Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system

MANOLO F. PEREZ1,2,*, ISABEL A. S. BONATELLI1,3,*, MONIQUE ROMEIRO-BRITO1, FERNANDO F. FRANCO1, NIGEL P. TAYLOR4, DANIELA C. ZAPPI5, EVANDRO M. MORAES1,*

1Departamento de Biologia, Universidade Federal de São Carlos, Rodovia João Leme dos Santos km 110, Sorocaba, SP, 18052780, Brazil;
2Departamento de Genética e Evolução, Universidade Federal de São Carlos, Rodovia Washington Luiz km 235, C.P. 676, São Carlos, SP, 13565905, Brazil;
3Departamento de Ecologia e Biologia Evolutiva, Universidade Federal de São Paulo, Diadema, SP 09972-270, Brazil;
4University of Gibraltar, Gibraltar Botanic Gardens Campus, The Alameda, PO Box 843, Gibraltar;
5Programa de Pós Graduação em Botânica, Instituto de Ciências Biológicas, Universidade de Brasília, PO Box 04457, Brasília, DF, 70910970, Brazil.

* Manolo F. Perez and Isabel A. S. Bonatelli contributed equally to this article.

* Correspondence to be sent to Departamento de Biologia, Universidade Federal de São Carlos, Rodovia João Leme dos Santos km 110, Sorocaba, SP, 18052780, Brazil;
E-mail: emarsola@ufscar.br

Abstract. - Delimiting species boundaries is a major goal in evolutionary biology. An increasing body of literature has focused on the challenges of investigating cryptic diversity within complex evolutionary scenarios of speciation, including gene flow and demographic fluctuations. New methods based on model selection, such as approximate Bayesian computation, approximate likelihood, and machine learning approaches, are promising tools arising in this field. Here, we introduce a framework for species delimitation using the multispaces coalescent model coupled with a deep learning algorithm based on
convolutional neural networks (CNNs). We compared this strategy with a similar ABC approach. We applied both methods to test species boundary hypotheses based on current and previous taxonomic delimitations as well as genetic data (sequences from 41 loci) in *Pilosocereus aurisetus*, a cactus species with a sky-island distribution and taxonomic uncertainty. To validate our proposed method, we also applied the same strategy on sequence data from widely accepted species from the genus *Drosophila*. The results show that our CNN approach has high capacity to distinguish among the simulated species delimitation scenarios, with higher accuracy than the ABC procedure. For *Pilosocereus*, the delimitation hypothesis based on a splitter taxonomic arrangement without migration showed the highest probability in both CNN and ABC approaches. The splits observed within *P. aurisetus* agree with previous taxonomic conjectures considering more taxonomic entities within currently accepted species. Our results highlight the cryptic diversity within *P. aurisetus* and show that CNNs are a promising approach for distinguishing divergent and complex evolutionary histories, even outperforming the accuracy of other model-based approaches such as ABC.

Keywords: Species delimitation, fragmented systems, recent diversification, deep learning, Convolutional Neural Networks, Approximate Bayesian Computation

**INTRODUCTION**

Recognizing species boundaries has long been a major challenge for biologists. The major difficulty is to some degree related to the numerous existing species concepts. The use of distinct concepts can lead to different strategies for identifying species boundaries in empirical datasets (de Queiroz 2007; Carstens et al. 2013). However, different species concepts can be considered elements related to diverse properties that are associated with the dynamics of the speciation continuum. After the proposal of the unified species concept...
(Queiroz, 2007), the roles of species concepts as a theory and species delimitation as a methodology become apparent. The view of species as independent segments of a metapopulation indicated lineage independence as the only criterion necessary for delimiting species boundaries, avoiding any disagreement purely related to the species concept.

Selecting a suitable approach for species delimitation has been difficult, especially in species complexes (Pinheiro et al. 2018). Identifying discontinuities among incipient stages of divergence, such as those found in these systems, demands great effort and a multidisciplinary approach to assess different sources of evidence supporting species limits (Carstens et al. 2013). In this context, estimates based on independent sources of data such as morphology, cytogenetics, anatomy, ecology, and genetics have been used to achieve a better resolution in species delimitation (Domingos et al. 2014; Alvarado-Sizzo et al. 2018; Denham et al. 2019). Some of these traits, such as phenotypes and geographic distributions, can be a starting point for hypotheses about species circumscriptions (Solís-Lemus et al. 2015; Luo et al. 2018).

