Testing Different Approaches to Construct an Olive (*Olea europaea* L.) Core Subset Suitable for Association Genetic Studies

A. El Bakkali[1,2,3,4 *], H. Haouane[1,2], P. Van Damme[4], B. Khadari[1,5]

[1]INRA, UMR 1334 Amélioration Génétique et Adaptation des Plantes (AGAP), F-34070 Montpellier, France

[2]Montpellier SupAgro, UMR 1334 AGAP, F-34070 Montpellier, France

[3]INRA, CRRA-Meknès, UR Amélioration des Plantes et Conservation des Ressources Phytogénétiques (APCRPG), B.P. 578, Meknès, Morocco

[4]University of Ghent, Faculty of Bioscience Engineering, Coupure links 653 B-9000, Ghent, Belgium

[5]CBNMED, UMR 1334 AGAP, F-34398 Montpellier, France

*Corresponding author: ahmed_elbakkali@yahoo.fr

Keywords: genetic resources, core collection, sampling methods, Simple Sequence Repeats (SSR)

Abstract

Evaluation of genetic diversity is of great interest for the management of germplasm collections and breeding programs. Management can be efficient when the evaluation is focused on a subset of accessions that represents the variability observed in the whole germplasm collection. Most core sets have been developed for seed crops using different approaches and sampling size to select entries on the basis of genetic and/or phenotypic data, while few studies on perennial crops have been published.

Here, we proposed a core collection for cultivated olive (*Olea europaea* L.) using both Simple Sequence Repeat (SSR) markers and phenotypic traits by testing different sampling approaches including stratified and non-stratified methods. Twelve SSR markers were used to construct a core subset from an initial collection of 505 single genotypes sourced from 14 Mediterranean countries. Among all the sampling methods, we showed that a sample size of 12.5% was most suitable in capturing all the observed alleles using the M-method approach. Based on both SSR and phenotypic data, we established an initial core set, including the main Mediterranean varieties, which displayed the highest genetic and phenotypic variability. No obvious genetic structure was indicated when the core subset was analyzed with Principal Coordinate analysis (PCoA). Our results gave an efficient basis as a first step for olive association mapping. The constructed core subset could be further evaluated for traits of agronomic interest, leading to association between the allelic variation and the phenotypic variability.

INTRODUCTION

Ex-situ conservation of genetic resources is important to preserve adaptive characters, to prevent erosion and extinction of local varieties as well as to enable the use of outstanding accessions with interesting genes in breeding and selection programs. Allelic diversity in germplasm has been used to understand the genetic basis of complex traits and to identify genes related to phenotypic variation for specific traits (Zhu et al., 2008). However, the management, characterization and use of large germplasm

collections are frequently unfeasible and inefficient due to cost and time constraints. Genetic resources can be more efficiently managed if they are focused on a subset of accessions known as a core collection (or core subset) which includes as much variability as possible from the whole collection (Frankel and Brown, 1984). Among all the methods proposed to construct core subsets, two are the most used: stratifying method (Brown, 1989; Franco et al., 1998) and maximizing method (M-method; Schoen and Brown, 1993) implemented in the MSTAT software package (Gouesnard et al., 2001). Recently, an advanced stochastic local search method attempting to optimize a single or multiple genetic parameters simultaneously has been proposed and implemented in the Core Hunter program (Thachuk et al., 2009).

During the last ten years, many core subsets have been published for diverse plant species, including some perennial fruit crops. These core subsets were made using different eco-geographical, agro-morphological or genetic data (McKhann et al., 2004; Franco et al., 2006; Le Cunff et al., 2008; Escribano et al., 2008; Chung et al., 2009; Santesteban et al., 2009;). However, despite the social-economic importance of the olive species (*Olea europaea L.*) and its broad genetic diversity (more than 1,200 varieties; Bartolini et al., 1998), no core subset has been developed so far.

