Statistical Approaches to Analyse Gene Bank Data Using a Lentil Germplasm Collection as a Case Study

Gaetano LAGHETTI (✉)
Domenico PIGNONE
Gabriella SONNANTE

Summary

Normally in a plant gene bank a large number of accessions per each crop and/or taxon is stored. During their characterization and preliminary evaluation, several quantitative and qualitative data are recorded and, usually, a wide intra accession variation is observed. The management of all this information becomes very difficult without effective statistical methods combining these different types of data. At the Institute of Plant Genetics, CNR, in Bari (Italy) this problem has been tackled by testing many statistical approaches. The present contribution describes one of these approaches, which to date has proven to be highly adequate; a case study describing a lentil germplasm collection has been used for demonstration. A valuable application of this method is the determination of core subsets important to increase the utilization and accessibility of plant genetic resources.

In the presented case study a subset of the lentil germplasm collection was chosen to perform molecular analysis based on ISSR markers. The samples were selected on the basis of both morpho-agronomic evaluation and geographical origin. These markers proved to be useful for distinguishing among closely related genotypes and for possibly substantiating the genetic peculiarity of some interesting material.

Key words

statistical method, ISSR markers, germplasm collection, lentil

Institute of Plant Genetics, Via Amendola 165/A, 70126 Bari, Italy
✉ e-mail: gaetano.laghetti@igv.cnr.it

Received: October 31, 2006 | Accepted: December 13, 2007

Introduction

Usually seed gene banks store a large number of accessions per each crop/taxon *ex situ*. During the characterization process of this material several quantitative and qualitative data are recorded. Usually, a wide variation is recorded at the intra accession level in addition to inter-accession one. The management of all this information becomes very difficult without effective statistical tools able to combine different types of data of this sort. At the Institute of Plant Genetics (IGV), National Research Council (Bari, Italy) this problem has been addressed by testing many statistical approaches (Laghetti et al., 1990; Perrino et al., 1984; Polignano et al., 2001). However they were old methods that studied separately the quantitative and qualitative data and, in addition, did not consider the intra accession variability using an average datum. This paper describes one approach, never used before at IGV, which proved to be highly adequate; a case study describing a lentil germplasm collection has been used for demonstration. This method may not only disclose the overall level of variation present in a collection, but also describe how variation is distributed in the collection. A further valuable application of this method is the determination of core subsets which play an important role in increasing the accessibility and utilization of crop genetic resources, and in improving the general management of a crop collection (Brown, 1989).

Generally, characterization and preliminary evaluation data are based on agronomic traits linked to yield performance, but they often give little information on the actual genetic constitution of the examined material. Conversely, molecular markers precisely define the genetic constitution of a sample, but give no information on yield attitude. This is particularly true in some crops like lentil (*Lens culinaris* Medik), in which it is reported by several authors that genetic variation as examined at a molecular level, does not go along with the level of variation assessed at the morpho-productive level. In fact, domestication pressure in lentil has fixed few Mendelian characters, e.g. absence of dormancy or pod shattering, and few quantitative traits, like seed size (Fuller, 2007). These characters account for a small proportion of the genome that is often not associated to molecular markers, most of which are therefore evolutionary neutral (Hammer, 1984; Grandillo et al., 1999). For this reason lentil was selected as a case study and a subset of the lentil collection was analysed also using molecular markers. The present contribution reports on the results of this study.

Materials and methods

In the present case study 133 accessions of lentil stored at the IGV (Table 1) were scored for the set of characters listed in Table 2. These characters were selected on the basis of IBPGR (1985) descriptors. Furthermore, a subset of this germplasm was selected to perform molecular analysis based on ISSR markers. The samples were chosen on the basis of both morpho-agronomic evaluation and geographical origin.

