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

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

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

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

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

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

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

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

[7)] University of Ljubljana, Biotechnical Faculty, Department of Agronomy, Jamnikarjeva 101, 1000 Ljubljana, Slovenia

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

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

ORCID ID of the authors

Paolo Boccacci: 0000-0001-8574-0478; Matthew Ordidge: 0000-0003-0115-5218; Theo van Hintum: 0000-0003-4953-4700; Daniela Torello Marinoni: 0000-0002-3679-4813; Nadia Valentini: 0000-0002-8820-9006; Merce Rovira: 0000-0002-0540-3752; Loretta Bacchetta: 0000-0001-8878-4054; Anita Solar: 0000-0001-9755-4998; Roberto Botta: 0000-0002-1952-8775.

Abstract

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

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

Introduction

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

Germplasm collections ensure the long-term conservation of genetic resources and provide easy access to plant breeders, researchers, and other users. The management and use of large germplasm collections requires significant economic costs for routine tasks, such as conservation, regeneration, duplication, documentation, and evaluation. Moreover, collections invariably contain duplicate and redundant accessions that may invalidate both the efficiency of the conservation and the effectiveness of germplasm evaluation and use (van Hintum et al. 2000). Consequently, the long-term conservation of collections can be endangered. Thus, a core collection concept was introduced in the 1980s to define

a limited set of accessions from the whole collection that represents, with a minimum of repetitiveness, the genetic diversity of a crop species and its wild relatives (van Hintum et al. 2000). Most core collections developed include 5-20% of the accessions present in the collection, capturing 70-90% of the diversity, and without redundant entries (van Hintum et al. 2000). Core collections do not replace the whole collections from which these are obtained, however they can optimize the characterization and evaluation efforts by focusing on a subset of accessions (van Hintum et al. 2000). Recognizing these objectives, core collections were recommended by the global plan of action for the conservation and sustainable utilization of plant genetic resources for food and agriculture as a necessary activity to progress the use of genetic resources (FAO 1996).

The development of core collections is traditionally based on passport data or phenotypic traits that are often unreliable and incomplete or influenced by environmental factors, respectively. DNA markers, such as simple sequence repeats (SSRs) and single-nucleotide polymorphisms (SNPs), are the tool of choice for the development of core collections. They can accurately represent the genetic diversity of the entire collection and have no problems related to incomplete data and environmental interactions, typically linked to passport and phenotypic markers. SSRs generally show a high level of polymorphism than SNPs, leading to population-specific alleles that are useful for revealing population structure. Nevertheless, SSRs are usually developed in small numbers for one species, and they may not reflect the genome-wide genetic diversity respect to SNPs (Bernard et al. 2020). The latter are much more frequent in the genomes and many SNPs can be identified using high-throughput genomics tools, allowing to develop panels of markers useful for genetic diversity and fine mapping. Thus, SSR and SNP markers bring different views of the population structure, and their other characteristics are reported by Guichoux et al. (2011). In hazelnut, a total of 718 SSR markers have been developed and more than 430 of them were used for the development of a reference linkage map (Mehlenbacher 2018; Botta et al. 2019). They have been used to fingerprint cultivars, to identify duplicate accessions and parents, to study genetic diversity in cultivated and wild populations, and in association mapping studies (Mehlenbacher 2018; Botta et al. 2019). On the contrary, SNPs have been used in hazelnut only recently and to develop two high-density genetic maps (Botta et al. 2019).

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

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

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

Material and methods

Plant material and microsatellite genotyping

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

Total genomic DNA was extracted from 0.20 g of young leaves or immature catkins using the modified procedure of Thomas et al. (1993). A total of 10 SSR loci selected by Boccacci and Botta (2010) were analysed: CaT-B107, CaT-B501, CaT-B502, CaT-B503, CaTB504, CaT-B505, CaT-B507, CaT-B508 (Boccacci et al. 2005), CaC-B020, and CaC-B028 (Bassil et al. 2005). PCR amplifications were performed in a volume of 15 µl containing 40 ng DNA, 0.5 U Taq-DNA polymerase (Bioline, Meridian Bioscience, OH, USA), 3 µl 5x PCR buffer (Bioline, Meridian Bioscience), 2.2 mM $MgCl_2$, 200 µM dNTPs, and 0.5 µM of each primer. The PCR conditions were: a first denaturation step at 95 °C for 9 min, followed by 26 cycles of denaturation (30 s at 95 °C), annealing (45 s at 55 °C and 50 °C for CaT-B502), and extension (90 s at 72 °C). The final elongation step was carried out at 72 °C for 30 min.

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

Amplification products were analysed using an ABI-PRISM 3130 Genetic Analyzer capillary electrophoresis instrument; UK samples were analysed using an ABI 3730xl capillary electrophoresis instrument (both, Applied Biosystems, Foster City, CA, USA). Results were processed with GeneMapper software (Applied Biosystems), and alleles were designated by their size in base pairs

using a GeneScan-500 LIZ standard (Applied Biosystems). UK data were aligned to the main dataset by applying a simple conversion based on a series of overlapping cultivars between the two datasets.

