The LINNEAN SOCIETY of London BOTANICAL Journal of the Linnean Society

Botanical Journal of the Linnean Society, 2013, ••, ••–••. With 7 figures

Study of the polymorphism of the Patagonian *Calceolaria polyrhiza* (Calceolariaceae) using decision tree and sequential covering rule induction

MARINA M. STRELIN[1]*, ANDREA COSACOV[1], MARTIN DILLER[2] and ALICIA N. SÉRSIC[1]

[1]*Laboratorio de Ecología Evolutiva y Biología Floral, Instituto Multidisciplinario de Biología Vegetal (IMBIV), CONICET-Universidad Nacional de Córdoba, Av. Velez Sarsfield 299, Córdoba, Córdoba 5000, Argentina*
[2]*Facultad de Matemática, Astronomía y Física (FaMAF), Universidad Nacional de Córdoba, Av. Medina Allende s/n, Córdoba, Córdoba 5000, Argentina*

Received 12 June 2012; revised 8 November 2012; accepted for publication 15 June 2013

Climate and landscape changes that occurred in Patagonia during the Pleistocene presumably led to the ample biodiversity that currently characterizes the region. Many Patagonian species constitute unresolved complexes that may be related to these environmental changes. Accordingly, discriminating among taxonomic entities within an endemic and widely distributed species in Patagonia would enable a better understanding of the diversification processes that occurred in the recent past. We explored the polymorphism of *Calceolaria polyrhiza,* a widely distributed species for which phylogeographic studies suggest a recent diversification, with the aim of disentangling the morphological variability patterns in this species. We employed quantitative and qualitative floral and vegetative traits, including geometric morphometric variables. We used two clustering algorithms, testing for correspondence between detected clusters and the species previously described for the complex. Finally, we described these clusters using decision trees and sequential covering rule induction. Major clusters were determined, which match the previously described species. Floral traits related to corolla design and shape were the most useful characters for distinguishing groups and the variation patterns in these traits might be associated with historical and ecological factors. The need for conducting plant systematic studies using techniques that ensure objective taxonomic delimitations and descriptions is stressed. © 2013 The Linnean Society of London, *Botanical Journal of the Linnean Society*, 2013, **00**, 000–000.

ADDITIONAL KEYWORDS: cluster analysis – endemic species – floral traits – flower size – geometric morphometrics – infraspecific variation – leaf morphology – species complexes.

INTRODUCTION

The vast Patagonian region displays rich biodiversity of probably recent origin associated with the important climatic and landscape changes that occurred during the Pleistocene, such as glacial advances and retractions, fluctuations of the shoreline, orographic processes and volcanism (Ruzzante & Rabassa, 2011). These events led to species diversification processes in the region, promoting the establishment of species complexes or polymorphic phenotypes within species. Discerning the boundaries among these taxonomic entities is sometimes difficult and may lead to misestimation of the current biodiversity (Hewitt, 2000, 2004; Baylac, Villemant & Simbolotti, 2003; Morando, Avila & Sites, 2003).

The most common approach for resolving species complexes has typically relied on linear (or angular) morphometric variables. Nonetheless, the recent use of geometric morphometric techniques (Baylac *et al*., 2003; Mitteroecker & Gunz, 2004; Zelditch *et al*., 2004) in biological studies has proved useful. These techniques allow for a better detection of differences in shape and display higher statistical power than

*Corresponding author. E-mail: marina.strelin85@gmail.com traditional morphometric techniques. Geometric morphometrics has been frequently used in anthropological studies (González-José *et al*., 2008; Baab & McNulty, 2009; De Groote, 2011) and in studies using other animals as model organisms (Cordeiro-Estrela *et al*., 2008; Álvarez, Perez & Verzi, 2011; Von Reumont *et al*., 2012). However, it is a relatively novel tool in plant evolutionary research (Shipunov & Bateman, 2005; Gomez, Perfectii & Camacho, 2006).

Data exploration to classify and describe species complexes in the field of plant systematics has been generally carried out using principal component analysis, discriminant analysis (Ferreira da Costa, Pena Rodrigues & Wanderlay, 2009; Li *et al*., 2010) or only one type of clustering algorithm (Erxu *et al*., 2009; Li *et al*., 2010; Hatziskakis, Tsiripidis & Papageorgiou, 2011). Clustering algorithms can follow different strategies, therefore it is often useful to compare the outputs obtained with different clustering schemes. In the detection of subtle differences and similarities within taxa affected by a recent diversification event, decision trees, classification rules and other techniques derived from machine learning can be an alternative to discriminant analysis. Moreover, decision trees and classification rules enable us to automatically infer qualitative as well as simple quantitative descriptions of the groupings under study. As the algorithms use different schemes for data exploration and various parameters can be adjusted, the taxonomist can explore the data using different strategies in a practical way (Witten & Frank, 2005). Therefore, these classification algorithms are an interesting tool in systematic research. Neural network-based classifiers are one of the machine learning techniques that have been used for taxon delimitation (Baylac *et al*., 2003; Clark, 2009; Seiler & Keeley, 2009); however, although they are often successful, they are not based on a structured representation of the classified instances, which can give further insight into the entities being studied.

Calceolaria L. (Calceolariaceae, a south Antarctic element; Andersson, 2006; Cosacov *et al*., 2009). *Calceolaria* comprises approximately 250 species (Molau, 1988; Correa, 1998; Ehrhart, 2000) that occur from southern Mexico to Tierra del Fuego, along the Andes. Of the 42 Argentine *Calceolaria* spp., 16 occur in Patagonian steppe and Andean forest regions (Correa, 1998; Zuloaga & Morrone, 1999). The Patagonian steppe is a large (673 000 km^2), dry, extra-Andean plain covered by grassland and scrubby vegetation, which extends from the eastern slopes of the southern Andes to the Atlantic coast. The Andean-Patagonian forest is a much smaller region (248 100 km^2), covered with woodlands and extending on the eastern and western slopes of the Andes. Many Patagonian *Calceolaria* spp. form complexes with an unresolved systematics (i.e. Ehrhart, 2000, 2005), probably attributable to the recent diversification of the genus (Cosacov *et al*., 2009).

Different morphological traits and criteria have been used to delimit species and sections in the genus. Molau (1988) based his classification on growth habit, leaves, flowers and seed morphology, whereas Ehrhart (2000) employed traits related mainly to floral and seed morphology, as she considered that vegetative traits did not contain enough phylogenetic information. In particular, Ehrhart (2000) proposed the use of corolla shape, colour pattern and throat length for differentiating southern *Calceolaria* spp. However, she employed these traits in a qualitative way.

Calceolaria polyrhiza Cav. (Calceolariaceae) is a perennial rosulate herb distributed in Argentina from southern Mendoza province (35°S) to southern Santa Cruz province (52°S), with a latitudinal range of *c.* 2000 km. It is found from sea level up to 3000 m a.s.l. and tolerates diverse climatic conditions. The species is more abundant in the southern than in the northwestern area of its distribution, where it occurs in small, isolated populations in the high mountains. In Chile, the species is less abundant and is found in scattered locations from 35°S to 45°S. In the monograph of the genus in Chile, Ehrhart (2000) joined four related taxa, formerly considered different species, *C. polyrhiza, C. prichardii* (Rendle) Kränz., *C. lanceolata* Cav. and *C. mendocina* Phil., (Descole & Borsini, 1954; Correa, 1998), into a single species, *C. polyrhiza,* based mainly on a qualitative analysis of floral phenotypes; thus, the species in this broad sense is highly polymorphic, showing remarkable variation throughout its distribution range. In this work, we explored polymorphism in *C. polyrhiza*, employing quantitative and qualitative floral and vegetative traits, using geometric morphometrics and classical taxonomic variables. We propose new approaches to data exploration using different algorithms and describe detected clusters using machine learning techniques. In particular, our goals were (1) to determine whether major clusters of individuals can be identified across the distribution range of *C. polyrhiza*, (2) to establish whether these groups correspond to the species previously identified by Descole & Borsini (1954) and (3) to identify a set of rules that allows the differentiation of the detected clusters.