Table 1. Accessions of the lentil germplasm collection stored in the IGV's gene bank used in this case study

Origin	Subspecies		
	Macrosperma	Microsperma	Total
Albania	–	3	3
Algeria	6	4	10
Cyprus	7	3	10
Egypt	–	10	10
Ethiopia	–	8	8
Greece	1	3	4
Iran	–	2	2
Italy	23	18	41
Libya	2	1	3
Morocco	6	4	10
Nepal	–	2	2
Pakistan	–	10	10
Spain	4	2	6
South Africa	–	1	1
Tunisia	6	7	13
Total	55	78	133

For ISSR analysis, 46 accessions (22 Macrosperma type and 24 Microsperma type) were analysed, 31 of which originated in Italy, the remaining mostly other Mediterranean countries. Six ISSR primers were chosen for DNA amplification: (AG)8YG, (CA)8RY, (AC)8YA, (GA)8YT, (GT)8YC, and BDB(CA)7. Amplificates were visualized on pre-cast polyacrilamide gels stained using silver staining. A total of 74 reliable bands were scored, 65% of which were polymorphic. A similarity matrix based on Jaccard's index was obtained, from which a UPGMA dendrogram was generated (Sonnante and Pignone, 2007).

The modified[1] statistical method by Cole-Rodgers et al. (1997) was used combining both qualitative and quantitative traits (see the example in Table 3) to calculate dissimilarity scores between each two landraces:

[1] A potential limitation to the original Cole-Rodgers et al.'s method is that it does not distinguish between variation in the proportions (evenness) of a subtrait among accessions (e.g. 80% violet and 20% pink 'flower ground colour' in one accession and a different proportion in another accession); this situation is acceptable from a plant breeding perspective but not for a curator and/or a researcher characterizing a germplasm collection, so that we modified the original algoritm adding some numerical coefficients taking into account these cases (mathematical details in a paper in preparation for a biometry journal).

Table 2. Traits used for the morphological and agronomic characterisation from 'Lentil Descriptors' by IBPGR (1985)

Trait	Type of item	Subtrait/Score
seedling stem pigmentation	qualitative	absent (0), present (1)
leaf pubescence	qualitative	absent (0), slight (3), dense (7)
leaflet size	qualitative	small (3), medium (5), large (7)
plant height	quantitative	cm
tendril length	qualitative	absent (0), rudimentary (1), prominent (2)
time to flowering	quantitative	days from sowing
time to maturity	quantitative	days from sowing
flower ground colour	qualitative	white (1), white with blue veins (2), blue (3), violet (4), pink (5), other (6)
pod pigmentation	qualitative	absent (0), present (1)
number of seeds per pod	quantitative	no.
100 seed weight	quantitative	g
ground colour of testa	qualitative	green (1), grey (2), brown (3), black (4), pink (5)
pattern of testa	qualitative	absent (0), dotted (1), spotted (2), marbled (3), complex (4)
colour of pattern on testa	qualitative	absent (0), olive (1), grey (3), brown (4), black (5)
cotyledon colour	qualitative	yellow (1), orange/red (2), olive–green (3)
lodging susceptibility	qualitative	none (0), low (3), medium (5), high (7)
number of flowers per peduncle	quantitative	no.
height of lowest pod	quantitative	cm
seed yield	quantitative	g/m^2
pod shedding	qualitative	none (0), low (3), medium (5), high (7)
pod dehiscence	qualitative	none (0), low (3), medium (5), high (7)
harvest index	quantitative	%

Table 3. Five example lentil accessions and three of their traits

Accession – Code	Flower ground colour*	Height of lowest pod (cm) **	Cotyledon colour
MG106699 – 124	1	12.0	1,2
MG107189 – 121	1,2	10.2	3
MG112118 – 105	1,2,4	12.4	1,3
MG112164 – 107	1	20.7	1
MG116052 – 32	2	9.6	3

* the codes 1, 2 etc. indicate the subtrait (reported in Table 2) observed in the accession; ** average value recorded in the accession

Table 4. 'Flower ground colour' subtrait scores for five lentil accessions

Accession – Code	'Flower ground colour' subtraits					
	(1)*	(2)	(3)	(4)	(5)	(6)
MG106699 – 124	1/√6	0	0	0	0	0
MG107189 – 121	1/√6	1/√6	0	0	0	0
MG112118 – 105	1/√6	1/√6	0	1/√6	0	0
MG112164 – 107	1/√6	0	0	0	0	0
MG116052 – 32	0	1/√6	0	0	0	0

*white (1), white with blue veins (2), blue (3), violet (4), pink (5), other (6)

First step of method, qualitative traits:

