ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization; 1–12
doi:10.1017/S1479262113000075

# Genetic characterization of sunflower breeding resources from Argentina: assessing diversity in key open-pollinated and composite populations

M. V. Moreno[1], V. Nishinakamasu[2], M. A. Loray[3], D. Alvarez[1], J. Gieco[1], A. Vicario[3], H. E. Hopp[2], R. A. Heinz[2,4], N. Paniego[2,4]† and V. V. Lia[2,4]*†

[1]*Estación Experimental Agropecuaria Manfredi, Instituto Nacional de Tecnología Agropecuaria (INTA), 5988 Manfredi, Córdoba, Argentina,* [2]*Instituto de Biotecnología, Centro de Investigación en Ciencias Veterinarias y Agronómicas (CICVyA), Instituto Nacional de Tecnología Agropecuaria (INTA), Nicolás Repetto y Los Reseros s/n B1686ICG, Hurlingham, Buenos Aires, Argentina,* [3]*Instituto Nacional de Semillas (INASE), 1095 Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina and* [4]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), 1033 Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina*

**Received 22 November 2012; Accepted 11 February 2013**

**Abstract**

Open-pollinated (OPs) and composite populations (CPs) represent a valuable resource for sunflower breeding programmes. However, little is known about the levels and distribution of genetic variation within each of these populations. In this study, quantitative and qualitative traits along with molecular markers were used to evaluate 14 populations from the Instituto Nacional de Tecnología Agropecuaria (INTA) sunflower germplasm collection. These populations were chosen to represent historically important accessions that still play a central role within the INTA sunflower breeding programme due to their extensive variation in diverse agronomically important traits. Nine quantitative and eight qualitative agro-morphological descriptors were recorded and compared with those of a larger set of accessions representative of the phenotypic diversity of the sunflower collection. Molecular characterization was conducted on a total of 311 individuals using 16 microsatellite markers. Overall, the average gene diversity was 0.56 and the average number of alleles per locus was 6.25. No statistically significant differences in genetic diversity were detected between the OPs and CPs. Global estimates of $F_{ST}$ revealed very high levels of differentiation among accessions ($F_{ST} = 0.413$, $P < 0.05$). Population structure analyses were consistent with the observed levels of differentiation and identified two major groups. The results of this work show that high global diversity is preserved within the accessions analysed here. Additionally, this study provides a set of reliable and discriminant markers for the cost-effective molecular characterization of sunflower accessions, along with the guidelines for the delineation of sampling strategies for OPs and CPs, thus aiding the efficient management and exploitation of sunflower germplasm collections.

**Keywords:** cultivated sunflower germplasm; genetic diversity; molecular markers; SSR

* Corresponding author. E-mail: vlia@cnia.inta.gov.ar
† These authors co-directed the work and share last authorship.

## Introduction

Cultivated sunflower (*Helianthus annuus* L.) represents one of the world's most important sources of edible oil, with a cultivated acreage of over 23 million ha (http://www.fao.org/; Bowers *et al.*, 2012).

Originally domesticated in central North America, sunflower became a well-established oil crop by the end of the 19th century when selection was practised by farmers in several parts of Russia (Skoric, 1992). By 1950, there was large-scale production in the Soviet Union and Argentina, where the crop first arrived via Jewish immigrants bringing small quantities of seeds from the south of Russia (Bertero de Romano and Vázquez, 2003). These initial populations gave rise to the first local commercial varieties (Klein, La Previsión and Massaux) and subsequently led to the development of several varieties at the Instituto Nacional de Tecnología Agropecuaria (INTA) by the introduction of early materials from Russia, Canada and Romania, as well as through introgression with wild *Helianthus* species. This diverse ensemble is collectively known as Argentinian germplasm; it has a distinct genetic constitution and is well adapted to local growing conditions. As in other sunflower-producing countries, hybrid seed had already gained wide acceptance among farmers by the mid-1970s, and they currently account for almost 80% of the total production of Argentina. Nowadays, Argentina is among the four largest sunflower producers worldwide and is the second oil exporter. Although the sunflower cultivated surface of the country has increased over the last 10 years, reaching at present ca. 2 million ha, its production has been shifted towards less favourable environments, emphasizing the need for the expansion of breeding efforts (De la Vega *et al.*, 2007).

The Active Germplasm Bank of INTA Manfredi (AGB-IM) has been evaluating and preserving sunflower genetic resources for more than 50 years. It holds ca. 1000 accessions of cultivated sunflower, which encompass a broad range of geographic origins. Cultivated materials have been generated by diverse breeding methods, with open-pollinated populations (OPs), composite populations (CPs) and inbred lines (ILs) accounting for 51, 9.8 and 39.2% of the collection, respectively.

Knowledge of the levels and distribution of genetic diversity in germplasm collections is of great importance for the conservation and utilization of genetic resources. However, exploitation of genetic variability preserved in germplasm collections has yet to be fully achieved, with sunflower being no exception to this pattern. Even though molecular markers have been extensively used for the genetic characterization of cultivated sunflower (e.g. Burke *et al.*, 2002; Paniego *et al.*, 2002; Garayalde *et al.*, 2011; Mandel *et al.*, 2011), the genetic diversity of Argentinian germplasm has been only poorly documented.

With the aim of uncovering the genetic diversity of sunflower OPs and CPs preserved at the INTA sunflower germplasm collection, this work intended to: (1) analyse its agro-morphological diversity based on 17 quantitative and qualitative traits; (2) assess its genetic diversity based on a set of microsatellite markers; (3) evaluate the relationship among the accessions; (4) define a set of highly informative microsatellites that can be routinely and efficiently applied in germplasm characterization.

## Materials and methods

### *Plant material*

The 14 OPs and CPs included in this study, hereafter the selected set, were chosen to represent historically important accessions that are currently used within the INTA sunflower breeding programme. The name of the accessions, breeding status, the country of origin and the number of individuals used for molecular analysis are presented in Table 1.

