Sabato, D.; Esteras Gómez, C.; Grillo, O.; Picó Sirvent, MB.; Bacchetta, G. (2015). Seeds morpho-colourimetric analysis as complementary method to molecular characterization of melon diversity. Scientia Horticulturae. 192:441-452. doi:10.1016/j.scienta.2015.06.006.

Original paper

Seeds morpho-colourimetric analysis as complementary method to molecular characterization of melon diversity

Diego Sabato[a], Cristina Esteras[b], Oscar Grillo[a-c], Belén Picó[b,*], Gianluigi Bacchetta[a]

[a] Centro Conservazione Biodiversità (CCB), Dipartimento di Scienze della Vita e dell'Ambiente, Università degli Studi di Cagliari, v.le Sant'Ignazio da Laconi 11-13, 09123 Cagliari, Italy.

[b] Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politécnica de Valéncia, Camino de Vera s/n, 46022 Valencia, Spain.

[c] Stazione Consorziale Sperimentale di Granicoltura per la Sicilia (SSGS), via Sirio 1, 95041 Borgo Santo Pietro, Caltagirone, Italy.

*Corresponding author: mpicosi@btc.upv.es, +34)963877000 Ext:79415

Abstract

Melon has undergone an intense process of selection and crossbreeding, resulting in many landraces distributed all over Europe, Africa and Asia. Due to this huge variability, the systematic position of this *taxon* has been reviewed many times in the last two decades. The goal of this article is to compare the phenotypic characterization achieved by seed features with the molecular analysis on melon genotypes. A set of 124 accessions of *Cucumis melo* has been selected for molecular and morpho-colourimetric analyses plus an additional selection of accessions of *Cucumis sativus*, *Citrullus lanatus* and *Citrullus colocynthis* used to highlight seed morphology distances among

genus and species. Genotyping was performed on the basis of 211 polymorphic SNPs and was executed using the iPLEX® Gold MassARRAY Sequenom technology. A total of 137 parameters were specifically designed to evaluate seed colour, size, shape and texture. Both molecular and seed morpho-colourimetrical analyses confirm the existence of two melon subspecies while an intermediate group has also been found. A non random allelic distribution in SNPs located in specific genomic regions suggests that some of these regions may account for a part of the observed variation in seed size. Six major groups of varieties can be discriminated on the basis on seed traits.

Keywords

Cucurbitaceae; *Cucumis melo*; genetic characterization; old landraces; seed image analysis; wild crop relatives.

1. Introduction

The Cucurbitaceae family includes about 130 genera and 800 *taxa* (Jeffrey, 2005; Jeffrey and De Wilde, 2006). Among them, the most economically important species are *Cucumis melo* L. (melon), *Cucumis sativus* L. (cucumber), *Citrullus lanatus* (Thunb.) Matsum & Nakai (watermelon) and *Cucurbita* L. spp. (gourds and squashes).

Melon is worldwide diffused and comprises wild, feral and cultivated varieties, including sweet melons used for dessert and non-sweet ones consumed raw, pickled or cooked (Kirkbride, 1993; Bates and Robinson, 1995). Africa has been traditionally thought to be the centre of origin of this

51 species. However, due to the high level of variation found in Asia, especially
52 in India, melon could have originated there and then reached Africa (Renner
53 et al., 2007; Sebastian et al., 2010). Other theories suggest that two
54 independent domestications took place (Jeffrey, 1980; Esquinas-Alcazar and
55 Guilick, 1983; Mallick and Mausi, 1986; Bates and Robinson, 1995).

56 In the Mediterranean area the presence of melon is recorded in Egypt
57 since the third millennium BC (Zohary et al., 2012), and in Greece and Italy
58 since at least the Late Bronze Age (Megaloudi, 2003; Sabato et al., 2015).
59 First representations show fruits likely belonging to the non-sweet melon
60 varieties *chate* and *flexuosus* (with long cucumber-like fruits) (Janick et al.,
61 2007). The presence of round sweet melons in the Mediterranean basin till the
62 Classical Age is uncertain, but it is well proven since the 11th century AD by
63 Arabian trade with Central Asia (Paris et al., 2012).

64 *C. melo* has been traditionally separated into two subspecies, *melo* and
65 *agrestis* (Naudin 1859), each one including different varieties (Munger and
66 Robinson, 1991). Pitrat et al. (2000) recognized 16 varieties: *cantalupensis*
67 Naudin, *reticulatus* Ser. (cantaloupes, muskmelons), *inodorus* H.Jac. (winter
68 melons, casaba melons), *flexuosus* L. (snake melons), *chate* Hasselq.
69 (cucumber melons), *adana* Pangalo*, chandalak* Gabaev*, ameri* Pangalo
70 (Asian melons)*, chito* C.Morren (American melons)*, dudaim* L. (pocket
71 melons), and *tibish* Mohamed within the subsp. *melo* L. and *acidulus* Naudin,
72 *conomon* Thunb., *makuwa* Makino and *chinensis* Pangalo (pickling melons),
73 and *momordica* Roxb. (snap melons) within subsp. *agrestis* Naudin. In later
74 revisions, Pitrat (2008) merged some varieties and Esteras et al. (2009,
75 2013), after further molecular studies, moved *tibish* and *chito* into the

subspecies *agrestis*. Some of these varieties are quite heterogeneous, and accessions displaying intermediate features are difficult to classify. The wild forms of melon, usually referred to as *C. melo* subsp. *agrestis* var. *agrestis*, are mainly distributed in North and Eastern Africa and in Asia, while free-living forms of small size fruited melons have been found in Northern Australia, Southern USA and Central America (Roy et al., 2012). The sweet *cantalupensis, reticulatus* and *inodorus* melons are the ones with the most commercial interest worldwide (Pitrat, 2008).

In order to establish the genetic relationships among subspecies and varieties, several molecular studies have been carried out in melon, employing different markers (reviewed in Esteras et al., 2012). Most of them support the division at the sub-specific level and have contributed to better reclassify some of the varieties (Stepansky et al., 1999; Deleu et al., 2009; Esteras et al., 2009; Esteras et al., 2013). SNPs (Single Nucleotide Polymorphisms) are high-quality markers mostly used for genome-wide surveys in high to medium-throughput genotyping platforms (Fan et al., 2006; Steermers and Gunderson 2007; Gabriel et al., 2009). The number of SNPs available in melon has largely increased in the last few years (Blanca et al. 2011, 2012; http://melogene.net/; Garcia-Mas et al., 2012). Esteras et al. (2013) reported the first application of a GoldenGate genotyping platform to analyse a melon core collection with SNPs distributed throughout the genome, demonstrating their usefulness for genetic diversity and population structure studies.

Also phenotypic variability has been studied with different core collections (Stepansky et al., 1999; Esteras et al., 2009; Leida et al., 2015),

and with germplasm from specific centres of diversity (reviewed in: Esteras et al., 2012; Raghami et al., 2014). These phenotyping assays have basically focused on fruit traits and on the responses to biotic and abiotic stress, and many QTLs controlling these traits have been mapped in the melon genome (Diaz et al., 2011). In contrast to other species for which extensive efforts have been made in mapping QTLs for seed properties (Cai et al., 2012), and even in cloning the underlying genes (Orsi and Tanksley, 2009), only some studies have included seed traits in melons. Some of them report a significant correlation between seed traits and botanical classification (Stepansky et al., 1999; Yashiro et al., 2005; Tanaka et al., 2007). Fujishita and Nakagawa (1973) pointed out that seed size is one important trait for the identification of melon varieties. Fujishita (1980) described *makuwa* and *conomon* varieties with seeds smaller than 9 mm, *reticulatus* with seeds larger than 9 mm and *momordica* with intermediate seeds. Also in a recent study Tanaka et al. (2013) associated the variation in seed length and weight to chloroplast genome variation. However, seed traits have not been extensively analysed in large collections, representing the whole diversity of the species. In other cucurbits like *Cucurbita pepo*, seed traits have been used as discriminating factors since early studies (Decker and Newsom, 1988), and correlation between seed and fruit traits has been reported (Paris and Nerson, 2003).

Since the inception of the taxonomy, hierarchical classifications have been constructed on the basis of morphology, and molecular analyses can support those parts of phylogeny for which morphological data is lacking (Scotland et al., 2003). Morpho-colourimetric evaluations are commonly employed as tools to assess shape, size and colour of objects (Bacchetta et

126 al., 2008, Venora et al., 2009; Grillo et al., 2010). Several works about the
127 application of image analysis to the diaspores of wild vascular flora have been
128 carried out, providing excellent results of classification within taxonomic units
129 close to infra-generic and infra-specific levels (Bacchetta et al., 2011a, 2011b;
130 Grillo et al., 2012; Pinna et al., 2014). Many studies have been focused also
131 on crop wild relatives and landraces (Venora et al., 2007a, Venora et al.,
132 2007b; Smykalova et al., 2011; Smykalova et al., 2013), and recently many
133 authors focused on the *Vitis vinifera* complex (Rivera et al., 2007; Terral et al.,
134 2010; Orrú et al., 2013a, 2013b).

135 The knowledge of the existing diversity in melon is important, not only
136 for its conservation, but also for its exploitation in commercial breeding, as this
137 species displays crossability problems with other species of the genus
138 *Cucumis*.

139 The goals of this research are to:

140 - compare the groups established using molecular analyses with those
141 achieved by seed characters;
142 - analyse the variability of morpho-colourimetric seed features;
143 - implement statistical classifiers able to discriminate among the studied
144 varieties;
145 - increase the knowledge about the variation of the current extant melon
146 seed collections.