In this study, we propose an initial core collection for cultivated olive species using 12 nuclear SSR markers and 38 phenotypic traits. Different sampling methods were compared and the representativeness of the overall genetic diversity of the constructed core subset was studied.

MATERIAL AND METHODS

Data set

A total of 561 accessions, maintained in the *ex-situ* world olive germplasm bank at the experimental orchard of Tassaoute, INRA Marrakech, Morocco (WOGB Marrakech), were characterised using 12 nuclear SSR loci (Haouane et al., 2011). Phenotypic data was taken from published data based on the variety name as an identification key (Bartolini et al., 1998; Bartolini, 2008). Data of 38 agronomic traits classified into 114 classes according to standards described by the International Olive Oil Council (IOOC) was compiled for 419 varieties from different countries.

Comparison of different sampling methods

Six approaches using only SSR data were evaluated to compare the performance of the current state-of-the-art methods to construct core sets: (i) Random sampling (R-method); (ii) Maximizing method (M-method): maximizing the number of alleles at each locus using the MSTRAT software; (iii) Simulated annealing method (SA-method): using the PowerMarker v3.25 software (Liu and Muse, 2005) under the criterion of optimizing the genetic diversity (GD) or the number of alleles (NA); (iv) Advanced stochastic local search method (SLS-method): using the pseudo index parameter α in the range of [0-1] implemented in Core Hunter software (Thachuk et al., 2009) [core sets were first formed by optimizing each of the four parameters independently of the others {single; average Modified Rogers distance (MR), Shannon diversity index (SH), coverage of alleles (CV) and expected proportion of heterozygous loci (He)}, then, the Core Hunter program was run to optimize both MR distance and SH diversity index (double) and finally, the four measures were optimized simultaneously (multiple) with equal weight assigned to each genetic parameter]; (v) Logarithmic method (L-method): 11 groups were defined based on genetic data set (data not shown), [for each group, individuals were selected

proportional to the logarithm of the number of accessions in that group by maximizing the number of alleles using MSTRAT software]; (vi) Proportional method (P-method): based on the 11 defined groups, selection of individuals was proportional to the number of accessions in each group by maximizing the number of alleles in that group. For each approach, 20 independent runs were performed.

To define the preliminary core collection, genotypic and phenotypic data were considered. For each core set sampled with each sampling method, the Shannon-Weaver phenotypic diversity index (H_c; Hutcheson, 1970) was calculated.

Representativeness of overall diversity genetic

Principal Coordinate Analysis (PCoA) was performed with the GenAlex 6.1 macro program (Peakall and Smouse, 2005) to provide a spatial distribution of the proposed core collection.

RESULTS AND DISCUSSION

Comparison of different sampling methods

Based on 12 nuclear SSR loci, the analysis of 561 accessions revealed 505 single genotypes with 210 alleles of which 24 alleles were observed only once (Haouane et al., 2011). Sixty four cultivars (12.5%) were necessary to capture all the alleles using the M-method and SLS method when optimizing the CV. The Core Hunter program was used to construct core sets by optimizing each genetic parameter independently of the others (single). For the 12.5% sample size (Table 1), the single SLS method was able to select better core sets than all the other methods with respect to the single measure being optimized, except for the M-method for the coverage of alleles (CV). While the Core Hunter program is able to select core subsets which meet or exceed those chosen by other programs for a particular genetic parameter, other parameters not considered during optimization were highly affected. For instance, when Core Hunter attempted to optimize CV, selected cores had lower values than M-method for the three other parameters. Moreover, Core Hunter reported the lowest score for CV when attempting to maximize any other genetic parameter. This means that selection of a core set with a higher number of alleles did not result in a higher MR genetic distance or SH diversity index.