MATERIAL AND METHODS

SAMPLING

One hundred and seventy-one individuals of *C. polyrhiza* were collected from 18 localities covering most of its distribution range (Table 1, Fig. 1). Approxi-

Table 1. Collection localities, ecoregions, morphs, coordinates, altitude and sample size (Nind) of the sampled populations of *Calceolaria polyrhiza* in Patagonia. Localities (Nloc) are numbered consecutively, as shown on the map in Figure 1

N_{loc}	Sample location	Ecoregion	Morph	Latitude	Longitude	Altitude	N_{ind}
1	Valle Hermoso	Forest	Mendocina	−35.10	−70.14	2783	10
2	Epulaufquen	Forest	Mendocina	−36.84	−71.01	1555	10
3	Trolopé	Forest	Mendocina	−37.82	−70.96	1517	6
4	Piedra del Águila	Steppe	Lanceolata	−39.99	−70.04	667	10
5	Piltriquitrón	Forest	Prichardii–Polyrhiza	−46.98	−70.69	760	10
6	Cholila	Forest	Mendocina	−42.46	−71.61	580	9
7	Cholila abajo	Forest	Prichardii	−42.46	−71.61	580	9
8	La Tapera	Forest	Prichardii	−44.65	−71.73	556	9
9	Laguna Escondida	Forest	Prichardii	−45.52	−71.81	687	10
10	Balmaceda	Forest	Prichardii	−45.86	−71.83	590	9
11	Los Antiguos	Steppe	Lanceolata	−46.61	−71.64	415	10
12	Sumich	Steppe	Polyrhiza	−41.97	−71.48	1200	10
13	Lago Roca	Steppe	Polyrhiza	−50.48	−72.65	222	10
14	Lochiel	Steppe	Polyrhiza	−44.71	−66.12	348	10
15	Road 26	Steppe	Polyrhiza	−45.82	−67.97	677	10
16	Puerto Deseado	Steppe	Polyrhiza	−47.75	−65.92	9	9
17	Puerto Deseado 12 km	Steppe	Polyrhiza–Lanceolata	−47.75	−65.92	9	10
18	San Julián	Steppe	Polyrhiza–Lanceolata	−49.32	−67.77	7	10

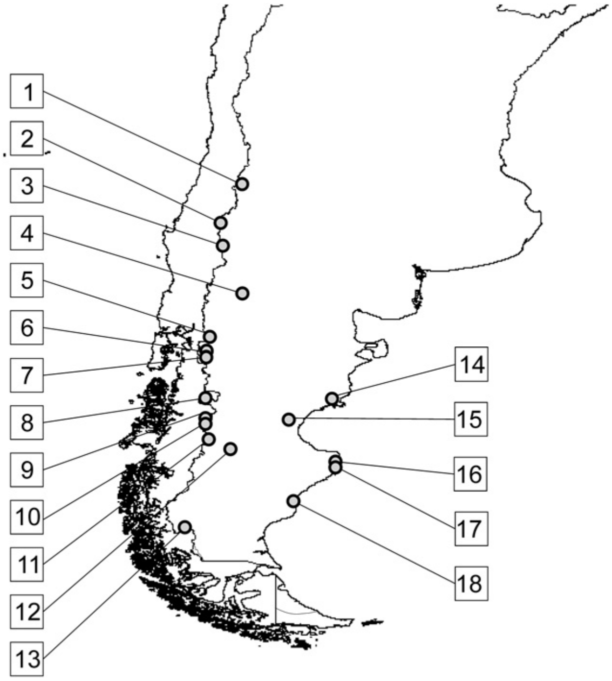


Figure 1. Geographical location of sampled *Calceolaria polyrhiza* populations. Numbers correspond to the localities shown in Table 1.

mately ten individuals were sampled from each of the 18 populations, with the aim of covering most of the variability in the species previously described by Descole & Borsini (1954). Throughout the text, we will refer to these previously described species as the morphs '*prichardii*', '*polyrhiza*', '*lanceolata*' and '*mendocina*'. Populations were selected, including almost all type localities of the above-mentioned morphs or sites close to them. Sampled individuals were at a minimum distance of 5 m in an attempt to avoid collecting clones or close relatives. One mature flower and the uppermost leaf were taken per individual sampled and kept in 70% ethanol and silica, respectively. Flowers were dissected to remove anthers, gynoecium and elaiophores. The dissected flowers and the leaves were mounted on a glass with dark background and scale photographed with a Nikon D80(B) digital camera. Fresh material corollas were photographed in lateral and frontal view; photographs of the frontal view of the corolla were used to describe corolla designs that are not visible in lateral view. Reproductive structures and adaxial view of leaves were photographed separately.

Traditional morphometrics

Six linear floral traits were measured (Fig. 2): instep length (IL), throat length (ThL), elaiophore width (EW), length of stamen filament (FL), length of the theca (TL) and style length (SL). These measurements were obtained from digitalized images using the ImageTool version 3 (Brent Dove, 1996–2002) software.

Geometric morphometrics

Geometric morphometrics allows the capture of spatial information through a Cartesian coordinate

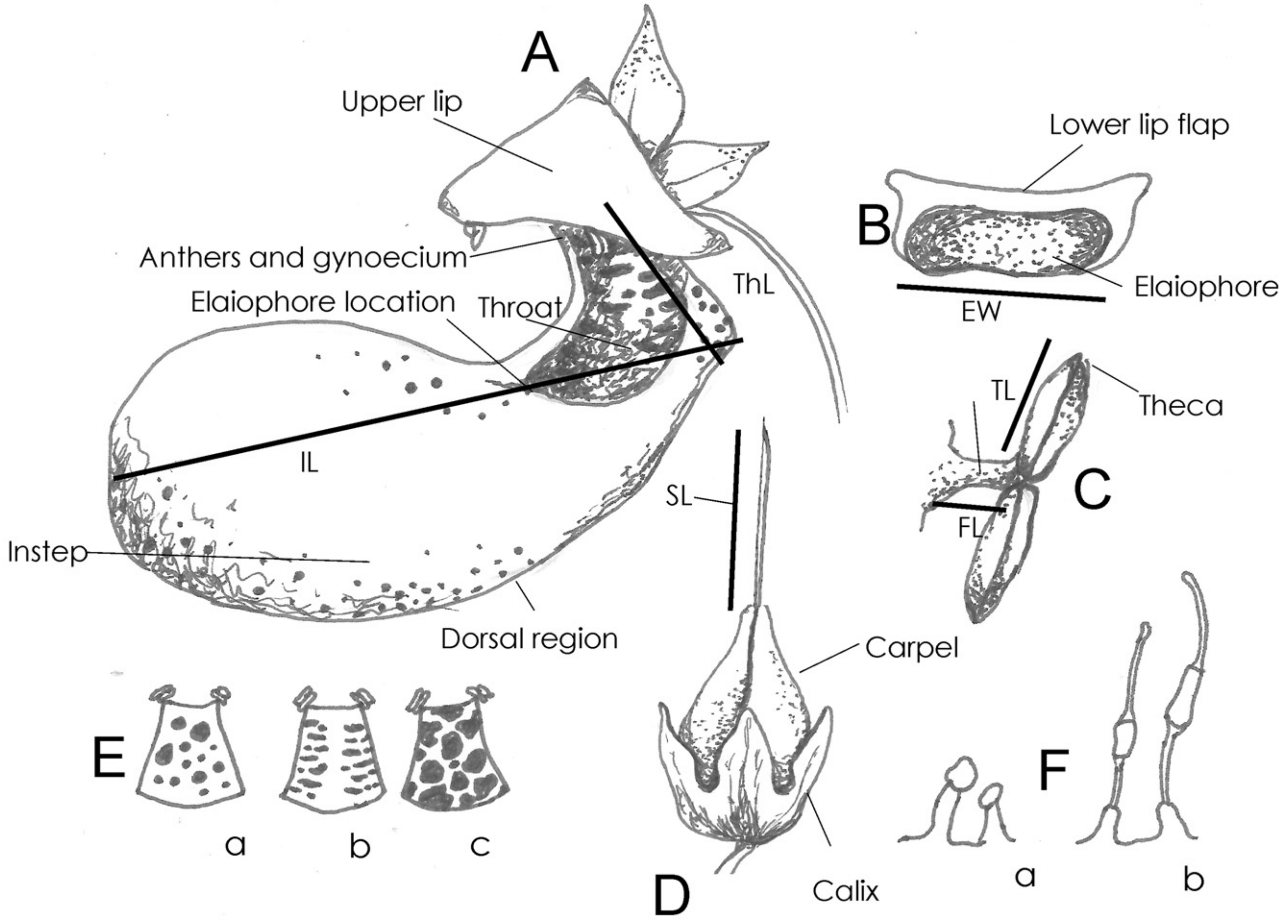


Figure 2. Drawing of a *Calceolaria polyrhiza* flower showing flower parts, traditional morphometric and qualitative traits. A, complete flower showing corolla measurements. B, elaiophore. C, stamen. D, gynoecium. E, corolla design: a, dots; b, strips; c, flecks. F, type of leaf trichomes: a, glandular; b, simple.

