RESEARCH ARTICLE

Molecular characterization of the cucumber (*Cucumis sativus* L.) accessions held at the COMAV's genebank

Jose V. Valcárcel, Ana Pérez-de-Castro, María José Díez, and Rosa Peiró

Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Poltècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

Abstract

The cucumber (*Cucumis sativus* L.) is an important crop worldwide. In the present study, the molecular genetic diversity of 131 Spanish accessions was analyzed using 23 simple sequence repeat (SSRs). Eighteen of these SSRs were polymorphic; the mean number of alleles, mean observed heterozygosity and mean polymorphic information content were 3.2, 0.065 and 0.229, respectively. Seven SSRs showed a polymorphic information content (PIC) ranging from 0.31 to 0.44, therefore they were reasonably informative. Around 60% of the alleles showed a frequency higher than 0.05, and only one allele in the SSR31399 showed a frequency lower than 0.01. In addition, three accession-specific alleles were found. A high proportion of variation among accessions was obtained. In no case all plants of any accession showed the same genotype and only 18 of 131 Spanish accessions had at least two plants with the same genotype. A cluster analysis did not show any relation with morphological types or geographical area. Therefore, these results demonstrated that molecular diversity of the cucumber did not reflect its phenotypic variability. Finally, this study provided information for the rationalization of the cucumber collection of the COMAV. Morphological traits, origin and molecular data were taken into account to select 47 accessions, six belonging to 'French' type, 15 to 'Long' type, and 24 to 'Short' type. Phenotypic and molecular variability contained in the complete collection was conserved in the selected accessions.

Introduction

Cucumber (*Cucumis sativus* L. var. *sativus*) is one of the most important cultivated cucurbits worldwide and ranks the fourth in global vegetables production (FAOSTAT, 2017). Spain is traditionally the first producer in the European Union and the second country in the world to export this crop (FAOSTAT, 2017). Most of the Spanish production comes from Andalusia and to a lesser extent from the Canary Islands, Madrid, Catalonia, Murcia, Extremadura, the Valencian Community and other communities (MAPAMA, 2017). Most of the cucumbers cultivated in Spain are directed to the fresh market and grown under greenhouse. Commercial types mainly cultivated are Dutch cucumber (81%), especially for the foreign trade, and Spanish and French types, for the domestic trade (Reche, 2011). Besides commercial varieties, landraces cultivated by farmers for self-consumption and for local markets are also produced. These plant materials are especially useful in breeding as a source of specific traits. Thus, conservation and use of this germplasm must be enhanced.

Spanish cucumber accessions are mainly held at the 'Banco de Hortícolas de Zaragoza', BGHZ (Vegetables Genebank, Zaragoza, Spain) and at the Genebank of the 'Instituto de Conservación y Mejora de la Agrodiversidad Valenciana', COMAV (Institute for Conservation and

Improvement of Agrodiversity, Valencia, Spain). All these accessions are landraces, collected from local farms. A morphological characterization of the collection maintained at the COMAV has been recently reported, including 195 Spanish accessions held in this genebank (Valcárcel *et al.*, 2018). The most represented types were the Spanish cucumber and the French cucumber. Variability was found within each of the groups for fruit traits and also for flowering and production related traits. The diversity found in the collection provides interesting resources in cucumber breeding. Approximately 40 of these accessions are conserved at the 'US National Plant Germplasm System' (NPGS, USDA) and have been characterized at the molecular level in a previous work which included 3,342 accessions from different origins (Lv *et al.*, 2012). However, molecular characterization of the complete collection has not been performed yet.

Initial molecular diversity studies in cucumber were carried out using isozyme loci, and included a high number of accessions from the NPGS (Knerr *et al.*, 1989; Meglic *et al.*, 1996; Staub *et al.*, 1997a, 1999). All these works confirmed that the genetic variability found in cucumber is relatively low as compared to other *Cucumis* species. Furthermore, the higher diversity was identified in accessions collected in India, the center of origin of cucumber, and China, secondary center of diversity (Staub *et al.*, 1997a, 1999). Subsequent analysis were carried out using DNA molecular markers, such as restriction fragment length polymorphisms (RFLPs) (Dijkhuizen *et al.*, 1996), random amplified polymorphic DNA (RAPDs) (Staub *et al.*, 1997b; Horejsi & Staub, 1999; Mliki *et al.*, 2003), amplified fragment length polymorphism (AFLPs) (Li *et al.*, 2004), and simple sequence repeat (SSR) (Mu *et al.*, 2008; Hu *et al.*, 2010a). Availability of the cucumber whole genome sequence for three different cucumber lines (Huang *et al.*, 2009; Wóycicki *et al.*, 2011; Yang *et al.*, 2012), together with the lowering costs of sequencing, has enabled the large scale development of SSRs (Ren *et al.*, 2009; Cavagnaro *et al.*, 2010; Liu *et al.*, 2015) as well as single nucleotide polymorphisms (SNPs) (Qi *et al.*, 2013). Moreover, the advances in sequencing make it affordable to resequence multiple genotypes revealing variation at the nucleotide level (Rubinstein *et al.*, 2015; Zhou *et al.*, 2015).

Given the availability of high-density molecular markers, the key point in diversity studies is the sampling of the genetic variability. In this sense, Lv *et al.* (2012) analyzed 3,342 accessions from different origins with the aim of elucidating the genetic diversity and structure of cucumber populations. A total of 23 SSRs evenly distributed in the genome and selected for their high level of polymorphism (Ren *et al.*, 2009) were used for the genotyping. Results showed that the accessions broadly represented the diversity of the species. Moreover, a core collection was defined including 115 accessions, which captured 77% of the alleles found in the accessions analyzed (Lv *et al.*, 2012). This core collection was later evaluated by deep resequencing, generating information regarding cucumber domestication and divergence among cultivated populations (Qi *et al.*, 2013).

Efficiency of the management of genebanks can be increased with rationalization of their collections (van Hintum *et al.*, 2000). Rationalization consists in the reduction in the number of accessions conserved, consequently decreasing the cost of maintenance. The first step to rationalize a collection relies in passport data and morphological characterization of the accessions. However, different levels of variability are identified with morphological and molecular data. Thus, the use of molecular data in rationalization of germplasm collections has increased in recent years (van Treuren, 2010).