147 **2. METHODS**

148 ***2.1 Seed lots***

149 The whole melon collection was established on the framework of a
150 previous project (MELRIP 2007-2010, Esteras et al., 2009, 2013; Leida et al.,

2015). Accessions were characterized for vine and fruit traits, multiplied and conserved at the COMAV Genebank (Institute for the Conservation and Breeding of the Agrobiodiversity). We selected 124 seed lots (103 for both molecular and seed analyses plus 21 only for morpho-colourimetric analysis of the seeds). Seeds belonged to accessions from 48 countries and represented all melon varieties. Fruits were collected at the optimum maturity stage, corresponding to the complete morphologic and chromatic seed development. To avoid over-representation of single plants and/or fruit features, seeds from the highest number of plants and fruits available for each accession were analysed. Undeveloped, deformed and sterile seeds were excluded. For further details about the composition of the analysed collection see Supplementary data.

With the purpose to evidence morphological distances at the genus and species levels, a small set of close relatives of melon were used: Twenty-one accessions of *Cucumis sativus*, 18 of *Citrullus lanatus* and 9 of *Citrullus colocynthis*, were selected (see Supplementary data). All accessions were supplied by COMAV Genebank and represent mainly Mediterranean, African and Asian landraces.

2.2 Molecular analysis

The DNA was extracted from young leaves using the CTAB method with minor changes (Esteras et al., 2013). Genotyping was done with a total of 211 polymorphic SNPs, evenly distributed throughout the genome, that were selected from the SNP melon collection available in the Melogene database (http://www.melogene.net/) and *in silico* identified in two previous re-sequencing analysis (Blanca et al., 2011, 2012). Genotyping was performed

using the iPLEX® Gold MassARRAY Sequenom technology at the Epigenetic and Genotyping Unit of the University of Valencia (Unitat Central d´Investigació en Medicina UCIM). The basis of this technology is described in Gabriel et al. (2009).

The genotyping results were employed to perform a cluster analysis using the PowerMarker software (Liu and Muse, 2005). Nei's genetic distance (Nei et al., 1983) was used, and the support values for the degree of confidence at the nodes of the dendrogram were analysed by bootstrap re-sampling 1,000 times. Phylip 3.69 software (Felsenstein, 1997) was employed to construct the consensus tree and TreeView32 (Page, 1996) to visualize it. Genotyping summary statistics such as the number of alleles, the frequency of the most common allele (MAF) and the polymorphism information content (PIC) for each *locus* is provided in Supplementary data. In addition, a Principal Coordinate Analysis (PCoA) was performed using GenAlEx 6.5 (Peakall and Smouse, 2012).

The genetic structure underlying the genotyped collection, that is the number of populations and the probability of each accession belonging to each inferred population, was previously analysed (Esteras et al., 2013; Leida et al., 2015) using STRUCTURE v2.2 (Pritchard et al., 2000).

Both the database of the melon genome, Melonomics (http://melonomics.net), and the SNP melon collection available in Melogene were used to select and analyse the variation of the melon orthologue of a gene underlying a major QTL associated to seed size in tomato, Seedweight 4.1 (Sw4.1) (Orsi and Tanksley, 2009). This is an ABC transporter

orthologous to the *Arabidopsis* ABC transporter gene At4g39850, also associated with variation in both seed length and width in this model species.

Most of the SNPs used in this study were employed in previous mapping experiments and their position in the genetic map is known (Esteras et al., 2013). This genetic position was used to check the allelic distribution in the germplasm collection of SNPs located in regions of the genetic map in which QTLs for fruit size were previously located (Diaz et al., 2011), and to confirm if differential allelic distributions were also related with differences in the seed traits measured in the present study.

2.3 Seed morpho-colourimetric analysis

Before image acquisition, the scanner was calibrated for colour matching, following the protocol set by Shahin and Symons (2003) as suggested by Venora et al. (2009). Using a flatbed scanner (Epson EU22), two images were acquired for each sample, with black and white background, with a resolution of 400 dpi and 24 bit-depth, in RGB colour model and stored in TIFF format. Sub-samples consisting of 100 seeds were randomly chosen from the original seed lots and arranged on the scanner tray, in such a way that they did not touch each other. When the original accession was numerically lower than 100 units, the analysis was executed on the whole seed lot. All images were analysed with KS-400 release 3.0 image analysis software by Carl Zeiss Vision GmbH (Oberkochen, Germany). A macro, expressly developed for the characterization of cultivated leguminous seeds (Venora et al., 2009), was partially modified to perform automatically all the analysis procedures, reducing the execution time and contextual mistakes in the analysis process.

A total of 20 parameters, specifically designed to evaluate seed colour, were measured together with 17 features descriptive of seed dimensions, 78 shape Elliptic Fourier Descriptors (EFDs) able to define seeds contour shape, and further 22 Haralik's features to assess seed surface texture, for an overall amount of 137 morpho-colourimetric parameters (Table 1).

Table 1

List of characters analysed in morpho-colourimetric analysis

Colour parameters	
Rmean	Red channel mean value of seed pixels (grey levels)
R_SD	Standard Deviation of Red channel value
Gmean	Green channel mean value of seed pixels (grey levels)
G_SD	Standard Deviation of Green channel value
Bmean	Blue channel mean value of seed pixels (grey levels)
B_SD	Standard Deviation of Blue channel value
Hmean	Hue channel mean value of seed pixels (grey levels)
H_SD	Standard Deviation of Hue channel value
Lmean	Lightness channel mean value of seed pixels (grey levels)
L_SD	Standard Deviation of Lightness channel value
Smean	Saturation channel mean value of seed pixels (grey levels)
S_SD	Standard Deviation of Saturation channel value
Dmean	Density channel mean value of seed pixels (grey levels)
D_SD	Standard Deviation of Density channel value
S	Skewness, asymmetry degree of intensity values distribution (grey levels)
K	Kurtosis, peakness degree of intensity values distribution (densitometric units)
H	Energy measure of the increasing intensity power (densitometric units)
E	Entropy Dispersion power (bit)
Dsum	Sum of Density values of the seed pixels (grey levels)
SqDsum	Sum of the Squares of density values (grey levels)
Shape parameters	
A	Area (mm^2)
P	Perimeter (mm)
Pconv	Convex Perimeter (mm)
PCrof	Crofton's Perimeter (calculated using the Crofton's formula) (mm)
Pconv/PCrof	Ratio between convex and Crofton's perimeters
Dmax	Maximum diameter of the seed (mm)
Dmin	Minimum diameter of the seed (mm)
Dmin/Dmax	Ratio between minimum and maximum diameters
Sf	Shape Factor = (4 X π X area)/$Perimeter^2$ (normalized value)
Rf	Roundness Factor = (4 X π X area)/max $diameter^2$) (normalized value)
Ecd	Diameter of a circle with an area equivalent to that of the seed (mm)
EAmax	Maximum axis of an ellipse with equivalent area (mm)
EAmin	Minimum axis of an ellipse with equivalent area (mm)
Cpt	Compact grade = (√2 (4/π) X area)/Dmax
C	Curl = ratio between maximum diameters and Fiber lengths
Fl	Fiber length (mm)
Cvx	Convexity = ratio between Crofton's Perimeters and real Perimeters
EFDs 1 to 78	Elliptic Fourier Descriptors
Texture parameters	
Haralik 1	Angular second moment
HaralikSD1	Standard Deviation of Angular second moment
Haralik 2	Contrast
HaralikSD2	Standard Deviation of Contrast
Haralik 3	Correlation
HaralikSD3	Standard Deviation of Correlation
Haralik 4	Sum of square: variance
HaralikSD4	Standard Deviation of Sum of square: variance
Haralik 5	Inverse difference moment
HaralikSD5	Standard Deviation of moment
Haralik 6	Sum average
HaralikSD6	Standard Deviation of Sum average
Haralik 7	Sum variance

HaralikSD7	Standard Deviation of Sum variance
Haralik 8	Sum Entropy
HaralikSD8	Standard Deviation of Sum Entropy
Haralik 9	Entropy
HaralikDS9	Standard Deviation of Entropy
Haralik 10	Difference variance
HaralikSD10	Standard Deviation of Difference variance
Haralik 11	Difference Entropy
Haralik SD11	Standard Deviation of Difference Entropy

232

233 Data were statistically elaborated applying the stepwise LDA (Linear
234 Discriminant Analysis) (Fukunaga, 1990) that finds the combination of
235 predictor variables with the aim of minimizing the within-class distance and
236 maximizing the between-class distance simultaneously, thus achieving
237 maximum class discrimination. This approach is commonly used to
238 classify/identify unknown groups characterized by quantitative and qualitative
239 variables (Fisher, 1936, 1940; Hastie et al., 2001; Holden et al., 2011). The
240 stepwise method selects the parameters most statistically significant among
241 the 137 measured on each seed, using three statistical variables: *Tolerance*,
242 *F-to-enter* and *F-to-remove*. The *Tolerance* value indicates the proportion of a
243 variable variance not accounted for by other independent variables in the
244 equation. *F-to-enter* and *F-to-remove* values define the power of each
245 variable in the model and are useful to describe what happens if a variable is
246 inserted and removed, respectively, from the current model. This method
247 starts with a model that does not include any of the variables, adding step by
248 step one more, until no remaining variables are able to increase the
249 discrimination ability, stopping the process (Grillo et al., 2012; Venora et al.,
250 2009). At each step, the predictor with the largest *F-to-enter* value that
251 exceeds the entry criteria ($F \geq 3.84$) is added to the model. The cross-
252 validation procedure, also called rotation estimation (Picard and Cook, 1984;
253 Kohavi, 1995), was applied, both to evaluate the performance and validate

any classifier and to avoid problems and/or mistakes that might arise on account of seed samples not enough numerically representative.

The correlation of some measured seed traits with the previously available fruit weight data was studied using Pearson Coefficient and the relation of genetic regions previously known to be involved in fruit weight variation with seed variation was studied as described above.