system; such information can be used to test statistical hypotheses about shapes (Zelditch *et al.*, 2004). Geometric morphometric variables were obtained after placing landmarks and semi-landmarks on photographs showing corollas in lateral view and leaves in adaxial view (see Appendix 1). Six homologous landmarks and nine semi-landmarks were placed on the flowers, and four homologous landmarks and eight semi-landmarks were placed on the leaves. Homologous landmarks were placed on the corolla and leaf locations that could be anatomically or mathematically identified, such as incision of the upper corolla lobe and point of maximum curvature in the leaf contour. Although the use of homologous landmarks is recommended over the use of semi-landmarks (Bookstein, 1996), the globular shape of corollas and the round contour of leaves required the use of a high number of semi-landmarks. The sliding semi-landmark method employed was the minimum bending energy criterion (Bookstein, 1996), which is more appropriate for obtaining thin-plate spline visualizations. Briefly, the minimum bending energy criterion is equivalent to the conservative assumption that the contour on a particular specimen, described with semi-landmarks, is the result of the smoothest possible deformation of the corresponding contour on a reference form (Perez, Bernal & Gonzalez, 2006). This reference form, called 'consensus', is the average landmark configuration of all the specimens included in the relative warp analysis (Zelditch *et al.*, 2004).

Both landmarks and semi-landmarks were digitalized using tpsDIG 2.10 (Rohlf, 2006), whereas tpsRelw 1.45 (Rohlf, 2006) was used to align the landmarks and semi-landmarks, and to obtain partial warps (PWs; components that explain part of the total deformation that affects some landmarks and not others) and relative warps (RWs; principal components of the PWs). Centroid size, the size measure used in geometric morphometrics (Zelditch *et al.*, 2004), was also obtained using the program

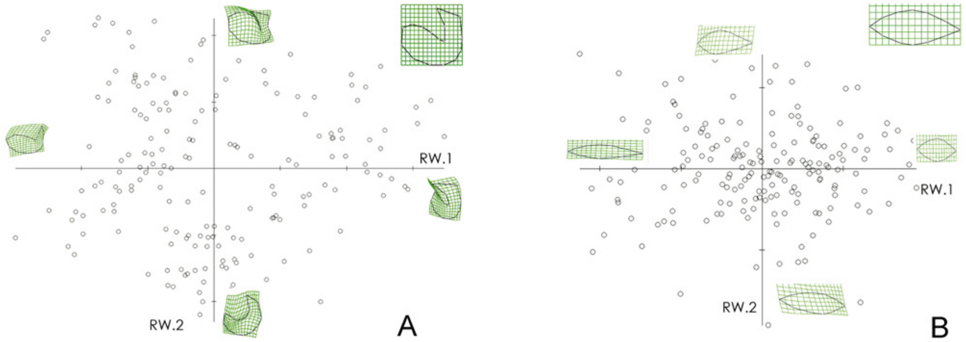


Figure 3. Variation along the two first relative warps (RWs) obtained from: (A) the corolla and (B) the leaf. Each circle represents an individual. The consensus (average) shape is displayed at the upper right corner. The first and second relative warp account for 46.20 and 26.75% of variation in corolla shape and for 44.70% and 25.42% of leaf shape variation, respectively.

tpsRelw 1.45 (Rohlf, 2006). Twenty-eight and 20 relative warps were captured for corolla shape and for leaf shape, respectively. Based on the scree test (Cattell, 1966), only the first three relative warps for corolla and leaf shape were used for subsequent analyses, which explained 72.95 and 87% of the total variance, respectively. Shape variation along the two first relative warps of the corolla and the leaf are shown in Figure 3.

Qualitative traits

Corolla design was described for the throat, the instep and the dorsal region of the lower lip (Fig. 2E). As a categorical vegetative trait, we selected the type of trichomes present on the abaxial side of leaves. Trichomes were uniform along the leaf lamina (Fig. 2F).

DATA ANALYSES

Delimitation of clusters in *C. polyrhiza*

Individuals were a priori classified as '*prichardii*', '*polyrhiza*', '*lanceolata*' and '*mendocina*' morphs before analyses, using the criteria of Descole & Borsini (1954), to check further for correspondence between the previously described species and the clusters retrieved by the different clustering schemes. Some individuals that could not be classified a priori were also included in the analyses.

To discard co-linearity between traditional and geometric morphometric variables, the condition number was calculated according to Quinn (2002). As this number suggested no redundancy between the traditional morphometric and geometric morphometric variables (condition number = 3.81; Quinn, 2002), all mentioned variables, for example, traditional and geometric morphometric variables and qualitative variables, were included in the subsequent analyses.

Two clustering algorithms, *k*-means (Witten & Frank, 2005) and expectation–maximization (EM; Witten & Frank, 2005) were employed to identify clusters in *C. polyrhiza*, including all above-mentioned variables and using the open source platform Weka (Hall *et al*., 2009). *k*-means (Witten & Frank, 2005) starts with *k* points chosen randomly in the hyperspace of instances in the data set and identifies the nearest neighbours to each of these points according to some distance measure (in this work, Euclidean distance). It then iteratively updates the *k* points to be the centroids of the partitions generated in the previous step and calculates the nearest neighbours to these points. The iteration continues until the same centroids are assigned to each of the clusters in consecutive steps of the execution of the algorithm.

EM assumes that the data set is generated by a mixture of *k* probability distributions (in this work, normal distribution) that represent the clusters and tries to estimate the parameters (mean and standard deviation) that describe these *k* distributions. EM begins by assigning random values to the parameters and then iterates the 'expectation' and 'maximization' steps: the parameter values used in the previous step are used to calculate the probability that each of the instances in the training set belongs to the clusters induced by the parameters and the distribution parameters are updated based on these assignments, maximizing the likelihood of the distributions given the data. The iteration continues until the increase in the overall log-likelihood of the data, given the estimated parameters become negligible for successive iterations.

For both clustering schemes, the number of clusters (*k*) was determined by increasing the number of clusters until the log-likelihood of the data given the clusters stabilizes or decreases. This was carried out running ten times tenfold validations and applying the corrected re-sampled *t*-test with 95% significance.

To determine whether clusters retrieved from the previous analyses differed significantly, we performed a multivariate analysis of variance (MANOVA) using the two axes derived from a non-metric multidimensional scaling (NMDS), which included the whole data set (qualitative and quantitative variables). The stress value of the analysis was 0.19, which means that the two-dimensional representation of the data is appropriate (Mendoça-Neto, Monteiro-Neto & Moraes, 2008). These analyses were performed using the PAST software (Hammer, Harper & Ryan, 2001). Finally, silhouette width values ('s'; Rousseeuw, 1987), a measure of cluster validation that indicates how tightly grouped all data in a given clusters are, were calculated for the cluster retrieved by k-means, $k = 5$ (see Results), using the R cluster package version 1.14.2 (Maechler, 2012). A high 's' value (almost 1) indicates that individuals highly match their respective cluster, whereas a low 's' value (almost 0) indicates the opposite.

Decision trees and classification rules

Classification algorithms, based on decision trees and classification rules, were used to find descriptions of the clusters retrieved in previous analyses. Although many classification algorithms are available in the field of machine learning, the following general structure can be summarized. The input data set (the 'training set') of this 'training phase' consists of the values associated with each attribute of interest for each of the individuals randomly sampled and the classification given to the individuals (in this case, 'cluster 1', 'cluster 2', etc.). From this 'training set' a hypothesis is generated that allows for the classification of the individuals with a certain degree of confidence. The accuracy of the hypothesis is usually evaluated on an independent set, the 'test set', using statistical methods. In the present work, the percentage of accurate classifications of the individuals in the 'test set' was used as the principal means of measuring performance.