Table 4 illustrates the 'flower ground colour' subtrait values for five lentil accessions. The dissimilarity for 'flower ground colour' between e.g. the accessions MG107189 and MG112118 is given by:

$(1/\sqrt{6} - 1/\sqrt{6})^2 + (1/\sqrt{6} - 1/\sqrt{6})^2 + (0 - 0)^2 + (0 - 1/\sqrt{6})^2 + (0 - 0)^2 + (0 - 0)^2 = 0 + 0 + 0 + 1/6 + 0 + 0 = \underline{0.17}$

The values for the other 12 qualitative traits are obtained from similar calculations and then they are added up all. In the case of the two example accessions MG107189 and MG112118 their dissimilarity value for the 13 qualitative traits is 2.68 [the theoretical range is: 0 (the two accessions are exactly alike) - 13 (the two accessions are entirely different)].

Second step of method, quantitative traits:

each quantitative trait recorded has an outer limit value depending of the specific data set considered; as an example the trait 'height of lowest pod' (HLP) has Min(HLP) = 9.6 (MG116052) and Max(HLP) = 20.7 (MG112164) cm. Let the difference between these two be termed dif(HLP) = Max(HLP) - Min(HLP) = 20.7 - 9.6 = 11.1

A calculation of HLP score (HLPscr) is needed to determine the distance between the two example accessions selected (MG107189 and MG112118): this is obtained first by subtracting Min(HLP) from the value of HLP for a

specific accession and this difference is then divided by dif(HLP):

$HLPscr_{MG107189} = [HLP_{MG107189} - Min(HLP)]/ dif(HLP) = [10.2 - 9.6]/ 11.1 = 0.05$

$HLPscr_{MG112118} = [HLP_{MG112118} - Min(HLP)]/ dif(HLP) = [12.4 - 9.6]/ 11.1 = 0.25$

The HLP dissimilarity between two accessions is the square of the difference between their HLPscr:

$HLP\ dissimilarity\ value_{MG107189/MG112118} = (0.05 - 0.25)^2 = \underline{0.04}$

The values for the other eight quantitative traits were obtained from similar calculations and then they are added up all. In the case of the two example accessions MG107189 and MG112118 their dissimilarity value for the nine quantitative traits is <u>0.97</u>

<u>Third step of method</u>, to incorporate qualitative and quantitative traits:

by summing the dissimilarity values for each individual trait between two accessions, a total dissimilarity for all measured traits, both qualitative and quantitative, can be calculated:

$total\ dissimilarity\ value_{MG107189/MG112118} = 2.68 + 0.97 = \underline{3.67}$

With 22 traits (13 qualitative and nine quantitative), two accessions that are exactly alike will have a total dissimilarity value of zero. If the two are entirely different, the total dissimilarity value will be 22.

<u>Fourth step of method</u>, matrix of dissimilarities:

with the method described above, a 'total dissimilarity value' was calculated for each pair of the 115 lentil accessions forming a matrix of dissimilarities (not shown).

<u>Fifth step of method</u>, cluster analysis:

after generating the matrix of dissimilarities, using these values, a standard cluster analysis can be used to group the accessions. To obtain the cluster of Fig. 1 the 'Cluster' procedure from the SAS® 9.1 statistical package (SAS® 2004) was adopted.

Results and discussion

The dendrogram shown in Fig. 1 is relative to dissimilarity matrix by modified Cole-Rodgers et al. (1997), based on 22 morpho-agronomical descriptors. A first clear separation appears between accessions belonging to the two morphogroups. Within the Macrosperma group a further division is present between 14 accessions from Italy and 36 from other Mediterranean countries except for three Italian ones as already occurred in a previous study (Laghetti et al., 2005). Inside the Microsperma group two subgroups are identified: one of 23 not Mediterranean lentils and one with 53 of Mediterranean origin. These two subgroups in their turn can be split in two, according to their high (> y) or low (< y) yield performance.

At level of 10 (Dleg./Dmax)*100 (see in Polignano et al., 2001 a review on the criteria to choose the optimum number of clusters) it is possible to cut the dendrogram in 12 clusters (that we can call 'core subsets') with average similarity values statistically different ($P< 0.01$). These 12 'core subsets' (formed by 1 to 42 accessions) are very homogenous on their inside but have specific morpho-agronomical characteristics.