Species delimitation methods based on multilocus data and the multispecies coalescent model (MSC) compare the probability of species trees with different numbers of operational taxonomic units (OTUs) to identify optimal partitions for the data (Ence and Carstens 2011). Highly fragmented systems impose potential caveats on such methods, potentially resulting in the oversplit of existing diversity, and thus, highly subdivided entities (Sukumaran and Knowles 2017). Jackson et al. (2017) proposed the incorporation of the genealogical divergence index ($gdi$) as an attempt to reduce the delimitation of population structure, instead of species limits. Furthermore, the developers of the bayesian delimitation method implemented in BPP (Yang and Rannala 2014) described an approach to integrate $gdi$ on the outputs of their software and showed that such a strategy was efficient in mitigating the effect of over splitting (Leaché et al. 2018). However, a number of concerns on the use of $gdi$ have
been raised. For instance, a large interval of \( gdi \) values (0.2 to 0.7) that corresponds to a 'grey zone' where delimitation is ambiguous, and datasets containing putative species with population sizes very different or extremely small (Rannala and Yang 2020).

Model selection methods using simulated datasets under competing delimitation hypotheses are a promising tool for taxa that have potentially experienced a complex evolutionary history, including recurrent gene flow and demographic fluctuations. These methods also allow us to easily test for different assignments and topologies of the analyzed samples into populations/species, a procedure that is not straightforward with other approaches (but see Leaché et al. 2014). Examples of such approaches include approximate Bayesian computation (ABC; Camargo et al. 2012), approximate likelihood analysis (Morales et al. 2017), machine learning coupled with ABC (Smith and Carstens 2020) and machine learning methods based on summary statistics (SuSt) (Pei et al. 2018). Among them, the ABC represents a class of flexible likelihood-free algorithms for performing Bayesian inference (Beaumont et al. 2002). The principle behind the method relies on the massive simulation of genetic data using parameter values drawn from a prior distribution, followed by the calculation of SuSt for each simulation. SuSt are values generated from the raw genetic data, which are expected to capture important information to differentiate among the simulated models. Simulations that produce genetic variation patterns (SuSt) close to the observed data are retained to form an approximate sample from the posterior distribution.

A recently developed approach for demographic model selection, based on deep learning image classification, uses convolutional neural networks (CNNs) to retrieve as much information as possible from genetic datasets by converting them to images without requiring user-specified SuSt (Flagel et al. 2019). Simulations have shown that CNNs maximize the use of available information from the data and increase the ability to distinguish divergent evolutionary histories, frequently overcoming the limitations of traditional approaches (Flagel...
et al. 2019). Despite the recent use of CNNs in population genetics (Flagel et al. 2019; Sanchez et al. 2020) and phylogeography (Souza et al. 2019; Oliveira et al. 2020; da Fonseca et al. 2020), this approach has not previously been tested in the context of species delimitation using genetic data.

Here, we introduce an approach for species delimitation that combines the use of MSC-based methods with model selection via CNN. We compared our model selection method with ABC to evaluate distinct species delimitation hypotheses in the cactus *Pilosocereus aurisetus*. We chose this taxon because (1) it presents a controversial taxonomy, and (2) it has a highly fragmented geographical distribution, comparable to other sky-island systems, imposing recognized difficulties upon MSC-based methods (Leaché et al. 2018).

To gain better insights into the species limits in *P. aurisetus*, we used the following protocol (Fig. 1): (1) apply the MSC approach implemented in BPP (Yang and Rannala 2014) with sequence data and OTUs based on sampling localities, (2) apply the *gdi* index to the BPP results, to obtain a hypothesis less prone to oversplitting, (3) simulate data under the obtained BPP results (with and without *gdi*) and for each taxonomic hypothesis, with and without gene flow among lineages (this step requires some previous knowledge about the study group, such as divergence times, effective population size, and number of migrants per generation, when applicable), (3) transform the simulated data into binary images, defining the major and minor state for each site, (4) train a CNN to recognize simulations generated from each model, (5) evaluate the accuracy of the trained CNN in comparison to an ABC approach, by using cross-validation tests based on simulated data (i.e. independent simulations not evaluated before), and (6) predict model probability with the trained network (CNN) or by approximating the posterior (ABC). We also validated our method by following this same procedure using a recently published dataset from two *Drosophila* species pairs.
(Campillo et al. 2020), which have information about pre- and postzygotic reproductive isolation and do not have controversial taxonomy.