In decision trees (Witten & Frank, 2005), each node corresponds to some attributes of the instance; the branches at the node correspond to the different values the attribute can take, in the case of categorical attributes, and the interval in which the value falls, in the case of numerical attributes. The algorithms employed for generating the decision trees used in the present work are implementations of variants of C4.5 (Quinlan, 1993), which select the attributes to be tested for each level of the decision tree, favouring those attributes that lead to a 'purer' partition of the training instances. For this, a quantitative measure of the degree to which the attributes reduce the 'entropy' of the training instances is used. As this criterion tends to favour attributes with many values over those with few values, a term is used to penalize these attributes.

In general terms, classification rules iteratively search the set of possible rules, from more general to more specific, for each class, adding attributes to be tested in the antecedent of the rule that maximize the p/t relation, where 'p' is the number of instances of the class correctly covered by the rule and 't' is the number of instances that satisfy the antecedent of the rule. This process continues until a rule is generated that correctly classifies all instances that satisfy the antecedent of the rule. In this work, the sequential covering rule generating algorithms RIPPER and PART (Witten & Frank, 2005) were used.

RESULTS

Delimitation of clusters in *Calceolaria polyrhiza*

The likelihood value for the number of clusters generated, both in EM and in k-means, did not increase significantly from four clusters onwards ($P < 0.05$). However, for k-means, likelihood increased to a considerable degree for $k = 5$, although this increase in likelihood value was not significant when comparing $k = 4$ and $k = 5$ configurations. For this reason, two k-means clustering configurations, $k = 4$ and $k = 5$, were considered in this study, together with the clustering configuration retrieved by EM. The correspondence between the clusters retrieved by these clustering schemes and the species (morphs) previously described by Descole & Borsini (1954) is shown in Figure 4.

Clusters obtained with the EM algorithm retrieved individuals that were mostly in correspondence with the morphs described by Descole & Borsini (1954) (e.g. individuals a priori classified as '*prichardii*', '*polyrhiza*', '*lanceolata*' and '*mendocina*' occurred in separate clusters, Fig. 4A). With k-means, $k = 4$ displays a different pattern, in which '*lanceolata*' and '*polyrhiza*' morphs are clustered together, and '*mendocina*' splits in two (Fig. 4B). k-means, $k = 5$ configuration summarizes to a large extent the patterns retrieved by k-means, $k = 4$ and EM (it distinguishes '*polyrhiza*', '*lanceolata*' and '*prichardii*' morphs, whereas it splits '*mendocina*' into two separate entities), as shown in Figure 4C. Independently of the clustering scheme employed, most of the individuals that could not be classified a priori were assigned to clusters composed mainly of '*mendocina*' individuals. Taking into account that the cluster configuration obtained with k-means, $k = 5$ summarizes to a large extent the cluster configuration retrieved by EM and k-means, $k = 4$, and that its likelihood value does not significantly differ from the one obtained for k-means, $k = 4$, result presentation and discussion will be mostly focused on k-means, $k = 5$ cluster configuration.

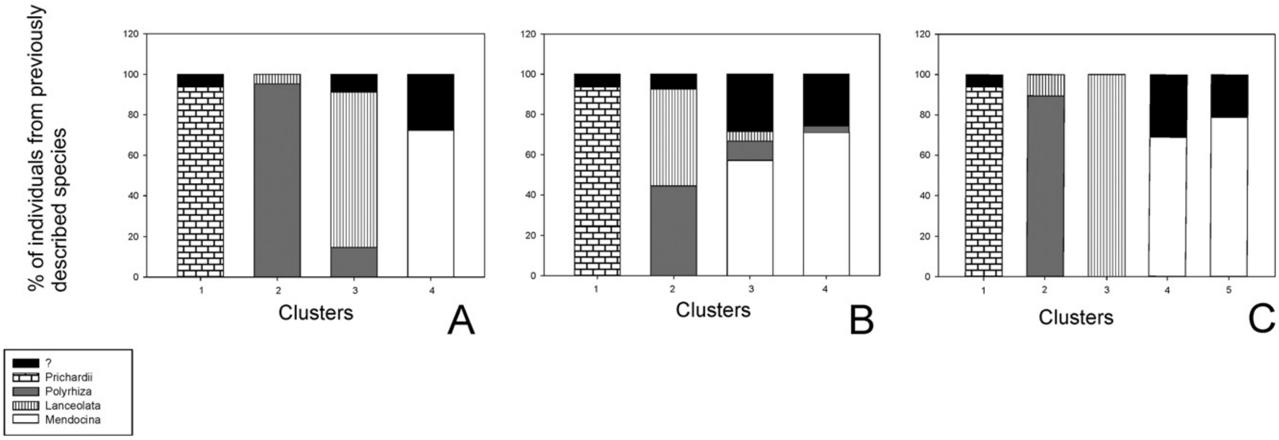


Figure 4. Correspondence between the previously described species and the clusters generated by (A) expectation–maximization (EM), (B) *k*-means, *k* = 4 and (C) *k*-means *k* = 5. '?' indicates non *a priori* classified instances.

MANOVA results and a posteriori Hotelling test (alpha = 0.05) indicated significant differences among groups for all clustering schemes ($F_{4;\,171}$ = 245.91, P < 0.0001; $F_{3;\,171}$ = 379.24, P < 0.0001; $F_{3;\,171}$ = 387.73, P < 0.0001 for *k*-means, *k* = 5, EM and *k*-means, *k* = 4 respectively).

Clusters obtained with *k*-means, *k* = 5 are plotted against the two axes of a NMDS in Figure 5. Silhouette width values were low ('s' < 0.4) for this cluster configuration (Fig. 6), indicating a low correspondence among individuals and their respective clusters.

Decision trees and classification rules

Decision trees and classification rules were highly efficient for *k*-means, *k* = 4, *k* = 5, and for EM clustering schemes. For *k*-means, *k* = 4 accuracy for the decision tree and classification rule was 96.33%, (SD = 4.51) and 95.32% (SD = 4.31), respectively. For *k*-means, *k* = 5 accuracy for the obtained decision tree and classification rule was 96.66% (SD = 4.35%) and 93.91%, (SD = 5.05%), respectively, and for EM accuracy was 94.92%, (SD = 5.59) and 94.31%, (SD = 5.35) for the decision tree and classification rule, respectively. Decision trees and classification rules obtained for *k*-means, *k* = 4; *k*-means, *k* = 5; and EM were congruent to a large extent (Appendix 2). The decision tree and the classification rule obtained for *k*-means, *k* = 4 clusters and those generated for *k*-means, *k* = 5 and EM selected unambiguously floral traits (Fig. 7, Appendix 2) instead of vegetative traits. Otherwise, only floral traits appeared in the decision trees and classification rules, although input data also included vegetative traits. Contingency tables generated for these classification algorithms are included in Appendix 3. Both classification rules and decision trees generated for *k*-means, *k* = 5 and EM retrieved coincident descriptions for clusters 1, 2 and 3. Cluster 1 (which contains all '*prichardii*' individuals) presents a corolla design with flecks; individuals retrieved in cluster 2 (mainly '*polyrhiza*' individuals) and cluster 3 (mainly '*lanceolata*' individuals) differ in corolla shape (e.g. individuals in cluster 3 display a more inflated instep than individuals in cluster 2). The descriptions generated by the classification algorithms retrieved

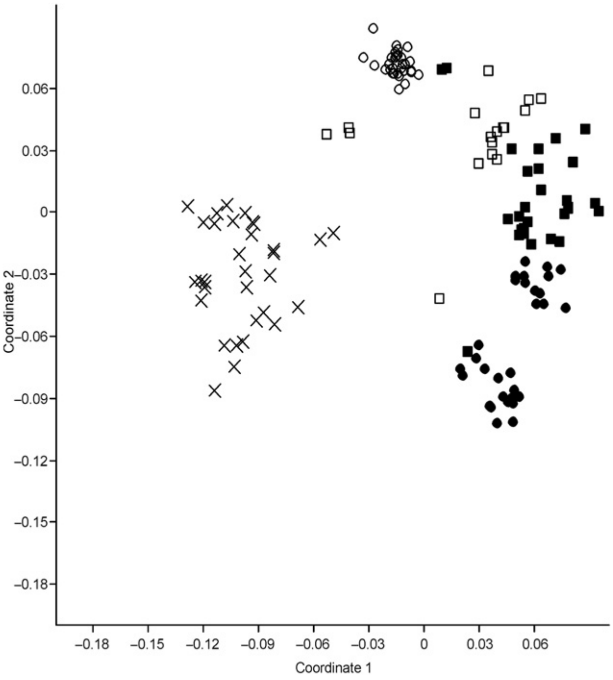


Figure 5. Clusters generated with *k*-means, *k* = 5, using axes 1 and 2 obtained with non-metric multidimensional scaling (NMDS). Individuals in clusters 1–5 in Figure 4C are represented with a cross, an open square, a solid square, a solid circle and an open circle, respectively.