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

Construction of the core collections by different M-strategies

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

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

optimizing each of the following measures independently: average Cavalli-Sforza and Edwards (Dce) and Modified Rogers (Mr) as genetic distance parameter, expected proportion of heterozygous loci (He) and Shannon's diversity index (Sh) as allelic diversity indices, and allele coverage (Cv); ii) optimizing simultaneously both Cv and Dce (Cv-Dce) with equal weight assigned to each parameter. Indeed, when a weight of 50% was assigned to Cv and 50% to Dce, all observed alleles were captured in the sampled subset (Online Resource 2, Fig. S1). For each strategy, five subsets with 10, 20, 30, 40, and 50 entries, respectively, were developed.

Characterization and validation of the representativeness of the core collections

In order to evaluate the ability of each sampling strategy in capturing the diversity and representativeness in the sampled subsets, as compared to the whole germplasm collection, different parameters were considered: i) no significant differences in the number of alleles (A), genetic diversity (GD), observed heterozygosity (Ho), and polymorphism information content (PIC), computed by the ANOVA analysis with the SPSS software (IBM, Armonk, NY, USA). *Post hoc* Dunnett's test was used to compare the means of diversity estimates from different core subsets with the entire collection used as control; ii) number of loci with significantly different allele frequencies (Fr). Each locus was analysed independently, comparing the frequency of each allele between the entire collection and each core subset by the chi-squared test (Escribano et al. 2008). A, GD, Ho, PIC, and Fr were calculated using PowerMarker v.3.25 (Liu and Muse 2005).

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

Comparison and quality of sampling strategies

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

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

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

In order to create a baseline to evaluate these criteria, 1,000 random subsets (rA-NE and rE-NE) were generated of the desired sizes. The criteria were determined for each random set and the standard deviation of the results per size were calculated. To indicate the potential value of the criteria (pA-NE and pE-NE), an optimisation was done for both criteria for all core sizes. All calculations were performed following the genetic distance optimisation (GDOpt) procedure described by Odong et al. (2011).

Genetic structure analysis

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

Results

Sets of synonyms

The genetic profiles of 480 accessions across 10 SSR loci were analysed using the Identity software to identify duplicates, synonyms, and mistakes. Among them, 430 accessions are conserved in 11 international *ex situ* germplasm repositories, while 50 accessions are landraces maintained *on-farm* in five southern European countries (Portugal, Spain, Italy, Slovenia, and Greece). Moreover, only 106 accessions were identified with a unique cultivar name, while the remaining 374 were groups of two or more accessions labelled with the same name.

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

Development of core collections and comparison to the whole collection

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

The results of the variability parameters (A, GD, Ho, and PIC) obtained from a total of 37 subsets compared with the initial collection are reported in Table 1. No SSR loci with significantly different allele frequencies (Fr) were observed and thus, the criterion of no significant differences ($P < 0.05$) in at least 95% of the loci was met in all subsets. The characterization of the WHGC showed 118 amplification fragments (A) with a mean GD, Ho and PIC of 0.79, 0.80 and 0.76, respectively. Among

the subsets obtained by the MSTRAT (MS) strategy, only MS10 showed a significant difference ($P < 0.05$) in the number of alleles (A). MS50 was deemed the best subset capturing all the alleles present in the WHGC, while the optimized MS19 subset given by the algorithm captured a total of 99 alleles (83.9%). The core collection obtained by the Power Core (PC) strategy, representing a full coverage of all the alleles existing in the WHGC, comprised 53 entries (29.3%) and no significant differences were observed with the entire collection. Among the subsets obtained with the Core Hunter (CH) single-strategy, optimizing the Dce, Mr, Cv, He, and Sh indices independently, significantly higher values ($P < 0.05$) were detected for the number of alleles (A) at Dce10, Dce20, Mr10, Mr20, Mr30, Cv10, He10, and Sh10 subsets, for Ho in all Mr subsets, and for GD and PIC in all Dce, He and Sh subsets. Thus, only by optimizing the Cv index was it possible to build a core collection that respected the criteria proposed by Escribano et al. (2008) and the best subset was recovered with a minimum of 30 entries (Cv30). The CH multi-strategy, where both Cv and Dce were optimizing simultaneously with equal weight (50%), only Cv-Dce30 respected the criteria proposed by Escribano et al. (2008) and was selected as the best subset.

Selection of the final core collection

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

In Table 2 all sampling strategies are compared with the whole collection (WHGC) and are listed the mean values of the independents runs for each of the following parameters: the genetic distances MR and CE, the genetic diversity indices SH, HE, and NE, and of the auxiliary values PN and CV (Thachuk et al. 2009). All core subsets showed higher average genetic distance between entries and higher allelic richness than the WHGC. Moreover, all sampling strategies were optimal in minimizing PN and maximizing CV (0.0 and 100.0, respectively). Among them, the CH multi-strategy (Cv-Dce30) showed slightly higher values at CE, SH, and HE and the highest NE value.