MATERIAL AND METHODS

The Studied System

The taxon *P. aurisetus* is a puzzling system regarding species delimitation. The taxonomy of *P. aurisetus* has been unstable over the years, which is a common attribute in the family Cactaceae, probably related to the absence of clear diagnostic characters for some groups due to convergent and parallel evolution (Hernández-Hernández et al. 2011; Copetti et al. 2017). The morphological variation observed within this taxon has led to repeated taxonomic evaluations, with species and subspecies being recurrently synonymized and re-established (Zappi, 1994; Taylor and Zappi, 2004; Hunt et al., 2006). Another complicating issue is the naturally fragmented distribution of *P. aurisetus*. The taxon is associated with the Espinhaço Mountain Range campo rupestre landscapes in eastern Brazil, occurring on mountaintops separated by valleys and lowland vegetation (Fig. 2). In line with this sky-island distribution, *P. aurisetus* shows high intraspecific genetic structure (Bonatelli et al. 2014; Perez et al. 2016a, b). Currently, two subspecies are recognized based on differences in the size and diameter of stems and the color and density of hairs on flower-bearing areoles: the more widespread *P. aurisetus* subsp. *aurisetus* (Werderm.) Byles & Rowley, and *P. aurisetus* subsp. *aurilanatus* (Ritter) Zappi, which occurs as few populations restricted to a disjunct mountain (Serra do Cabral) west of the Espinhaço Mountain Range. We also included in our study plants from three heterotypic synonyms of *P. aurisetus* subsp. *aurisetus* (Fig. S1): *P. aurisetus* subsp. *densilanatus* (Ritter) Braun & Esteves, located in the easternmost populations from the Serra Negra Mountains and distinguished by their densely wooly stems, reaching up to 4 cm diameter; *P. aurisetus* subsp. *supthutianus* (Braun) Braun & Esteves, a northern outlier of the species that also present densely hairy stems, which can
reach more than 5 cm diameter; and *P. aurisetus* subsp. *werdermannianus* (Buining & Brederoo) Braun & Esteves, located in the central populations and showing more slender stems, fewer ribs and sometimes green epidermis.

**Sampling and DNA extraction**

We sampled four individuals of *P. aurisetus* in each of nine localities in eastern Brazil, totaling 36 sampled individuals and covering the entire known distribution of the species. Hereafter, we refer to these sampled localities as populations justified by the clear limits of the *P. aurisetus* habitat patches and the marked genetic structure in this taxon (Bonatelli et al. 2014). This sampling included representatives of all accepted and synonymized taxa mentioned above (Fig. 2; Table 1). Root tissue was stored in silica gel and then transferred to a -80°C freezer until DNA extraction. We maintained a distance of approximately 10 m between sampled individuals to avoid collecting clones. Total DNA was extracted using the Extract All kit (Applied Biosystems) and purified with 95% ethanol and 3 M sodium acetate buffer. The DNA concentration was measured with a Qubit fluorometer (Invitrogen).

**Microfluidic PCR and sequencing**

Massive parallel target amplification of 41 loci in 36 samples was carried out in microfluidic PCR reactors (Access Array System, Fluidigm). The selected loci included 26 anonymous nuclear markers developed for *P. aurisetus* (Perez et al. 2016c) and another five nuclear, eight plastid and two mitochondrial genic regions selected from the available data in GenBank (Supplementary Table S1). Development of these new markers was based on available sequences from Cactaceae or from other related plant family species to detect conserved regions for primer design. All primers were developed with PRIMER3 v4.0.0 (Untergrasser et al. 2012), with parameters suggested by the Fluidigm Assay Design Group.
as follows: primer size from 18–23 bases, annealing temperature between 59°C and 61°C,
maximum poli-X of 3. The selected markers were synthesized according to Fluidigm
specifications and applied to a single 48×48 access array chip with reactions performed
according to the manufacturer’s protocol. The obtained PCR products were pooled and
purified with 0.6X AMPure beads (Agencourt). The quality and amplification range were
visualized in a BioAnalyzer, and the concentration was estimated via real-time PCR using
KAPA qPCR (Kapa Biosystems). The samples were subjected to a MiSeq (Illumina) 300 bp
paired-end run along with 156 other samples from other projects.
After sequencing, a first filtering step was performed with cutadapt (Martin 2011), excluding
sequences smaller than 60 bp and removing 3’ bases with a PHRED score <Q22. PHRED
quality for all reads was then visualized in FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) and graphically compiled in MultiQC (http://multiqc.info/). Reference sequences for each marker were generated from the
alignments used to design the primers. Mapping reads for each marker was performed with
BWA-MEM (Li 2013), which is suited for pair-end long reads with indels, while SAMtools
(Li et al. 2009) was used to retain only sequences that were mapped with high fidelity. SNP
calling was carried out with GATK (McKenna et al. 2010) by transforming low-quality bases
in missing data, identifying possible indels with realignment, applying polymorphism filters
for quality and coverage, and generating FASTA files for each sample. To confirm the
obtained polymorphisms, each detected SNP was then mapped against reference files for
each marker that were built from reads from each analyzed sample. The phasing of SNPs in
nuclear markers was achieved with a Python script, modified from Harvey et al. (2016). The
aligned sequences from each marker were obtained with MAFFT v.7 (Katoh and Standley
2013) with default parameters. Recombination within each marker was tested with the DSS
method (McGuire and Wright 2000) implemented in TOPALi v2 (Milne et al. 2008) and
topological incongruences were tested in KDETrees (Weyenberg et al. 2014). The most likely substitution model for each locus was obtained with PartitionFinder v.1.1.1 (Lanfear et al. 2012).