3. RESULTS

3.1 Molecular relationships among accessions

The 211 polymorphic SNPs used in this study were quite informative with PIC average values ranging from 0.01 to 0.48 (see Supplementary data). Table 2 shows the polymorphism detected in each group of accessions. The highest degree of polymorphism was found in the *ameri* group (91.00%) (that included *ameri*, *adana* and *chandalack* accessions), followed by *flexuosus* (85.78%) which also displays the highest genetic diversity (0.32), being *cantalupensis* and *inodorus* less variable (63.51% and 63.98% respectively). Within subsp. *agrestis* the highest polymorphism level was found in *momordica* (73.93%), with a genetic diversity of 0.30, being the wild *agrestis* less variable (46.92%).

Table 2

Polymorphism level and gene diversity in the different groups according to SNP analysis.

C. melo subsp. *Melo*		*C. melo* subsp. *agrestis*	
ameri	91.00% / 0.2559[a]	*agrestis*	46.92% / 0.1535
cantalupensis	63.51% / 0.2022	*conomon*	42.65% / 0.1243
dudaim [b]	-	*chito* [b]	-
flexuosus	85.78% / 0.3193	*acidulus*	42.65% / 0.1504
inodorus	63.98% / 0.1609	*makuwa*	5.21%/ 0.0197
reticulatus	53.55%/ 0.1873	*momordica*	73.93% / 0.3044
chate [b]	-	*chinensis*	53.35%/0.1715
		tibish[b]	-

[a] Nei's gene diversity (1973).

[b] not calculated, less than three genotypes analysed in these groups

The relationships among the varieties assayed in this study are shown in the NJ tree, constructed with the polymorphic SNPs (Fig. 1). *C. melo* subsp. *melo* varieties, are clustered apart, with two differentiated clusters within *inodorus* (one containing mainly the Spanish types and a more disperse one with African and Eastern Europe types), and two clusters of *cantalupensis* (commercial Charentais types, and other *cantalupensis* and *reticulatus* types). Accessions of the *ameri* group appear mixed with *inodorus* and *cantalupensis*. In the *C. melo* subsp. *agrestis* cluster, the African *agrestis* and *tibish* varieties, and the Asian *conomom, makuwa* and *chinensis* ones, could be clearly separated in two groups. The other varieties were intermediate between both subspecies. These results are coherent with the PCoA (see Graphical Abstract), whose 3 first coordinates explained the 53.4% of the total variation. These were also coincident with the Structure analysis conducted in Esteras et al., 2013 and Leida et al., 2015 (Supplementary data), who reported two structured populations within both the *cantalupensis* (Charentais French types and American *reticulatus*), and the *inodorus* groups (Spanish casaba and European/African *inodorus*), a population of the Asian *ameri group,* showing admixture with other populations, and two separated populations within the subsp. *agrestis* (exotic Eastern *conomon-makuwa-chinensis*and African wild *agrestis* plus *acidulus* and *tibish*), with some Asian *acidulus,* and Asian and American wild *agrestis* showing population admixture. Structure analysis also supported the intermediate position of the remainder varieties, *momordica, flexuosus,* and *chate.*

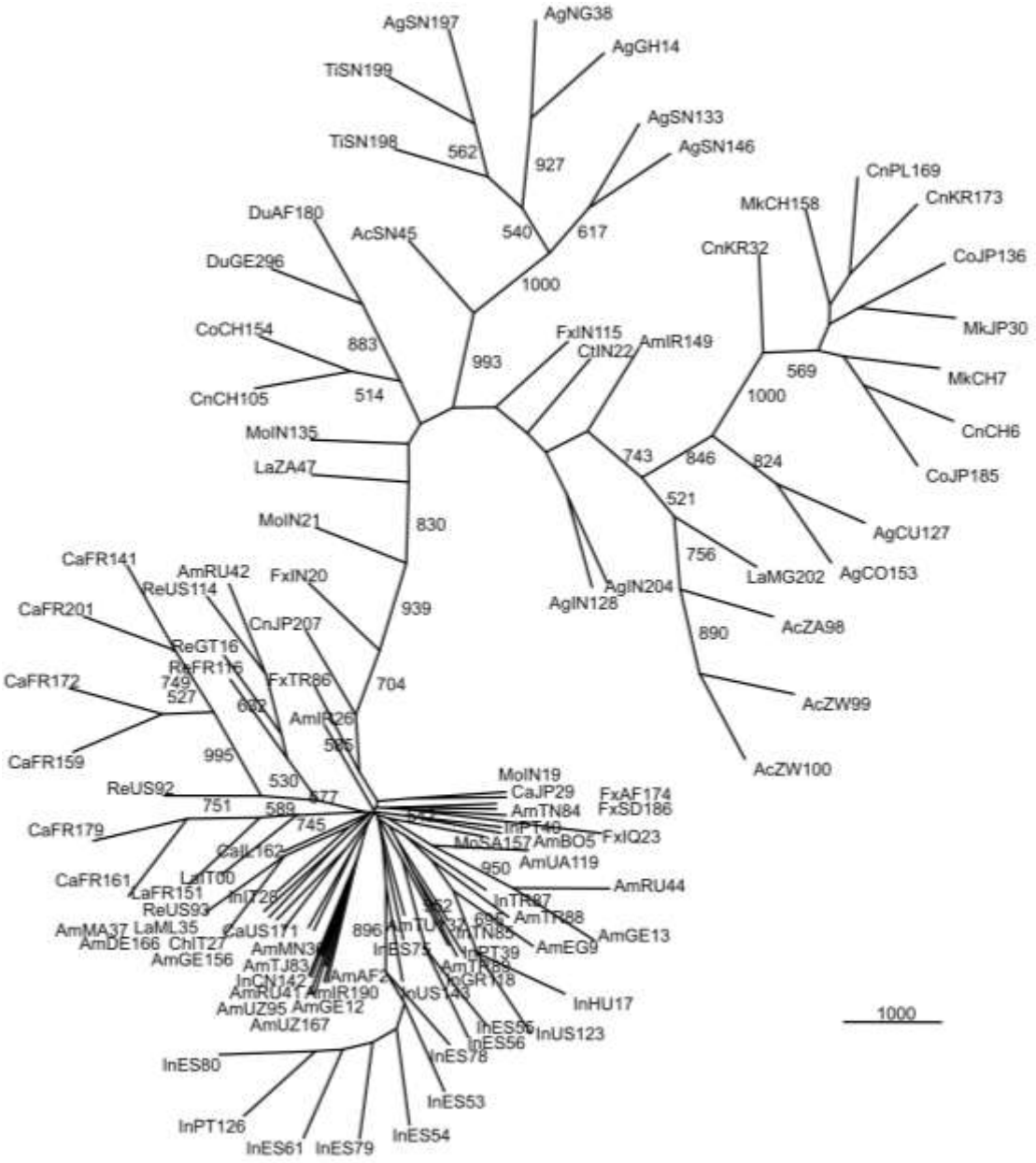

Fig. 1. NJ tree constructed with SNPs results, only Bootstrap values higher than 500 are showed. Code abbreviations: Ca= *cantalupensis*, Re=*reticulatus* In= *inodorus*, Am= *ameri* (including *ameri*, *adana* and *chandalack*), Fx= *flexuosus*, Ch= *chate*, Du= *dudaim*, Co= *conomon*, Cn= *chinensis*, Mk= *makuwa*, Mo=*momordica*, Ct= *chito*, Ac= *acidulus*, Ti= *tibish*, Ag= *agrestis*, La= intedeterminate landraces. The third and fourth letter indicates the provenience according to ISO 3166-1 alpha-2 country code.

3.2 Morpho-colorimetry analysis

In order to assess the phenotypic differences among genus and species hierarchy, the LDA was applied, considering all accessions of *Cucumis melo, C. sativus, Citrullus lanatus* and *C. colocynthis*. The discrimination of these species, based on 16,096 seeds, was quite clear, with an overall correct identification of 97.6% (Table 3). Misclassification between the two genera, *Cucumis* and *Citrullus*, was fairly close to zero. Also classification errors between the two respective species within the same genus were not significant (1.1%-5.4% between *C. melo* and *C. sativus*, 0.7%-5.2% between *C. lanatus* and *C. colocynthis*).

Table 3

Results of cross validated LDA analysis on *Cucumis melo*, *C. sativus*, *Citrullus lanatus* and *C. colocynthis*. First part of the table reports the amount of analysed seeds, the second part the respective percentage. The value of the number of an item crossed with itself and the other items indicates the number/percentage of seeds correctly classified as the same group, e.g. among the 11,595 analyzed *C. melo* seeds, 11,472 (98.9%) have been correctly classified as melon, 123 (1.1%) as *C. sativus* and none as *C. colocynthis* or *C. lanatus*.

	Cucumis melo	*Cucumis sativus*	*Citrullus lanatus*	*Citrullus colocynthis*	Total
97.6% overall classification					
Seeds number					n°
Cucumis melo	**11,472**	123	-	-	11,595
Cucumis sativus	112	**1,950**	-	-	2,062
Citrullus lanatus	76	1	**1,412**	10	1,499
Citrullus colocynthis	16	-	49	**875**	940
Percentage					%
Cucumis melo	**98.9**	1.1	-	-	100.0
Cucumis sativus	5.4	**94.6**	-	-	100.0
Citrullus lanatus	5.1	0.1	**94.2**	0.7	100.0
Citrullus colocynthis	1.7	-	5.2	**93.1**	100.0

Fig. 2 reports the 2D scatter-plot graph of the discrimination among *Cucumis* and *Citrullus* species. Seeds of each of the four *taxa* are clearly grouped and the distance between the two genera, *Cucumis* and *Citrullus*, is higher than that between the two species within the same genus. *C. colocynthis* and *C. lanatus* seeds are more distant than *C. melo* and *C. sativus* seeds.