In Table 3 are reported the A-NE and E-NE quality parameters for each strategy: the realized values (A-NE and E-NE), the potential optimal value (pA-NE and pE-NE), the average value from 1,000 random sets and the corresponding standard deviation (rA-NE and rE-NE). All four realized values for the A-NE criterion were higher than their respective potential optimal values but were not different from those of the random sets. Therefore, no strategies improved the A-NE criterion when compared to the random sets. On the contrary, all four E-NE values were considerably higher than random values and considerably less than potential values. In proportion, none of them reached more than 80% (PC53 and Cv-Dce30) of the maximum achievable, while the random sets reached 67-69% of the maximum achievable. Thus, all four strategies did improve the E-NE criterion, as compared to a random set, and the CH multi-strategy (Cv-Dce30) outperformed the others.

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

Genetic population structure

The estimation of ΔK (Online Resource 2, Fig. S3) from the analysis of 181 unique genotypes revealed the highest value for K = 3 (ΔK = 246.82), but high values were also obtained for K = 2 (ΔK = 205.22) and K = 5 (ΔK = 184.24). In K =2, genotypes were grouped in two gene pools (Fig. 1): one composed mainly by cultivars from the Central Europe (CEU) and the British Islands (BI), and another composed by cultivars from the Iberian Peninsula (IbeP), the Italian Peninsula (ItaP), and the Balkans-Black Sea (BBS). A total of 63 cultivars were not clearly placed in these groups (Q < 0.8) and were classified as admixed. In K = 3 were observed three groups composed mainly by cultivars from CEU and BI, ItaP, and BBS, respectively (Fig. 1). Cultivars from the Iberian Peninsula (IbeP) were widespread within all three groups, while 62 cultivars were classified as admixed. In K = 5 genotypes were classified into five groups (Fig. 1). Cultivars from CEU and BI were placed in two separate groups: Q1 was composed by 9 cultivars from CEU and 2 accessions of unknown origin (‘Mogulnuss’ and ‘Pallagrossa’), while Q2

included 8 cultivars from BI, 2 from CEU, and 3 accessions of unknown origin ('Apolda', 'Bearn', and 'Sodlinger'). 18 cultivars from IbeP showed the tendency to constitute a separate group (Q3), together with 4 cultivars and 3 landraces from ItaP, 1 landrace from BBS, and 1 accession of unknown origin ('Comen'). Q4 grouped 13 cultivars and 10 landraces from ItaP, 4 cultivars and 1 landrace from IbeP, and 1 cultivar from CEU. Q5 clustered 20 cultivars from BBS, 5 landraces from ItaP, and 2 accessions of unknown origin ('Fructo rubro' and 'Jann's'). A total of 74 genotypes were classified as admixed ($Q < 0.8$) and were deemed "mosaics" (M group).

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

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

Discussion

Building the core collection

Mislabeling and duplication are important challenges for germplasm conservation. In addition, the existence of synonyms is a characteristic challenge in vegetatively propagated woody perennial species. Thus, SSR markers have become very valuable tools in the management of *ex situ* and *in situ* hazelnut collections. In our work, as duplicates were indicated the genotypes that showed the same SSR profile, and they were 52.5 % of the total number of accessions analysed. Among them were reported two types of duplicates: i) accessions labelled with the same name and collected from different collection fields. In this first case was possible to define the true-to-type SSR genotype of most cultivars by comparing the profiles obtained from several accessions or identify some mislabeling among them; and ii) accessions labelled with a different name, conserved in different collection fields or in the same collection field. In this second case, the results allowed us to identify some mislabeling (9.7% of the total number of accessions analysed) due to planting or labeling mistakes, as happened to Bassil et al. (2009) during a backup of the USDA collection. Moreover, it was also possible to confirm several sets of synonyms reported in the literature (Boccacci et al. 2006, 2008, 2013; Gökirmak et al. 2009; Gürcan et al. 2010; Bacchetta et al. 2015) and identify new ones. Among them, the local cultivars 'Negret primerenc' and 'Negret primerenc cort' revealed the same genetic profile of 'Negret', although are more productive and their fruits mature earlier (Rovira et al. 2017). They represented a possible case of clonal mutation, and a similar result was observed between 'Tonda di Biglini' and 'Tonda Gentile delle Langhe' by Valentini et al. (2014). Consequently, to construct our core collection it was important to identify mislabeling, duplicates, and synonyms from the whole hazelnut germplasm collection (WHGC), to delete a significant source of redundancy and build the core collection only from true-to-type genotypes.