In this study, the molecular characterization of the cucumber collection held at the COMAV has been carried out. The final objective is to obtain information to rationalize the collection. The 23 highly polymorphic SSRs described by Lv *et al.* (2012) have been used to assess the genetic diversity of the collection and to reveal the population structure.

Material and methods

Material

Young, healthy, and fully-expanded leaves from the 627 plants representing 131 Spanish accessions as well as leaves belonging to three foreign accessions, were analyzed (Table S1 [suppl.]). Four or five plants per accession were used. This number was considered enough as populations in cucumber (and other cucurbits) are generally small, not only at farms but also in wild populations, favoring endogamy. The Spanish accessions included in the study were all landraces, collected from local farms. Thus, these accessions derive from a low number of plants, since local farmers usually grow small populations for self-consumption.

One hundred and twenty-seven Spanish accessions come from the COMAV and the other four accessions were provided by the BGHZ. Some of the accessions conserved at the COMAV genebank were collected by the holder institution and some others were collected by the BGHZ and maintained as duplicates in the COMAV. All these landraces are routinely reproduced using 10 plants per accession. Considering the origin of the Spanish accessions, 31 accessions came from Andalusia,

26 from Valencian Community, 21 from Castilla-La Mancha, 16 from Aragon, nine from Castilla-Leon, seven from Extremadura, six from Murcia, four from Canary Islands and three from the Basque Country, two accessions from Cantabria, Catalonia and Navarra, and one accession from Galicia and Asturias. Besides, three accessions provided by the Center for Genetic Resources (CGN, The Netherlands) and by the Chinese Academy of Agricultural Science (CAAS, China) were used as outgroups in the analysis. The phenotypic characterization of the complete collection of cucumber hold at the genebank of the COMAV has been recently conducted (Valcárcel *et al.*, 2018). They established five groups ('White', 'French', 'Short', 'Long' and 'Very long') after a visual inspection, according to their similarity to commercial types or the phenotypic traits of the fruits, mainly skin colour and fruit length. Using this classification, 131 accessions were selected and used in the present study to perform the molecular characterization. Seventy of the accessions belonged to the 'Short' type, 48 to the 'Long' type, 11 to the 'French' type, one to the 'Very long' type, and the other to the 'White' type. The selection of the 131 accessions was based on the criteria of excluding accessions with very similar fruit characteristics and maximizing the diversity of origins. Still, the proportion of selected accessions corresponding to each group was similar to those in the complete collection.

SSR analysis

DNA was extracted using the protocol proposed by Doyle & Doyle (1987) with some modifications and DNA quality and quantity was assessed using gel electrophoresis and spectrophotometry. A total of 23 SSR previously described by Lv *et al.* (2012) were analyzed in three sets of multiplex polymerase chain reactions (PCR) reactions (Table S2 [suppl.]). Each multiplex was carefully assembled according to the compatibility of the SSRs during PCR and the molecular size of their amplicons. The forward primer of the SSR markers was labelled with one of the three fluorescent dyes, carboxy fluorescein (FAM), carboxytetramethylrhodamine (TAMRA) and hexachloro-6-carboxyfluorescein (HEX). Multiplex PCR was carried out in a 11 µL volume using 5 µL of commercial Master Mix PCR Multiplex (Takara Multiplex Hot Short PCR, Takara), 20-40 ng of genomic DNA and labeled multiplexed SSR primers (0.4 pmol). The amplification was performed in an Mastercycler® personal thermocycler, and the amplification conditions were 95 °C for 5 min followed by 35 cycles of 95 °C for 20 s, specific annealing temperature for 90 s, and 72 °C for 30 s, and a final extension of 72 °C for 8 min. Multiplex PCR product was visualized using gel electrophoresis and then PCR fragment sizes were determined using capillary electro-

phoresis. The capillary electrophoresis was carried out by the Sequencing Service at the Institute for Plant Molecular and Cell Biology (IBMCP, Valencia, Spain), on an ABI 3100® platform (Appl Biosyst, Foster City, CA, USA), and the allele sizes were calculated using GeneScan 3.7 (Appl Biosyst). For PCR fragment size determinations, 0.05 µL of an internal size standard (Rox-500, ROX) was mixed with 1 µL of diluted PCR product (1/100) and 6 µL formamide. The mixture was heated at 94 °C for 3 min and then cooled within icy water. The size of the SSR fragments was determined with the software packages GeneScan 3.7 (Appl Biosyst).

Data analysis

Polymorphic SSR markers were used to analyze the genetic diversity of 131 Spanish cultivars and the three foreign accessions considered as outgroups. The number of alleles (Na), the number of genotypes (Ge) and the effective number of alleles (Ne) was determined for each SSR locus in the Spanish collection using the PowerMaker software (Liu & Muse, 2005). To estimate the discriminatory power of the microsatellite loci, polymorphic information content (PIC) value for each locus was estimated by $PIC = 1 - \sum_i p_i^2 - \sum_{i=i}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$ where p_i and p_j are the frequency of the ith and j^{th} alleles, and n is the number of alleles (Botstein *et al.*, 1980). The observed heterozygosity (H_0) and expected heterozygosity (H_e) were computed for each SSR locus using GenAlEx version 6.501 (Peakall & Smouse, 2012).

Nei (1978) genetic similarities were calculated and an unweighted pair group method with arithmetic mean (UPGMA) phenogram was performed using genetic distances with the PowerMaker software (Liu & Muse, 2005) and plotted using TreeView software (Page, 1996). The reliability and robustness of the dendrogram was tested by bootstrap analysis with 1,000 replicates to assess branch support. Genetic distances were also used to graphically represent genetic relationships among accessions by principal coordinates analysis (PCoA) using GenAlEx 6.5 software (Peakall & Smouse, 2012).

Pearson's correlation coefficients were computed between the number of alleles per locus found by Lv *et al.* (2012) and in our study using Statgraphics Centurion XVI software (Statistical Graphics, Rockville, MD, USA).