Briefly, KLM 209 OP was obtained from KLM selection, which was a multiple cross between Klein × local cultivars (a pool of local varieties from the INTA EEA Pergamino breeding programme) × 'Manfredi' (a pool of varieties from the INTA EEA Manfredi breeding programme). IAC-Anhandy is a Brazilian OP selected due to its high oil content (OC) and good performance in local environments. The Canadian variety Sunrise was one of the main contributors to the genetic background of Guayacán INTA and was also involved in the origin of Forestal cambá. Jupiter is a German OP that shows extensive variation for diverse agronomical traits such as yield, resistance to abiotic stress, etc. Puntano × Smena was derived from a cross between Puntano (Manfredi, Argentina) and Smena (Russia). Puntano was developed from cultivated and wild species of *Helianthus*, showing resistance to *Puccinia helianthi*. VNIIMK 8931-1 is a Russian OP. This accession was used as the background for the HA89 IL, which is one of the parental lines of a reference population for quantitative trait loci (QTL) mapping of resistance to *Sclerotinia sclerotiorum* (Maringolo, 2007), and also has commercial use as oilseed maintainer (Paniego *et al.*, 2002). Prao-Co was developed at INTA EEA Manfredi. It was created by intercrossing between a wide range of high-oleic commercial hybrids from Italy. Forestal cambá is a derivative of the variety 'Pehuén INTA', which was developed at INTA EEA Pergamino to obtain an early variety with a high OC and resistance to *P. helianthi*. HAR1 and HAR4 CP were derived from a genetic pool

**Table 1.** Sunflower accessions from the Active Germplasm Bank of INTA Manfredi included in the present study

| Accession | Category | Country of origin | Number of individuals |
|---|---|---|---|
| KLM 209 (KLM) | OP | Argentina | 24 |
| IAC-Anhandy (IAC) | OP | Brazil | 24 |
| Sunrise | OP | Canada | 19 |
| Jupiter | OP | Germany | 24 |
| Puntano × Smena (P × S) | OP | Argentina | 20 |
| VNIIMK 8931-1 (VNIIMK) | OP | Russia | 24 |
| Prao-Co | OP | Argentina | 24 |
| Forestal cambá (F. camba) | OP | Argentina | 15 |
| HAR1 | CP | Argentina | 24 |
| HAR2 | CP | Argentina | 20 |
| HAR3 | CP | Argentina | 14 |
| HAR4 | CP | Argentina | 32 |
| Comangir | CP | Argentina | 24 |
| Colliguay | CP | Chile | 23 |

OP, open-pollinated population; CP, composite population.

developed at the EEA Pergamino (INTA, Argentina) known as 'Mezcla Precoz' (Early Mix). The latter came from an intercross between lines from several Russian varieties and the wild species *H. annuus* ssp. *annuus*, *Helianthus petiolaris* and *Helianthus argophyllus*. The HAR1 and HAR4 lines derived from the respective CP are used as international differential lines for *P. helianthi*, and for *P. helianthi* and *Plasmopara halstedii*, respectively. HAR2 CP was derived from the variety Impira INTA (EEA Manfredi). The associated line HAR2, registered at the USDA, Fargo, ND (Gulya, 1985), was developed from this CP and is currently used as the international differential line for *P. helianthi* (Gulya and Maširevic, 1995). HAR3 CP was developed based on the 'Charata INTA' selection, which was obtained by interspecific crossings between Russian varieties and wild germplasm belonging to the species *H. annuus* ssp. *annuus*, *H. petiolaris* and *H. argophyllus*. 'Charata INTA' also showed resistance to *P. halstedii*. Comangir was derived from local germplasm and shows extensive variation for diverse agronomical traits such as yield and resistance to biotic stresses. Originally provided by Pioneer, it was subsequently developed at the INTA EEA Manfredi, Argentina. Colliguay is a Chilean accession that also shows extensive variation for agronomical traits and is well adapted to Argentinian growing conditions.

Furthermore, 309 accessions, hereafter the global set, were included in the study as a representative panel of the agro-morphological diversity of the AGB-IM.

### *Genomic DNA extraction and microsatellite genotyping*

DNA was extracted from the young leaves of 5-week-old plants grown in the greenhouse, using the NucleoSpin Plant II (Macherey-Nagel, Germany) genomic extraction kit. Sixteen microsatellite loci (simple sequence repeat (SSR)) were selected from a preliminary survey of 35 based on reproducibility, straightforward interpretation, the frequency of null alleles and the distribution across different linkage groups (Paniego *et al*., 2002; Poormohammad Kiani *et al*., 2007; Supplementary Table S1, available online only at http://journals.cambridge.org). PCR amplification and detection methods were performed as described previously (Poormohammad Kiani *et al*., 2007, Talia *et al*., 2010). A loading mix for capillary electrophoresis detection was prepared by multiplexing two SSR loci, as reported in Tang *et al*. (2003), based on allele fragment size compatibility, genotyping performance, allelic length range and map position. GeneMapper 4.0 software (Applied Biosystems, Foster City, CA, USA) was used to score SSR alleles.

### *Agro-morphological diversity analysis*

Eight trial sets were conducted during 2002–2008 at five experimental locations (Manfredi, Córdoba (31°49′12″S; 63°46′00″W); Balcarce, Buenos Aires (37°83′33″S; 58°25′63″W); Pergamino, Buenos Aires (33°51′00″S; 60°33′00″W), Reconquista, Santa Fé (29°14′48″S; 59°64′35″W) and El Colorado, Formosa (26°03′00″S; 59°36′67″W) as part of the sunflower germplasm bank activity. The experimental design used was alpha lattice with three replications. Each plot consisted of two 5.10 m rows with 70 cm apart between the rows. Seventeen seeds were sown per row; passport data were recorded for each accession.

Data for 17 morphological and agronomical traits were retrieved from the AGB-IM evaluation records for the 14 accessions included in this study (selected set)

and for 309 accessions representative of the agro-morphological diversity of the AGB-IM (global set). Nine quantitative traits were included in the data matrix: leaf width (LW); leaf length (LL); leaf area (LA); height to flowering (HF); head diameter (HD); stem diameter (StD); days to maturity (DM); weight of 100 seeds in g (100-SW); OC. Eight qualitative traits were also used for the analysis: leaf surface (LS); leaf shape (LSH); leaf margin (LM); head inclination or obliquity (HO), anthocyanin presence (AP); branching pattern (BP); seed colour (SC); seed stripes (SS). To account for the variation within the accessions, HO was decomposed into five binary characters according to the inclination angle of sunflower heads (ICD1 45°, ICD2 90°, ICD3 135°, ICD4 180°, ICD5 225°). Similarly, BP was divided into four binary characters (WB, without branching; BB, basal branching; AB, apical branching; TB, total branching). Agro-morphological diversity of the selected set of accessions was compared with that of the global set using descriptive statistics (i.e. mean, standard deviation, coefficient of variation) and principal component analysis (PCA) for quantitative traits, and the Shannon and Weaver (1963) diversity index and principal coordinate analysis (PCO) based on the simple matching index for qualitative traits. PCA and PCO were conducted using InfoGen (Balzarini *et al.*, 2010). Affiliations among the accessions were also assessed using cluster analysis. Mean values were used for quantitative characters and standardization was applied to remove unequal weights imposed by the use of different scales of measurement. A taxonomic distance matrix and an unweighted pair group method with the arithmetic average (UPGMA) dendrogram were obtained using NTSYS-pc 2.11 software (Rohlf, 2004).