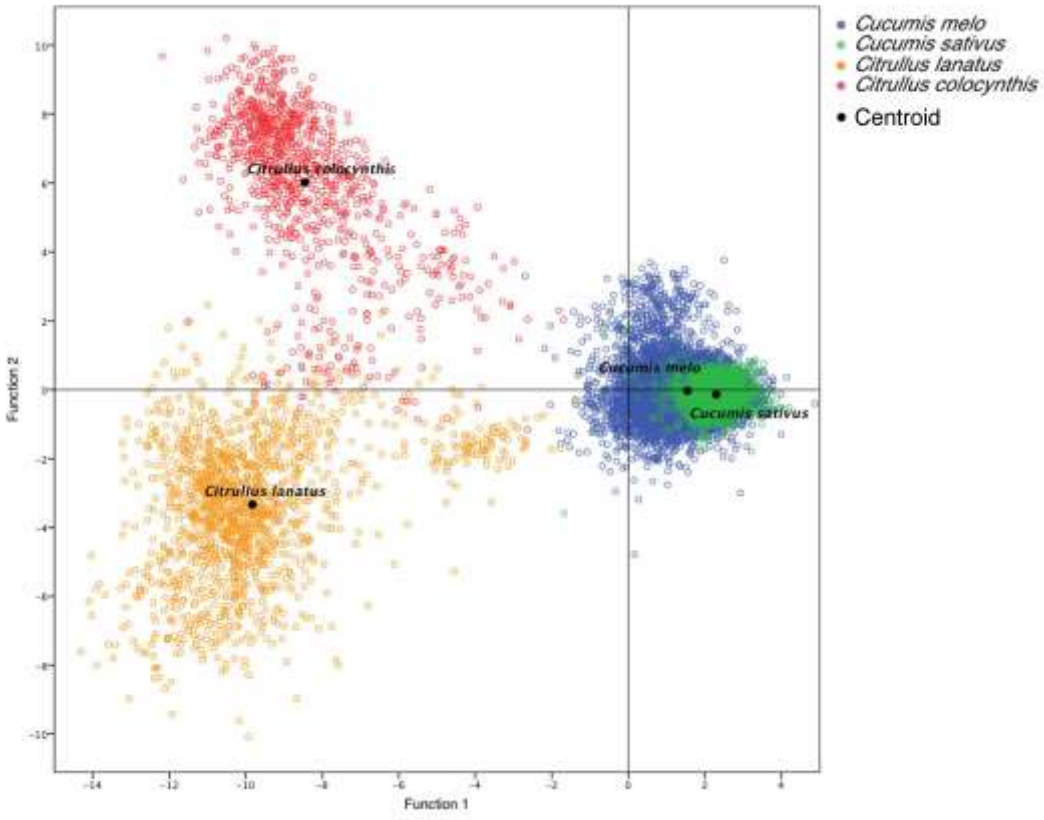

Fig. 2. Scatter plot graph based on LDA analysis discrimination of *Cucumis melo*, *C. sativus*, *Citrullus lanatus* and *C. colocynthis*. Small points represent single seed data, black points represent their average (centroid). Spatial arrangement of points suggests similarity and dissimilarity of groups, but just first two functions of 3 available have been used for the graphical representation. The variance of Function 1 is 74.4% and Function 2 is 15.1%, the remaining 10.5% is distributed on the non represented third function.

Table 4 shows the first 10 discriminant parameters according to the *F-to-remove* value. As expected, the most important seed character was related to seed colour, mainly the mean red channel value (Rmean). It means that colour, and in particular the wavelengths related to the red light, is the most reliable feature to distinguish seeds of these *taxa*. Seed dimension also plays an important role in discrimination. Minimum diameter (Dmin), Area (A) and some derived measures, such as the maximum axis of the ellipse with equivalent area (EAmax), were of great importance. Also two texture parameters (Haralik11 and 5), resulted useful for *taxa* identification, mainly related to differences between the spotted and wrinkled watermelon seeds and the smooth and monochromatic melon and cucumber seeds.

Table 4

First 10 factors used for discrimination among genera and species in order of decreasing *F-to-remove*, that describes the power of each variable in the model. The *Tolerance* indicates the proportion of a variable variance not accounted by other independent variables in the equation. *Wilks' lambda* is a direct measure of the proportion of variance in the combination of dependent variables that is unaccounted for by the independent variable.

	Parameter	*F-to-remove*	*Tolerance*	*Wilks' lambda*
1	Rmean	3526.6	0.045	0.008
2	Dmin	914.2	0.020	0.005
3	A	755.6	0.019	0.005
4	Haralik11	651.4	0.197	0.005
5	Haralik5	595.1	0.001	0.005
6	FD18	593.8	0.290	0.005
7	SDsum	570.9	0.008	0.005
8	EAmax	532.5	0.126	0.005
9	Sf	526.4	0.009	0.005
10	Cpt	404.2	0.005	0.005

Discrimination of melon seeds between the two *C. melo* subspecies was also clear with an overall correct identification of 93.2% (Table 5). Seeds of

362 the subspecies *melo* were correctly classified in 98.2% of cases, with only 146
363 seeds out of the 8,125 analysed seeds misattributed to the subspecies
364 *agrestis*. Wild melons were also correctly classified in 96.4% of cases, with
365 the remaining 3.6% of the cases misclassified as seeds belonging to the
366 same subspecies and not to subspecies *melo*. Subspecies *agrestis* was
367 correctly discriminated in 76.7% of cases. Unlike the inter-genera
368 classification, infraspecific classification is mainly due to seed size parameters
369 (Table 6). In fact, Area and Ecd are the first two parameters used to
370 distinguish between subspecies, the colour traits being less important.
371 Nevertheless, R_SD, S_SD, B_SD, L_SD colour values were useful to
372 differentiate *melo* seeds, with a darker cream colour, from lighter *agrestis*
373 seeds.

374 **Table 5**

375 Results of cross validated LDA analysis on melon subspecies (see legend of
376 table 3)

	C. melo subsp. *melo*	*C. melo* subsp. *agrestis*	*C. melo* subsp. *agrestis* (wild)	
93.2% overall classification				
Seed number				n°
C. melo subsp. *melo*	**7,979**	146	-	8,125
C. melo subsp. *agrestis*	486	**2,008**	124	2,618
C. melo subsp. *agrestis* (wild)	-	31	**821**	852
Percentage				%
C. melo subsp. *melo*	**98.2**	1.8	-	100.0
C. melo subsp. *agrestis*	18.6	**76.7**	4.7	100.0
C. melo subsp. *agrestis* (wild)	-	3.6	**96.4**	100.0

377

378 **Table 6**

379 First 10 factors used for discrimination among subspecies in order of
380 decreasing *F-to-remove* (see legend of table 4).

	Parameter	*F-to-remove*	*Tollerance*	*Wilks' lambda*
1	Ecd	532.1	0.006	0.112
2	A	333.0	0.003	0.108

3	Gmean	224.8	0.018	0.106
4	R_SD	188.1	0.006	0.106
5	S_SD	170.4	0.039	0.106
6	B_SD	144.7	0.023	0.105
7	L_SD	118.6	0.003	0.105
8	H	107.0	0.032	0.104
9	E	99.4	0.017	0.104
10	HaralikSD7	99.0	0.022	0.104

Table 7 shows the cross-validated results of melon varieties classification. Correct classification percentages of varieties belonging to subspecies *agrestis* (ranging from 55.8% to 93.3%) were greater than that of *melo* varieties (ranging from 37.1% to 86.1%). Overall correct identification was of 64.9%. Wild melons formed a homogeneous group, with only a 6.7% of misclassification with other related types of the *agrestis* subspecies. The other two varieties of subspecies *agrestis* more genetically related to the wild types are *tibish* and *acidulus*. The accessions of these two varieties included in this study were all from Africa, except one from Sri Lanka, and their seeds were quite well discriminated, 79.5% and 80.6% respectively. A lower degree of discrimination (ranging from 54.2% to 70.6%) was found in accessions of the *conomon* and related varieties, *chinensis* and *makuwa*, misclassifications mostly occurred among them. Within the subspecies *melo*, seed analysis gave a 19% of *ameri* classified as *inodorus* and vice versa. Also a 24.4% and 16.4% of misclassification with the *ameri* type was found in *cantalupensis* and *reticulatus,* respectively. *Momordica* was misidentified in several other varieties, mainly belonging to subspecies *melo* (*ameri*, *flexuosus* and *reticulatus*). It was not possible to define this variety as a determined group, since seeds can be correctly classified as *momordica* only in 55.8% of cases.