Species Tree and Coalescent Delimitation

The phylogenetic relationships among populations were recovered with a species tree in BEAST v2.1.3 (Bouckaert et al. 2014), assigning the sequences from each population to a single OTU. For all markers, we used the selected substitution model (Table S1) and a Relaxed clock Log-Normal, with the mean sampled from a broad lognormal distribution (according to Perez et al. 2016b). We performed three runs with $5 \times 10^8$ MCMC iterations each, sampling trees every $10^4$ generations with 75% burn-in. Convergence was observed in TRACER 1.6 (Rambaut et al. 2014) and a Maximum Clade Credibility (MCC) tree with median heights was obtained in TreeAnnotator (Drummond and Rambaut 2007).

We used the obtained species tree topology and the genomic sequences obtained from microfluidic PCR as input to test the species limits in *P. aurisetus* with the method implemented in BPP v3.4 (Yang and Rannala 2014). For this analysis, we concatenated all mitochondrial (*cox1* and *cox2*) and chloroplast (*atpB-rbcL*, *petL-psbE* and *rpl16*) markers. The BPP method uses a reversible jump Bayesian (rjMCMC) framework to estimate the posterior probability of existence at each node of the phylogeny. To ensure convergence, five independent runs were performed to estimate the best delimitation scheme, with parameters selected according to the values suggested by Leaché et al. (2018) combined with previous estimates for the species from Perez et al. (2016a). $\theta$ and $\tau$ priors followed an inverse gamma (IG) distribution with $\alpha=3$ and a $\beta$ of 0.01 and 0.002, respectively (hereafter BPP 'specific prior values'). We evaluated the potential effect of priors in the delimitation results by also performing the same number of runs with broad prior values $\theta \sim \text{IG}(3, 0.002)$ and $\tau \sim \text{IG}(3,$.
0.03) as suggested in the BPP manual (hereafter 'diffuse prior values'). We considered a node as valid only if both analyses support its existence with a posterior probability (PP) of 0.99 or higher. The runs were carried out for $10^5$ iterations sampled every five steps after discarding the first $10^4$ iterations as burn-in. The obtained delimitations were mapped to the tips of the species tree topology. As a more conservative alternative, we also evaluated our BPP results using the gdi index, as suggested by Leaché et al. (2018). We considered the threshold values suggested by Jackson et al. (2017) and Leaché et al. (2018), in which pairwise gdi values smaller than 0.2 indicate a single species and above 0.7, two species.