Results and discussion

Diversity in the collection based on SSRs

A total of 627 plants belonging to 131 accessions were analyzed using 23 SSRs. Two of them, SSR02895 and SSR07543, yielded amplification in less than

50% of the plants. SSR20852 and SSR23370 were monomorphic in all samples analyzed, while SSR16068 was monomorphic except for two plants of outgroup R41, homozygotes for the alternative allele. These five markers were not included in subsequent analysis.

The remaining 18 polymorphic SSR markers generated a total of 58 alleles in the Spanish collection (Table 1). The mean number of alleles per locus was 3.2, ranging from two to six. Significant correlation was found between the number of alleles per locus found by Lv *et al.* (2012) and in our study for the 18 polymorphic SSRs (r=0.512, p=0.03). Lv *et al.* (2012) reported a number of alleles between six and 16 for these 18 SSRs when analyzing the mega-collection of 3,342 accessions with representation from the five continents. The lower number of alleles per locus obtained in our study may be due not only to the lower number of accessions analyzed but also to their limited origin. In fact, other authors found ranges for the number of alleles per locus similar to our study, using different SSRs, when analyzing smaller collections of cucumber accessions from specific origins (Kong *et al.*, 2006; Watcharawongpaiboon & Chunwongse, 2008; Hu *et al.*, 2010a, 2010b, 2011; Pandey *et al.*, 2013; Yang *et al.*, 2015).

Table 1. Molecular diversity of 131 Spanish accessions of *Cucumis sativus* L. using 18 SSRs.

Chr	SSR	Na	Ge	Ne	H_0	H_e	PIC
1	SSR10018	4	8	1.65	0.09	0.39	0.33
1	SSR11340	2	3	1.85	0.06	0.46	0.35
1	SSR05723	2	2	1.09	0.00	0.08	0.08
1	SSR16695	3	5	1.16	0.04	0.14	0.13
2	SSR16226	3	5	1.17	0.03	0.15	0.14
2	SSR23220	2	3	1.07	0.01	0.06	0.06
2	SSR22653	5	9	1.23	0.04	0.18	0.18
3	SSR01738	2	3	1.36	0.08	0.27	0.23
3	SSR16056	6	15	2.36	0.22	0.58	0.49
3	SSR05012	2	3	1.60	0.14	0.38	0.31
4	SSR11043	4	10	1.86	0.14	0.46	0.42
4	SSR05125	3	5	1.19	0.04	0.16	0.15
4	SSR07543	3	4	1.10	0.03	0.09	0.09
5	SSR19998	4	7	1.04	0.01	0.04	0.04
6	SSR31399	3	4	1.26	0.02	0.20	0.18
6	SSR20218	3	5	1.23	0.03	0.19	0.18
7	SSR29260	3	6	1.95	0.15	0.49	0.44
7	SSR14861	4	8	1.58	0.04	0.37	0.33
	Mean	3.2	5.8	1.431	0.065	0.261	0.229

Chr: chromosome; Na: number of alleles; Ge: number of genotypes; Ne: effective number of alleles; H_0: observed heterozygosity; H_e: expected heterozygosity; PIC: polymorphic information content.

Polymorphism levels and mutation rate were correlated with the number of repeat units, with different results depending on the species. A positive correlation was found in grapevine (Thomas & Scott, 1993), tomato (Areshchenkova & Ganal, 1999) and watermelon (Zhu *et al.*, 2016), meanwhile negative correlation was found in rice (Panaud *et al.*, 1996), brassica (Szewc-McFadden *et al.*, 1996) and barley (Struss & Plieske, 1998). No correlation was found in our study between the nucleotide repeats and the level of polymorphism (p>0.05). However, Watcharawongpaiboon & Chunwongse (2008) found a positive correlation between these traits. This different result could be explained because different SSRs were used, and therefore different regions of the genome were analyzed.

Differences were also found depending on the study when comparing the relative frequencies of the alleles. In our study, among the 58 alleles found in the Spanish collection, 34 (58%) were 'common' alleles, as their frequency was higher than 5% (Table 2). In fact, frequency of the most frequent allele for the 18 polymorphic SSR markers ranged between 50 and 98%, being for 12 of them higher that 75%. Out of the total 58 alleles, 15 (26%) were denoted as 'less common' alleles, with frequencies between 1 and 5%, while eight (14%) were 'rare' alleles (frequency between 0.1 and 1%). One of the alleles of SSR31399 was classified as 'very rare' allele (frequency lower than 0.1%) as it was only present in one plant, in heterozygous state. Among the 316 SSR alleles found by Lv *et al.* (2012) when analyzing the collection of 3,342 accessions, only 20% were classified as 'common' alleles. Again, differences may be due in part to the different size of the collections, but mainly to the most diverse origins of the accessions analyzed by Lv *et al.* (2012). The fact that the accessions included in their analysis represented different origins explains the higher variability, which conditions the lower frequency of 'common' alleles. Moreover, 59% of the alleles identified by Lv *et al.* (2012) showed a frequency below 1%, suggesting a broad representation of the diversity of cucumber. The Spanish collection merely represents part of the diversity of the species, thus the lower percentage of 'rare' and 'very rare' alleles is justified.

Accession-specific alleles were found for SSR19998, SSR22653 and SSR31399. SSR19998 allele 339 bp was present in two of the plants of accession Sty17, in homozygosis in one of them and in heterozygote state for the other one. Allele 401 bp for SSR22653 was specific for three plants of accession Lti113, in homozygosis in one of them and combined with allele 411 bp in the other two. SSR31399 allele 202 bp was the only 'very rare' allele of the collection. As previously stated, it was only present in heterozygote state in one

Table 2. Alleles (in bp) and frequencies (Freq.) obtained in 18 polymorphic microsatellites (sorted by chromosome) obtained in 131 Spanish cucumber accessions.

