selected as the best strategy to
13 build our final core collection. The ability of each sampling strategy to capture the diversity and
14 representativeness, and the effectiveness and quality of each sampling method, were performed using
15 various well-known approaches. Thus, our final core collection, representing most of the diversity
16 conserved in the European hazelnut germplasm repositories, could be used as a base for new research
17 into genotype x environment interactions focused on a minimum number of accessions. However,
18 reducing the number of selected accessions inevitably increases the probability of discarding genotypes
19 with agronomical traits of interest. Thus, it will be important to consider core collections combining
20 molecular markers, morphological and phenotypical traits, as well as resistance to biotic and abiotic
21 stresses. It is also pragmatic to include cultivars that are considered more influential in a determined
22 cultivation area. Finally, any approach toward core collections should remain dynamic and be revised
23 periodically, to include new accessions and information about new characterization methods (e.g.,
24 functional markers), as well as new methodologies aimed at increasing their representativeness.

25

26

27

1 **Declarations**

2 **Acknowledgments** In memory of my dad, Ugo Boccacci (October 6, 1946 - March 19, 2021)

3 **Funding** This work was funded by AGRI GEN RES Community Program (European Commission, Directorate-
4 General for Agriculture and Rural Development, under Council Regulation (EC) No. 870/2004) – SAFENUT
5 project (“Safeguard of almond and hazelnut genetic resources: from traditional uses to modern agro-industrial
6 opportunities”), AGRI GEN RES 068

7 **Conflicts of interest/Competing interests** Authors declare that they have no conflicts of interest/competing
8 interests

9 **Ethics approval** Not applicable

10 **Consent to participate** Not applicable

11 **Consent for publication** Not applicable

12 **Data archiving statement/Availability of data and material** The Online Material 1 (EMS1, “xlsx” format) and
13 Online Material 2 (EMS2, “pdf” format) are available in the Dryad Repository (<https://datadryad.org/stash>) as
14 "Hazelnut SSR database: genetic profiles of the accessions, list of synonyms, and true-to-type genotypes"
15 (doi:10.5061/dryad.cz8w9gj45). For private access during the review period, reviewers may share the unpublished
16 dataset using this temporary link:
17 https://datadryad.org/stash/share/fG-4pHenR4hQ_z5G7y18b43ytHnTLpy3w-F2lFJqljk

18 **Code availability** Not applicable

19 **Authors' contributions** Paolo Boccacci conceived the study and written the manuscript. Paolo Boccacci, Maria
20 Aramini, Daniela Torello Marinoni, Nadia Valentini, Mercè Rovira, Anita Solar, and Jean-Paul Sarraquigne
21 collected vegetal material from the collection fields. Paolo Boccacci, Maria Aramini, Matthew Ordidge, and
22 Daniela Torello Marinoni performed SSR analyses. Paolo Boccacci and Theo van Hintum performed data
23 elaborations. Loretta Bacchetta coordinated and Roberto Botta co-coordinated the SAFENUT project. Paolo
24 Boccacci and Matthew Ordidge revised the manuscript. All authors read and approved the final version of the
25 manuscript

26

27 **References**

28 Bacchetta L, Rovira M, Tronci C, Aramini M, Drogoudi P, Silva AP, Solar A, Avanzato D, Botta R, Valentini N,
29 Boccacci P (2015) A multidisciplinary approach to enhance the conservation and use of hazelnut *Corylus*
30 *avellana* L. genetic resources. Genet Resour Crop Evol 62:649–663