### *Genetic diversity analysis*

Allele frequencies, the mean number of alleles per locus (*A*), allelic richness (Rs) (El Mousadik and Petit, 1996) and gene diversity (He) (Nei, 1972, 1987) were computed using FSTAT (Goudet, 2001). Estimates of observed heterozygosis (Ho) were obtained by direct count from the raw data matrix. The presence of private alleles was examined for each OP and CP. Null allele frequencies were estimated by direct count from raw data. Comparisons of genetic diversity indices between categories were conducted using the permutation test implemented in FSTAT.

To evaluate the discriminant power of our SSR panel, the probability of identity (PID) and the PID considering genetic similarity among siblings (PIDsib) were calculated across the complete data matrix, according to Waits *et al.* (2001), using GenAlEx (Peakall and Smouse, 2006). The polymorphism information content (PIC) (Anderson *et al.*, 1993) was computed using PowerMarker 3.25 (Liu and Muse, 2005).

To assess whether our sampling intensity allowed accurate estimation of genetic variability within each OP and CP, rarefaction curves were constructed for the number of alleles and gene diversity indices using the bootstrap resampling procedure implemented in InfoGen (Balzarini *et al.*, 2003) with 250 replicates per size category.

The relative contribution of each OP and CP to the pools of individuals maximizing genetic variation (i.e. the number of alleles) was estimated using the annealing algorithm implemented in the software PowerMarker 3.25 to build core sets of different sizes (10–50 individuals), with five repetitions and 1000 replicates for each size.

### *Population structure and genetic relationships*

Estimates of Wright's (1978) fixation indices were obtained according to Weir and Cockerham (1984) using FSTAT software (Goudet, 2001). Significance was determined using the randomization test implemented in the same package. Departures from Hardy–Weinberg proportions (HW) at individual loci were tested within each accession.

The Bayesian model-based approach of Pritchard *et al.* (2000) was used to infer population structure using the software Structure 2.2. The number of clusters evaluated ranged from 1 to 20. All simulations were run with a burn-in period length of $1 \times 10^5$ and a run length of $5 \times 10^5$ under the admixture model and correlated allele frequencies, with no prior information on the origin of individuals (Falush *et al.*, 2003). Ten replicate runs were performed for each $K$ to assess the variation of likelihood values. The most likely number of clusters was determined using the *ad hoc* criteria described in the software documentation (Pritchard and Wen, 2003) and the $\Delta K$ method of Evanno *et al.* (2005).

Population structure was also examined by applying the discriminant analysis (DA) of principal components (DAPC; Jombart *et al.*, 2010), a recently developed multivariate method designed to identify and describe clusters of genetically related individuals, which is free of assumptions about HW or linkage equilibrium, and thus more appropriate for the analysis of breeding materials. Briefly, the method relies on allele data transformation using PCA as a prior step to DA. DA defines a model in which genetic variation is partitioned into a 'between-group' and a 'within-group' component. Groups can be defined *a priori* (i.e. populations, collection sites, temporal affiliations, etc.) or can be inferred using first sequential $K$-means (Legendre and Legendre, 1998) and model selection. DAPC was performed using the

Adegenet package (Jombart, 2008) for R 2.10.1 software (R development Core Team, 2009). The function DAPC was executed using the clusters identified by *K*-means. The number of clusters was assessed using the function 'find.clusters', evaluating a range from 1 to 40. The optimal number of clusters was chosen on the basis of the lowest associated Bayesian information criterion. Three PCs were retained to construct discriminant functions.

Genetic distances were calculated between pairs of accessions using Nei's genetic distance (1972, 1987). A dendrogram was constructed using the neighbour-joining method (NJ; Saitou and Nei, 1987) and branch support was estimated by bootstrapping (1000 pseudoreplicates). Genetic distance calculations, cluster analyses and resampling were performed with PowerMarker 3.25. The resulting tree was visualized and edited with the program Fig-Tree v. 1.3.1 (Rambaut, 2006–2009).

## Results

### *Agro-morphological variation*

Descriptive statistics for the nine quantitative agro-morphological traits evaluated in this study are presented in Table 2. A large phenotypic variation was observed for all traits. The highest coefficients of variation were found for LA, 100-SW and HD. No statistically significant differences were detected between the selected and global sets of accessions in any of the evaluated traits (*t*-test, $P > 0.05$). Mean values of the accessions included in the selected set are presented in Supplementary Table S2 (available online only at http://journals.cambridge.org).

The Shannon–Weaver diversity index ($H'$) was used to measure phenotypic richness in qualitative characters. A low $H'$ indicates an extremely unbalanced frequency of classes for an individual trait and a lack of diversity. In the present study, most of the $H'$ values of the selected set were similar to those of the global set, with no statistical differences between the means of the two groups (Wilcoxon rank test, $P > 0.05$; Table 3). The traits ICD1 and AB showed no variation in the selected set. Details of the qualitative trait variation for the selected set are given in Supplementary Table S3 (available online only at http://journals.cambridge.org).

PCA revealed that the first three PCs accounted for 50.6, 13.7 and 11% of the total variability, respectively. The variation in PC1 was mainly associated with LW, LL, LA and OC, whereas the variation in PC2 was primarily determined by HD, DM and OC (Table 2). The PCA biplot showed no clear groupings, with the accessions included in the selected set being evenly distributed along both PCs (Fig. 1). The PCO of qualitative traits allowed discrimination of three main groups along the first axis. Again, the accessions of the selected set were interspersed along both axes and were represented in all the three groups.

The UPGMA dendrogram also showed the accessions of the selected set to be scattered among the accessions of the global set (Supplementary Fig. S1, available online only at http://journals.cambridge.org).

### *Genetic diversity*

A total of 311 sunflower individuals from the selected set were genotyped with 16 SSR markers, revealing the presence of 100 alleles. Missing data accounted for 4.86% of the data matrix. The average null allele frequency was 0.027 for OP and 0.020 for CP. Among the 100 alleles