For the SLS method where the MR genetic distance and the SH diversity index were optimized simultaneously, a trade-off between the two parameters was observed relative to their respective weight assigned to each measure proportional to the pseudo-index α (Fig. 1). When optimizing both MR and SH (double) and four parameters (multiple) using Core Hunter program, we selected core subsets that simultaneously had better average MR distance and SH diversity index than core sets chosen by others approaches. For M-method, the selected core subsets showed higher MR and SH than the subset sampled by SLS method when optimizing only CV (Fig. 1).

Construction of initial core subset

Core sets selected by M-method had higher values of Shannon-Weaver phenotypic diversity index (H_c) than those sampled by other approaches, indicating that more traits' classes were retained when using M-method (data not shown). Eighteen out of the twenty cores selected by the M-method showed no significant differences in frequency of trait classes compared to the entire collection.

An initial core subset was proposed that included 64 entries sampled by the M-method showing the highest H_c among the twenty sampled subsets (100% of allelic

diversity and 84.2% of phenotypic diversity), a further eight entries capturing the remaining phenotypic variation and an additional eight other cultivars considered as the most cultivated in the Mediterranean basin. The eighty olive entries (16% of the entire collection; Table 2) included in the final core collection correspond to the following countries: six from Algeria (14% of country's accessions conserved in the WOGB collection), two from Croatia (12.5%), three from Cyprus (10.7%), five from Egypt (26.3%), one from France (8.3%), four from Greece (30.7%), 24 from Italy (14.3%), one from Lebanon (6.2%), five from Morocco (12.5%), two from Portugal (14.2%), 12 from Spain (12.3%), nine from Syria (12.6%) and six from Tunisia (25%). All the countries in the whole collection are represented in the selected core collection except Slovenia that has nine accessions in the WOGB Marrakech collection.

Representativeness of overall genetic diversity

The spatial distribution on the first two axes of the principal coordinate analysis (PcoA; Fig. 2) indicates that not only does the proposed core subset span the range of all the accessions of the whole collection but also shows an even distribution along the two main axes. This result indicates a lack of obvious genetic structure in the proposed core collection.

CONCLUSION

This work is a preliminary step towards optimized conservation and management of olive genetic resources, and subsequently for association mapping studies by assessing the linkage disequilibrium (LD) extent between loci and the relatedness within the formed core. As the first core subset proposed for cultivated olive species, the current unstructured core provides a working collection of cultivated olive germplasm that can be used to help to design association-mapping experiments to identify the genetic basis of the most economically important traits.

ACKNOWLEDGEMENTS

This study was supported by PRAD 08-01 and Merit Scholarship Program for High Technology 1430H/2009 of the Islamic Development Bank (IDB).

Literature cited

Bartolini, G., Prevost, G., Messeri, C. and Carignani, G. 1998. Olive germplasm: Cultivars and world-wide collections. FAO Library. Rome, Italy.

Bartolini, G. 2008. Olive germplasm (*Olea europaea* L.), cultivars, synonyms, cultivation area, collections, descriptors. http://www.oleadb.it/olivodb.html.

Brown, A.D.H. 1989. Core collection: a practical approach to genetic resources management. Genome 31:818-824.

Chung, H.K., Kim, K.W., Chung, J.W., Lee, J.R., Lee, S.Y., Dixit, A., Kang, H.K., Zhao, W., McNally, K.L., Hamilton, R.S., Gwag, J.G. and Park, Y.J. 2009. Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. J. Integr. Plant Biol. 51:1116-25.

Escribano, P., Viruel, M.A. and Hormaza, J.I. 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.

Franco, J., Crossa, J., Villaseñnor, J., Taba, S. and Eberhart, S.A. 1998. Classifying genetic resources by categorical and continuous variables. Crop Sci. 38:1688-1696.

Franco, J., Crossa, J., Warburton, M.L. and Taba S. 2006. Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci. 46:854-864.

Frankel, O.H. and Brown, A.H.D. 1984. Plant genetic resources today: a critical appraisal. In crop genetic resources: conservation and evaluation (Holden JHW and Williams JT. eds). London 249-257.