SSR10018	Freq.	SSR11340	Freq.	SSR05723	Freq.	SSR16695	Freq.	SSR16226	Freq.	SSR23220	Freq.
134	0.24	378	0.64	287	0.96	172	<0.01	262	0.92	298	0.03
137	0.01	387	0.36	298	0.04	180	0.93	267	0.05	311	0.97
143	0.74					192	0.07	269	0.03		
147	0.01										

SSR22653	Freq.	SSR01738	Freq.	SSR16056	Freq.	SSR05012	Freq.	SSR11043	Freq.	SSR05125	Freq.
401	<0.01	126	0.16	124	0.04	205	0.75	203	0.08	237	0.91
405	0.07	129	0.84	130	0.50	213	0.25	214	0.17	242	0.06
407	0.01			143	0.01			228	0.71	245	0.03
409	0.02			157	0.41			231	0.04		
411	0.90			162	0.03						
				167	0.01						

SSR07543	Freq.	SSR19998	Freq.	SSR31399	Freq.	SSR20218	Freq.	SSR29620	Freq.	SSR14861	Freq.
124	0.03	339	<0.01	182	0.11	202	0.01	361	0.13	378	0.76
126	0.01	344	0.01	185	0.89	228	0.89	367	0.67	382	0.02
128	0.96	347	0.98	202	<0.01	252	0.10	372	0.20	390	0.03
		353	0.01							394	0.19

plant, which belonged to accession Sty157. Moreover, some alleles appeared either exclusively or almost exclusively in some cucumber types. SSR16695 allele 172 bp and SSR10018 allele 137 bp were only present in accessions belonging to the short type. SSR10018 allele 147 bp, SSR19998 allele 344 bp and SSR22653 allele 409 bp were mainly present in short type accessions. On the other hand, some of the 'rare' alleles and the 'very rare' alleles were in many cases specific of concrete geographical origins. As an example, allele 143 bp of SSR16056 was only identified in accessions from Andalusia or SSR14861 allele 382 bp was exclusive of accessions from Andalusia or Castilla-La Mancha. All these results must be considered when using this information to rationalize management of the collection.

Most of the possible genotypes for each of the SSRs were identified in the Spanish collection, concretely, 105 out of the 133 genotypes (Table 1). As previously stated, the very rare allele identified in the Spanish collection (allele 202 bp for SSR31399) was only present in heterozygote state. The remaining homozygote combinations were present, except for homozygotes for allele 202 bp for SSR20218. This allele is 'rare' in the population (frequency=0.0088), thus the lack of homozygotes could be expected. The rest of genotypes absent in the Spanish population were heterozygotes, in no case formed by two 'common alleles'.

The values of H_0 ranged from 0.00 (SSR05723) to 0.22 (SSR16056), with a mean of 0.07. All individuals analyzed in the present study presented the SSR05723 in homozygosis, corresponding to the lack of heterozygosity observed for this marker. The highest value corresponded to SSR16056, the SSR with the highest number of alleles identified in our collection, two of them 'common' (with similar frequency), other two 'less common' and only the other two with frequencies lower than 1%. Results for this SSR were similar for the three groups that include more than one accession, $i.e.$, 'Short', 'French' and 'Long' type. Using the same SSRs, Lv $et\ al.$ (2012) obtained the maximum value of H_0 (0.36) for SSR05012. This is one of the SSRs with higher H_0 in the present study (0.14 as an average) although in this case differences were observed among groups, given that all 'French' type plants were homozygote for this $locus$. Two alleles were identified for this SSR in our collection (with frequencies 75 and 25%, respectively), while six different alleles were found in the samples analyzed by Lv $et\ al.$ (2012). In both assays other SSRs presented a higher number of alternative alleles, so the high values for heterozygosity observed for this marker must correspond to the fact that none of the alleles appear at very low frequencies. The minimum value obtained by Lv $et\ al.$ (2012) was 0.07, for SSR23220, also corresponding to low heterozygosity in our collection. In fact, only plants of four accessions of the 'Short' type were heterozygous for this marker. However, Kong $et\ al.$ (2006) obtained higher values of the H_0 (0.41), similar to the results obtained by Watcharawongpaiboon & Chunwongse (2008). Their results could be explained by the fact that some of the materials included were hybrids.

The values of H_e ranged from 0.04 (SSR19998) to 0.58 (SSR16056), with a mean of 0.261. These values were lower than those obtained by Lv et al (2012) analyzing the same SSRs in their mega-collection (0.58). Similar H_e values have been obtained by Kong *et al.* (2006) and by Watcharawongpaiboon & Chunwongse (2008). The H_0 was relatively low if compared with the H_e value for all SSRs, suggesting certain degree of inbreeding of these materials. Besides, the relatively low number of alleles and gene diversity estimates for cucumber reflect its narrow genetic base (Dijkhuizen *et al.*, 1996). This result was also obtained in other *Cucurbitaceae* like *C. melo* L. (Raghami *et al.*, 2014). It should be stated that endogamy depression is not important in the cucurbit family (Fehet, 1992; McCreight *et al.*, 1992; Tatlioglu, 1992). Besides, populations in cucumber, and other cucurbits, are generally derived from seeds obtained from one or few fruits in small populations, not only at farms but also in wild populations, favoring endogamy. The Spanish accessions included in the study were all landraces, collected from local farms. Thus, these accessions derived from a low number of plants, since local farmers usually grow small populations for self-consumption.

Taking into account the classification performed by Botstein *et al.* (1980), seven SSRs were reasonably informative (0.5>PIC>0.25), and the rest were only slightly informative (PIC<0.25). Similar to our results, Pandey *et al.* (2013) and Yang *et al.* (2015) analyzing around 40 cucumber cultivars each other from India and China obtained similar PIC values.

Cluster analysis based on SSR markers

The AMOVA of the distance matrix for all the analyzed plants permitted a partitioning of the overall variation into two levels. The proportion of variation attributable to within-accessions differences was high, 45%, whereas 55% occurred among accessions. The intra and inter-accession variability's distribution has implications for the management of conserved accessions when regeneration has to be done in the genebank. When similar intra and inter-accessions variability is found, the number of plants to be regenerated must be high, in order not to lose the variability among plants of the same accession and to avoid the influence of genetic drift and selection, resulting in changes in genetic composition. The effective population size and the methods of pollination and harvesting seeds have to be also taken into account. However, when variability is higher among accessions, the effort must be devoted to increase the number of regenerated accessions.