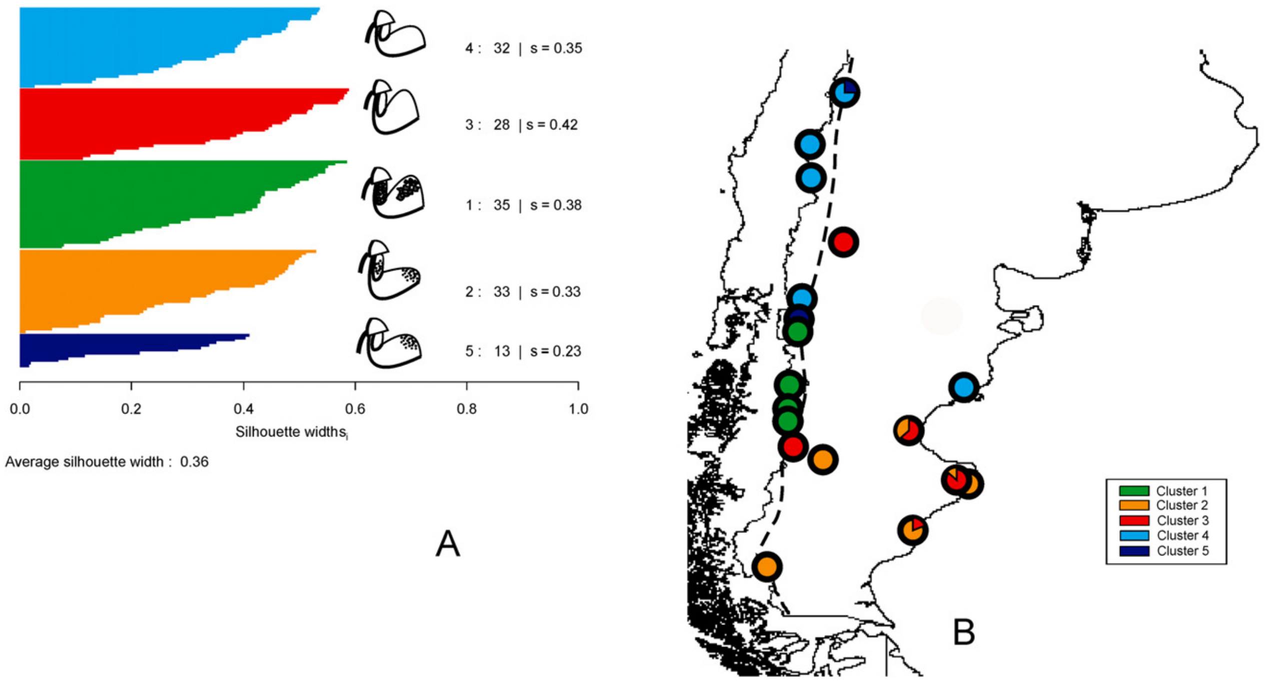
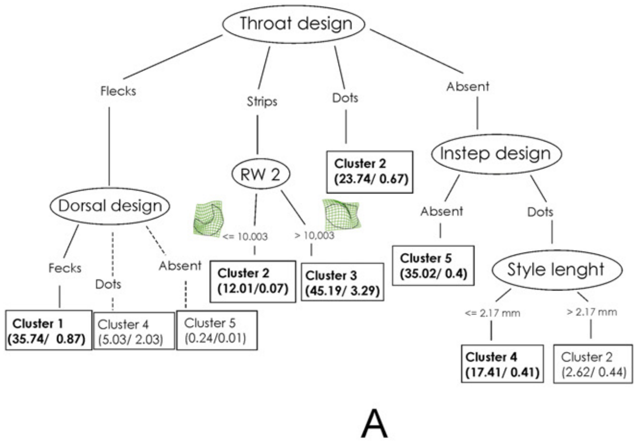


Figure 6. Silhouette plot for *k*-means, *k* = 5 clusters. The length of the bar on the y-axis is the silhouette width of each individual and represents the degree to which each individual is clustered to the other individuals in the cluster. On the right, a main characterization of flower phenotype, cluster number, number of individuals and average silhouette value for each cluster are indicated. Observations with a large ‘s’ value (almost 1) are very well clustered, a small ‘s’ value (*c.* 0) means that the observation lies between different clusters (A). Geographical distribution of the clusters generated with *k*-means, *k* = 5. The proportion of individuals assigned to each cluster within each population is indicated on the map. Individuals belonging to clusters 1, 4 and 5 predominate in Andean forest, whereas individuals in cluster 2 and 3 predominate in the steppe ecoregion (B). The border between the steppe and the Andean forest is indicated with a dotted line.



IF dorsal design is absent AND instep design is flecks THEN cluster 4 (17.0/0.0)
IF instep design is dots AND dorsal design is dots THEN cluster 4 (3.0/0.0)
IF throat design is flecks THEN cluster 1 (35.0/0.0)
IF instep design is absent THEN cluster 5 (38.0/2.0)
IF dorsal design is dots THEN cluster 2 (41.0/5.0)
IF throat design is dots THEN cluster 2 (3.0/0.0)
IF none of these conditions is met THEN cluster 3 (40.0/1.0)

B

Figure 7. Decision tree generated for *k*-means *k* = 5 clusters. Dotted lines represent those branches leading to a low number of classified individuals. Groups in which most individuals were classified are in bold. Numerators indicate the number of correctly classified instances; denominators indicate wrongly classified instances. Note that these numbers are not integers: the trees are evaluated in many classification instances, these numbers being averages of the results obtained in all these iterations (A). Classification rules generated for 5-means. Groups in which most individuals were classified are in bold. Numerators indicate the number of correctly classified instances; denominators indicate incorrectly classified instances (B).

for the different cluster configurations also presented some inconsistencies. For example, the decision tree and the classification rule generated for *k*-means, *k* = 5 recognizes the presence of spots as a distinctive trait of cluster 2 (mainly '*polyrhiza*' individuals), whereas the classification algorithms generated for EM do not distinguish this cluster based on this trait. In *k*-means, *k* = 5, according to decision trees, clusters 4 and 5 (mainly '*mendocina*' individuals) differ from the other clusters in the lack of design in the throat and differ from each other in the presence or absence of a dot design on the instep; classification rules generated for EM indicated the absence of design on different parts of the corolla as a distinctive characteristic for cluster 4 (e.g. clusters 4 and 5 in *k*-means, *k* = 5), without specifically indicating the absence of throat design. According to classification rules generated for the EM, cluster 4 also displays the shortest throat, whereas none of the classification algorithms generated for *k*-means, *k* = 5 makes this distinction. In some cases, style length was the trait selected for distinguishing among clusters. Nevertheless, in these distinctions only a small group of individuals is involved (Fig. 7, Appendix 2). Classification rules and decision trees generated for 4-mean clusters (Appendix 2) display some consistencies with those generated for EM and *k*-means, *k* = 5 clustering schemes. For example, a fleck corolla design is also attributed to '*prichardii*' individuals, whereas most of the individuals in clusters 3 and 4 ('*mendocina*' individuals) lack throat design. A long throat was assigned to cluster 2, which is composed of '*polyrhiza*' and '*lanceolata*' individuals.

A summary of the main characterizations obtained for the *k*-means, *k* = 5, *k*-means, *k* = 4 and for EM cluster configurations is shown in Figure 6, along with the geographic distribution of *k*-means, *k* = 5 clusters.

DISCUSSION

The results obtained with different clustering algorithms (*k*-means and EM) display clusters with a similar composition of individuals. This composition was recovered in all or in two of the three clustering schemes used (Fig. 4) Otherwise, clustering methods that make different assumptions point to the same result, allowing us to draw objective conclusions about partitions in the *C. polyrhiza*. complex.