As for ISSR analysis, according to the similarity matrix obtained, the highest likeness was observed between accessions n. 21 and n. 22 (Jsi=0.982), from two small Sicily Channel islands, Pantelleria and Lampedusa respectively. In addition these accessions were clustered together with other accessions collected from two other small islands near Sicily, Linosa and Ustica. Clustering based on the Jaccard's index showed also other patterns of similarity based on geographic distance. Samples from more-or-less neighbouring areas tend to cluster under the same node. Overall, with only few exceptions from Sicily, the Italian material is quite differentiated from the remaining Mediterranean samples, independent of the seed size character (Sonnante and Pignone, 2007).

It is interesting to notice that the Italian material shares a fairly large amount of similarity with respect to the samples from the rest of the Mediterranean. This might be related to the isolation due to reduction of food trade over the sea during the Middle Ages as a consequence of the expansion of the Arabs in the Mediterranean. The Arabs dominated the Southern part of the Mediterranean for a few centuries starting 7th century AD, and possibly helped the homogenization of lentil germplasm from North Africa and part of Sicily, which was under their domination (Barone and Caruso, 1996). The most differentiated accession in the examined set was N. 45 from Ethiopia, a country reported to be a secondary centre of diversity for many crops (Polignano and Sonnante, 1992; Alemayehu and Parlevliet, 1997).

The comparison of the two dendrograms obtained from morpho-agronomic (Fig. 1) and molecular (Fig. 2) data is not straightforward, since it is evident that in the former one the weight of seed size is predominant over other characters: as a matter of facts the two main clusters separate the Macrosperma types from the Microsperma ones. This trait has also an influence on plant vigour. In the tree based on ISSR markers, on the contrary, the geographical origin of the material has a much stronger influence and samples belonging to both seed morpho-types are interspersed. This occurrence confirms that the selec-