The mean similarity coefficient of the 131 cucumber accessions was 0.14, ranging for 0.01 (between Ltp8 and Ltp48) to 0.44 (between Shs101 and Fty183). Diversity in Spanish cucumber is lower than that found by Yang *et al.* (2015) analyzing 42 Chinese accessions (mean similarity coefficient=0.76). These results were expected and probably due to the fact that China is a secondary center of diversity.

None of the accessions showed the same genotype for all the analyzed plants. Sty122 accession showed the highest number of plants sharing the same genotype, four plants. The other plant showed three different alleles. Shs101 and Sty116 accessions had three plants with the same genotypes, and the other plants showed different genotype inter and intra-accessions. Lti47, Lhs121 and Stp144 accessions had two groups of two plants with the same genotype differing among genotypes in two or three SSRs, depending on the accessions. A total of 12 accessions (Ltp8, Stp89, Lti131, Ltp147, Sty165, Sty180, Sel79, Sty146, Stp56, Sty94, Stp59 and Stp62) showed two plants with the same genotype. The rest of the Spanish accessions showed plants with a unique genotype within accession.

Discrimination between closely related individuals has been reported using SSRs, even when few loci were employed (Powell *et al.*, 1996). Therefore, SSR markers could be useful molecular markers in order to estimate the cucumber genetic diversity and also to perform marker assisted breeding programs. The dendrogram based on Nei coefficient revealed that the accessions were grouped into three unbalanced major clusters consisting of 1, 131 and 2 accessions, respectively (Fig. 1). As expected, accession Shs101 was separated from the rest of the accessions since it showed the lower similarity coefficient with most of the rest of accessions. Shs101 accession showed low variability since three plants had the same genotype with all the SSRs in homozygosity. For four of the SSRs, the plants of this accession were homozygote for alleles appearing at low frequency in the Spanish population. This accession was collected in Canary Islands. Cluster II included almost all accessions of the Spanish collection and one of the outgroups, and the subclusters obtained were not supported by high bootstrap values. Two of the outgroup accessions were grouped to the cluster III. Both accession showed different phenotype but they shared their Asiatic origin. Moreover, they both accessions showed specific alleles (alleles 202 bp and 292 bp for SSR31399 and SSR23220, respectively).

The clustering pattern of cucumber genotypes based on SSR markers was not in consonance with the groupings based on morphological traits or geographical area either including (Fig. 1) or not the outgroup (data not shown). Similar results were obtained when a PCoA (Principal Coordinates
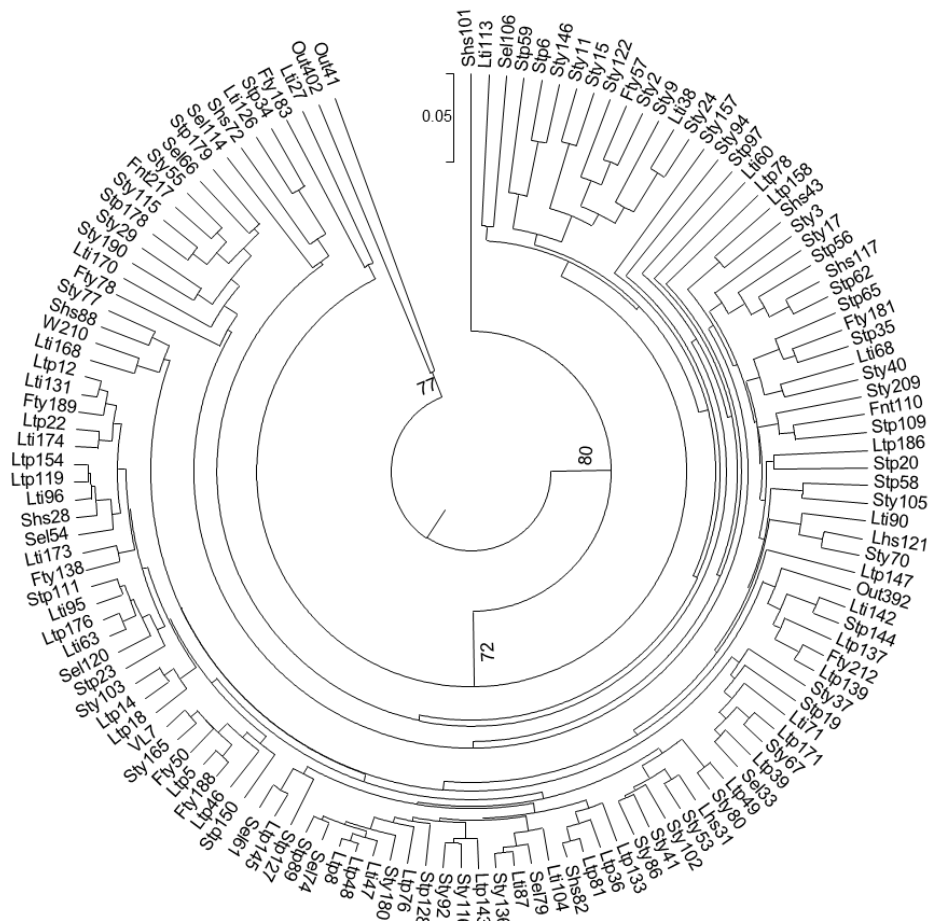
Figure 1. Unweighted pair group method with arithmetic (UPGMA) mean dendrogram based on Nei's distance of the 131 Spanish *Cucumis sativus* L. accessions, and also three foreign accessions, based on SSR markers. Values at the nodes indicate percentage of 1000 bootstrap runs supporting a particular node. Only values higher than 60% were included.