Gouesnard, B., Bataillon, T.M., Decoux, G., Rozale, C., Schoen, D. J. and David, J. L. 2001. MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. 92:93-94.

Haouane, H., El Bakkali, A., Moukhli, A., Tollon, C., Santoni, S., Oukabli, A., El Modafar, C. and Khadari B. 2011. Genetic structure and core collection of the World Olive Germplasm Bank of Marrakech: towards the optimised management and use of Mediterranean olive genetic resources. Genetica. DOI 10.1007/s10709-011-9608-7.

Hutcheson, K. 1970. A test for comparing diversities based on the Shannon formula. J. Theor. Biol. 29:151 -154.

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

Liu, K. and Muse, S.V. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129.

McKhann, H.I., Camilleri, C., Berard, A., Bataillon, T., David, J.L., Reboud, X., Le Corre, V., Caloustian, C., Gut, I.G. and Brunel, D. 2004. Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J. 38:193-202.

Peakall, R. and Smouse P.E. 2005. GenAlEx 6: genetic analysis in Excel. population genetic software for teaching and research. Mol. Ecol. Notes 6:288-295.

Santesteban, L.G., Miranda, C. and Royo, J.B. 2009. Assessment of genetic and phenotypic diversity maintained in apple core collections constructed by using either agro-morphologic or molecular marker data. Spanish J. Agri. Res. 7:572-584.

Schoen, D.J. and Brown. A.H.D. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proceeding of the National Academy of Science. USA. 38:10623-10627.

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M. and Davenport, G.F. 2009. Core Hunter: an algorithm for sampling genetic resources based on multiple genetics measures. Bioinformatics 10:243.

Zhu, C., Gore, M., Buckler, ES. and Yu, J. 2008. Status and prospects of association mapping in plants. Plant Genome 1:5-20.

Tables

Table 1. Genetic parameters of core subsets selected by different sampling methods at a sample size of 12.5%. MR: average Modified Rogers distance, SH: Shannon diversity index, HE: expected proportion of heterozygosity, CV: coverage of alleles (%), NA: number of alleles, GD: genetic diversity.

Sampling methods	**MR**	**SH**	**HE**	**CV**
Stockastic local search method (single)[1]	**0.704**	**4.668**	**0.839**	**100**
Stockastic local search method (multiple)[2]	0.665	4.653	0.828	96.7
Maximizing method[3]	0.651	4.57	0.809	**100**
Simulated annealing method (NA)[3]	0.607	4.214	0.745	63.5
Simulated annealing method (GD)[3]	0.608	4.211	0.752	62.1
Logarithmic method[3]	0.632	4.474	0.799	79.4
Proportional method[3]	0.631	4.48	0.799	79.5
Random method[3]	0.609	4.239	0.753	64.7
Whole collection	0.609	4.294	0.756	100

[1]Each selection parameter was attempted to be optimized independently by performing 20 runs with 100% weight given to the respective parameter during each run. Results reported for each measure are independent of results reported for all other measures.
[2]Twenty independent runs were performed with equal weight given to each of the four parameters to maximize all genetic parameters simultaneously.
[3]Each genetic parameter is the mean value of 20 independent runs for each method

Table 2. Cultivar name, code, and origin of 80 entries included in the constructed core subset representing the 16% of the World Olive Germplasm Bank of Marrakech (WOGB).