The main strategies used to construct core collections from molecular marker data can be classified into two groups. The first methods are based on genetic distance, with or without stratified sampling techniques, that cluster the accessions and then select entries from each cluster using different allocation approaches (van Hintum et al. 2000; De Beukelaer et al. 2012). The second methods are based on the M-strategy that construct cores with high allelic richness, maximizing the number of observed alleles at each marker locus (Schoen and Brown 1993). M-methods are the only approaches that recover

all the alleles of the whole collection, including rare alleles, and keep the original allele frequencies at each marker, favouring smaller subsets, reducing redundancy, and capturing most of the genetic diversity (Marita et al. 2000; Gouesnard et al. 2001). In fruit and nut tree genera, such as *Annona*, *Ficus* and *Castanea*, the M-strategy was the most efficient method to develop core collections, outperforming other strategies (Escribano et al. 2008; Balas et al. 2014; Pereira-Lorenzo et al. 2017); and for this reason was largely used by many other authors (Le Cunff et al. 2008; Miranda et al. 2010; Belaj et al. 2012; Díez et al. 2012; El Bakkali et al. 2013; Štajner et al. 2014; Liang et al. 2015; Liu et al. 2015; Öztürk et al. 2017; Bernard et al. 2020). Nevertheless, the choice of the most appropriate evaluation measures depends on the purpose of the core collection. Methods based on the allele representativeness (genetic distances) are preferred by plant breeders, while methods based on the allele richness, including rare and localized alleles, interest taxonomists and geneticists (Marita et al. 2000).

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

The main advantage of the Core Hunter software is its ability to build core collections using different allocation strategies by optimizing one parameter or many parameters simultaneously. Generally, core subsets optimized using multiple criteria perform worse than these obtained using individual measures (Thachuk et al. 2009; Díez et al. 2012). Nevertheless, our core subset obtained from CH multi-strategy (Cv-Dce30) showed slightly higher values at MR, CE, SH, and HE and the highest

NE value, compared to that obtained from CH single-strategy (Cv30) (Table 2). Thus, CH multi-strategy approach, optimizing Cv and Dce simultaneously with equal weight, was selected as the best strategy to build our final core collection. It satisfied both the breeders' and geneticists'/taxonomists' perspectives despite including only 16.6 % of the WHCG genotypes. This value sits within the 5–20 % proposed by van Hintum et al. (2000) and is lower than the 19 % obtained from the Turkish national hazelnut collection (Öztürk et al. 2017). Other studies on fruit tree crops reported a minimum requirement of 4 % inclusion in grape (Le Cunff et al. 2008; Štajner et al. 2014), a common range from 13 % in fig (Balas et al. 2014) to 15-19 % in olive (Belaj et al. 2012; Díez et al. 2012; El Bakkali et al. 2013), and a maximum of 28.6 % in apple (Liang et al. 2015) and 29.8 % in chestnut (Pereira-Lorenzo et al. 2017).

According to Ondong et al. (2013), the criterion of choice for evaluating the quality of core collections should be determined by the objectives or type of the core collection. Thus, they proposed two genetic distance-based criteria, A–NE and E–NE, for evaluating the quality of two important types of core collections, respectively: i) a core collection (CC-I) where each entry represents one (itself) or more accessions of the whole collection, in order to maximize the representativeness of genetic diversity (A-NE); ii) a core collection (CC-X) where the diversity of the traits of the entries is maximized, in order to represent the total genetic diversity (E-NE). Using these criteria, our results indicated that all the best subsets obtained from each sampling strategy (MS50, PC53, Cv30, and Cv-Dce30) aimed at covering the range of the genetic diversity, rather than representing the accessions from WHGC. All values for the A-NE criterion were not significantly different from that of a random set. On the contrary, all four strategies improved the E-NE criterion, as compared to a random set, and the CH multi-strategy (Cv-Dce30) outperformed the others (Table 3). Since the objective of our final core collection was the maximisation of the allelic richness, including rare and localized alleles, the E-NE criterion was the most appropriate to evaluate the quality of the four core collections obtained. Nevertheless, the fact that the size of the core collections varied made E-NE comparison difficult (Table 3). In the case of the Cv30 and Cv-Dce30 subsets (30 entries), the CH multi-strategy outperformed the CH single-strategy in all aspects of the E-NE criterion, but not by a large factor (80% vs 77% of the maximum achievable). Comparing these two subsets with the larger core collections MS50 (50 entries) and PC53 (53 entries),

all four were in a similar range. Nevertheless, PC53 did best in terms of standard deviation from the random E-NE, whereas Cv-Dce30 did best in terms of approaching the potential maximum.

Characteristics of the core collection

The population structure and relatedness among SSR genotypes of 181 cultivars and landraces from WHGC, indicated the existence of three levels of genetic structure (Fig 1, Fig. 2, and Fig. 3). In the first level (K =2) was observed a geographic pattern with one gene pool dominating the western and central Europe (BI and CEU) and another gene pool dominating the southern Europe (IbeP, ItaP, and BBS). In the second level (K = 3), there were still the gene pools frequent in the western and central Europe (BI and CEU) and in the southern Europe (IbeP and ItaP), but there appeared a third gene pool (BBS) most frequent in the Balkans and the Anatolian Peninsula. Finally, in the third level (K = 5) there was still the BBS gene pool, while genotypes from northern Europe were further subdivided into the BI and CEU gene pools and those from southern Europe into the IbeP and ItaP gene pools.