Analysis) based on SSR markers was done (Fig. 2); the first three principal coordinates accounted for 40.8% of the variation. These results seem to be in agreement with the absence of relationship between morphological traits and molecular markers found in cucumber (Yang *et al.*, 2015) and other plant species like perennial ryegrass (*Lolium perenne*), durum wheat (*Triticum durum*), sorghum (*Sorghum bicolor*), potato (*Solanum tuberosum*), garlic (*Allium sativum*) or tall fescue (*Festuca arundinacea*) (Baker *et al.*, 1998; Roldan-Ruiz *et al.*, 2001; Geleta *et al.*, 2006; Elameen *et al.*, 2011; Chen *et al.*, 2014; Sun *et al.*, 2015). A higher variability in morphological traits was usually found. Neutral molecular markers such as SSRs are commonly used in molecular diversity studies and they may sample diversity in non-coding regions of the genome. Therefore, they have limited use in predicting the phenotypic diversity of individuals, especially in complex traits, which usually have a polygenic control. Consequently,

low molecular variability may not be accompanied by low phenotypic variability for important traits (Collard *et al.*, 2005). Null or low correlation between phenotypical and molecular diversity could also be due to low marker saturation (Sun *et al.*, 2015). Therefore, more SSRs could improve the accuracy of cucumber genetic diversity estimation.

Dendrogram for the specific main groups ('Short', 'Long' and 'French') were performed, and a lack of agreement among subgroups were also found. Similar results were found when the geographical origin was taken into account (data not shown). Although environmental conditions and market orientation in a specific geographical area may select a specific morphological appearance, which can explain the high similarity among accessions from the same area, several accessions collected in different geographical areas could have at least some common genetic bases as a consequence of gene flow.
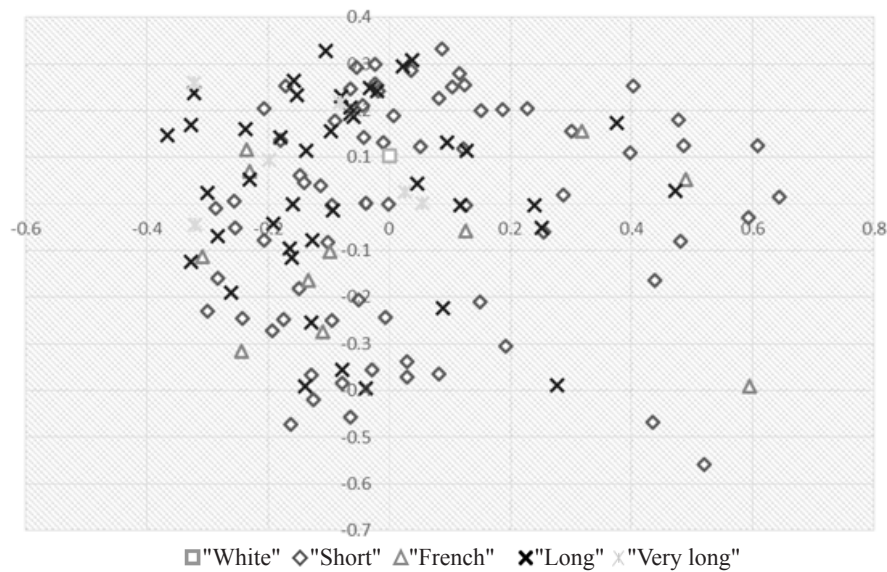
□"White" ◇"Short" △"French" ✖"Long" ✕"Very long"

Figure 2. Principal coordinate analysis (PCoA) plot based on 18 SSRs of the 131 Spanish cucumber accessions. The two first coordinates are shown.

Rationalization of the collection

Given the low molecular diversity among the Spanish cucumber accessions found in the present study, a combined strategy was followed to perform the rationalization of the collection, including phenotypic traits, origin and molecular data. According to the groups established using morphological traits (Valcárcel *et al.*, 2018), accessions of all of them were selected trying to conserve the phenotypic variation, to include accessions of most of the Spanish Autonomous Communities (Table S3 [suppl.]). Moreover, the genetic distances calculated from molecular data among accessions inside each group were taken into account. Priority was given to accessions carrying rare alleles, and phenotypic groups and Autonomous Communities with only one accession. The final set of selected accessions (Table S1 [suppl.]) included 47 accessions, six belonging to 'French' type, 15 to 'Long' type, and 24 to 'Short' type. The unique accessions molecularly analyzed belonging to 'Very long' and 'White' types were also selected. These accessions came from 14 Autonomous Communities. The phenotypic variation of the selected group was similar to the one of the complete collection, including mean, range and coefficient of variation for the studies quantitative traits and also for the qualitative ones (data not shown). The parameters used to explore the molecular variability (N_a, N_e, H_0 and H_e) were comparable for each marker between the complete and the selected collection, validating the selection made. However, further validation of the selection made should be done with data not previously used for the selection of the accessions, such as other relevant phenotypic traits and or a different type of molecular markers, as suggested by van Hintum *et al.* (2000), Parra-Quijano *et al.* (2011) and Odong *et al.* (2013).

According to the results obtained, the use of the stratified method, first constructing groups based on phenotypic traits, maximizing the origin and uniqueness of the accessions, and checking then the genetic distances among the groups and subgroups established, seem to be useful for rationalizing the collections of this crop. The rationalization of the collection done will be useful for optimizing the management of the cucumber collection conserved at the COMAV genebank.

Acknowledgements

References

Areshchenkova T, Ganal MW, 1999. Long tomato microsatellites are predominantly associated with centromeric regions. Genome 42: 536-544. https://doi.org/10.1139/g98-155

Baker RH, Yu XB, DeSalle R, 1998. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. Mol Phylogenet Evol 9: 427-436. https://doi.org/10.1006/mpev.1998.0519

Botstein D, White RL, Skolnick M, Davis RW, 1980. Construction of genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331.

Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, Huang S, Weng Y, 2010. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 11: 569. https://doi.org/10.1186/1471-2164-11-569

Chen SX, Chen WF, Shen XQ, Yang YT, Qi F, Liu Y, Meng H, 2014. Analysis of the genetic diversity of garlic (*Allium sativum* L.) by simple sequence repeat and inter simple sequence repeat analysis and agro-morphological traits. Biochem Syst Ecol 55: 260-267. https://doi.org/10.1016/j.bse.2014.03.021

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK, 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142: 169-196. https://doi.org/10.1007/s10681-005-1681-5

Dijkhuizen A, Kennard WC, Havey MJ, Staub JE, 1996. RFLP variation and genetic relationships in cultivated cucumber. Euphytica 90: 79-87.