The classifications offered by the decision trees and the classification rules were highly efficient and to a large extent congruent when comparing the rules and the trees obtained for different clustering schemes (Fig. 7, Appendix 2). This stresses the usefulness of this set of methodologies when dealing with taxonomic entities that may have experienced recent diversification events and therefore display only subtle morphological differences. At the same time, considering that both classification algorithms led independently to similar results, the set of traits selected by them may be considered objective and therefore reliable for classifying and describing taxonomic entities in *C. polyrhiza*.

In the classifications performed with the whole data set, classification algorithms selected floral traits unambiguously (Fig. 7, Appendix 2). This highlights the usefulness of floral traits when distinguishing entities in *C. polyrhiza*. The most recurrent set of traits used in the classifications included corolla design and the variable shape, which indicates the degree of inflation of the instep (RW2). Ehrhart (2000) proposed the above-mentioned traits for making a new classification of some southern sections of *Calceolaria*. Although her proposal only consisted of making qualitative descriptions of the sections in *Calceolaria* using these characters, our findings support the usefulness of these traits when making classifications within a species. The fact that corolla centroid size was never selected by classification algorithms is also remarkable. This variable is not expected to have a strong phylogenetic signal because it may be more subjected to ontogeny or microenvironmental factors than floral shape or coloration (i.e. Hodgins & Barret, 2008; Gómez *et al*., 2009). When classification rules and decision trees were performed without including floral traits (results not shown), vegetative traits also generated efficient classifications, the type of trichome being the most appropriate trait for distinguishing clusters and previously described species. Ehrhart (2005) proposed this trait, among a few others, to distinguish taxonomic entities in the *C. integrifolia* L. complex. However, in the present study it is clear that floral traits were selected over vegetative ones, suggesting that they are more appropriate to resolve the *C. polyrhiza* complex. In a broader context of plant biology, the efficiency of floral over vegetative traits to discriminate entities within *C. polyrhiza* is consistent with previous findings (i.e. Berg, 1959; Méndez & Traveset, 2003; Mascó, Noy-Meir & Sérsic, 2004; Medrano, Castellanos & Herrera, 2006) regarding the environmental stability of floral over vegetative traits.

A high correspondence (EM clusters) or at least some correspondence (*k*-means clusters, *k* = 4) was detected between the retrieved clusters and the four species previously described (Descole & Borsini, 1954) (Fig. 4); however, according to *k*-means, *k* = 5, which summarizes to a large extent the pattern exhibited in EM and *k*-means, *k* = 4, notably the entities that constitute *C. polyrhiza* are five instead of four.

Evidently, those individuals classified a priori as '*prichardii*' were the ones most clearly differentiated within *C. polyrhiza*. This entity differs from the others in bearing a fleck design on different parts of the corolla. Morphs '*lanceolata*' and '*polyrhiza*' tended to split into separate clusters in EM and in *k*-means, *k* = 5 (Fig. 4A, C). Nevertheless, individuals a priori classified under these morphs were assigned to a single cluster in *k*-means, *k* = 4 (Fig. 4B). Therefore, there is no evidence showing '*lanceolata*' and '*polyrhiza*' to be clearly differentiated entities. The entities '*polyrhiza*' and '*lanceolata*' differ in corolla shape, '*polyrhiza*' displaying a less inflated instep than '*lanceolata*'. Although the previously described species '*mendocina*' was recovered as a single entity in EM (Fig. 4A), it was split in two in *k*-means, *k* = 4 and in *k*-means, *k* = 5 (Fig. 4A, C). Therefore, the possibility of a slightly differentiated additional entity within '*mendocina*' could be suggested ('*mendocina* 1' and '*mendocina* 2'). Both '*mendocina*' groups differ from the other entities in not having a throat design, whereas they differ in instep designs, '*mendocina* 1' having a dot design and '*mendocina* 2' lacking a design (Fig. 7, Appendix 2B).

The clusters detected in this study display a geographical pattern (steppe vs. Andean forest distribution), as can be seen in Figure 6. Only one cluster occurs in both regions. According to classification rules generated for EM (Appendix 2A), individuals inhabiting the Andean forest (except for '*prichardii*' individuals, nested in cluster 1) have a shorter throat than those inhabiting the steppe. Classification rules and decision trees generated for *k*-means, *k* = 4 attributed a long throat to individuals inhabiting the steppe. Throat length is critical for the pollination process; an association between the most frequent pollinators and mean throat length of *C. polyrhiza* populations from the forest and the steppe has been reported (Cosacov, 2010). Moreover, pollinators and geographical isolation were reported as important drivers of diversification in *Calceolaria* (Molau, 1988; Sérsic, 2004; Cosacov *et al*., 2009). Accordingly, in their phylogeographical study of the *C. polyrhiza* complex, Cosacov *et al*. (2010) suggested that the evolutionary history of the species shows a phylogeographical footprint consistent with past fragmentations and allopatric differentiation during Pleistocene glaciations.

Silhouette width values were low for all clusters obtained with different clustering schemes (Fig. 6) and, despite the high efficiency of classification rules and decision trees in generating classification in *C. polyrhiza*, a certain percentage of error was associated with the classifications. This could suggest that the entities in *C. polyrhiza* would still not be clearly differentiated, being in an incipient differentiation process, or that they would have been clearly differentiated in the recent past because of historical processes (e.g. Pleistocene glaciations) and might be currently interbreeding again. Further ongoing studies combining molecular markers, morphological information and niche modelling are necessary for disentangling the evolutionary processes underlying the patterns found in this study.

Finally, we expect that future studies based on different approaches, such as multi-locus molecular markers, reciprocal transplant experiments and reproductive isolation tests among detected morphotypes, will provide a better understanding of the speciation mechanism in the *C. polyrhiza* complex and other angiosperm species complexes sharing similar historical and geographical contexts.

CONCLUSIONS

The existence of major clusters in *C. polyrhiza* has been determined; this is largely consistent with results obtained with clustering methods that make different assumptions. Those clusters are also highly consistent with the species proposed by Descole & Borsini (1954). Floral traits performed well in the delimitation of these groups. The pattern of phenotypic variation retrieved in the present study could be related to historical and ecological factors. Further research is required to support those assertions and to understand the evolutionary processes underlying the observed variation patterns. This is the first study carried out using geometric morphometrics and machine learning techniques for a terrestrial species inhabiting the Patagonian steppe. We hope that additional morphological studies of Patagonian plants and animals will allow us, through a comparative approach, to disentangle the evolutionary processes underlying biodiversity patterns in this largely unexplored region of the world.

ACKNOWLEDGEMENTS

We thank G. Oliva, Rosa Kofalt, Gervasio Humano and the staff of INTA-Santa Cruz for field assistance, Santiago Benitez-Vieyra for suggesting the use of Silhouette width values, Jorgelina Brasca for editing the English version of the manuscript and two anonymous reviewers for comments on previous versions of the manuscript. We also thank APN Argentina for granting permits to work in populations located in their parks and reserves. A.N.S. is a researcher for the National Research Council of Argentina (CONICET) and M.M.S. and A.C. are doctoral and postdoctoral fellows, respectively. Financial support was provided by the following grants: CONICET (PIP 11220080101264), FONCyT (PICT 01-10952 and 01-33755), MINCYT (PID 000113/2011) and SeCyT-UNC (ASN).

REFERENCES

Álvarez A, Perez SI, Verzi DH. 2011. Ecological and phylogenetic influence on mandible shape variation of South American caviomorph rodents (Rodentia:Hystricomorpha). *Biological Journal of the Linnean Society* **102:** 828–837.

Andersson S. 2006. On the phylogeny of the genus *Calceolaria* (Calceolariaceae) as inferred from ITS and plastid matK sequences. *Taxon* **55:** 125–137.

Baab K, McNulty KP. 2009. Size, shape, and asymmetry in fossil hominids: the status of the LB1 cranium based on 3D morphometric analyses. *Journal of Human Evolution* **57:** 608–622.

Baylac M, Villemant C, Simbolotti G. 2003. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. *Biological Journal of the Linnean Society* **80:** 89–98.

Berg RL. 1959. A general evolutionary principle underlying the origin of developmental homeostasis. *The American Naturalist* **93:** 103–105.

Bookstein FL. 1996. Applying landmark methods to biological outline data. In: Mardia KV, Gill CA, Dryden IL, eds. *Proceedings in image fusion and shape variability techniques*. Leeds: University of Leeds Press, 59–70.

Cattell RB. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* **1:** 245–276.