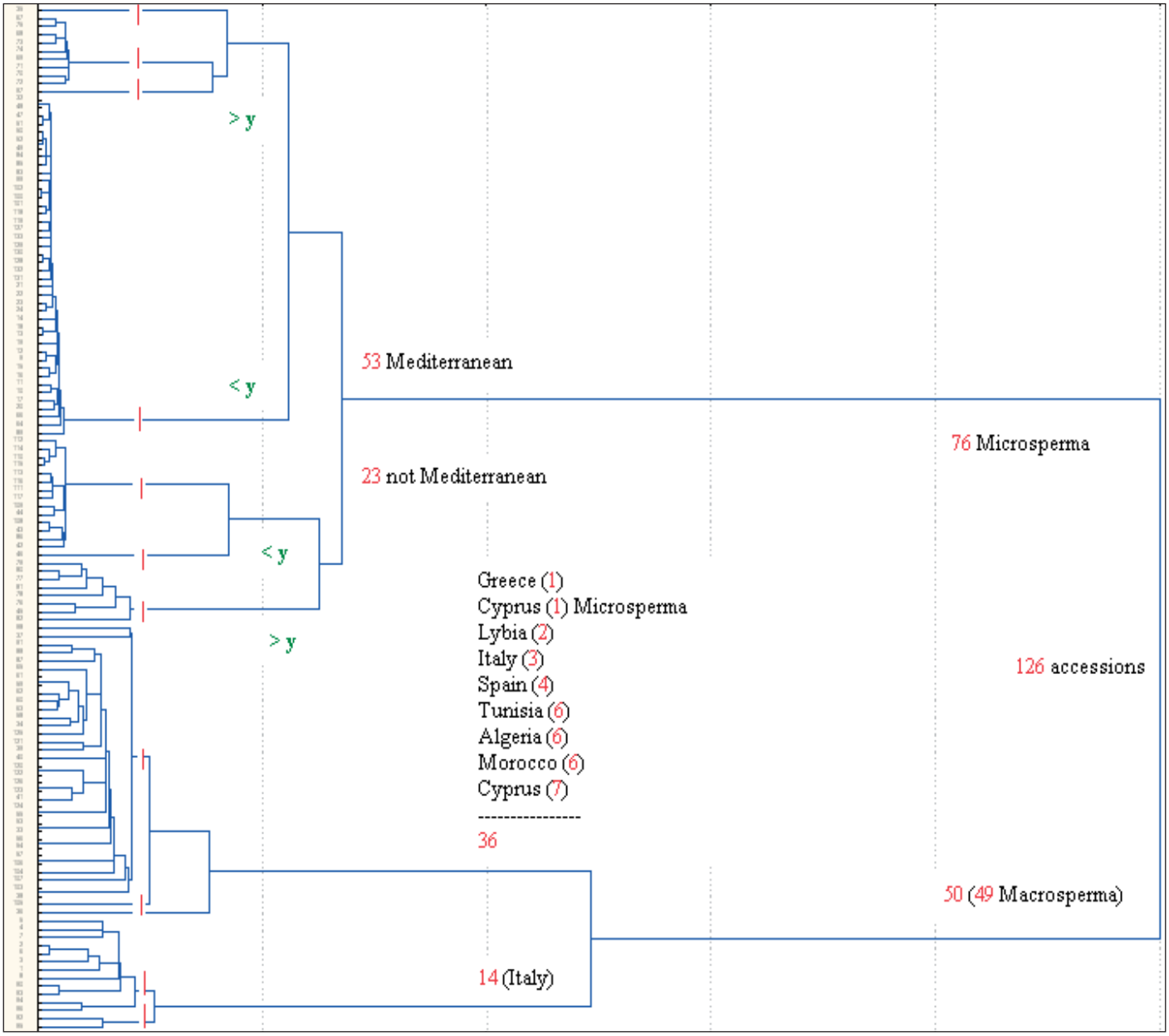

Figure 1.
Dendrogram relative to dissimilarity matrix by Cole-Rogers et al. (1997) modified, based on 22 morpho-agronomical descriptors

tive pressure that consequent the domestication process has actually acted on few isolated areas of the genome, the so called "domestication islands" (Papa et al., 2007), leaving the great part of it free of human pressure. The only selective pressure on the genome regions outside the domestication islands was due to the spreading of this crop outside its area of origin, implying mostly adaptive forces, and therefore these regions of the genome contain markers that are mostly neutral.

Nevertheless, it is interesting to notice that there are some congruent aspects between the two dendrograms. For example, most Italian accessions are clustered under specific nodes in both the Macrosperma and Microsperma clusters of the dendrogram of Figure 1. This might be due to the fact that a proportion of the molecular traits set is in fact associated to morphological characters, as also molecular maps demonstrate (Hamwieh et al., 2005), thus providing useful markers for Marker Assisted Selection.

Table 5. Code of the accessions reported in Fig. 1 with their geographical origin and subspecies