The high level of genetic similarity between cultivars grown in the Iberian and Italian Peninsula, observed in K = 2 and K = 3 (Fig. 1), was already reported by other authors (Boccacci et al. 2006; Gökirmak et al. 2009; Gürcan et al. 2010) and was a consequence of a high gene flow between western and central Mediterranean basin (Boccacci and Botta 2010). Nevertheless, in K = 5 (Fig. 1) cultivars from the Iberian Peninsula were separated from Italian ones (Boccacci et al. 2013) and a significant genetic differentiation between the Spanish and Italian gene pools was reported in subsequent studies (Boccacci and Botta 2010; Boccacci et al. 2013), indicating that northern Spain and southern Italy were two independent hazelnut domestication areas (Boccacci et al. 2013). On the contrary, the genetic similarity 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 true-to-type 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 were included as "kernel" in our core collections (van Hintum et al. 2000), but the most popular hazelnut cultivar 'Tonda Gentile delle Langhe' (TGL), particularly appreciated by the industry for the morphological, organoleptic, and nutritional characteristics of its nuts and kernels, was included in all MS50, PC53, Cv30, and Cv-Dce30 subsets. Different reference cultivars could be added on a case-by-case basis in different places where this core collection could be studied, such as: 'Negret' and

'Casina' in Spain, 'Barcelona' (syn. 'Fertile de Coutard') in France, 'Tonda Gentile Romana' and 'Tonda di Giffoni' in Italy, and 'Tombul' in Turkey.

Conclusions

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

Declarations

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

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

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

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

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

Code availability Not applicable

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

References

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

Balas FC, Osuna MD, Domínguez G, Pérez-Gragera F, López-Corrales M (2014). *Ex situ* conservation of underutilised fruit tree species: establishment of a core collection for *Ficus carica* L. using microsatellite markers (SSRs). Tree Genet Genomes 10:703–710

Bassil NV, Botta R, Mehlenbacher SA (2005) Microsatellite markers in the hazelnut: isolation, characterization, and cross-species amplification in *Corylus*. J Am Soc Hort Sci 130:543-549

Bassil NV, Postman J, Hummer K, Botu M, Sezer A (2009) SSR fingerprinting panel verifies identities of clones in backup hazelnut collection of USDA genebank. Acta Hortic 845: 95-102

Belaj A, del Carmen Dominguez-García M, Atienza SG, Martín Urdíroz N, De la Rosa R, Satovic Z, Martín A, Kilian A, Trujillo I, Valpuesta V, Del Río C (2012). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. Tree Genet Genomes 16:76

Bernard A, Barreneche T, Donkpegan A, Lheureux F, Dirlewanger E (2020). Comparison of structure analyses and core collections for the management of walnut genetic resources. Tree Genet Genomes 8:365–378

Boccacci P, Botta R (2010) Microsatellite variability and genetic structure in hazelnut (*Corylus avellana* L.) cultivars from different growing regions. Sci Hortic 124:128-133

Boccacci P, Akkak A, Bassil NV, Mehlenbacher SA, Botta R (2005) Characterization and evaluation of microsatellite loci in European hazelnut (*Corylus avellana* L.) and their transferability to other *Corylus* species. Mol Ecol Notes 5:934-937

Boccacci P, Akkak A, Botta R (2006) DNA-typing and genetic relationships among European hazelnut (*Corylus avellana* L.) cultivars using microsatellite markers. Genome 49:598-611

Boccacci P, Rovira M, Botta R (2008) Genetic diversity of hazelnut (*Corylus avellana* L.) germplasm in northeastern Spain. HortScience 43:667-672

Boccacci P, Aramini M, Valentini N, Bacchetta L, Rovira M, Drogoudi P, Silva AP, Solar A, Calizzzano F, Erdorğan V, Cristofori V, Ciarmiello LF, Contessa C, Ferreira JJ, Marra FP, Botta R (2013) Molecular and morphological diversity of *on-farm* hazelnut (*Corylus avellana* L.) landraces from southern Europe and their role in the origin and diffusion of cultivated germplasm. Tree Genet Genomes 9:1465–1480

Botta R, Molnar TJ, Erdorğan V, Valentini N, Torello Marinoni D, Mehlenbacher S (2019). Hazelnut (*Corylus* spp.) Breeding. In: Al-Khayri JM, Jain SM, Johnson DV (eds.) Advances in Plant Breeding Strategies: Nut and Beverage crops. Springer Nature, Switzerland, Volume 4, pp 157-219

De Beukelaer H, Smýkal P, Davenport GF, Fack V (2012) Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search. BMC Bioinformatics 13:312

Di Guardo M, Scollo F, Ninot A, Rovira M, Hermoso JF, Distefano G, La Malfa S, Batlle I (2019) Genetic structure analysis and selection of a core collection for carob tree germplasm conservation and management. Tree Genet Genomes 15: 41

Díez CM, Imperato A, Rallo L, Barranco D, Trujillo I (2012). Worldwide core collection of olive cultivars based on simple sequence repeat and morphological markers. Crop Sci 52:211-221

El Bakkali A, Haouane H, Moukhli A, Costes E, Van Damme P, Khadari B (2013) Construction of core collections suitable for association mapping to optimize use of Mediterranean olive (*Olea europaea* L.) genetic resources. PLoS ONE 8:e61265