Table 7

Results of cross validated LDA analysis on melon varieties (see legend of table 3)

	ameri	*inodorus*	*cantalupensis*	*reticulatus*	*chate*	*flexuosus*	*dudaim*	*momordica*	*acidulus*	*tibish*	*chinensis*	*conomon*	*makuwa*	*chito*	*agrestis*	indet. *melo*	indet. *agrestis*	Total
64.9% overall classification																		
Seed number																		n°
ameri	**1569**	479	164	114	27	54	6	45	12	-	-	-	-	1	-	41	3	2515
inodorus	456	**1516**	64	93	12	11	1	23	1	1	-	-	-	-	-	20	1	2199
cantalupensis	233	57	**354**	182	51	37	-	7	4	1	10	-	-	-	-	9	10	955
reticulatus	112	24	68	**444**	-	1	18	-	-	-	-	-	-	-	-	14	-	681
chate	6	-	-	-	**85**	5	-	2	-	-	-	-	-	-	-	-	-	98
flexuosus	147	2	16	12	21	**637**	22	47	1	1	-	-	-	-	-	1	-	907
dudaim	42	-	-	13	1	-	**229**	1	7	1	-	-	-	1	-	-	-	295
momordica	68	31	2	18	2	32	1	**198**	-	-	-	-	-	-	-	3	-	355
acidulus	14	2	15	-	-	7	1	1	**458**	10	19	1	19	3	-	1	17	568
tibish	-	-	-	-	-	-	4	-	18	**124**	-	1	1	-	-	-	8	156
chinensis	23	1	24	-	-	-	4	-	16	1	**317**	80	46	30	42	-	1	585
conomon	-	-	-	-	-	-	-	-	10	14	52	**195**	24	-	-	-	1	296
makuwa	-	-	-	-	-	-	-	-	8	17	55	30	**274**	-	4	-	-	388
chito	-	-	-	-	-	-	1	-	-	2	-	-	-	**83**	4	-	-	90
agrestis	-	-	-	-	-	-	-	-	-	15	13	-	22	7	**795**	-	-	852
indet. *melo*	60	126	44	60	-	-	-	-	27	5	5	-	-	-	-	**114**	34	475
indet.*agrestis*	-	-	6	2	-	6	-	1	29	1	1	-	-	-	-	4	**130**	180
Percentage																		%
ameri	**62.4**	19.0	6.5	4.5	1.1	2.1	0.2	1.8	0.5	-	-	-	-	-	-	1.6	0.1	100.0
inodorus	20.7	**68.9**	2.9	4.2	0.5	0.5	-	1.0	-	-	-	-	-	-	-	0.9	-	100.0
cantalupensis	24.4	6.0	**37.1**	19.1	5.3	3.9	-	0.7	0.4	0.1	1.0	-	-	-	-	0.9	1.0	100.0
reticulatus	16.4	3.5	1-	**65.2**	-	0.1	2.6	-	-	-	-	-	-	-	-	2.1	-	100.0
chate	6.1	-	-	-	**86.7**	5.1	-	2.0	-	-	-	-	-	-	-	-	-	100.0
flexuosus	16.2	0.2	1.8	1.3	2.3	**70.2**	2.4	5.2	0.1	0.1	-	-	-	-	-	0.1	-	100.0
dudaim	14.2	-	-	4.4	0.3	-	**77.6**	0.3	2.4	0.3	-	-	-	0.3	-	-	-	100.0
momordica	19.2	8.7	0.6	5.1	0.6	9.0	0.3	**55.8**	-	-	-	-	-	-	-	0.8	-	100.0
acidulus	2.5	0.4	2.6	-	-	1.2	0.2	0.2	**80.6**	1.8	3.3	0.2	3.3	0.5	-	0.2	3.0	100.0
tibish	-	-	-	-	-	-	2.6	-	11.5	**79.5**	-	0.6	0.6	-	-	-	5.1	100.0
chinensis	3.9	0.2	4.1	-	-	-	0.7	-	2.7	0.2	**54.2**	13.7	7.9	5.1	7.2	-	0.2	100.0
conomon	-	-	-	-	-	-	-	-	3.4	4.7	17.6	**65.9**	8.1	-	-	-	0.3	100.0
makuwa	-	-	-	-	-	-	-	-	2.1	4.4	14.2	7.7	**70.6**	-	1.0	-	-	100.0
chito	-	-	-	-	-	-	1.1	-	-	2.2	-	-	-	**92.2**	4.4	-	-	100.0
agrestis	-	-	-	-	-	-	-	-	-	1.8	1.5	-	2.6	0.8	**93.3**	-	-	100.0
indet. *melo*	12.6	26.5	9.3	12.6	-	-	-	-	5.7	1.1	1.1	-	-	-	-	**24.0**	7.2	100.0
indet.*agrestis*	-	-	3.3	1.1	-	3.3	-	0.6	16.1	0.6	0.6	-	-	-	-	2.2	**72.2**	100.0

Figure 3 shows the spatial position occupied by each variety where affinity distances can be deduced. Varieties belonging to subspecies *agrestis*, except from *momordica* variety, are distributed mainly in the left quadrant, rather than subspecies *melo* that occupies the right side. An intermediate group across the two subspecies, formed with *momordica*, *dudaim*, *flexuosus*

and *chate* varieties, can be easily determined. Both dimension and colour traits proved to be key parameters for the varieties discrimination (Table 8). Area and Ecd were again the most discriminant factors followed by colour parameters.

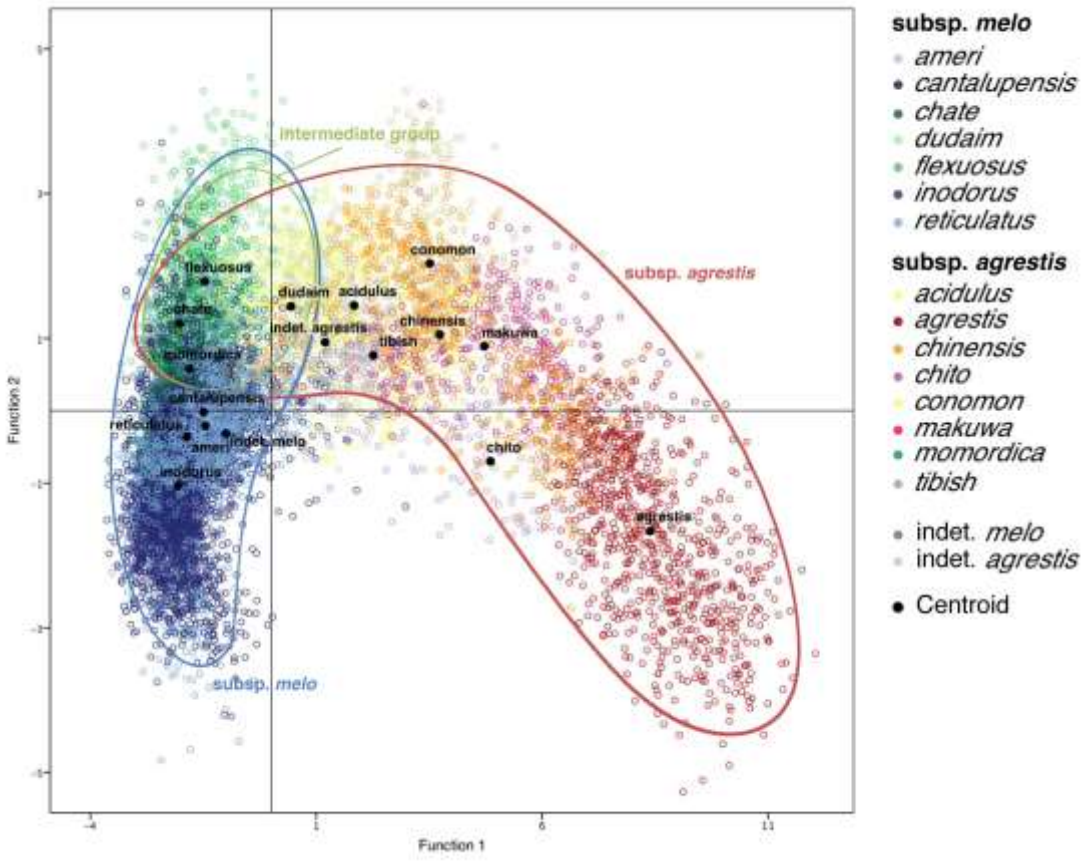

Fig. 3. Scatter plot graph based on melon varieties. Small points represent single seed data, black points represent their average (centroid). Spatial arrangement of points suggests similarity and dissimilarity of groups, but just first two functions of 16 available can be used for the graphical representation. The variance on Function 1 is 70.1% and on Function 2 is 8.2%, the remaining 21,7% is distributed over the 16 non represented functions. Major areas occupied by subspecies *agrestis* and *melo* varieties and their intermediate forms are marked.

Table 8

First 10 factors used for discrimination among varieties in order of decreasing *F-to-remove* (see legend of table 4).

	Parameter	*F-to-remove*	*Tolerance*	*Wilks' lambda*
1	A	221.7	0.008	0.005
2	Ecd	156.4	0.004	0.004
3	SqDsum	100.5	0.018	0.004
4	Gmean	97.1	0.023	0.004
5	H_SD	86.7	0.254	0.004
6	Rf	74.7	0.047	0.004
7	B_SD	69.8	0.024	0.004
8	R_SD	57.4	0.006	0.004
9	Pconv	54.6	0.001	0.004
10	Dmax	49.0	0.002	0.004

In order to set up statistically solid groups, according to morpho-colourimetric data, varieties with similar phenotypic characters were clustered. Six different macro-groups were isolated: the ameri/inodorus group, the cantaloupe group (*cantalupensis* and *reticulatus*), the intermediate group (*dudaim*, *chate*, *flexuosus* and *momordica*), the African agrestis group (*tibish* and African *acidulus*), the conomon group (*conomon*, *chinensis*, *makuwa* and Asian *acidulus*) and the wild types group (*agrestis* and *chito*). In Table 9, cross validated results of LDA are shown, while Table 10 reports the main features that contribute to *taxa* discrimination. Again Area was one of the most important discriminatory parameters, together with Eecd and some colour descriptors. Overall correct identification was of 78.3%. Most macro-groups resulted in being correctly classified, with percentages up to 73.5%, except for cantaloupe group that reached 61.9% of correct identification, confirming a high overlapping with the *ameri*/*inodorus* group, which anyway can be correctly isolated in 85.1% of cases. Indeterminate *melo* and *agrestis* groups, formed with non-classifiable accessions, were totally scattered in their respective subspecies and intermediate forms