Clark JY. 2009. Neural networks and cluster analysis for unsupervised classification of cultivated species of *Tilia* (Malvaceae). *Botanical Journal of the Linnean Society* **159:** 300–314.

Cordeiro-Estrela P, Baylac M, Denys C, Polop J. 2008. Combining geometric morphometrics and pattern recognition to identify interspecific patterns of skull variation: case study in sympatric Argentinean species of the genus *Calomys* (Rodentia:Cricetidae:Sigmodontinae). *Biological Journal of the Linnean Society* **94:** 365–378.

Correa MN. 1998. *Flora patagónica. Parte I*. Buenos Aires: Colección Científica del INTA.

Cosacov A. 2010. Patrones de variación de caracteres fenotípicos y frecuencias génicas en el rango de distribución de la especie endémica de Patagonia *Calceolaria polyrhiza*: su relación con variables ambientales, polinizadores y factores históricos. PhD Thesis. Universidad Nacional de Córdoba, Argentina.

Cosacov A, Sérsic AN, Sosa V, De-Nova JA, Nylinder S, Cocucci AA. 2009. New insights into the phylogenetic relationships, character evolution, and phytogeographic patterns of *Calceolaria* (Calceolariaceae). *American Journal of Botany* **96:** 2240–2255.

Cosacov A, Sérsic AN, Sosa V, Johnson LA, Cocucci AA. 2010. Multiple periglacial refugia in the Patagonia steppe and post-glacial colonization of the Andes: the phylogeography of *Calceolaria polyrhiza*. *Journal of Biogeography* **37:** 1463–1477.

De Groote I. 2011. Femoral curvature in Neanderthals and modern humans: a 3D geometric morphometric analysis. *Journal of Human Evolution* **60:** 540–548.

Descole H, Borsini O. 1954. Scrophulariaceae. In: Descole H, ed. *Genera et species plantarum argentinarum*. Tucumán: Guillermo Kraft, V, I: 1–167.

Ehrhart C. 2000. Die Gattung *Calceolaria* (Scrophulariaceae) in Chile. *Bibliotheca Botanica* **153:** 1–283.

Ehrhart C. 2005. The Chilean *Calceolaria integrifolia s.l.* species complex (Scrophulariaceae). *Systematic Botany* **30:** 383–411.

Erxu P, Qiufa P, Hongfei L, Jingbo S, Yueqiang D, Feila H, Hui H. 2009. Leaf morphology and anatomy of *Camellia* section *Camellia* (Theaceae). *Botanical Journal of the Linnean Society* **159:** 456–476.

Ferreira da Costa A, Pena Rodrigues PJF, Wanderlay M. 2009. Morphometric analysis and taxonomic revision of the *Vriesea paraibica* complex (Bromeliaceae). *Botanical Journal of the Linnean Society* **159:** 163–181.

Gomez JM, Perfectii F, Camacho JPM. 2006. Natural selection on *Erysimum mediohispanicum* flower shape: insights into the evolution of zygomorphy. *The American Naturalist* **168:** 531–545.

Gómez JM, Abdelaziz M, Muñoz-Pajares J, Perfectti F. 2009. Heritability and genetic correlation of corolla shape and size in *Erysimum mediohispanicum*. *Evolution* **63:** 1820–1831.

González-José R, Bortolini MC, Santos FR, Bonatto SL. 2008. The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *American Journal of Physical Anthropology* **137:** 175–187.

Hall M, Frank E, Holmes G, Pfahringer B, Reurtemann P, Witten H. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* **11:** 10–18.

Hammer Ø, Harper DAT, Ryan PD. 2001. PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4:** 1–9. Available at: http://nhm2.uio.no/norlex/past/download.html

Hatziskakis S, Tsiripidis I, Papageorgiou AC. 2011. Leaf morphological variation in beech (*Fagus sylvatica* L.) populations in Greece and its relation to their post-glacial origin. *Botanical Journal of the Linnean Society* **165:** 422–436.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405:** 907–913.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Proceedings of the Royal Society of London Series B: Biological Sciences* **359:** 183–195.

Hodgins KA, Barret SCH. 2008. Geographic variation in floral morphology and style-morph ratios in a sexually polymorphic daffodil. *American Journal of Botany* **95:** 185–195.

Li D, Liu Y, Zhong C, Huang H. 2010. Morphological and cytotype variation of wild kiwifruit (*Actinidia chinensis* complex) along an altitudinal and longitudinal gradient in central-west China. *Botanical Journal of the Linnean Society* **164:** 72–83.

Maechler M. 2012. R.2.14.2 The R Cluster package version 1.14.2. Available at: http://cran.r-project.org/web/packages/cluster/index.html. Accessed February 2012.

Mascó M, Noy-Meir I, Sérsic AN. 2004. Geographic variation in flower color patterns within *Calceolaria uniflora* Lam. in Southern Patagonia. *Plant Systematic and Evolution* **244:** 77–91.

Medrano M, Castellanos MC, Herrera CM. 2006. Comparative floral and vegetative differentiation between two European *Aquilegia* taxa along a narrow contact zone. *Plants Systematic and Evolution* **262:** 209–224.

Méndez M, Traveset A. 2003. Sexual allocation in single-flowered hermaphroditic individuals in relation to plant and flower size. *Oecologia* **137:** 69–75.

Mendoça-Neto JP, Monteiro-Neto C, Moraes LE. 2008. Reef fish community structure on three islands of Itaipu, Southeast Brazil. *Neotropical Ichthyology* **6:** 267–274.

Mitteroecker P, Gunz P. 2004. Advances in geometric morphometrics. *Evolutionary Biology* **36:** 235–247.

Molau U. 1988. Scrophulariaceae. Part I. Calceolarieae. In: Zanoni T, ed. *Flora Neotropica. Monograph*. New York: New York Botanical Garden, 1–326.

Morando M, Avila LJ, Sites JJW. 2003. Sampling strategies for delimiting species: genes, individuals, and populations in the *Liolaemus elongatus-kriegi* complex (Squamata:Liolaemidae) in Andean-Patagonian South America. *Systematic Biology* **52:** 159–185.

Perez IS, Bernal V, Gonzalez PN. 2006. Differences between sliding semi-landmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. *Journal of Anatomy* **208:** 769–784.

Quinlan JR. 1993. *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Quinn GP. 2002. Multiple and complex regression. In: Quinn GP, Keough MJ, eds. *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press, 101–174.

Rohlf FJ. 2006. TPS series software. Available at: http://life.bio.sunysb.edu/morph/

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computation and Applied Mathematics* **20:** 53–65.

Ruzzante DE, Rabassa J. 2011. Palaeogeography and palaeoclimatology of Patagonia: effects on biodiversity. *Biological Journal of the Linnean Society* **103:** 221–228.

Seiler MB, Keeley ER. 2009. Intraspecific taxonomy and ecology characterize morphological divergence among cutthroat trout (*Oncorhynchus clarkii* ssp. Richardson) populations. *Biological Journal of the Linnean Society* **96:** 266–281.

Sérsic AN. 2004. Reproductive biology of the genus Calceolaria. *Stapfia* **82:** 1–121.

Shipunov AB, Bateman RM. 2005. Geometric morphometrics as a tool for understanding *Dactylorhiza* (Orchidaceae) diversity in European Russia. *Biological Journal of the Linnean Society* **85:** 1–12.

Von Reumont BM, Struwe J, Schwarzer J, Miso B. 2012. Phylogeography of the burnet moth *Zygaena transalpina* complex: molecular and morphometric differentiation suggests glacial refugia in southern France, western France and micro-refugia within the Alps. *Journal of Zoological Systematics and Evolutionary Research* **50:** 30–50.

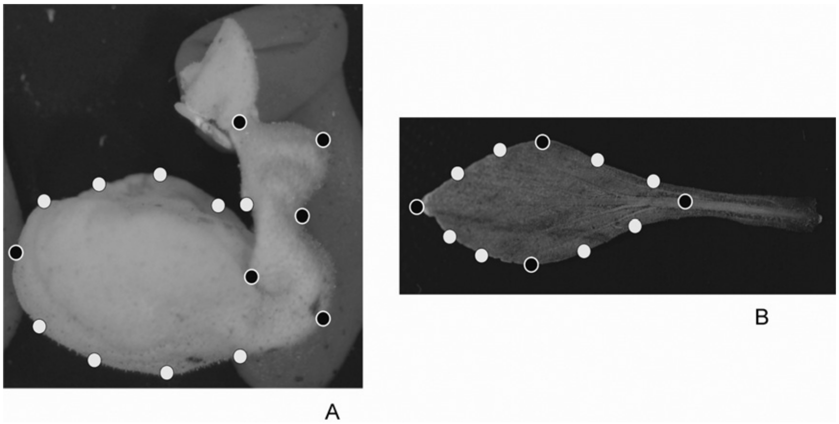
Witten IH, Frank E. 2005. *Data mining. Practical machine learning tools and techniques (second edition)*. San Francisco: Morgan Kaufmann Publishers Inc.