Escribano P, Viruel MA, Hormaza JI (2008) Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (*Annona cherimola*, Annonaceae), an underutilised subtropical fruit tree species. Ann Appl Biol 153:25–32

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620

FAO (1996) Global plan of action for the conservation and sustainable utilization of plant genetic resources for food and agriculture. Food and Agriculture Organization, Rome

FAOSTAT (2021) http://www.fao.org/faostat/en/?#data. Accessed 06 May 2021

Franco J, Crossa J, Taba S, Shands H (2005) A sampling strategy for conserving genetic diversity when forming core subsets. Crop Sci 45:1035–1044

Franco J, Crossa J, Warburton ML, Taba S (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci 46:854–864

Ferreira JJ, Garcia-González C, Tous J, Rovira M (2010) Genetic diversity revealed by morphological traits and ISSR markers in hazelnut germplasm from northern Spain. Plant Breed 129:435–441

Gökirmak T, Mehlenbacher SA, Bassil NV (2009) Characterization of European hazelnut (*Corylus avellana* L.) cultivars using SSR markers. Genet Resour Crop Evol 56:147-172

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Guichoux E, Lagache L, Wagner S et al (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591–611

Gürcan K, Mehlenbacher SA, Erdoğan V (2010) Genetic diversity in hazelnut (*Corylus avellana* L.) cultivars from Black Sea countries assessed using SSR markers. Plant Breed 129:422–434

Hummer KE (2001) Hazelnut genetic resources at the Corvallis repository. Acta Hortic 556:21–24

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

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

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

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

Lassois L, Denancé1 C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, Poncet C, Lasserre-Zuber P, Feugey L, Durel CE (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. Plant Mol Biol Rep 34:827–844

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

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

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

Liu Q, Song Y, Liu L, Zhang M, Sun J, Zhang S, Wu1 J (2015) Genetic diversity and population structure of pear (*Pyrus* spp.) collections revealed by a set of core genome-wide SSR markers. Tree Genet Genomes 11:128

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

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

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

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

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

Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk FA (2011) Statistical techniques for defining reference sets of accessions and microsatellite markers. Crop Sci 51(6):2401–2411

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 126:289–305

Öztürk SC, Balık Hİ, Balık SK, Kızılcı G, Duyar Ö, Doğanlar S, Frary A (2017) Molecular genetic diversity of the Turkish national hazelnut collection and selection of a core set. Tree Genet Genomes 13:113

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 4:347–354

Pereira-Lorenzo S, Ramos-Cabrer AM, Barreneche T, Mattioni C, Villani F, Díaz-Hernández MB, Martín LM, Martín A (2017) Database of European chestnut cultivars and definition of a core collection using simple sequence repeats. Tree Genet Genomes 13:114

Perrier X, Jacquemoud-Collet J (2006) DARwin software. Available from: http://darwin.cirad.fr/. Accessed 5 May 2021

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Ramasamy RK, Sumathy Ramasamy S, Bindroo BB, Naik VG (2014) STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus 3:431

Rovira M, Hermoso JF, Romero AJ (2017) Performance of hazelnut cultivars from Oregon, Italy, and Spain, in Northeastern Spain. HortTechnology 27(5):631-638

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

Štajner N, Tomić L, Ivanišević D, Korać N, Cvetković-Jovanović T, Beleski K, Angelova E, Maraš V, Javornik B (2014). Microsatellite inferred genetic diversity and structure of Western Balkan grapevines (*Vitis vinifera* L.). Tree Genet Genomes 10:127–140

Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10:243

Thomas MR, Matsumoto S, Cain P, Scott NS (1993) Repetitive DNA of grapevine: classes present and sequences suitable for cultivar identification. Theor Appl Genet 86:173-180

Thompson MM, Lagerstedt HB, Mehlenbacher SA (1996) Hazelnuts. In: Janick J, Moore JN (eds) Fruit breeding: nuts, vol 3. Wiley, New York, pp 125–184

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

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

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

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

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

Sampling strategy	Subset code	Subset size	*A*	*GD*	*Ho*	*PIC*
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]

A, number of alleles; *GD*, genetic diversity; *Ho*, observed heterozygosity; *PIC*, polymorphism information content
[a] Statistically 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	*MR*	*CE*	*SH*	*HE*	*NE*	*PN*	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	*A-NE*	*pA-NE*	*rA-NE*	*std dev*	*E-NE*	*pE-NE*	*rE-NE*	*std dev*
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*)