Doyle JJ, Doyle JL, 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19: 11-15.

Elameen A, Larsen A, Klemsdal SS, Fjellheim S, Sundheim L, Msolla S, Masumba E, Rognli OA, 2011. Phenotypic diversity of plant morphological and root descriptor traits within a sweet potato, *Ipomoea batatas* (L.) Lam., germplasm collection from Tanzania. Genet Resour Crop Evol 58: 397-407. https://doi.org/10.1007/s10722-010-9585-1

FAOSTAT, 2017. Food and agriculture data. Cucumber production. http://faostat3.fao.org

Fehet T, 1992. Watermelon. In: Genetic improvement of vegetable crops; Kaloo G, Bergh BO (eds.). pp: 295-314. Pergamon Press, NY.

Geleta N, Labuschagne MT, Viljoen CD, 2006. Genetic diversity analysis in sorghum germplasm as estimated by AFLP, SSR and morpho-agronomical markers. Biodivers Conserv 15: 3251-3265. https://doi.org/10.1007/s10531-005-0313-7

Horejsi T, Staub JE, 1999. Genetic variation in cucumber (*Cucumis sativus* L.) as assessed by random amplified polymorphic DNA. Genet Res Crop Evol 46: 337-350. https://doi.org/10.1023/A:1008650509966

Hu J, Zho X, Lia J, 2010a. Development of novel EST-SSR markers for cucumber (*Cucumis sativus*) and their transferability to related species. Sci Hortic 125: 534-538. https://doi.org/10.1016/j.scienta.2010.03.021

Hu J, Liang F, Liu L, Si S, 2010b. Genetic relationship of a cucumber germplasm collection revealed by newly developed EST-SSR markers. J Genet 89: 28-32.

Hu J, Wang L, Li J, 2011. Comparison of genomic SSR and EST-SSR markers for estimating genetic diversity in cucumber. Biol Plant 55: 577-580. https://doi.org/10.1007/s10535-011-0129-0

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, *et al.*, 2009. The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41: 1275-1281. https://doi.org/10.1038/ng.475

Knerr LD, Staub JE, Holder DJ, May BP, 1989. Genetic diversity in *Cucumis sativus* L. assessed by variation at 18 allozyme coding loci. Theor Appl Genet 78: 119-128. https://doi.org/10.1007/BF00299764

Kong Q, Xiang C, Yu Z, 2006. Development of EST-SSRs in *Cucumis sativus* from sequence database. Mol Ecol Notes 6: 1234-1236. https://doi.org/10.1111/j.1471-8286.2006.01500.x

Li X, Zhu D, Du Y, Shen D, Kong Q, Song J, 2004. Studies on genetic diversity and phylogenetic relationship of cucumber (*Cucumis sativus* L.) germplasm by AFLP technique. Acta Hortic Sin 34: 309-314.

Liu J, Qu J, Hu K, Zhang L, Li J, Wu B, Luo C, Wei A, Han Y, Cui X, 2015. Development of genomewide simple sequence repeat fingerprints and highly polymorphic markers in cucumbers based on next-generation sequence data. Plant Breed 134: 605-611. https://doi.org/10.1111/pbr.12304

Liu K, Muse SV, 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 1: 2128-2129. https://doi.org/10.1093/bioinformatics/bti282

Lv J, Qi J, Shi Q, Shen D, Zhang S, Shao G, Li H, Sun Z, Weng Y, Shang Y, *et al.*, 2012. Genetic diversity and population structure of cucumber (*Cucumis sativus* L.). PLoS One 7: 1-9. https://doi.org/10.1371/journal.pone.0046919

MAPAMA, 2017. Statistical Yearbook 2015. Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente, Gobierno de España. http://www.mapama.gob.es

McCreight JD, Nerson H, Grumet R, 1992. Melon. In: Genetic improvement of vegetable crops; Kaloo G, Bergh BO (eds.). pp. 267-294. Pergamon Press, NY.

Meglic V, Serquen F, Staub JE, 1996. Genetic diversity in cucumber (*Cucumis sativus* L.): I. A reevaluation of the U.S. germplasm collection. Genet Resour Crop Evol 43: 533-546. https://doi.org/10.1007/BF00138830

Mliki A, Staub JE, Zhangyong S, Ghorbel A, 2003. Genetic diversity in African cucumber (*Cucumis sativus* L.) provides potential for germplasm enhancement. Genet Resour Crop Evol 50: 461-468. https://doi.org/10.1023/A:1023957813397

Mu S, Gu X, Zhang S, Wang X, Wang Y, 2008. Genetic diversity of cucumber (*Cucumis sativus* L.) germplasm by SSR. Acta Hortic Sin 35: 1323-1330.

Nei M, 1978. Estimation of average heterozygosity and genetic distance from small number of individuals. Genetics 89: 583-590.

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL, 2013. Quality of core collections for effective utilization of genetic resources review, discussion and interpretation. Theor Appl Genet 126: 289-305. https://doi.org/10.1007/s00122-012-1971-y

Page RDM, 1996. Treeview: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358.

Panaud O, Chen SR, McCouch R, 1996. Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.). Mol Gen Genet 252: 597-607. https://doi.org/10.1007/BF02172406

Pandey S, Ansari WA, Mishra VK, Singh AK, Singh M, 2013. Genetic diversity in Indian cucumber based on microsatellite and morphological markers. Biochem Syst Ecol 51: 19-27. https://doi.org/10.1016/j.bse.2013.08.002

Parra-Quijano M, Iriondo JM, Torres E, De la Rosa L, 2011. Evaluation and validation of ecogeographical core collection using phenotypic data. Crop Sci 51: 694-703. https://doi.org/10.2135/cropsci2010.05.0273

Peakall R, Smouse P, 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28: 2537-2539. https://doi.org/10.1093/bioinformatics/bts460

Powell W, Machray GC, Provan J, 1996. Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1: 215-222. https://doi.org/10.1016/S1360-1385(96)86898-0

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, *et al.*, 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45: 1510-1515. https://doi.org/10.1038/ng.2801

Raghami M, López-Sesé AI, Hasandokht MR, Zamani Z, Moghadam MR, Kashi A, 2014. Genetic diversity among melon accessions from Iran and their relationships with melon germplasm of diverse origins using microsatellite markers. Plant Syst Evol 300: 139-151. https://doi.org/10.1007/s00606-013-0866-y

Reche J, 2011. Cultivo del pepino en invernadero. Ministerio de Medio Ambiente, Medio Rural y Marino, Gobierno de España.