Table 9

Results of cross validated LDA analysis on groups of variety with higher similarity: cantaloupe grp. (*cantalupensis* and *reticulatus*), ameri/inodorus grp., intermediate grp. (*chate, dudaim, flexuosus* and *momordica*), African agrestis grp. (African *acidulus* and *tibish*), conomon grp. (*conomon*, *chinensis*, *makuwa* and Asian *acidulus*), wild types grp. (*agrestis* and *chito*) and *agrestis* and *melo* indeterminate landraces (see legend of table 3)

	cantaloupe grp.	ameri/inodorus grp.	intermediate grp.	African *agrestis* grp.	conomon grp.	wild types	indet. *melo*	indet. *agrestis*	Total
78.3% overall classification									
Seed number									n°
cantaloupe grp.	**1027**	443	135	3	10	-	13	5	1,636
ameri/inodorus grp.	462	**4,006**	193	2	1	-	42	8	4,714
intermediate grp.	76	306	**1,230**	32	5	-	2	4	1,655
African *agrestis* grp.	16	13	17	**610**	51	-	2	15	724
conomon grp.	27	28	1	39	**1,096**	73	3	2	1,269
wild types	-	-	-	24	46	**870**	-	2	942
indet. *melo*	84	232	3	31	10	-	**95**	20	475
indet. *agrestis*	15	1	10	27	1	-	2	**124**	180
Percentage									%
cantaloupe grp.	**61.9**	27.9	8.3	0.2	0.6	-	0.9	0.4	100.0
ameri/inodorus grp.	9.9	**85.1**	4.1	-	-	-	0.8	0.1	100.0
intermediate grp.	5.6	18.5	**73.5**	1.1	0.7	-	0.2	0.4	100.0
African *agrestis* grp.	2.1	1.3	3.2	**87.9**	1.8	-	0.3	3.5	100.0
conomon grp.	1.8	2.3	0.1	2.2	**87.6**	5.4	0.4	0.1	100.0
wild types	-	-	-	0.7	5.6	**93.3**	-	0.3	100.0
indet. *melo*	16.2	51.4	0.8	7.8	1.9	-	**18.5**	3.4	100.0
indet. *agrestis*	6.1	0.6	5.6	13.9	1.7	-	1.1	**71.1**	100.0

Table 10

First 10 factors used for discrimination among macro-groups in order of decreasing *F-to-remove* (see legend of table 4).

	Parameter	*F-to-remove*	*Tollerance*	*Wilks' lambda*
1	Gmean	159.86	0.022	0.015
2	A	125.42	0.001	0.015
3	Ecd	122.42	0.001	0.015
4	H_SD	106.21	0.234	0.015
5	Dmin/Dmax	78.28	0.024	0.014
6	B_SD	71.56	0.023	0.014

7	SqDsum	68.77	0.011	0.014
8	Cpt	67.38	0.004	0.014
9	HaralikSD5	59.32	0.041	0.014
10	EAmax	57.08	0.003	0.014

458

459 ***3.3 Integration of molecular data and seed/fruit phenotypes***

460 The seed parameter Area was one of the most relevants to discriminate
461 melon varieties. High positive correlation (r= 0.921) was found between the
462 parameter Area and fruit weight, measured in a previous phenotyping assay
463 (Leida et al., 2015). The coefficient of determination (R^2) amounted to 0.849.

464 The analysis of the allelic frequencies of some SNPs located in genomic
465 regions in which QTLs involved in fruit weight have been reported (Diaz et al.,
466 2011), reveals non-random allelic distribution in different seed size groups of
467 accessions. Specific alleles were more frequent in groups of accessions with
468 small/large seeds. For example, accessions with low values of the Seed
469 parameter Area have high frequencies of one of the two alleles of 3 SNPs
470 located in Linkage group I (LGI), in regions in which QTLs for fruit weight have
471 been reported, CMPSNP711 and AI_17-E07 (located at 45.2 and 46.8 cM,
472 respectively) and CMPSNP731 (located at 80.4 cM), whereas similar
473 frequencies for the two alleles are observed in the SNPs located in other
474 regions of this LGI. In fact, the ANOVA shows significant differences in the
475 average Seed Area for these three markers between accessions belonging to
476 the two homozygous genotypic classes (mean ± standard deviation of
477 homozygous a, allele more frequent in large seed accessions = 39.9±10.9,
478 39.1± 10.8 and 39.2±11.4; homozygous b, allele more frequent in small seed
479 accessions = 25.0±14.3, 19.4±10.4, and 18.9±10.5).

480 The best hit of the *Arabidopsis thaliana* gene At4g39850 with the melon

unigene collection of Melonomics was found with MELO3C018991, located in the CM3.5_scaffold00035 from 2610482 to 2613142. This scaffold is anchored to the melon genetic map in LGVII at 32,1cM. A differential allelic distribution in small *versus* large seed size accessions, similar to that found markers of LGI, was found in SNP CMPSNP262 (located in LGVII at 30,5cM).

Information about the natural genetic variation of the gene MELO3C018991 was obtained from the Melogene database of SNPs. A single SNP in this gene was found. It was a non-synonym mutation (C/T aa 249 S/N). According to the sequence information provided by the Melogene data base, one allele had been sequenced in three pools composed of *inodorus*, *momordica* and agrestis *acidulus* accessions (including most of the accessions of these groups analysed in this study that have large or intermediate seed size), whereas the alternative allele had been sequenced in the *conomon* pool of accessions (also including many of the *conomon* accessions analysed in the present study, all with small seed size).

4. Discussion

Seed image analysis has proved to be successful to discriminate among genera and species in the selected set of cucurbits. *Citrullus lanatus* and *C. colocyntis* show higher heterogeneity in seeds morpho-colourimetric features than *Cucumis melo* and *C. sativus*, which is consistent with their origin and taxonomic relationships. Despite of *C. colocynthis* being traditionally considered the wild ancestor of *C. lanatus*, genetic analysis showed that the cultivated and Egusi watermelon (var. *lanatus*) and the citron type (var. *citroides*) diverged into separate lineages appearing independently evolved from a common ancestor, possibly *C. ecirrhosus* (Dane and Liu, 2007).

506 Furthermore, whereas cucumber and melon are two cultivated crops
507 phylogenetically close related and surely deriving from the same ancestor
508 (Sebastian et al., 2010), *C. colocynthis* is a perennial (rarely annual) wild
509 species growing on sandy habitats in desert and semi-desert areas of North
510 Africa, the Near East and South-West Asia as far as India (Jeffrey, 2001).
511 These differences are reflected on seed morphology, where species of the
512 same genus are more related than species of different genus.

513 The division of melon in two subspecies, *C. melo* subsp. *melo* and *C.*
514 *melo* subsp. *agrestis*, already described elsewhere (Silberstein et al., 1999;
515 Monforte et al., 2003; Deleu et al., 2009; Esteras et al., 2009; Blanca et al.,
516 2011), was confirmed by both molecular and seed morpho-colourimetric
517 analyses. The intermediate position found with SNPs for the *flexuosus*, *chate*
518 and *momordica* varieties is in agreement with the reported idea that from
519 these varieties evolved most of the current melon populations (Esteras et al.,
520 2013; Leida et al., 2015).In fact, in the largest melon re-sequencing assay
521 performed by Blanca et al. (2012), *momordica* was the most heterogeneous
522 variety, and shared the highest percentage of SNPs with other varieties of
523 both subspecies. All of these non-sweet varieties have limited diffusion
524 through Africa and Near East. The same pattern can be evidenced by
525 morpho-colourimetric analysis which shows the intermediate position of these
526 varieties between the two subspecies. Consistency between molecular and
527 morpho-colourimetric results was also found for *C. melo* subsp. *agrestis* var.
528 *agrestis*, the wild forms of melons, which were isolated from the cultivated
529 accessions.

Within the subspecies *melo*, *ameri* accessions show high degrees of admixture with all other *melo* varieties. This group of accessions is the most heterogeneous within the cultivated melon and include quite different landraces. The high crossbreeding of the *ameri* group with *inodorus* and *cantalupensis* might have produced a wide range of intermediate forms that does not always permit to isolate *ameri* from the others. These results also agree with the hypothesis that modern *inodorus* and *cantalupensis* derived from these variable Asian melons (Pitrat et al., 2000). Few reports describe the variability of Asian types of the subsp. *melo*. In a recent study on Iranian melons, Raghami et al. (2014) reported the high diversity in melons from this area and remarked their differences with European/American *inodorus* and *cantalupensis*. Seed image analysis results agree with molecular data and show a high degree of misclassification of *ameri* seeds with *inodorus*, *cantalupensis* and *reticulatus*, Molecular analysis reveals four differentiated groups within *inodorus* and *cantalupensis* that could not be distinguished on the basis of seed traits.

Within the subspecies *agrestis*, the accessions of the *conomon* group (*conomon*, *chinensis* and *makuwa*) were quite similar molecularly, also according to previous studies (Blanca et al., 2012), and presented closely related seed traits. Despite *acidulus* and *tibish* being molecularly similar to wild *agrestis*, the bigger size of their seeds allows to separate these two varieties from the wild form. However, *acidulus* and *tibish* are quite hardly differentiated on the basis of seed traits having a significant degree of misclassification. Old classification models placed *tibish* in subspecies *melo* (Pitrat et al., 2000), but molecular analyses demonstrated its greater similarity

555 to *agrestis* (Esteras et al., 2009). Seed morphology agrees with this
556 classification of this primitive melon belonging to subspecies *agrestis*. In line
557 with molecular data, the unclassified landraces seem to be mostly mixed
558 types of different varieties of the subspecies *melo*, mostly *inodorus*, but with
559 some traces of subspecies *agrestis*.

560 According only to seed morphology it was possible to isolate six different
561 groups of varieties: a group of accessions of *ameri* and *inodorus,* closely
562 related to the other group of *cantalupensis* and *reticulatus*, an intermediate
563 group between the two subspecies (*dudaim*, *chate*, *flexuosus* and
564 *momordica*), a group of African *agrestis* varieties (*tibish* and *acidulus*), the
565 conomon group formed with *conomon*, *chinensis*, *makuwa* and Asian
566 *acidulus*, and a group formed with wild melon types (*agrestis* and *chito*).