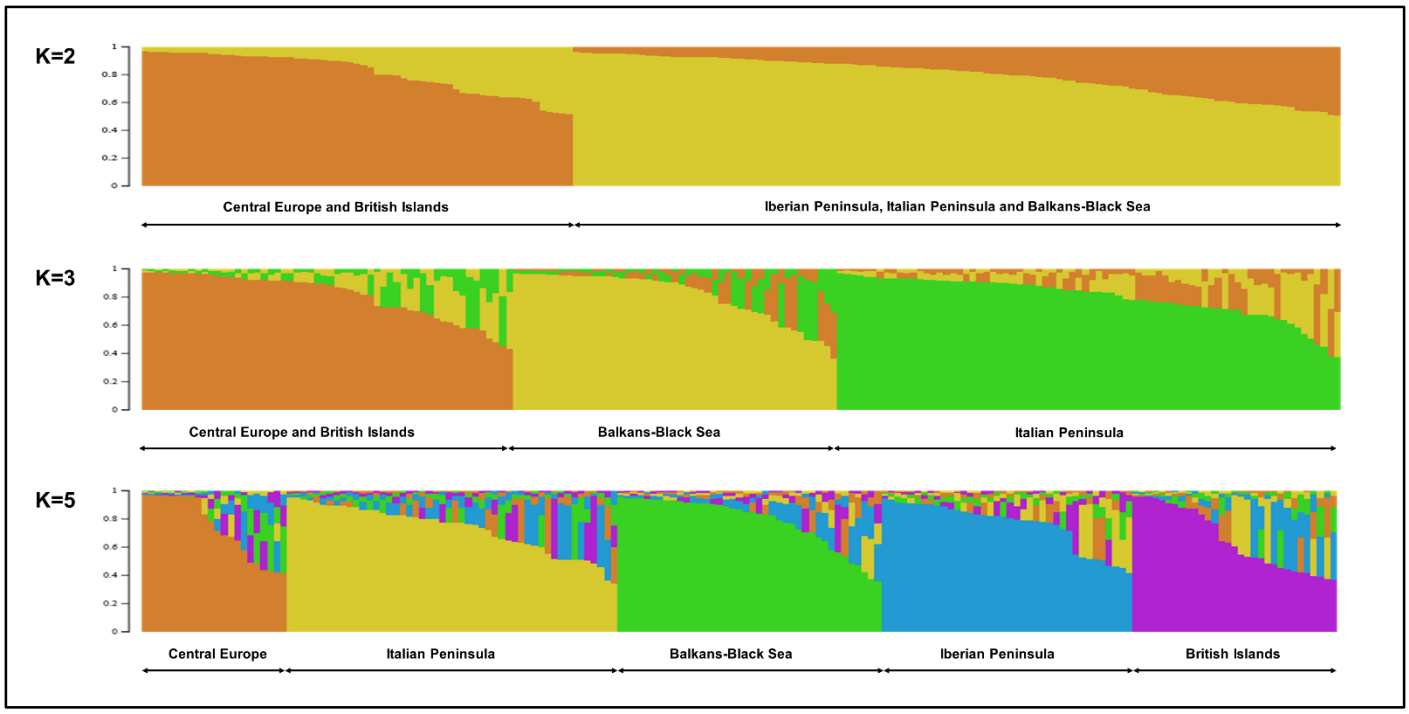

Fig. 1 Population structure and hierarchical organization of genetic relatedness of 181 genotypes from the whole hazelnut germplasm collection (WHGC) at K = 2, K = 3, and K = 5 as inferred by STRUCTURE software