- 1 Balas FC, Osuna MD, Domínguez G, Pérez-Gragera F, López-Corrales M (2014). *Ex situ* conservation of
2 underutilised fruit tree species: establishment of a core collection for *Ficus carica* L. using microsatellite
3 markers (SSRs). *Tree Genet Genomes* 10:703–710
- 4 Bassil NV, Botta R, Mehlenbacher SA (2005) Microsatellite markers in the hazelnut: isolation, characterization,
5 and cross-species amplification in *Corylus*. *J Am Soc Hort Sci* 130:543-549
- 6 Bassil NV, Postman J, Hummer K, Botu M, Sezer A (2009) SSR fingerprinting panel verifies identities of clones
7 in backup hazelnut collection of USDA genebank. *Acta Hort* 845: 95-102
- 8 Belaj A, del Carmen Dominguez-García M, Atienza SG, Martín Urdíroz N, De la Rosa R, Satovic Z, Martín A,
9 Kilian A, Trujillo I, Valpuesta V, Del Río C (2012). Developing a core collection of olive (*Olea europaea* L.)
10 based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet Genomes* 16:76
- 11 Bernard A, Barreneche T, Donkpegan A, Lheureux F, Dirlwanger E (2020). Comparison of structure analyses
12 and core collections for the management of walnut genetic resources. *Tree Genet Genomes* 8:365–378
- 13 Boccacci P, Botta R (2010) Microsatellite variability and genetic structure in hazelnut (*Corylus avellana* L.)
14 cultivars from different growing regions. *Sci Hortic* 124:128-133
- 15 Boccacci P, Akkak A, Bassil NV, Mehlenbacher SA, Botta R (2005) Characterization and evaluation of
16 microsatellite loci in European hazelnut (*Corylus avellana* L.) and their transferability to other *Corylus* species.
17 *Mol Ecol Notes* 5:934-937
- 18 Boccacci P, Akkak A, Botta R (2006) DNA-typing and genetic relationships among European hazelnut (*Corylus*
19 *avellana* L.) cultivars using microsatellite markers. *Genome* 49:598-611
- 20 Boccacci P, Rovira M, Botta R (2008) Genetic diversity of hazelnut (*Corylus avellana* L.) germplasm in
21 northeastern Spain. *HortScience* 43:667-672
- 22 Boccacci P, Aramini M, Valentini N, Bacchetta L, Rovira M, Drogoudi P, Silva AP, Solar A, Calizzano F,
23 Erdorğan V, Cristofori V, Ciarmiello LF, Contessa C, Ferreira JJ, Marra FP, Botta R (2013) Molecular and
24 morphological diversity of *on-farm* hazelnut (*Corylus avellana* L.) landraces from southern Europe and their
25 role in the origin and diffusion of cultivated germplasm. *Tree Genet Genomes* 9:1465–1480
- 26 Botta R, Molnar TJ, Erdorğan V, Valentini N, Torello Marinoni D, Mehlenbacher S (2019). Hazelnut (*Corylus*
27 spp.) Breeding. In: Al-Khayri JM, Jain SM, Johnson DV (eds.) *Advances in Plant Breeding Strategies: Nut*
28 *and Beverage crops*. Springer Nature, Switzerland, Volume 4, pp 157-219
- 29 De Beukelaer H, Smýkal P, Davenport GF, Fack V (2012) Core Hunter II: fast core subset selection based on
30 multiple genetic diversity measures using Mixed Replica search. *BMC Bioinformatics* 13:312

- 1 Di Guardo M, Scollo F, Ninot A, Rovira M, Hermoso JF, Distefano G, La Malfa S, Batlle I (2019) Genetic structure
2 analysis and selection of a core collection for carob tree germplasm conservation and management. *Tree Genet*
3 *Genomes* 15: 41
- 4 Díez CM, Imperato A, Rallo L, Barranco D, Trujillo I (2012). Worldwide core collection of olive cultivars based
5 on simple sequence repeat and morphological markers. *Crop Sci* 52:211-221
- 6 El Bakkali A, Haouane H, Moukhli A, Costes E, Van Damme P, Khadari B (2013) Construction of core collections
7 suitable for association mapping to optimize use of Mediterranean olive (*Olea europaea* L.) genetic resources.
8 *PLoS ONE* 8:e61265
- 9 Escribano P, Viruel MA, Hormaza JI (2008) Comparison of different methods to construct a core germplasm
10 collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya
11 (*Annona cherimola*, Annonaceae), an underutilised subtropical fruit tree species. *Ann Appl Biol* 153:25–32
- 12 Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software
13 STRUCTURE: a simulation study. *Mol Ecol* 14:2611-2620
- 14 FAO (1996) Global plan of action for the conservation and sustainable utilization of plant genetic resources for
15 food and agriculture. Food and Agriculture Organization, Rome
- 16 FAOSTAT (2021) <http://www.fao.org/faostat/en/?#data>. Accessed 06 May 2021
- 17 Franco J, Crossa J, Taba S, Shands H (2005) A sampling strategy for conserving genetic diversity when forming
18 core subsets. *Crop Sci* 45:1035–1044
- 19 Franco J, Crossa J, Warburton ML, Taba S (2006) Sampling strategies for conserving maize diversity when
20 forming core subsets using genetic markers. *Crop Sci* 46:854–864
- 21 Ferreira JJ, Garcia-González C, Tous J, Rovira M (2010) Genetic diversity revealed by morphological traits and
22 ISSR markers in hazelnut germplasm from northern Spain. *Plant Breed* 129:435–441
- 23 Gökirmak T, Mehlenbacher SA, Bassil NV (2009) Characterization of European hazelnut (*Corylus avellana* L.)
24 cultivars using SSR markers. *Genet Resour Crop Evol* 56:147-172
- 25 Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for
26 building germplasm core collections by maximizing allelic or phenotypic richness. *J Hered* 92:93–94
- 27 Gürcan K, Mehlenbacher SA, Erdoğan V (2010) Genetic diversity in hazelnut (*Corylus avellana* L.) cultivars from
28 Black Sea countries assessed using SSR markers. *Plant Breed* 129:422–434
- 29 Hummer KE (2001) Hazelnut genetic resources at the Corvallis repository. *Acta Hort* 556:21–24