Ren Y, Zhang Z, Liu J, Staub JE, Han Y, Cheng Z, Li X, Lu J, Miao H, Kang H, *et al.*, 2009. An integrated genetic and cytogenetic map of the cucumber genome. PLoS One 4 (6): e5795. https://doi.org/10.1371/journal.pone.0005795

Roldán-Ruiz I, van Euwijk FA, Gilliland TJ, Dubreuil P, Dillmann C, Lallemand J, De Loose M, Baril CP, 2001. A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (*Lolium perenne* L.) varieties. Theor Appl Genet 103: 1138-1150. https://doi.org/10.1007/s001220100571

Rubinstein M, Katzenellenbogen M, Eshed R, Rozen A, Katzir N, Colle M, Yang L, Grumet R, Weng Y, Sherman A, Ophir R, 2015. Ultrahigh-density linkage map for cultivated cucumber (*Cucumis sativus* L.) using a single-nucleotide polymorphism genotyping array. PLoS One 10 (4): e0124101. https://doi.org/10.1371/journal.pone.0124101

Staub JE, Serquen F, McCreight JD, 1997a. Genetic diversity in cucumber (*Cucumis sativus* L): III. An evaluation of Indian germplasm. Genet Resour Crop Evol 44: 315-326. https://doi.org/10.1023/A:1008639103328

Staub JE, Box J, Meglic V, Horejsi TF, McCreight JD, 1997b. Comparison of isozyme and random amplified polymorphic DNA data for determining intraspecific variation in *Cucumis*. Genet Resour Crop Evol 44: 257-269. https://doi.org/10.1023/A:1008639616331

Staub JE, Serquen FC, Horejsi T, Chen J, 1999. Genetic diversity in cucumber (*Cucumis sativus* L.): IV. An evaluation of Chinese germplasm. Genet Res Crop Evol 46: 297-310. https://doi.org/10.1023/A:1008663225896

Struss D, Plieske J, 1998. The use of microsatellite markers for detection of genetic diversity in barley populations. Theor Appl Genet 97: 308-315. https://doi.org/10.1007/s001220050900

Sun X, Xie Y, Bi Y, Liu J, Amombo E, Hu T, Fu J, 2015. Comparative study of diversity based on heat tolerant-related morpho-physiological traits and molecular markers in tall fescue accessions. Sci Rep 5: 18213. https://doi.org/10.1038/srep18213

Szewc-McFadden AK, Kresovich SK, Bliek SM, Mitchell SE, McFerson JR, 1996. Identification of polymorphic, conserved simple sequence repeats (SSRs) in cutlivated *Brassica* species. Theor Appl Genet 93: 534-538. https://doi.org/10.1007/BF00417944

Tatlioglu TP, 1992. Cucumber. In: Genetic improvement of vegetable crops; Kaloo G, Bergh BO (eds.). pp: 197-234. Pergamon Press, NY.

Thomas MR, Scott NS, 1993. Microsatellite repeats in grapevine reveal DNA polymorphisms when analysed as sequence-tagged sites (STSs). Theor Appl Genet 86: 985-990. https://doi.org/10.1007/BF00211051

Valcárcel JV, Peiró R, Pérez-de-Castro A, Díez MJ, 2018. Morphological characterization of the cucumber (*Cucumis sativus* L.) collection of the COMAV's genebank. Genet Res Crop Evol 65 (4): 1293-1306. https://doi.org/10.1007/s10722-018-0614-9

van Hintum TJL, Brown AH, Spillane C, Hodgkin T, 2000. Core collections of plant genetic resources, IPGRI technical bulletin no. 3. IPGRI, Rome, Italy. 51 pp.

van Treuren R, de Groot EC, Boukema IW, van de Wiel CCM, van Hintum TJL, 2010. Marker-assisted reduction of redundancy in a genebank collection of cultivated lettuce. Plant Genet Resour 8: 95-105. https://doi.org/10.1017/S1479262109990220

Watcharawongpaiboon N, Chunwongse J, 2008. Development and characterization of microsatellite markers from an enriched genomic library of cucumber (*Cucumis sativus*). Plant Breed 127: 74-81.

Wóycicki R, Witkowicz J, Gawroński P, Dabrowska J, Lomsadze A, Pawełkowicz M, Siedlecka E, Yagi K, Plader W, Seroczyńska A, *et al.*, 2011. The genome sequence

of the North-European cucumber (*Cucumis sativus* L.) unravels evolutionary adaptation mechanisms in plants. PLoS One 6 (7): e22728. https://doi.org/10.1371/journal.pone.0022728

Yang L, Koo D, Li Y, Zhang X, Luan F, Havey M, Jiang J, Weng Y, 2012. Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. Plant J 71: 895-906. https://doi.org/10.1111/j.1365-313X.2012.05017.x

Yang YT, Liu Y, Qi F, Xu LL, Li XZ, Cong LJ, Guo X, Chen SX, Fang YL, 2015. Assessment of genetic diversity of cucumber cultivars in China based on simple sequence repeats and fruit traits. Genet Mol Res 14: 19028-19039. https://doi.org/10.4238/2015.December.29.10

Zhou Q, Miao H, Li S, Zhang S, Wang Y, Weng Y, Zhang Z, Huang S, Gu X, 2015. A sequencing-based linkage map of cucumber. Mol Plant 8: 961-963. https://doi.org/10.1016/j.molp.2015.03.008

Zhu H, Song P, Koo DH, Guo L, Li Y, Sun S, Weng Y, Yang L, 2016. Genome wide characterization of simple sequence repeats in watermelon genome and their application in comparative mapping and genetic diversity analysis. BMC Genomics 17: 557. https://doi.org/10.1186/s12864-016-2870-4