#	Cultivar name	Country	#	Cultivar name	Country
1	Aîmel (Alg002)	Algeria	41	Azeitonera Azeitira (Por001)	Portugal
2	Chemlal (Alg003)	Algeria	42	Acebuchera (Es001)	Spain
3	Ifiri (Alg026)	Algeria	43	Alameno blanco (Es002)	Spain
4	Khadraïa (Alg038)	Algeria	44	Arbequina (Es006)	Spain
5	Souidi (Alg041)	Algeria	45	Blanqueta (Es009)	Spain
6	Zeletni (Alg030)	Algeria	46	Corbelia (Es017)	Spain
7	Istarska crnica (Croi008)	Croatia	47	Llometa (Es077)	Spain
8	Sitnica (Croi016)	Croatia	48	Mollar de cieza (Es082)	Spain
9	Flasoy (Ch014)	Cyprus	49	Sevillenca (Es091)	Spain
10	Lithrodontas (Ch025)	Cyprus	50	varudo (Es063)	Spain
11	Menikon 1 (Ch027)	Cyprus	51	Vera (Es064)	Spain
12	Aggizi Oshime (Egy003)	Egypt	52	Bent Alkade (Syr061)	Syria
13	Baid Lhmam (Egy011)	Egypt	53	Idleb 3 (Syr002)	Syria
14	Hamed (Egy019)	Egypt	54	Janude 2 (Syr047)	Syria
15	Wateken (Egy004)	Egypt	55	Khashabi (Syr041)	Syria
16	Lucques de l'Herault (Fr009)	France	56	Killin (Syr015)	Syria
17	Amphisis (Gr001)	Greece	57	Mesyaf 1 (Syr051)	Syria
18	Karolia (Gr005)	Greece	58	Oum Kanane (Syr064)	Syria
19	Arancino (It028)	Italy	59	Chetoui (Tun019)	Tunisia
20	Brandofino (It085)	Italy	60	Doukar (Tun005)	Tunisia
21	Castricianella (It096)	Italy	61	Gerboui Nord (Tun022)	Tunisia
22	Cavalieri (It098)	Italy	62	Jemri bouchouka (Tun023)	Tunisia
23	Cucca (It032)	Italy	63	Neb jmel (Tun008)	Tunisia
24	Frantoio (It041)	Italy	64	Zalmati nord (Tun018)	Tunisia
25	Grossa di spagna (It021)	Italy	65	Ascolana tenera (It076)	Italy
26	Lastrino (It023)	Italy	66	Calatina (It086)	Italy
27	Leccio Marmmuono (It108)	Italy	67	Cariasina (It090)	Italy
28	Mignolo Cerretanolo Carretano (It047)	Italy	68	Gordale Sevillana (It106)	Italy
29	Moraiolo (It053)	Italy	69	Morchiaio (It049)	Italy
30	Moresca (It117)	Italy	70	Piangente (It059)	Italy
31	Ogliarola Bradano (It130)	Italy	71	Manzanilla de Sevilla (Es037)	Spain
32	Ogliarola vulture (It129)	Italy	72	Humaissi (Syr003)	Syria
33	Olivo di Mandanese (It131)	Italy	73	Aggizi Sham (Egy001)	Egypt
34	Ottobratica (It001)	Italy	74	Koroneiki (Gr010)	Greece
35	Sinopolese (It017)	Italy	75	Mastoidis (Gr013)	Greece
36	Bissani (Li050)	Lebanon	76	Leccino (It015)	Italy
37	Bouchouika (Mac024)	Morocco	77	Picholine marocaine (Mac002)	Morocco
38	Meslala (Mac004)	Morocco	78	Galega vulgar (Por006)	Portugal
39	VS3 (Mac037)	Morocco	79	Picual (Es054)	Spain
40	ZDH 6 (Mac015)	Morocco	80	Zaiti (Syr013)	Syria