1 Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label
2 switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806

3 Kim KW, Chung HK, Cho GT, Ma KH, Gwag CD, Kim TS, Cho EG, Park YJ (2007) PowerCore: a program
4 applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23:2155–
5 2162

6 Köksal AI (2000) Inventory of hazelnut research, germplasm and references. REU technical series. FAO-CIHEAM

7 Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying
8 clustering modes and packaging population structure inferences across K. *Mol Ecol Resour* 15: 1179-1191

9 Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, Poncet C, Lasserre-Zuber P,
10 Feugey L, Durel CE (2016). Genetic diversity, population structure, parentage analysis, and construction of
11 core collections in the French apple germplasm based on SSR markers. *Plant Mol Biol Rep* 34:827–844

12 Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon AF, Boursiquot JM, This P
13 (2008). Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis*
14 *vinifera* L. subsp. *sativa*. *BMC Plant Biol* 8:31

15 Liang W, Dondini L, De Franceschi P, Paris R, Sansavini S, Tartarini S (2015) Genetic diversity, population
16 structure and construction of a core collection of apple cultivars from Italian germplasm. *Plant Mol Biol Rep*
17 33:458–473

18 Liu KJ, Muse SV (2005). PowerMarker: An integrated analysis environment for genetic marker analysis.
19 *Bioinformatics* 21:2128–2129

20 Liu Q, Song Y, Liu L, Zhang M, Sun J, Zhang S, Wu J (2015) Genetic diversity and population structure of pear
21 (*Pyrus* spp.) collections revealed by a set of core genome-wide SSR markers. *Tree Genet Genomes* 11:128

22 Marita JM, Rodriguez JM, Nienhuis J (2000) Development of an algorithm identifying maximally diverse core
23 collections. *Genet Resour Crop Evol* 47:515–526

24 Mehlenbacher SA (2018) Advances in genetic improvement of hazelnut. *Acta Hort* 1226:1-12

25 Miranda C, Urrestarazu J, Santesteban LG, Royo JB, Urubina V (2010) Genetic diversity and structure in a
26 collection of ancient Spanish pear cultivars assessed by microsatellite markers. *J Am Soc Hortic Sci* 135:428-
27 437

28 Muehlbauer MF, Honig JA, Capik JM, Vaiciunas JN, Molnar TJ (2014) Characterization of eastern filbert blight-
29 resistant hazelnut germplasm using microsatellite markers. *J Am Soc Hortic Sci* 139:399–432