**Table 2.** Quantitative trait diversity in the global and selected sets of accessions from the Active Germplasm Bank of INTA Manfredi

| | Global set (309 accessions) | | | Selected set (14 accessions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trait | Mean | SD | CV | Mean | SD | CV | *t*-Test | PC1 loadings | PC2 loadings |
| LW | 24.13 | 5.82 | 24.11 | 25.16 | 6.42 | 25.52 | Ns | 0.43 | 0.12 |
| LL | 23.36 | 4.84 | 20.73 | 25.3 | 4.6 | 18.2 | Ns | 0.43 | 0.11 |
| LA | 577.32 | 238.3 | 41.28 | 639.21 | 251.8 | 39.39 | Ns | 0.43 | 0.08 |
| HF | 143.35 | 28.83 | 20.11 | 155.33 | 30.25 | 19.48 | Ns | 0.34 | 0.1 |
| HD | 15.78 | 5.37 | 34.01 | 16.2 | 5.08 | 31.34 | Ns | 0.34 | −0.35 |
| StD | 237.45 | 76.98 | 32.42 | 236.53 | 52.46 | 22.18 | Ns | 0.32 | 0.05 |
| DM | 95.83 | 10.2 | 10.64 | 100.36 | 12 | 11.95 | Ns | 0.12 | 0.65 |
| 100-SW | 6.61 | 2.63 | 39.88 | 6.54 | 2.12 | 32.43 | Ns | 0.26 | −0.17 |
| OC | 38.7 | 6.22 | 16.07 | 40.16 | 5.14 | 12.79 | Ns | −0.14 | 0.62 |

SD, standard deviation; CV, coefficient of variation; PC1, principal component 1; PC2, principal component 2; LW, leaf width; LL, leaf length; LA, leaf area; HF, height to flowering; HD, head diameter; StD, stem diameter; DM, days to maturity; 100-SW, weight of 100 seeds in g; OC, oil content; Ns, non-significant ($P > 0.05$).

identified, 17 (17%) were private or exclusive of a given accession. The total number of alleles detected for OP and CP was 91 and 74, respectively. The average *A* across the whole set of individuals was 6.25 and the average He was 0.56. A summary of genetic diversity estimates per category and accession is presented in Table 4. The statistical comparison of genetic diversity indices (Rs, He and Ho) showed no significant differences between the OPs and CPs ($P > 0.05$).

In order to evaluate the adequacy of our SSR panel for a reliable characterization of the AGB-IM collection, PIC, PID and PIDsib values were calculated across the complete data matrix. The PIC values varied from 0.80 (HA1848 and HA2077) to 0.20 (HA3581), with an average of 0.50 (Supplementary Table S1, available online only at http://journals.cambridge.org). The single-locus PID values ranged from 0.052 (HA1848) to 0.62 (HA3581), whereas the PIDsib values ranged from 0.358 (HA1848) to 0.795 (HA3581). The PID product over all loci was $4.2 \times 10^{-11}$ and PIDsib was $3.4 \times 10^{-5}$. Reducing the SSR panel to the ten loci with the highest PID values yielded cumulative PID and PIDsib values of $5.26 \times 10^{-9}$ and $3.9 \times 10^{-4}$, respectively.

Core sets of individuals capturing the maximum number of alleles for a given sample size were obtained in order to assess the relative contribution of each accession to the overall variation. The 100 alleles identified at the 16 loci used in this study were fully represented by 90 individuals (28.93%). Of the 14 accessions analysed here, seven contributed to the core sets of ten individuals. For the core sets having more than 40 individuals, the distribution of accessions was relatively even, except for HAR3 and VNIIMK, which showed only a minor contribution for core sets under 80 individuals (Supplementary Table S4, available online only at http://journals.cambridge.org).

Rarefaction curves were constructed to represent the number of alleles and gene diversity as a function of sampling effort within each accession (Supplementary Fig. S2, available online only at http://journals.cambridge.org). In all cases, the number of alleles showed an incremental tendency with the increase of sample size, although slopes were markedly different depending on the accession. The curves of the OP Prao-Co, Jupiter, IAC, P × S, VNIIMK and F. camba reached a plateau at a number of individuals that ranged from 5 to 13, whereas 5 to 17 individuals were needed to enter the curvilinear phase for the CP Comangir, Colliguay, HAR1, HAR3 and HAR4. For KLM, Sunrise and HAR2 a steady increment was still apparent for the sample sizes considered in this analysis. Rarefaction curves of genetic diversity showed that all accessions reached a plateau around five to ten individuals.

### *Population structure*

Global fit to HW proportions was only observed for the composite HAR1. $F_{IS}$ values per locus and accession are presented in supplementary Table S5 (available online only at http://journals.cambridge.org). In most cases (KLM, IAC, HAR4, Colliguay, HAR3, P × S, Prao-Co,

**Table 3.** Shannon–Weaver diversity index (*H′*) for 15 qualitative descriptors in the global and selected sets of accessions

| Trait | Global set (309 accessions) | Number of variants | Selected set (14 accessions) | Number of variants |
|---|---|---|---|---|
| LS | 0.945 | 3 | 0.509 | 3 |
| LSH | 0.331 | 4 | 0.393 | 2 |
| LM | 0.214 | 2 | 0.245 | 2 |
| ICD1 | 0.55 | 2 | 0.000 | 1 |
| ICD2 | 0.289 | 2 | 0.500 | 2 |
| ICD3 | 0.665 | 2 | 0.500 | 2 |
| ICD4 | 0.663 | 2 | 0.691 | 2 |
| ICD5 | 0.213 | 2 | 0.393 | 2 |
| AP | 0.565 | 3 | 0.730 | 3 |
| WB | 0.499 | 2 | 0.257 | 2 |
| BB | 0.195 | 2 | 0.257 | 2 |
| AB | 0.223 | 2 | 0.000 | 1 |
| TB | 0.381 | 2 | 0.257 | 2 |
| SC | 0.653 | 4 | 0.730 | 3 |
| SS | 0.426 | 2 | 0.637 | 2 |
| Mean | 0.421 | 2.4 | 0.407 | 2.067 |

LS, leaf surface; LSH, leaf shape; LM, leaf margin; ICD1–ICD5, range of head obliquity; AP, anthocyanin presence; WB, without branching; BB, basal branching; AB, apical branching; TB, total branching; SC, seed colour; SS, seed stripes.