Code	Type*	Origin	Code	Type	Origin	Code	Type	Origin	Code	Type	Origin
1	M	Italy	35	m	Egypt	69	m	Egypt	103	M	Morocco
2	M	Italy	36	M	Greece	70	m	Egypt	104	M	Morocco
3	M	Italy	37	M	Lybia	71	m	Egypt	105	M	Morocco
4	M	Italy	38	M	Morocco	72	m	Egypt	106	M	Morocco
5	M	Italy	39	M	Spain	73	m	Egypt	107	M	Morocco
6	M	Italy	40	M	Spain	74	m	Egypt	108	m	Nepal
7	M	Italy	41	M	Tunisia	75	m	Egypt	109	m	Pakistan
8	M	Italy	42	m	Iran	76	m	Ethiopia	110	m	Pakistan
9	m	Italy	43	m	Nepal	77	m	Ethiopia	111	m	Pakistan
10	m	Italy	44	m	Pakistan	78	m	Ethiopia	112	m	Pakistan
11	m	Italy	45	m	Ethiopia	79	m	Ethiopia	113	m	Pakistan
12	m	Italy	46	m	South Africa	80	m	Ethiopia	114	m	Pakistan
13	m	Italy	47	m	Albania	81	m	Ethiopia	115	m	Pakistan
14	m	Italy	48	m	Albania	82	m	Ethiopia	116	m	Pakistan
15	m	Italy	49	m	Algeria	83	m	Greece	117	m	Pakistan
16	m	Italy	50	m	Algeria	84	m	Greece	118	m	Spain
17	m	Italy	51	m	Algeria	85	m	Greece	119	m	Spain
18	m	Italy	52	m	Algeria	86	m	Iran	120	M	Spain
19	m	Italy	53	M	Algeria	87	M	Italy	121	M	Spain
20	m	Italy	54	M	Algeria	88	M	Italy	122	M	Tunisia
21	m	Italy	55	M	Algeria	89	m	Italy	123	M	Tunisia
22	m	Italy	56	M	Algeria	90	M	Italy	124	M	Tunisia
23	m	Italy	57	M	Algeria	91	M	Italy	125	M	Tunisia
24	m	Italy	58	M	Cyprus	92	M	Italy	126	M	Tunisia
25	M	Italy	59	M	Cyprus	93	M	Italy	127	m	Tunisia
26	M	Italy	60	M	Cyprus	94	M	Italy	128	m	Tunisia
27	M	Italy	61	M	Cyprus	95	M	Italy	129	m	Tunisia
28	m	Italy	62	M	Cyprus	96	M	Italy	130	m	Tunisia
29	M	Italy	63	M	Cyprus	97	m	Lybia	131	m	Tunisia
30	M	Italy	64	m	Cyprus	98	M	Lybia	132	m	Tunisia
31	M	Italy	65	m	Cyprus	99	m	Morocco	133	m	Tunisia
32	m	Albania	66	m	Cyprus	100	m	Morocco			
33	M	Algeria	67	m	Egypt	101	m	Morocco			
34	M	Cyprus	68	m	Egypt	102	m	Morocco			

*M = Macrosperma, m = Microsperma

Conclusions

A worthy application of this method is the determination of 'core subsets' important to increase the utilization and accessibility of plant genetic resources. This information is also economically very important for the management of a germplasm collection, since the number of accessions to be grown can be limited thus reducing the high costs of seed storage, characterization and increase. By choosing, for instance, only one accession from each 'core subset', it is possible to set up a working 'core collection' still conserving most of the genetic diversity as based on phenotypic evaluation. In addition, because this core collection is considerably reduced in size as compared to the whole collection, future screenings for traits such as disease resistance could be facilitated. Therefore, the application of a statistical method able to cluster similar accessions in a reliable way can be the starting point for further analyses on the germplasm available in genebanks. Molecular markers proved to further increase the sensibility of the analysis providing a tool for distinguishing among closely related genotypes and for possibly substantiating the genetic peculiarity of some interesting material.

References

Alemayehu, F. and Parlevliet J.E., 1997. Variation between and within Ethiopian barley landraces. Euphytica 94:183–189.

Barone, E. and Caruso T., 1996. Genetic diversity within *Pistacia vera* in Italy. *In*: S. Padulosi, T. Caruso & E. Barone (Eds.), Taxonomy, Distribution, Conservation and Uses of *Pistacia* genetic resources, pp. 20–28. IPGRI, Rome, Italy.

Brown, A.H.D., 1989. The case for core collections. *In*: The use of plant genetic resources. Brown A.H.D. et al. (ed.), Cambridge Univ. Press, Cambridge, England, pp. 136-156.

Cole-Rodgers, P., Smith D.W. and Bosland P.W., 1997. A novel statistical approach to analyze genetic resource evaluations using *Capsicum* as an example. *Crop Sci.* 37:1000-1002.

Fuller, D.Q., 2007. Contrasting patterns in crop domestication rates: recent archaeobotanical insights from the Old World. Ann Bot 100: 903-924.