30 Nei M (1987) *Molecular evolutionary genetics*. Columbia Univ. Press, New York, NY

- 1 Odong TL, van Heerwaarden J, Jansen J, van Hintum TJJ, van Eeuwijk FA (2011) Statistical techniques for
2 defining reference sets of accessions and microsatellite markers. *Crop Sci* 51(6):2401–2411
- 3 Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJJ (2013) Quality of core collections for effective utilisation
4 of genetic resources review, discussion and interpretation. *Theor Appl Genet* 126:289–305
- 5 Öztürk SC, Balık Hİ, Balık SK, Kızılcı G, Duyar Ö, Doğanlar S, Frary A (2017) Molecular genetic diversity of
6 the Turkish national hazelnut collection and selection of a core set. *Tree Genet Genomes* 13:113
- 7 Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian
8 polar bears. *Mol Ecol* 4:347–354
- 9 Pereira-Lorenzo S, Ramos-Cabrer AM, Barreneche T, Mattioni C, Villani F, Díaz-Hernández MB, Martín LM,
10 Martín A (2017) Database of European chestnut cultivars and definition of a core collection using simple
11 sequence repeats. *Tree Genet Genomes* 13:114
- 12 Perrier X, Jacquemoud-Collet J (2006) DARwin software. Available from: <http://darwin.cirad.fr/>. Accessed 5 May
13 2021
- 14 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data.
15 *Genetics* 155:945-959
- 16 Ramasamy RK, Sumathy Ramasamy S, Bindroo BB, Naik VG (2014) STRUCTURE PLOT: a program for
17 drawing elegant STRUCTURE bar plots in user friendly interface. *SpringerPlus* 3:431
- 18 Rovira M, Hermoso JF, Romero AJ (2017) Performance of hazelnut cultivars from Oregon, Italy, and Spain, in
19 Northeastern Spain. *HortTechnology* 27(5):631-638
- 20 Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of
21 genetic markers. *Proc Natl Acad Sci USA* 90:10623–10627
- 22 Štajner N, Tomić L, Ivanišević D, Korać N, Cvetković-Jovanović T, Beleski K, Angelova E, Maraš V, Javornik
23 B (2014). Microsatellite inferred genetic diversity and structure of Western Balkan grapevines (*Vitis vinifera*
24 L.). *Tree Genet Genomes* 10:127–140
- 25 Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF (2009) Core Hunter: an algorithm
26 for sampling genetic resources based on multiple genetic measures. *BMC Bioinformatics* 10:243
- 27 Thomas MR, Matsumoto S, Cain P, Scott NS (1993) Repetitive DNA of grapevine: classes present and sequences
28 suitable for cultivar identification. *Theor Appl Genet* 86:173-180
- 29 Thompson MM, Lagerstedt HB, Mehlenbacher SA (1996) Hazelnuts. In: Janick J, Moore JN (eds) *Fruit breeding:*
30 *nuts*, vol 3. Wiley, New York, pp 125–184

1 Valentini N, Calizzano F, Boccacci P, Botta R (2014) Investigation on clonal variants within the hazelnut (*Corylus*
2 *avellana* L.) cultivar ‘Tonda Gentile delle Langhe’. *Sci Hortic* 165:303-310

3 van Hintum, T.J.L, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. IPGRI
4 Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome

5 Wagner HW, Sefc KM (1999) IDENTITY 4.0. Centre for Applied Genetics, University Agricultural Sciences,
6 Vienna

7 Wang Y, Zhang J, Sun H, Ning N, Yang L (2011) Construction and evaluation of a primary core collection of
8 apricot germplasm in China. *Sci Hortic* 128:311–319

9

10 **Figure’s legend:**

11

12 **Fig. 1** Population structure and hierarchical organization of genetic relatedness of 181 genotypes from the whole
13 hazelnut germplasm collection (WHGC) at $K = 2$, $K = 3$, and $K = 5$, as inferred by STRUCTURE software

14

15 **Fig. 2** Neighbor-joining dendrogram based on the Dice similarity index showing the relationships among 181
16 hazelnut genotypes from WHGC. Genotypes are colored according to their assignment to the different gene pools,
17 as inferred by STRUCTURE software at $K = 5$: Central Europe (Q1), British Islands (Q2), Iberian Peninsula (Q3),
18 Italian Peninsula (Q4), Balkans-Black Sea (Q5), and mosaic group (M). Entries of the final core collection (Cv-
19 Dce30) are reported as CC

20

21 **Fig. 3** Two-dimensional PCoA scatterplot of 181 hazelnut genotypes from WHGC based on Dice’s distance.
22 Genotypes are colored according to their assignment to the different gene pools, as inferred by STRUCTURE
23 software at $K = 5$: Central Europe (Q1), British Islands (Q2), Iberian Peninsula (Q3), Italian Peninsula (Q4),
24 Balkans-Black Sea (Q5), and mosaic group (M). Entries of the final core collection (Cv-Dce30) are reported as
25 CC

26

27

28

1 **Table 1** Variability parameters for different core subsets compared with the whole collection. In bold are indicated
 2 the best core subset obtained from each sampling strategy