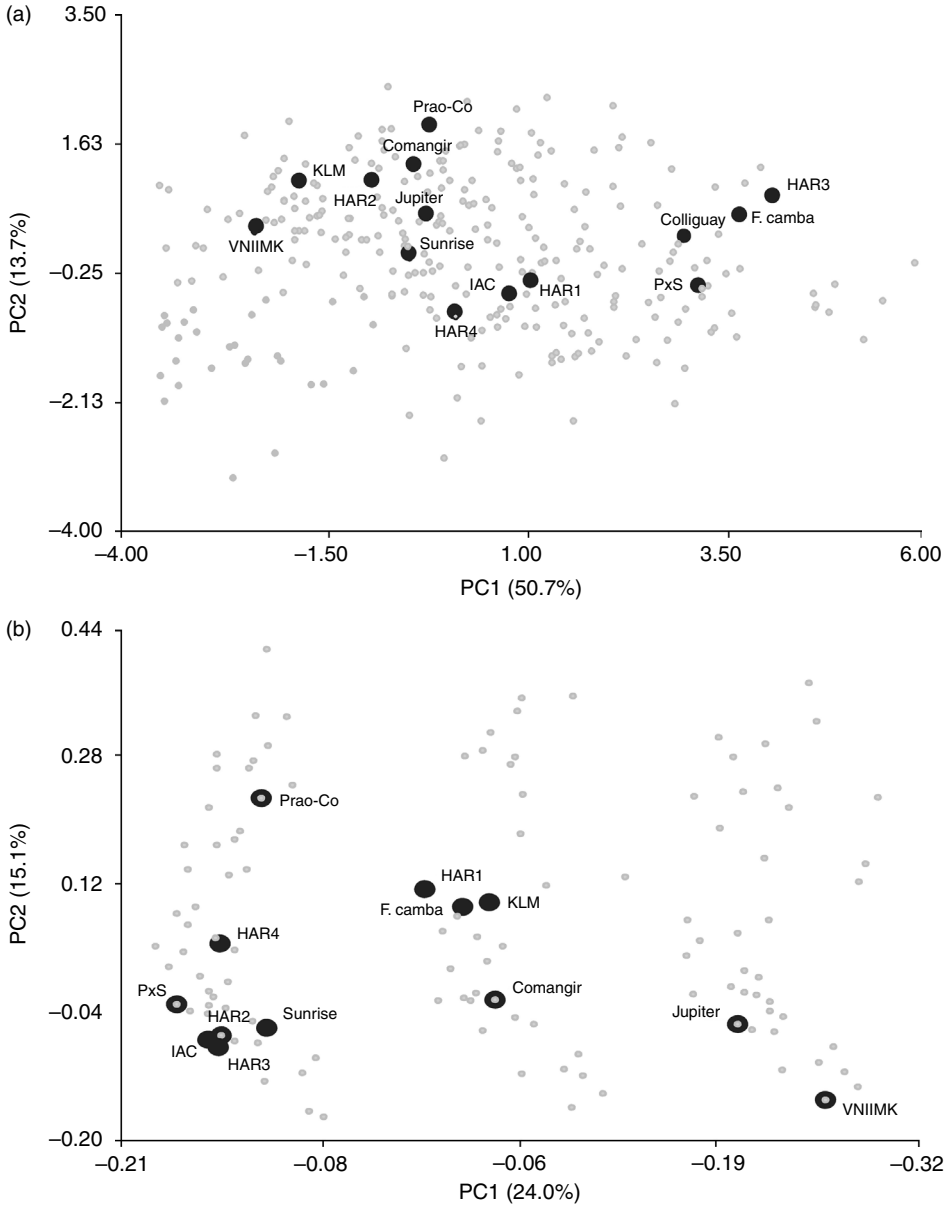


**Fig. 1.** Morphological variation of the sunflower accessions from the AGB-IM. (a) PCA based on the nine quantitative traits. (b) PCO based on the eight qualitative traits. Accessions from the selected set are indicated by large black circles.

VNIIMK, Jupiter, Sunrise, Comangir, HAR2 and F. camba), deviations were detected under the alternative hypothesis of heterozygote defect, with HA2077 being the only locus showing an excess of heterozygotes ($F_{IS} = -0.7388$; $P < 0.001$; Supplementary Table S5, available online only at http://journals.cambridge.org).

Global estimates of $F_{ST}$ revealed a considerable and statistically significant degree of differentiation among accessions ($F_{ST} = 0.413$; CI = 0.386–0.440). The largest pairwise $F_{ST}$ was found for VNIIMK-HAR1 ($F_{ST}$ 0.6394), whereas the least differentiated pair was Colliguay-Jupiter ($F_{ST}$ 0.1795) (Supplementary Table S6, available online only at http://journals.cambridge.org).

Bayesian analysis of population structure using the model-based Bayesian approach of Pritchard *et al.* (2000) provided further support to the existence of genetic structure in the set of accessions examined here. The log-likelihood values, Ln P(D), reached a plateau at $K = 14$, suggesting that genetic diversity is structured into 14 subpopulations. The method of Evanno *et al.* (2005) was also applied as a criterion to infer the most likely $K$ value. The maximum $\Delta K$ was detected at $K = 2$ (32.02), with a second maximum at $K = 14$ (1.55). Inspection of Structure outputs showed that at $K = 2$, the sample was divided into a group consisting of individuals from KLM, HAR4, HAR1 and HAR2, and a second group

**Table 4.** Genetic diversity estimates in the open-pollinated populations and composite populations from the Active Germplasm Bank of INTA Manfredi

| Accession | *A* | Rs | He | Number of private alleles |
|---|---|---|---|---|
| OP | | | | |
| KLM | 2.1875 | 1.475 | 0.26 | 0 |
| IAC | 3.3125 | 1.841 | 0.409 | 1 |
| Sunrise | 2.3125 | 1.558 | 0.298 | 1 |
| Jupiter | 3.25 | 1.923 | 0.456 | 3 |
| P × S | 2.8125 | 1.759 | 0.39 | 1 |
| VNIIMK | 2.0625 | 1.481 | 0.265 | 2 |
| Prao-Co | 3.0625 | 1.946 | 0.476 | 1 |
| F. camba | 1.75 | 1.491 | 0.277 | 3 |
| Mean | 2.5937 | 1.684 | 0.354 | |
| CP | | | | |
| HAR1 | 1.75 | 1.385 | 0.21 | 0 |
| HAR2 | 2.5625 | 1.661 | 0.353 | 3 |
| HAR3 | 1.9375 | 1.630 | 0.339 | 0 |
| HAR4 | 2.6875 | 1.651 | 0.336 | 1 |
| Comangir | 2.75 | 1.760 | 0.38 | 1 |
| Colliguay | 3.0625 | 1.884 | 0.433 | 0 |
| Mean | 2.4583 | 1.662 | 0.342 | |

*A*, mean number of alleles per locus; Rs, allelic richness; He, gene diversity.

composed of individuals from the remaining accessions (Fig. 2). At $K = 14$, the groupings detected at $K = 2$ are further subdivided into different clusters, matching almost exactly the accessions of origin, with high average membership coefficients. As an exception to this pattern, HAR4 seems to have received contributions from two different gene pools, which are the main constituents of HAR1 and HAR2, respectively. Similarly, two different gene pools were detected for Jupiter, although none of them is present in the other accessions. Prao-Co was the only accession showing substantial levels of admixture.