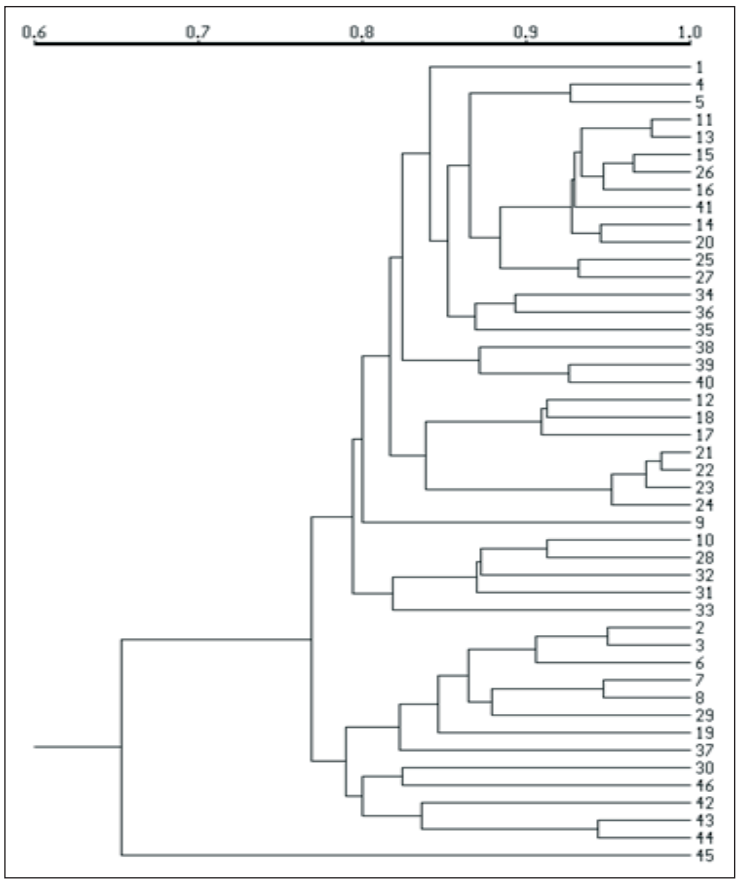

Figure 2.
Dendrogram based on ISSR data analysis using Jaccard's similarity index

Grandillo, S., Ku H.M. and Tanksley S.D., 1999. Identifying the loci responsible for natural variation in fruit size and shape in tomato. *Theor. Appl. Genet.* 99: 978-987.

Hammer, K. 1984. The domestication syndrome (Germ., Engl. summary). *Kulturpflanze* 32:11-34.

Hamwieh, A., Udupa S. M., Choumane W., Sarker A., Dreyer F., Jung C. and Baum M. 2005. A genetic linkage map of *Lens* sp. based on microsatellite and AFLP markers and the localization of fusarium vascular wilt resistance. Theor Appl Genet 110: 669–677.

IBPGR, 1985. Lentil Descriptors, Secretariat, Rome.

Laghetti, G., Padulosi S., Hammer K., Cifarelli S. and Perrino P., 1990. Cowpea (*Vigna unguiculata* [L.] Walp.) germplasm collection in southern Italy and preliminary evaluation. *In*: Cowpea Genetic Resources. Ng, N.Q. and Monti, L.M. (eds.), Thailand: Amarin Printing Group Co. pp.46-57, ISBN 978 131 0537.

Laghetti, G., Volpe N., Sonnante Gi., Pignone D. and Sonnante Ga., 2005. Salvaguardia, caratterizzazione e valorizzazione dell'antico agro-ecotipo pugliese 'Lenticchia di Altamura'. *VII Convegno Nazionale sulla Biodiversità*, Catania, 31 March – 2 April 2005, p.148.

Papa, R., Bellucci E., Rossi M., Leonardi S., Rau D., Gepts P., Nanni L. and Attene G. 2007. Tagging the signatures of domestication in common bean (*Phaseolus vulgaris*) by means of pooled DNA samples. Annals of Botany 100: 1039-1051.

Perrino, P., Yarwood M. and Hanelt P., 1984. Variation of seed characters in selected *Vicia* species. Kulturpflanze 32:103-122.

Polignano, G.B. and Sonnante G., 1992. Characterization of faba bean populations by isoenzyme patterns of GOT and PGI. FABIS Newsletter 31:12–16.

Polignano, G.B., Uggenti P. and Scippa G., 2001. Diversity analysis and core collection formation in Bari Faba Bean germplasm. Plant Genetic Resources Newsletter 125:33-38.

SAS®, 2004. Base SAS 9.1 Procedures Guide, Volumes 1-4 (Print on Demand). SAS Institute Inc., SAS Publishing

Sonnante, G. and Pignone D., 2007. The major Italian landraces of lentil: their molecular diversity, similarity and possible origin. *Genet. Resour. Crop Evol.*, 54: 1023-1031.

acs73_28