Figures

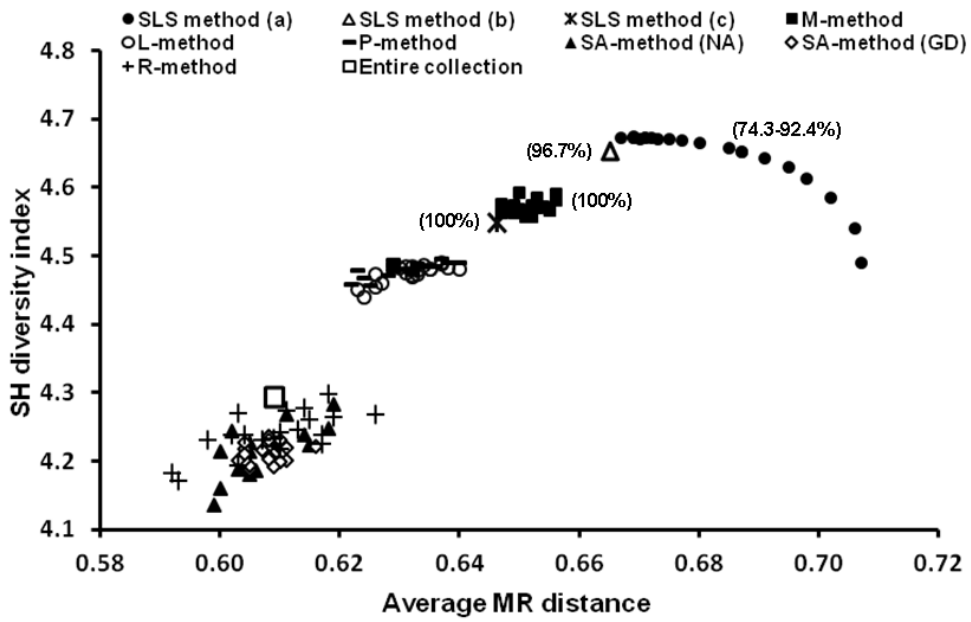

Fig. 1. Plot of values of average MR distance and SH diversity index of all the core subsets of each of the different sampling methods and the entire collection. (a) Twenty independent core sets when optimizing both the average MR genetic distance and SH diversity index according to weight given to each parameter in the range of [0-1] (double), (b) when optimizing the four parameters (CV, MR, SH and HE) simultaneously (multiple), (c) when optimizing only CV (single). Values in brackets indicate the coverage of alleles.

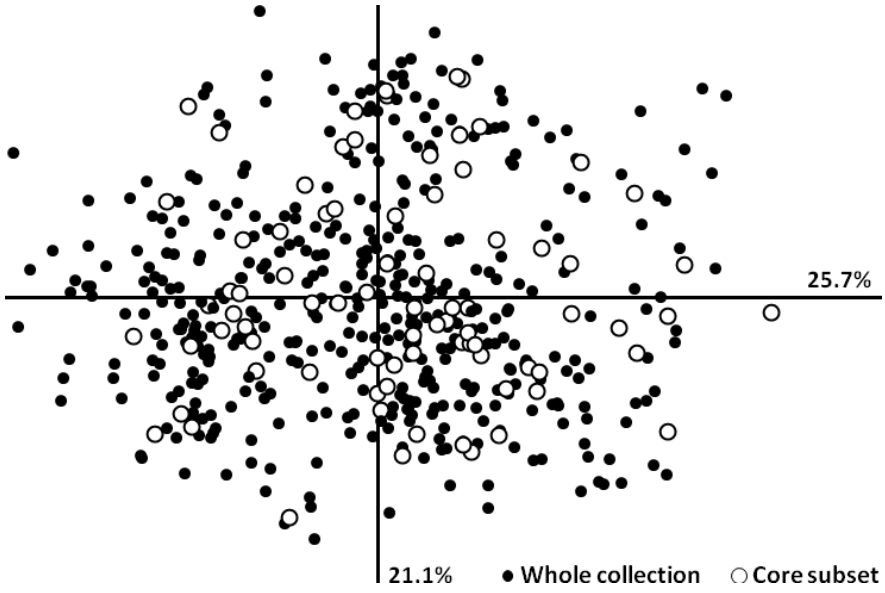

Fig. 2. Spatial distribution of the accessions selected as core subset and those of the entire collection using the main two principal coordinates (PCoA). The percentage of the variation explained by each axis is indicated.