567 Despite the importance of seed size in plant evolution and crop
568 domestication, relatively little is known about the genetic and molecular
569 processes underlying natural variation in seed size and morphology.
570 Integration of SNPs and phenotypic data sets provide the opportunity to
571 obtain information about the genetics of seed traits. A strong correlation was
572 found between the seed Area, the most discriminant seed trait among melon
573 accessions, and fruit weight, which agree with the results reported in other
574 Cucurbits (Paris and Nerson, 2003). Moreover, the evidence of a non random
575 allelic distribution in large/small seed groups of accessions of SNPs located in
576 some genomic regions in which QTLs involved in fruit weight had been
577 previously located (Diaz et al., 2011) suggests that some of these regions
578 might also account for part of the observed variation in seed size. This non
579 random distribution of alleles could be also due to an effect of the structure of

580 the population, then associations of alleles to seed traits must be proved in
581 larger unstructured populations or in populations specifically designed, such
582 as introgression lines with which we are currently working. For example,
583 populations derived from the cross of *acidulus*/*tibish* and wild *agrestis* could
584 be suited to study the genetics of seed traits as these varieties are
585 molecularly similar, but significantly differ in seed traits. The use of the
586 available genomic tools can also facilitate the identification of candidate
587 mutations involved in seed traits, such as that found in the melon orthologue
588 of the tomato *SW4* (Orsi and Tanksley, 2009). Our results raised the
589 possibility that the melon orthologue of *SW4* might also underlie natural
590 variations in seed size in melon, but the association needs to be
591 demonstrated in appropriate populations.

592 **5. Conclusions**

593 Molecular analysis recognized some differentiated populations, but
594 also a wide range of mixed types. Despite this molecular admixture, seed
595 image analysis revealed six major groups that can be discriminated on the
596 basis of specific phenotypic traits, mainly associated to seed size and
597 morphology and less to seed colour. The obtained seed groupings are in
598 agreement with the molecular relationships and with the history of the melon
599 varieties. In fact, wild *agrestis*, Far eastern *conomon* and African *tibish* and
600 *acidulus* could be clearly distinguished. Discrimination of the cultivated types
601 of the *melo* subspecies (*inodorus, cantalupensis, reticulatus* and *ameri*) was
602 also possible, although less clear, probably due to a more intense crossing
603 and breeding process undergone by these commercial groups. The
604 intermediate position of *momordica, flexuosus, chate* and *dudaim* groups

across the two subspecies is also detected by seed features. The identification of the more discriminant specific traits allows the development of a method to classify new seeds in any of the reported groups. A great deal of the extant melon variation is maintained in different seed collections, so this tool would be of great utility to manage their variation and optimize their conservation and use. Also the integration of molecular and seed data would be a useful tool to study the genetics of seed traits.

Funding

The Italian University and Research Ministry (MIUR) financed the PhD scholarship (Diego Sabato, Univeristy of Cagliari), making this research possible. Molecular analysis was carried out with contributions of the PLANT KKBE project PIM2010PKB-00691 and the complementary grant from the Generalitat Valenciana ACOMP/2013/141.

Acknowledgements

We wish to acknowledge Dr. Gianfranco Venora for his support and for enabling the use of the laboratory at Stazione Consorziale Sperimentale di Granicoltura per la Sicilia (CT). Authors thank the National Plant Germplasm System of the USDA, for providing some accessions of their melon collection. Also we would like to thank M. Pitrat who within the MELRIP project provided some of the melon accessions used in this study. Cucumber and Watermelon collections were provided by Dr. M.J. Diez from the COMAV Genebank. We thank Dr. R. Peiró from COMAV for her support on statistical analysis and E. Martinez for her technical support. Thanks to Sidney Goïame for the language review.

630 **References**

631 Bacchetta G, Grillo O, Mattana E, Venora G. 2008. Morpho-colorimetric
632 characterization by image analysis to identify diaspores of wild plant species.
633 *Flora* 203:669-682.

634 Bacchetta G, Escobar García P, Grillo O, Mascia F, Venora G. 2011*a*. Seed
635 image analysis provides evidence of taxonomical differentiation within the
636 *Lavatera triloba* aggregate (Malvaceae). *Flora* 206:468-472.

637 Bacchetta G, Fenu G, Grillo O, Mattana E,Venora G. 2011*b*. Identification of
638 Sardinian species of *Astragalus* section *Melanocercis* (Fabaceae) by seed
639 image analysis. *Annales Botanici Fennici* 48:449-454.

640 Bates DM, Robinson RW. 1995. Cucumbers, melons and water-melons. In:
641 Smartt J, Simmonds NW, eds. *Evolution of Crop Plants. 2nd edn*. Harlow:
642 Longman Scientific, 89-96.

643 Blanca J, Cañizares J, Ziarsolo P, et al. 2011. Melon transcriptome
644 characterization: Simple Sequence Repeats and Single Nucleotide
645 Polymorphisms discovery or high throughput genotyping across the species.
646 *The Plant Genome* 4:118-131.

647 Blanca J, Esteras C, Ziarsolo P, et al. 2012. Transcriptome sequencing for
648 SNP discovery across *Cucumis melo*. *BMC Genomics* 13:280.

649 Cai G, Yang Q, Yang Q, Zhao Z, Chen H, Wu J, Fan C, Zhou Y. 2012.
650 Identification of candidate genes of QTLs for seed weight in *Brassica napus*
651 through comparative mapping among *Arabidopsis* and *Brassica* species. *BMC*
652 *Genomics* 13:105.

653 Dane F, Liu J. 2007. Diversity and origin of cultivated and citron type
654 watermelon (*Citrullus lanatus*). *Genetic Resources and Crop Evolution*
655 54:1255-1265.
656 Deleu W, Esteras C, Roig C, *et al.* 2009. A set of EST-SNPs for map
657 saturation and cultivar identification in melon. *BMC Plant Biology* 9:90.
658 Decker DS, Newsom LA. 1988. Numerical analysis of archaeological
659 *Cucurbita pepo* seeds from Hontoon island, Florida. *Journal of Ethnobiology*
660 8:35-44.
661 Diaz A, Fergany M, Formisano G, *et al.* 2011. A consensus linkage map for
662 molecular markers and Quantitative Trait Loci associated with economically
663 important traits in melon (*Cucumis melo* L.). *BMC Plant Biology* 11:111.
664 Esquinas-Alcázar JT, Gulick PJ. 1983. *Genetic Resources of Cucurbitaceae:*
665 *A Global Report.* Rome: IBPGR Secretariat.
666 Esteras C, Formisano G, Roig C, *et al.* 2013. SNP genotyping in melons:
667 genetic variation, population structure, and linkage disequilibrium. *Theoretical*
668 *Applied Genetics* 126:1285-1303.
669 Esteras C, Lunn J, Sulpice R, et al. 2009. Phenotyping a highly diverse core
670 melon collection to be screened using Ecotilling. In: *8th Plant Genomics*
671 *European Meetings*, Lisbon, 7-10 October, 214.
672 Esteras C, Nuez F, Picó B. 2012. Genetic diversity studies in Cucurbits using
673 molecular tools. In: Wang Y, Behera TK, Kole C, eds. *Cucurbits: Genetics,*
674 *Genomics and Breeding of Cucurbits.* New Hampshire: Science Publishers
675 Inc, 140-198.
676 Fan JB, Chee MS, Gunderson KL. 2006. Highly parallel genomic assays.
677 *Nature Reviews Genetics* 7:632-644.

678 Felsenstein J. 1997. An alternating least squares approach to inferring
679 phylogenies from pairwise distances. *Systematic Biology* 46:101-111.
680 Fisher RA. 1936. The use of multiple measurements in taxonomic problems.
681 *Annals of Human Genetics* 7:179-188.
682 Fisher RA. 1940. The precision of discriminant functions. *Annals of Human*
683 *Genetics* 10:422-429.
684 Fujishita N. 1980. About comparison *Cucumis melo* seeds excavated from the
685 Ikegami ruin with the contemporary melon seeds and the melon seeds
686 excavated from other ruins. *Osaka Culture Center* 6:105-124.
687 Fujishita N, Nakagawa K. 1973. About seed (embryo) and fruit development
688 of *Cucumis* species. Effects of growth regulator affected apomixix and
689 parthenocarpy of *C. melo*. *Japanese Society for Horticultural Science* 42:186-
690 187.
691 Fukunaga K. 1990. *Introduction to statistical pattern recognition. 2nd edn.* San
692 Diego: Academic Press.
693 Gabriel S, Ziaugra L, Tabbaa D. 2009. SNP genotyping using the Sequenom
694 MassARRAY iPLEX platform. *Current Protocols in Human Genetics* 60:11-18.
695 Garcia-Mas J, Benjak A, Sanseverino W, *et al.* 2012. The genome of melon
696 (*Cucumis melo* L.). Genome amplification in the absence of recent duplication
697 in an old widely cultivated species. *Proceedings of the National Academy of*
698 *Sciences* 109:11872-11877.
699 Grillo O, Mattana E, Venora G, Bacchetta G. 2010. Statistical seed classifiers
700 of 10 plant families representative of the Mediterranean vascular flora. *Seed*
701 *Science and Technology* 38: 455-476.

Grillo O, Draper D, Venora G, Martìnez-Laborde JB. 2012. Seed image analysis and taxonomy of *Diplotaxis* DC (Brassicaceae, Brassiceae). *Systematics and Biodiversity* 10:57-70.

Hastie T, Tibshirani R, Friedman J. 2001. *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Holden JE, Finch WH, Kelly K. 2011. A comparison of two-group classification methods. *Educational and Psychological Measurement* 71:870-901.

Janick J, Paris HS, Parrish DC. 2007. The cucurbits of Mediterranean antiquity: identification of *taxa* from ancient images and descriptions. *Annals of Botany* 100:1441-1457.

Jeffrey C. 1980. A review of the Cucurbitaceae. *The Botanical Journal of the Linnean Society* 81:233-247.