Sampling strategy	Subset code	Subset size	<i>A</i>	<i>GD</i>	<i>Ho</i>	<i>PIC</i>
Whole collection	WHGC	181	118	0,79	0,80	0,76
MSTRAT	MS10	10	86 ^a	0,82	0,76	0,80
	MS20	20	100	0,83	0,81	0,81
	MS30	30	108	0,82	0,80	0,80
	MS40	40	109	0,82	0,82	0,80
	MS50	50	118	0,81	0,81	0,79
	MS19	19	99	0,83	0,81	0,81
Power Core	PC53	53	118	0,81	0,81	0,79
Core Hunter single - Dce	Dce10	10	86 ^a	0,85 ^a	0,77	0,83 ^a
	Dce20	20	93 ^a	0,85 ^a	0,78	0,83 ^a
	Dce30	30	101	0,84 ^a	0,77	0,83 ^a
	Dce40	40	103	0,84 ^a	0,78	0,82 ^a
	Dce50	50	103	0,84 ^a	0,77	0,82 ^a
Core Hunter single - Mr	Mr10	10	77 ^a	0,82	0,63 ^a	0,79
	Mr20	20	87 ^a	0,82	0,65 ^a	0,79
	Mr30	30	94 ^a	0,82	0,66 ^a	0,79
	Mr40	40	98	0,82	0,69 ^a	0,80
	Mr50	50	103	0,82	0,70 ^a	0,80
Core Hunter single - Cv	Cv10	10	94 ^a	0,83	0,88	0,81
	Cv20	20	109	0,80	0,83	0,77
	Cv30	30	118	0,81	0,85	0,79
	Cv40	40	118	0,78	0,82	0,76
	Cv50	50	118	0,79	0,81	0,76
Core Hunter single - He	He10	10	87 ^a	0,85 ^a	0,87	0,83 ^a
	He20	20	97	0,85 ^a	0,84	0,84 ^a
	He30	30	102	0,85 ^a	0,88	0,83 ^a
	He40	40	104	0,85 ^a	0,86	0,83 ^a
	He50	50	107	0,84 ^a	0,84	0,83 ^a
Core Hunter single - Sh	Sh10	10	89 ^a	0,85 ^a	0,86	0,83 ^a
	Sh20	20	103	0,85 ^a	0,85	0,83 ^a
	Sh30	30	107	0,85 ^a	0,86	0,83 ^a
	Sh40	40	109	0,85 ^a	0,85	0,83 ^a
	Sh50	50	112	0,84 ^a	0,84	0,82 ^a
Core Hunter multi - Cv-Dce	Cv-Dce10	10	94 ^a	0,83 ^a	0,87	0,82 ^a
	Cv-Dce20	20	108	0,84 ^a	0,83	0,82 ^a
	Cv-Dce30	30	118	0,82	0,83	0,80
	Cv-Dce40	40	118	0,83 ^a	0,82	0,81 ^a
	Cv-Dce50	50	118	0,83 ^a	0,79	0,81 ^a

3 *A*, number of alleles; *GD*, genetic diversity; *Ho*, observed heterozygosity; *PIC*, polymorphism information content

4 ^aStatistically significant difference, Dunnett's test ($P < 0.05$)

Table 2 Comparison of the best core subsets selected by each sampling method

Sampling strategy	Subset code	Subset size	<i>MR</i>	<i>CE</i>	<i>SH</i>	<i>HE</i>	<i>NE</i>	<i>PN</i>	Cv (%)
Whole collection	WHGC	181	0.62	0.80	4.15	0.79	4.94	0.00	118 (100)
MSTRAT	MS50	50	0.64	0.82	4.24	0.81	5.48	0.00	118 (100)
Power Core	PC53	53	0.64	0.83	4.26	0.81	5.58	0.00	118 (100)
Core Hunter single	Cv30	30	0.63	0.82	4.26	0.81	5.50	0.00	118 (100)
Core Hunter multi	Cv-Dce30	30	0.64	0.84	4.29	0.82	5.73	0.00	118 (100)

MR, Modified Rogers distance; *CE*, Cavalli-Sforza and Edwards distance; *SH*, Shannon's diversity index; *HE*, expected proportion of heterozygous loci; *NE*, number of effective alleles; *PN*, proportion of non-informative alleles; *CV*, allele coverage

Table 3 Quality evaluation of each sampling method based on the average distance between each accession and the nearest entry (A-NE) and average distance between each entry and the nearest neighbouring entry (E-NE)

Sampling strategy	Subset code	Subset size	<i>A-NE</i>	<i>pA-NE</i>	<i>rA-NE</i>	<i>std dev</i>	<i>E-NE</i>	<i>pE-NE</i>	<i>rE-NE</i>	<i>std dev</i>
Whole collection	WHGC	181	0.000	0.000	0.000	0.000	0.306	0.306	0.306	0.000
MSTRAT	MS50	50	0.273	0.232	0.279	0.006	0.433	0.563	0.387	0.018
Power Core	PC53	53	0.268	0.225	0.270	0.006	0.444	0.555	0.383	0.017
Core Hunter single	Cv30	30	0.356	0.290	0.348	0.008	0.480	0.627	0.420	0.024
Core Hunter multi	Cv-Dce30	30	0.354	0.290	0.348	0.008	0.499	0.627	0.420	0.024

A-NE and *E-NE*, realized values; *pA-NE* and *pE-NE*, potential optimal values; *rA-NE* and *rE-NE*, average values from 1,000 random sets and the corresponding standard deviation (*std dev*)