The results from DAPC also revealed the presence of genetically distinct groups among the accessions studied. When clustering individuals into the 14 groups identified by the successive *K*-means algorithm, the eigenvalues of the analysis showed that the genetic structure was captured by the first three PCS. KLM (cluster 8), HAR1 (cluster 12), HAR2 (cluster 1) and HAR4 (clusters 3 and 13) were clearly differentiated from the remaining composite and OPs, whereas no such distinction was apparent for the remaining accessions (Supplementary Fig. S3, available online only at http://journals.cambridge.org). Clusters 5 and 6 received contributions from seven and six accessions, respectively. Jupiter, Prao-Co, P × S and Sunrise were the most heterogeneous in terms of cluster assignment, with individuals distributed in three to four clusters (Supplementary Fig. S3, available online only at http://journals.cambridge.org).

The NJ reticulum based on Nei's genetic distances showed one major partition (bootstrap 81%; Supplementary Fig. S4, available online only at http://journals.cambridge.org). The two resulting clusters are in complete agreement with the uppermost level of structure detected by Bayesian analysis.

## Discussion

Argentina is one of the pioneers of sunflower production worldwide and has a long tradition of crop breeding. Over the years, valuable varieties, cultivars and ILs have been obtained and preserved at private and government institutions, among which the AGB-IM is a national active seed bank that holds both wild and cultivated sunflower accessions. As has been increasingly recognized for many crops, knowledge of the molecular and morphological diversity of germplasm collections is of fundamental importance for the efficient development of breeding programmes since it provides a reliable classification system and facilitates the identification of accessions with potential utility for specific traits (De la Vega *et al.*, 2007; Prada, 2009; Tanksley and McCouch, 1997).

The OPs and CPs included in the present work are an essential part of the INTA sunflower breeding programme and have contributed to the bank's diversity since its origins almost 50 years ago. Indeed, considerable levels of morphological variability were detected for the 14 accessions studied here. Moreover, comparison of quantitative and qualitative characters revealed no statistical differences between these accessions (selected set) and the global set of 309 accessions for which phenotypic data were available, suggesting that the former can be considered representative of a large proportion of the diversity spectrum of the entire collection. Consistent with this notion, the results from ordination and cluster analysis also showed the selected set to be evenly distributed among the remaining accessions (Table 2; Supplementary Fig. S1, available online only at http://journals.cambridge.org).

Most of the agro-morphological traits analysed here have been routinely included in many diversity surveys (i.e. Seiler, 1984; Alvarez *et al.*, 1992a, b; Belhassen *et al.*, 1994, Nooryazdan *et al.*, 2010, Kholghi *et al.*, 2011). In agreement with previous reports for wild sunflower populations (Seiler, 1984; Nooryazdan *et al.*, 2010), LA, 100-SW and HD exhibited the highest coefficients of variation. Accordingly, Nooryazdan *et al.* (2010) also found LL and HD to be among the main determinants of the distribution of accessions along the first and second PCs, respectively, with similar levels of variability accounted for by the first three components (ca. 80%).

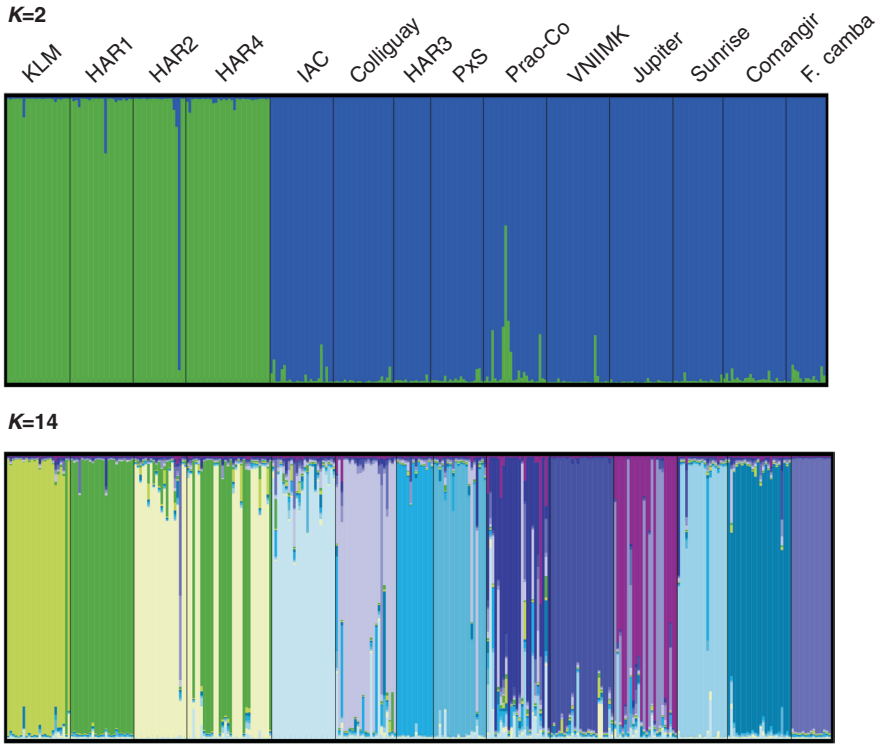


**Fig. 2.** Bayesian inference of the population structure. Each individual is represented by a thin vertical segment which can be partitioned into *K*-coloured segments that represent the individual estimated membership to the *K* cluster.

Molecular characterization of seed bank collections has become a fundamental tool for the evaluation of genetic resources. However, prior to the present study, no such information was available for the sunflower germplasm preserved at the INTA collection. The results from the SSR analysis revealed high levels of genetic diversity, particularly when considered in the context of variability estimates reported by Mandel *et al.* (2011) for a much larger array of sunflower cultivars. The average number of alleles per locus and gene diversity estimates obtained here across 311 individuals from 14 OPs and CPs (*A*: 6.25; He: 0.56) are remarkably similar to those found by Mandel *et al.* (2011) in the analysis of 1729 plants corresponding to 433 accessions from Europe and North America (*A*: 6.8; He:0.47). Although these figures might not be strictly comparable since they were derived from different sets of SSR, with the latter corresponding to expressed sequence tag (EST)-SSR which are expected to show lower levels of variation, they are still indicative of the allelic diversity of the OP and CP preserved at the AGB-IM.