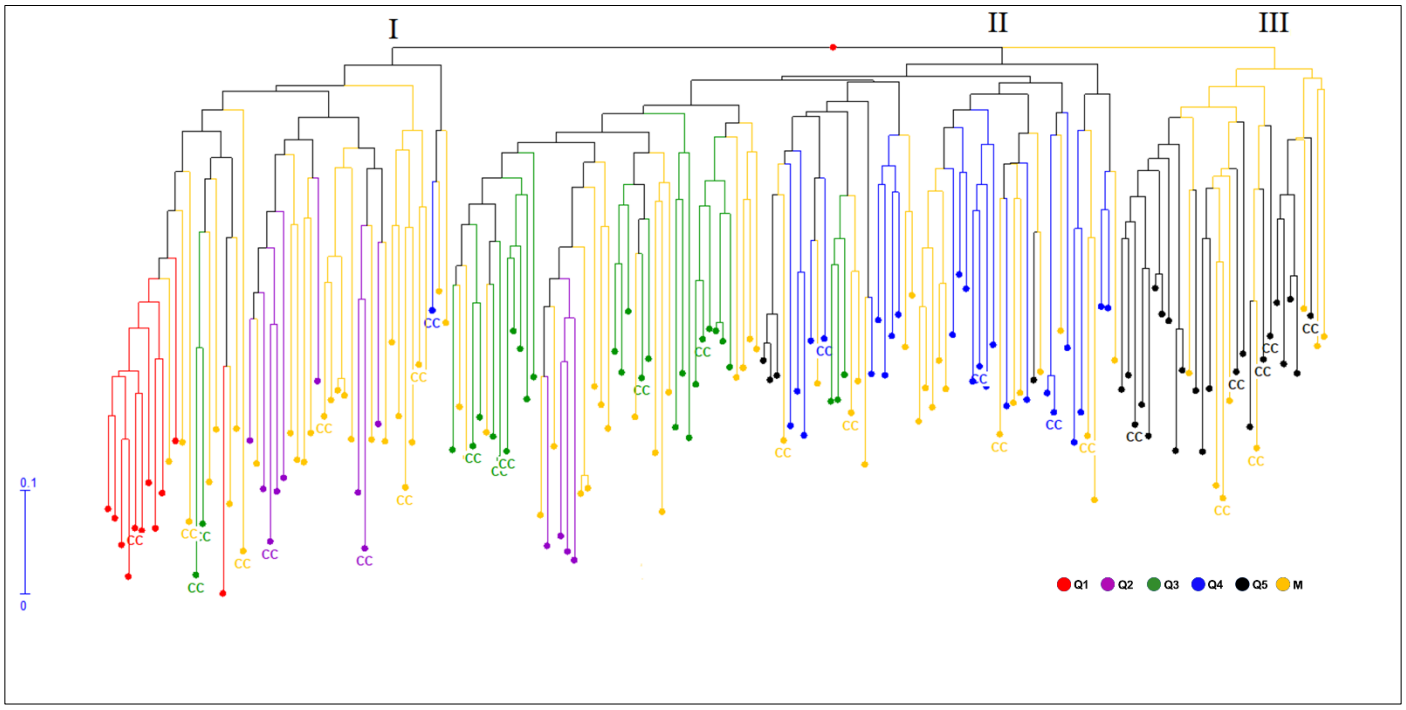

Fig. 2 Neighbor-joining dendrogram based on the Dice similarity index showing the relationships among 181 hazelnut genotypes from WHGC. Genotypes are colored according to their assignment to the different gene pools, as inferred by STRUCTURE software at K = 5: Central Europe (Q1), British Islands (Q2),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

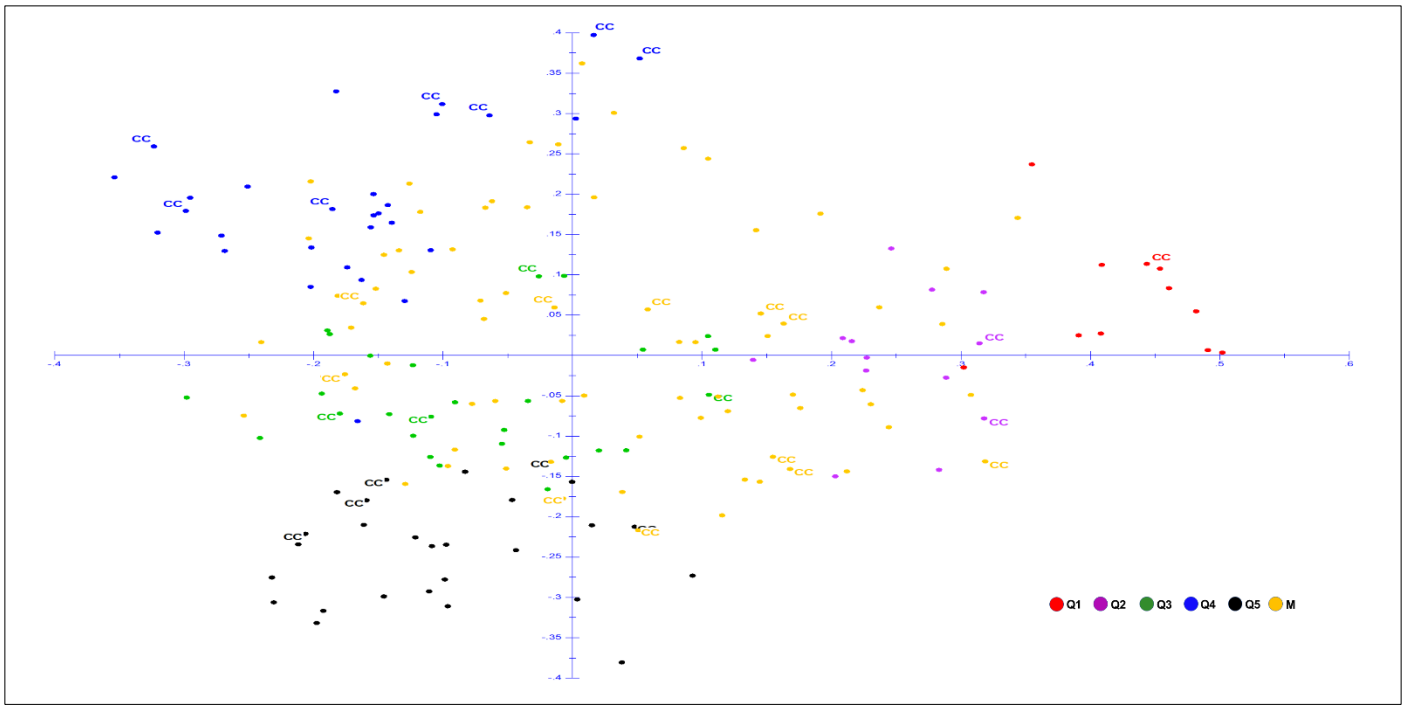

Fig. 3 Two-dimensional PCoA scatterplot of 181 hazelnut genotypes from WHGC based on Dice's distance. Genotypes are colored according to their assignment to the different gene pools, as inferred by STRUCTURE software at K = 5: Central Europe (Q1), British Islands (Q2), Iberian Peninsula (Q3), Italian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