Zelditch ML, Swiderski DL, Sheets HD, Fink WL. 2004. *Geometric morphometrics for biologists. A primer*. New York and London: Elsevier Academic Press.

Zuloaga FO, Morrone O. 1999. Catálogo de las plantas vasculares de la República Argentina. II. Fabaceae–Zygophyllaceae (Dicotyledoneae). *Monographs in Systematic Botany Missouri Botanical Garden* **74:** 1040–1044.

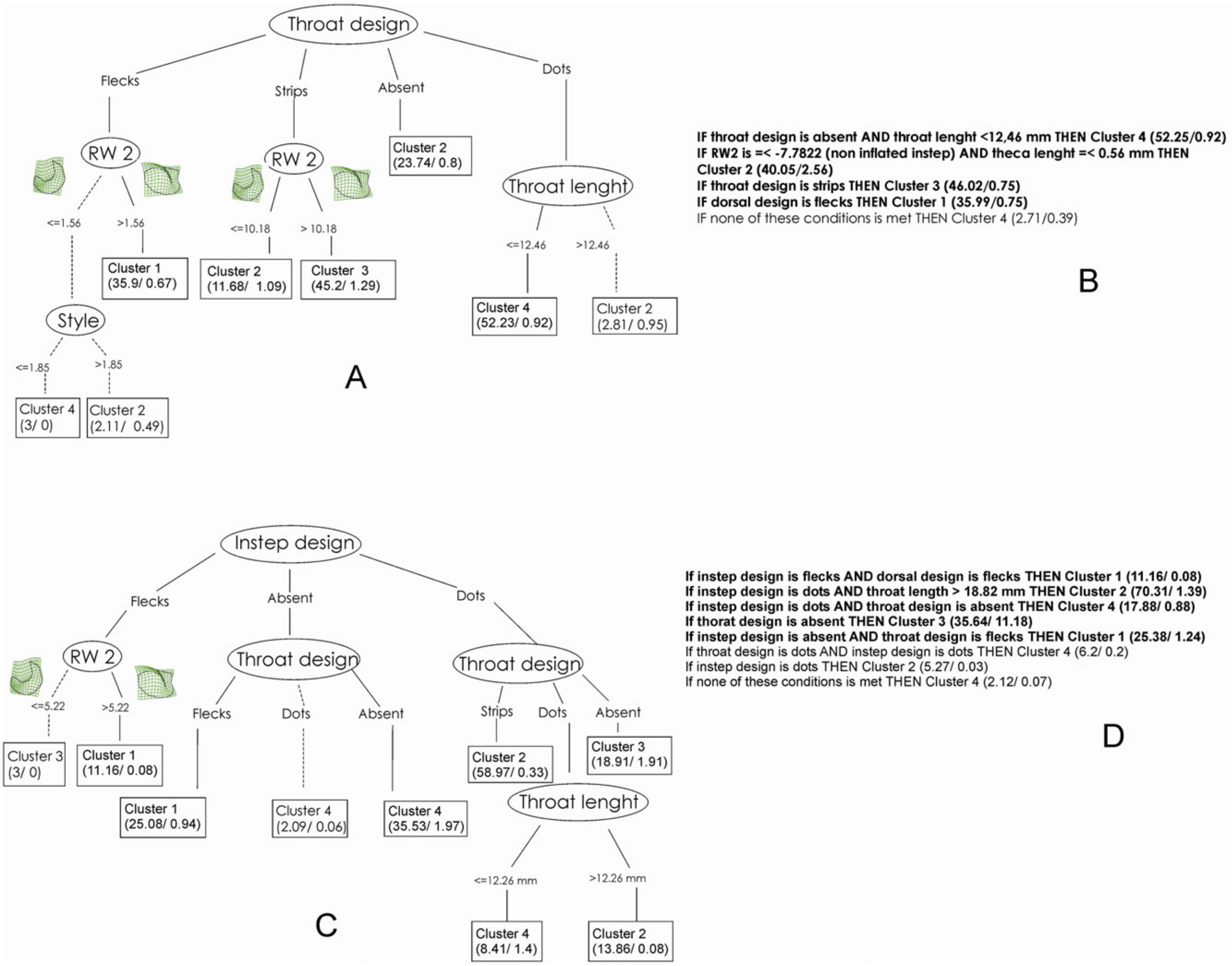
APPENDIX 1

Landmarks captured on: dorsal view of the corolla (A) and adaxial side of the leaf (B). Landmarks and semi-landmarks are represented by circles with and without outline, respectively.



APPENDIX 2

Decision trees generated for expectation–maximization (EM) and *k*-means, *k* = 4 clusters. Dotted lines represent those branches leading to a low number of classified individuals (A, C). Classification rules generated for EM *k*-means, *k* = 4 clusters (B, D). In all cases, groups in which most individuals were classified are in bold. Numerators indicate the number of correctly classified instances; denominators indicate wrongly classified instances. Note that these numbers are not integers: the trees are evaluated in many classification instances, these numbers being averages of the results obtained in all these iterations.



IF throat design is absent AND throat lenght <12,46 mm THEN Cluster 4 (52.25/0.92)
IF RW2 is =< -7.7822 (non inflated instep) AND theca lenght =< 0.56 mm THEN Cluster 2 (40.05/2.56)
IF throat design is strips THEN Cluster 3 (46.02/0.75)
IF dorsal design is flecks THEN Cluster 1 (35.99/0.75)
IF none of these conditions is met THEN Cluster 4 (2.71/0.39)

If instep design is flecks AND dorsal design is flecks THEN Cluster 1 (11.16/ 0.08)
If instep design is dots AND throat length > 18.82 mm THEN Cluster 2 (70.31/ 1.39)
If instep design is dots AND throat design is absent THEN Cluster 4 (17.88/ 0.88)
If thorat design is absent THEN Cluster 3 (35.64/ 11.18)
If instep design is absent AND throat design is flecks THEN Cluster 1 (25.38/ 1.24)
If throat design is dots AND instep design is dots THEN Cluster 4 (6.2/ 0.2)
If instep design is dots THEN Cluster 2 (5.27/ 0.03)
If none of these conditions is met THEN Cluster 4 (2.12/ 0.07)

APPENDIX 3

Contingency tables generated for decision trees and classification rules.

5-means

k-means, *k* = 5 clusters. Contingency table derived from decision tree.

Classified as.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	35	0	0	0	0
Cluster 2	0	39	0	1	0
Cluster 3	0	1	43	0	0
Cluster 4	0	1	0	19	0
Cluster 5	0	0	2	0	34

k-means, *k* = 5 clusters. Contingency table derived from classification rule.

Classified as.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	35	0	0	0	0
Cluster 2	0	38	3	0	1
Cluster 3	0	5	39	0	0
Cluster 4	0	0	0	20	0
Cluster 5	0	0	1	0	34

4-means

k-means, *k* = 4 clusters. Contingency table derived from decision tree.

Classified as . . .	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	35	0	0	1
Cluster 2	0	74	0	0
Cluster 3	2	0	43	1
Cluster 4	0	1	2	54

k-means, *k* = 4 clusters. Contingency table derived from classification rule.

Classified as . . .	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	35	1	1	0
Cluster 2	0	74	2	1
Cluster 3	0	0	36	0
Cluster 4	0	4	0	23

Expectation–maximization (EM)

EM clusters. Contingency table derived from decision tree.

Classified as . . .	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	35	1	1	0
Cluster 2	0	35	0	2
Cluster 3	0	3	43	2
Cluster 4	0	1	4	54

EM Clusters. Contingency table derived from classification rule.

Classified as . . .	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	35	0	0	0
Cluster 2	1	34	1	2
Cluster 3	1	2	44	1
Cluster 4	0	0	1	52