In spite of their prominent role in most breeding programmes, little is known about the levels and distribution of genetic variation within OPs and CPs. The main difference between these categories lies in the degree of heterogeneity of the breeding materials used to generate them. While an OP might be the result of combining individuals from different varieties, races, ILs or even species, the production of a CP is often restricted to the use of highly homozygous individuals from different ILs, rendering them theoretically less diverse than an OP. Although the genetic diversity indices obtained for the OP studied here were indeed slightly higher than those of CP, no statistically significant differences were detected between categories, suggesting that lower or higher levels of allelic diversity are not an intrinsic attribute of the CP or OP, but rather depend on the divergence of the breeding materials used to generate them. In line with these observations, no detectable differences in allelic diversity were found among the 12 cultivar classes (HA-Non-Oil, HA-Oil, RHA-Non-Oil, RHA-Oil, etc.) delimited by Mandel *et al.* (2011), in which accessions with different improvement status were also included. As mentioned earlier, the OP and CP constitute ca. 60% of the AGB-IM collection and are thus a main concern in the delineation of sampling strategies for future molecular characterization efforts, particularly in terms of the number of individuals that need to be screened to fully represent within-accession diversity. As evidenced by the rarefaction curves, the sample sizes used in this work were large enough to reach a plateau for both the number of alleles and gene diversity in most of the accessions analysed, with no differences being apparent between the OPs and CPs

(Supplementary Fig. S2, available online only at http://journals.cambridge.org). A detailed inspection of the curves suggests that a number of individuals around 15 would suffice to capture most of the alleles present within accessions and that approximately ten individuals would be needed to reach maximum gene diversity. The difference in the number of individuals required to maximize the two genetic diversity estimates indicates that the frequencies of those alleles captured with sampling efforts larger than ten individuals are too low to produce noticeable increments in gene diversity. From a different perspective, core set analyses also suggest that a lower number of individuals per accession could be used in future studies of AGB-IM molecular diversity. Core sets including only 29% of the 311 individuals examined have been shown to harbour the same number of alleles as the total sample and 74% of the alleles were retrieved in the core sets of ten individuals. Furthermore, the distribution of accessions in the core sets of different sizes indicates that all of them have a fairly similar contribution to the overall variation.

Another relevant aspect of future molecular analyses is selecting a set of markers that allows for a rapid and cost-effective evaluation of large numbers of individuals. The average PIC value of the SSR included in this work (0.50) was similar to the estimates reported by Zhang *et al.* (2005) for 78 loci and 124 sunflower ILs (0.51) and by Paniego *et al.* (2002) for 170 SSR and 16 ILs. This last congruence was not unexpected given that the SSR used in this study are a subset of those developed by Paniego *et al.* (2002). Considering the 16 loci analysed here, the probability of two individuals having the same profile was less than one in 23 billion, with this probability rising to one in 30,000 under the extremely conservative assumptions of the PIDsib index. Interestingly, reducing the number of markers to ten by choosing the ones with the highest individual PID provided a discrimination power of one in 190 million when calculated from PID and of one in 2500 when calculated from PIDsib, which is largely within the range accepted in population genetic studies as an indication of the minimum number of loci required for reliable genetic tagging (Waits *et al.*, 2001, Peakall and Smouse, 2006). In sum, the ten SSR panel represents a fast and informative system to routinely characterize AGB-IM accessions in the context of applications such as fingerprinting, seed purity determination and contamination monitoring. From a broader perspective, the Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular, from the International Union for the Protection of New Varieties of Plants, has been discussing the use of DNA-based techniques over the last 18 years and provided the guidelines for their utilization in the management of reference collections, variety identification and examination of essentially derived varieties. The SSR panel described here may also serve to fulfil any of these purposes.

Population structure analysis of the accessions included in this study was consistent with the morphological differentiation evidenced by ordination and cluster analyses. High levels of genetic differentiation among accessions were apparent from both global and pairwise $F_{ST}$. The global $F_{ST}$ calculated here (0.43) almost doubles the estimate reported by Mandel *et al.* (2011) in the comparison between wild and cultivated sunflower (0.22). Moreover, the pairwise $F_{ST}$ between the 12 broad categories recognized by these authors were also noticeably lower (range 0.016–0.183) than those found for the set of accessions examined here (range 0.1795–0.6394). High $F_{ST}$ values indicate large differences in the allelic frequencies of the entities being compared. Demographic processes, such as sudden population contractions often experienced by breeding materials, may lead to an increase of $F_{ST}$ estimates, which is generally accompanied by a reduction of genetic diversity. The remarkable population structure of the accessions included in this study does not seem to be the consequence of this type of phenomenon since genetic variability estimates were rather high and similar to those of previous studies showing lower $F_{ST}$ values (e.g. Mandel *et al.*, 2011). In contrast, the observed pattern of differentiation is in agreement with the singularity of the so-called Argentinian germplasm and is most probably ascribed to its divergent origins, in both time and space.

In agreement with the observed levels of differentiation, Bayesian analysis of individual multilocus genotypes supported the existence of 14 ideal populations, matching almost exactly the original accessions. A more inclusive level of structure was also detected by the *ad hoc* procedure proposed by Evanno *et al.* (2005), revealing the presence of two main groups, which are completely concordant with the partition showing the highest bootstrap value in the NJ dendrogram. The first group is exclusively composed of accessions that are the product of local breeding programmes (KLM, HAR1, HAR2 and HAR4), whereas the second also includes germplasm of European and Canadian origin. Thus, no apparent geographic structure was discernable, neither were known pedigree relationships (Bertero de Romano and Vázquez, 2003). Because breeding populations often violate HW equilibrium, as demonstrated for the accessions included in this study, inferences drawn solely from model-based methods can be misleading. The results from DAPC were consistent with the existence of the two groups identified by Structure, although the DAPC approach provided a more detailed picture of within-group heterogeneity. The clusters representing the accessions

KLM, HAR1, HAR2 and HAR4 were clearly differentiated along the scatterplot axes. In contrast, the accessions of the second group, although assigned to different clusters, were substantially overlapped, suggesting higher genetic similarity among them.

The availability of a genetically structured sunflower germplasm collection might be of relevance for many breeding applications. Recently, Reif *et al.* (2012) examined several strategies to predict hybrid performance using phenotypic and molecular data from two groups of sunflower lines adapted to Central Europe and their respective inter- and intra-group hybrids. They concluded that prediction of hybrid performance for crosses based on genetically distant parents remains challenging given that the association between heterosis and genetic distance was restricted to related lines, and that genomic selection methods failed to provide accurate predictions for crosses among unrelated lines. Considering the moderate genetic differentiation among the lines studied by Reif *et al.* (2012) and that population structure, i.e. the relative contribution of intra- and inter-group components to the overall genetic variation, can have a profound impact on hybrid performance predictions, the set of accessions examined here may provide a valuable resource to assess the generality of the findings presented by these authors.

In summary, the accessions included in this study have shown to harbour significant levels of morphological and genetic variation. The results presented here provide a reliable and discriminant set of markers for the cost-effective molecular characterization of sunflower collections and offer the guidelines for the delineation of sampling strategies for OPs and CPs, thus aiding the efficient management and exploitation of these valuable genetic resources. Further efforts to integrate molecular and morphological data are currently underway to help establish a baseline for quality standards of sunflower germplasm collections based on objective indicators of genotypic and phenotypic diversity.