Jeffrey C. 2001. Cucurbitaceae. In: Hanelt P, ed. *Mansfeld's encyclopedia of agricultural and horticultural crops*. Berlin: Springer, 1510-1557.

Jeffrey C. 2005. A new system of Cucurbitaceae. *Botanicheskii Zhurnal* 90:332-335.

Jeffrey C, De Wilde WJJO. 2006. A review of the subtribe *Thladianthinae* (Cucurbitaceae). *Botanicheskii Zhurnal* 91:766-776.

Kirkbride JHJr 1993. *Biosystematic monograph of the genus* Cucumis *(Cucurbitaceae)*, Boone: Parkway Publ.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137-1143.

Leida C, Moser C, Esteras C, Sulpice R, Lunn JE, de Langen F, Monforte AJ, Picó B. 2015. Variability of candidate genes and genetic association for sugar

727 accumulation and climacteric behavior in melon (*Cucumis melo* L.). *BMC*
728 *Genetics* 16:28
729 Liu K, Muse SV. 2005. Powermarker: integrated analysis environment for
730 genetic marker data. *Bioinformatics* 21:2128-2129.
731 Mallick MFR, Masui M. 1986. Origin, distribution and taxonomy of melons.
732 *Scientia Horticulturae* 28:251-261.
733 Megaloudi F. 2006. Plants and diet in Greece from Neolithic to Classical
734 Period: the archaeobotanical remains. British Archaeological Reports
735 International Series 1516. Oxford: Archaeopress.
736 Monforte AJ, Garcia-Mas J, Arús P. 2003. Genetic variability in melon based
737 on microsatellite variation. *Plant Breeding* 122:153-157.
738 Munger HM, Robinson RW. 1991. Nomenclature of *Cucumis melo* L. *Cucurbit*
739 *Genetics Cooperative* 14:43-44.
740 Naudin C. 1859. Essais d'une monographie des espèces et des variétés du
741 genre *Cucumis*. *Annales des Sciences Naturelles* 11:5-87.
742 Nei M. 1973. Analysis of gene diversity in subdivided populations.
743 *Proceedings of the National Academy of Sciences* 70:3321-3323.
744 Nei M, Tajima F, Tateno Y. 1983. Accuracy of estimated phylogenetic trees
745 from molecular data. II Gene frequency data. *Journal of Molecular Evolution*
746 19:153-170.
747 Orrù M, Grillo O, Venora G, Bacchetta G. 2013*a*. Computer vision as a
748 method complementary to molecular analysis: Grapevine cultivar seeds case
749 study. *Comptes Rendus Biologies* 335:602-615.
750 Orrù M, Grillo O, Lovicu G, Venora G, Bacchetta G. 2013*b*. Morphological
751 characterisation of *Vitis vinifera* L. seeds by image analysis and comparison

752 with archaeological remains. *Vegetation History and Archaeobotany* 22:231-
753 242.
754 Orsi CH, Tanksley SD. 2009. Natural Variation in an ABC Transporter Gene
755 Associated with Seed Size Evolution in Tomato Species. PLoS Genet
756 5:e1000347. doi:10.1371/journal.pgen.1000347.
757 Page RDM. 1996. Tree View: An application to display phylogenetic trees on
758 personal computers. *Computer Applications in the Biosciences* 12:357-358.
759 Paris HS, Amar Z, Lev E. 2012. Medieval emergence of sweet melons,
760 *Cucumis melo* (Cucurbitaceae). *Annals of Botany* 110:23-33.
761 Paris HS, Nerson H. 2003. Seed Dimensions in the Subspecies and Cultivar-
762 groups of *Cucurbita pepo. Genetic Resources and Crop Evolution* 50:615-
763 625.
764 Peakall R, Smouse PE. 2012 GenAlEx 6.5: genetic analysis in Excel.
765 Population genetic software for teaching and research – an update.
766 *Bioinformatics* 28:2537-2539.
767 Picard R, Cook D. 1984. Cross-Validation of Regression Models. *Journal of*
768 *the Acoustical Society of America* 79:575-583.
769 Pinna S, Grillo O, Mattana E, Cañadas E, Bacchetta G. 2014. Inter- and
770 intraspecific morphometric variability in *Juniperus* L. seeds (Cupressaceae).
771 *Systematics and Biodiversity* 12:211–223.
772 Pitrat M. 2008. Melon (*Cucumis melo* L.). In: Prohens J, Nuez F, eds.
773 *Handbook of Crop Breeding, vol I: Vegetables*. New York: Springer, 283-315.
774 Pitrat M, Hanelt P, Hammer K. 2000. Some comments on infraspecific
775 classification of cultivar of melon. In: Katzir N, Paris HS, eds. *Proceeding of*

776 *Cucurbitaceae 2000, Máaleh Hahamisha, Israel, 19–23 March 2000*. *Acta*
777 *Horticulturae* 510:29-36.
778 Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure
779 using multilocus genotype data. *Genetics* 155:945-959.
780 Raghami M, López-Sesé AI, Reza Hasandokht M, Zamani Z, Reza Fattahi
781 Moghadam M, Kashi A. 2014. Genetic diversity among melon accessions
782 from Iran and their relationships with melon germplasm of diverse origins
783 using microsatellite markers. *Plant Systematics and Evolution* 300:139-151.
784 Renner SS, Schaefer H, Kocyan A. 2007. Phylogenetics of *Cucumis*
785 (Cucurbitaceae): Cucumber (*C. sativus*) belongs in an Asian/Australian clade
786 far from melon (*C. melo*). *Evolutionary Biology* 7:58.
787 Rivera D, Miralles B, Obón C, Carreño E, Palazón JA. 2007. Multivariate
788 analysis of *Vitis* subgenus *Vitis* seed morphology. *Vitis* 46:158-167.
789 Roy A, Bal SS, Fergany M, *et al.* 2012. Wild melon diversity in India (Punjab
790 State). *Genetic Resources and Crop Evolution* 59:755-767.
791 Sabato D, Masi A, Ucchesu M, Peña-Chocarro L, Usai A, Giachi G, Capretti
792 C, Bacchetta G. 2015. Archaeobotanical analysis of a Bronze Age well from
793 Sardinia: a wealth of knowledge. *Plant Biosystems*. Doi:
794 10.1080/11263504.2014.998313.
795 Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny Reconstruction:
796 The Role of Morphology. *Systematic Biology* 52:539-548.
797 Sebastian P, Schaefer H, Telford IRH, Renner SS. 2010. Cucumber (*Cucumis*
798 *sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and
799 Australia, and the sister species from melon is from Australia. *Proceedings of*
800 *the National Academy of Sciences* 107:14269-14273.

801 Shahin MA, Symons SJ. 2003. Colour calibration of scanners for scanner
802 independent grain grading. *Cereal Chemistry* 80:285-289.
803 Silberstein L, Kovalski I, Huang R, Anagnostou K, Jahn M.K, Perl-Treves R.
804 1999. Molecular variation in melon (*Cucumis melo* L.) as revealed by RFLP
805 and RAPD markers. *Scientia Horticulturae* 79:101-111.
806 Smykalova I, Grillo O, Bjelkova M, Hybl M, Venora G. 2011. Morpho-
807 colorimetric traits of *Pisum* seeds measured by an image analysis system.
808 *Seed Science and Technology* 39:612-626.
809 Smykalova I, Grillo O, Bjelkova M, Pavelek M, Venora G. 2013. Phenotypic
810 evaluation of flax seeds by image analysis. *Industrial Crops and Products*
811 47:232–238.
812 Steermers FJ, Gunderson KL. 2007. Whole genome genotyping technologies
813 on the Bead Array TM platform. *Biotechnology Journal* 2:41-49.
814 Stepansky A, Kovalski I, Perl-Treves R. 1999. Intraspecific classification of
815 melons (*Cucumis melo* L.) in view of their phenotypic and molecular variation.
816 *Plant Systematic Evolution* 217:313-333.
817 Tanaka K, Akashi Y, Fukunaga K, *et al.* 2013. Diversification and genetic
818 differentiation of cultivated melon inferred from sequence polymorphism in the
819 chloroplast genome. *Breeding Science* 63:183-96.
820 Tanaka K, Nishitani A, Akashi Y, *et al.* 2007. Molecular characterization of
821 South and East Asian melon, *Cucumis melo* L., and the origin of Group
822 Conomon var. *makuwa* and var. *conomon* revealed by RAPD analysis.
823 *Euphytica* 153:233-247.
824 Terral J, Tabard E, Bouby L, *et al.* 2010. Evolution and history of grapevine
825 (*Vitis vinifera* L.) under domestication: new morphometric perspectives to

826 understand seed domestication syndrome and reveal origins of ancient
827 European cultivars. *Annals of Botany* 105:443-455.
828 Venora G, Grillo O, Shahin MA, Symons SJ. 2007*a*. Identification of Sicilian
829 landraces and Canadian cultivars of lentil by image analysis system. *Food*
830 *Research International*, 40:161-166.
831 Venora G, Grillo O, Ravalli C, Cremonini R. 2007*b*. Tuscany beans landraces,
832 on-line identifications from seeds inspection by image analysis and Linear
833 Discriminant Analysis. *Agrochimica*, 51:254-268.
834 Venora G, Grillo O, Saccone R. 2009. Quality assessment of durum wheat
835 storage centres in Sicily: evaluation of vitreous, starchy and shrunken kernels
836 using an image analysis system. *Journal of Cereal Science* 49:429-440.
837 Yashiro K, Iwata H, Akashi Y, *et al.* 2005. Genetic relationship among East
838 and South Asian melon (*Cucumis melo* L.) revealed by AFLP analysis.
839 *Breeding Science* 55:197-206.
840 Zohary D, Hopfand M, Weiss E. 2012. *Domestication of plants in the Old*
841 *World. The origin of cultivated plants in West Asia, Europe and the*
842 *Mediterranean Basin.* Oxford: Oxford University Press.