## Acknowledgements

The authors wish to thank Ing. Diego Cordes for technical assistance and two anonymous reviewers for their valuable comments on the manuscript. Several grants from the INTA (AEBIO 241001, 241352, 245001 and 245711) and the Agencia Nacional de Promoción Científica y Tecnológica (PAE 37100 PID 073) are gratefully acknowledged.

## References

Alvarez D, Ludueña P, Frutos E (1992a) Correlation and causation among sunflower traits. In: *Proceedings of the 13th International Sunflower Conference*, 7–11 September 1992, Pisa, Italy, pp. 957–962.

Alvarez D, Ludueña P, Frutos E (1992b) Variability and genetic advance in sunflower. In: *Proceedings of the 13th International Sunflower Conference*, 7–11 September 1992, Pisa, Italy, pp. 963–968.

Anderson JA, Churchill GA, Autrique JE, Tanksley SD and Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. *Genome* 36: 181–186.

Balzarini M, Di Rienzo J (2003) *InfoGen: Software para análisis estadístico de datos genéticos*. Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Córdoba.

Balzarini M, Bruno C, Peña A, Teich I and Di Rienzo J (2010) *Estadística en Biotecnología. Aplicaciones en InfoGen*. Córdoba: Encuentro Grupo Editor.

Bertero de Romano A and Vázquez AN (2003) Origin of the Argentine sunflower varieties. *Helia* 26: 127–136.

Belhassen E, Augé G, Ji J, Billot C, Fernandez-Martinez J, Ruso J and Vares D (1994) Dynamic management of genetic resources: first generation analysis of sunflower artificial populations. *Genetics, Selection, Evolution* 26: 241–253.

Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ and Burke JM (2012) Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *Genes-Genomes-Genetics* 2: 721–729.

Burke J, Tang S, Knapp SJ and Rieseberg LH (2002) Genetic analysis of sunflower domestication. *Genetics* 161: 1257–1267.

De la Vega AJ, De Lacy IH and Chapman SC (2007) Progress over 20 years of sunflower breeding in central Argentina. *Field Crops Research* 100: 61–72.

El Mousadik A and Petit R (1996) High level of genetic differentiation for allelic richness among populations of the argan tree (*Argania spinosa* (L.) Skeels) endemic to Morocco. *Theoretical and Applied Genetics* 92: 832–839.

Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14: 2611–2620.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.

Garayalde AF, Poverene M, Cantamutto M and Carrera AD (2011) Wild sunflower diversity in Argentina revealed by ISSR and SSR markers: an approach for conservation and breeding programmes. *Annals of Applied Biology* 158: 305–317.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available at http://www2.unil.ch/popgen/

Gulya TJ (1985) Registration of five disease resistant sunflower germplasm. *Crop Science* 25: 719–720.

Gulya TJ and Masirevic S (1995) Proposed methodologies for inoculation of sunflower with *Puccinia helianthi* and for disease assessment. In: *Studies on Common Methodologies of Artificial Inoculation and Population Dynamics of Sunflower Pathogens*. Rome: FAO European Research Network on Sunflower.

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405.

Jombart T, Devillard S and Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94.

Kholghi M, Bernousi I, Darvishzadeh Maleki R, Pirzad A and Hatami Maleki H (2011) Collection, evaluation and classification of Iranian confectionary sunflower (*Helianthus annuus* L.) populations using multivariate statistical techniques. *African Journal of Biotechnology* 10: 5444–5451.

Legendre P and Legendre L (1998) *Numerical Ecology*. 2nd English edn. Amsterdam: Elsevier.

Liu K and Muse S (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128–2129.

Mandel JR, Dechaine JM, Marek LF and Burke JM (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theoretical and Applied Genetics* 123: 693–704.

Maringolo CA (2007) Regiones cromosómicas asociadas a resistencia a Podredumbre Húmeda del Capítulo de girasol (*Sclerotinia sclerotiorum* (Lib.) de Bary). Magister Scientiae Thesis, Universidad Nacional de Mar del Plata.

Nei M (1972) Genetic distance between populations. *American Naturalist* 106: 283–292.

Nei M (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nooryazdan H, Serieys H, Baciliéri R, David J and Bervillé A (2010) Structure of wild annual sunflower (*Helianthus annuus* L.) accessions based on agro-morphological traits. *Genetics Resources in Crop Evolution* 57: 27–39.

Paniego N, Echaide M, Muñoz M, Fernandez L, Torales S, Faccio P, Fuxan I, Carrera M, Zandomeni R, Syarez EY and Hopp EH (2002) Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). *Genome* 45: 34–43.

Peakall R and Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6: 288–295.

Poormohammad Kiani S, Talia P, Maury P, Grieu P, Heinz R, Perrault A, Nishinakamasu V, Hopp E, Gentzbittel L, Paniego N and Sarrafi A (2007) Genetic analysis of plant water status and osmotic adjustment in recombinant inbred lines of sunflower under two water treatments. *Plant Science* 172: 773–787.

Prada D (2009) Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack? *Journal of Experimental Botany* 60: 2541–2552.

Pritchard JK and Wen W (2003) *Documentation for STRUCTURE Software: Version 2*. Chicago, IL: University of Chicago Press.

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

Rambaut A (2006–2009). Tree Fig. Drawing Tool Version 1.3.1. Institute of Evolutionary Biology, University of Edinburgh. Available at http://tree.bio.ed.ac.uk/

Reif JC, Zhao Y, Würschum T, Gowda M and Hahn V (2012) Genomic prediction of sunflower hybrid performance. *Plant Breeding*. doi: 10.1111/pbr.12007 132: 107–114.

Rohlf JF (2004) *NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System, Version 2.11*. Setauket, NY: Exeter.

R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406–425.

Seiler GJ (1984) Variation in agronomic and morphologic characteristics of several populations of wild annual sunflower. *Helia* 7: 29–33.

Shannon CE and Weaver W (1963) *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Skoric D (1992) Achievements and future directions of sunflower breeding. *Field Crops Research* 30: 231–270.

Talia P, Nishinakamasu V, Hopp HE, Heinz RA and Paniego NB (2010) Genetic mapping of EST-SSR, SSR and InDel to improve saturation of genomic regions in a previously developed sunflower map. *Electronic Journal of Biotechnology* 13: 6.

Tang S, Kishore VK and Knapp SJ (2003) PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci in cultivated sunflower. *Theoretical and Applied Genetics* 107: 6–19.

Tanksley SD and McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277: 1063–1066.

Waits LP, Luikart G and Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology* 10: 249–256.

Weir B and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Wright S (1978) *Variability Within and Among Natural Populations in Evolution and the Genetics of Populations*. Chicago, IL: University of Chicago Press.

Zhang LS, Le Clerc V, Li S and Zhang D (2005) Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. *Canadian Journal of Botany* 83: 66–72.