Comparing Delimitation Hypotheses with Deep Learning and ABC

We adopted an integrative approach to establish putative species limits by considering information on *P. aurisetus* from different sources and analyses, as suggested by Carstens et al. (2013). For that, we compared the two delimitation hypotheses generated by the BPP results (with and without the gdi index), as well as two morphological hypotheses based on one lumper and one splitter taxonomic arrangement (Table 1; Fig. 1). The genetic hypotheses, based on the BPP results, were composed of one species when gdi was considered and four species when it was not incorporated (see results below); the ‘splitter’ morphological hypothesis, considering both currently accepted and previously synonymized taxa, was composed of four species; and the ‘lumper’ morphological hypothesis, considering only the currently recognized subspecies, was composed of two species (Hunt et al. 2006). To test these competing delimitation hypotheses, we considered scenarios including the number of species and the potential migration between the delimited entities. Therefore, seven scenarios (Fig. 1) were tested: (1) ‘splitter’, (2) ‘splitter’ combined with migration, (3) ‘lumper’, (4) ‘lumper’ with migration, (5) BPP combined with gdi (BPP_GDI), (6) BPP without gdi (BPP_noGDI), and (7) BPP_noGDI with migration. As the BPP_GDI scenario consists of a
single species, an additional model including migration was not necessary. We performed
model comparison using both a deep learning approach based on CNN and a regular ABC
method. For these two strategies, we simulated genetic datasets under each scenario with a
modified version of the scripts from Perez et al. (2016a). To simplify our simulations and
avoid using an overly heterogeneous dataset, we included only nuclear markers and simulated
data mirroring sample sizes with the same number of segregating sites per loci of our
empirical dataset. For the deep learning approach, the simulated data were converted to
images that were used to train a CNN to recognize simulations generated by different
scenarios (for details on the simulation and training procedure, see Supplementary File 1).
The comparison of delimitation scenarios using regular ABC analysis (Fig. 1) was performed
following the approach of Perez et al. (2016a) with simulations conducted under the same
conditions used for the CNN (details in Supplementary File 1).

To validate our newly proposed deep learning method for species delimitation, we also
carried out the same analyses using a dataset recently published by Campillo et al. (2020).
These authors compared the results of reproductive isolation and coalescent-based (BPP)
delimitation methods to infer species boundaries in several species pairs of the genus
Drosophila, a classic model system in speciation research. We focused our approach on data
from two species pairs of the Melanogaster group: D. melanogaster - D. sechellia and D.
melanogaster - D. simulans. We chose these species because they are widely accepted and
corroborated by both BPP and reproductive isolation metrics (Campillo et al. 2020).
Furthermore, while the D. melanogaster - D. sechellia pair is allopatric, the D. melanogaster
- D. simulans pair is sympatric, allowing us to explore whether geographic context might
affect our ability to identify these species. Therefore, we tested a model with all specimens
belonging to the same species against a model in which the specimens were divided into two
species, which was the expected outcome according to the current classification. Again, our
simulations were based on empirical sample sizes and the number of segregating sites (details in Supplementary File 1), with priors based on the values suggested by Campillo et al. (2020).

RESULTS

Microfluidic PCR and sequencing

Four markers showed no amplification in any sample (PaANL_46, petA, rps16, and ycf1). To further minimize missing data and ensure that at least one individual was sequenced from each population, we discarded 16 loci (PaANL_8, PaANL_10, PaANL_50, PaANL_82, PaANL_123, PaANL_134, PaANL_142, PaANL_155, PaANL_160, PaANL_187, PaANL_196, its2, ppc, nhx1, psbA-psbB, psbC). No recombination was detected at any of the remaining loci, but kdetrees result suggested an incongruent topology in the isi1 marker, which was removed. Therefore, the final dataset included 20 markers that were considered in the subsequent analyses for P. aurisetus (PaANL_15, PaANL_17, PaANL_28, PaANL_35, PaANL_80, PaANL_87, PaANL_96, PaANL_126, PaANL_140, PaANL_147, PaANL_164, PaANL_165, PaANL_182, PaANL_205, apkl, atpB-rbcL, petL-psbE, rpl16, coxl and cox2), with a total of 8,908 bp.

Species Tree and Coalescent Delimitation

The species tree recovered most nodes with high support, especially the ones defining previously and currently valid subspecies (Fig. 3). Population JFE of P. aurisetus subsp. aurilanatus was the most external clade. Other populations were arranged in two subclades. One of them contained populations in the center (INA and MEN, previously classified as P. a. subsp. werdermannianus; ITA and PMN, formerly P. a. subsp. densilanatus) and south (COC)
of the *P. a.* subsp. *aurisetus* distribution. The second subclade contained populations from northern distribution (EDB, BOV and ODA, formerly classified as *P. a.* subsp. *supthutianus*).

All BPP runs under each prior set showed the same results, therefore we provided the mean posterior probability of each node in Fig. 3. Results with 'specific' and 'diffuse' priors were similar, though the latter was less splitter. The congruent delimited species were *P. a.* subsp. *aurisetus* (COC), *P. a.* subsp. *werdermannianus* (EDB, BOV and ODA) and the currently recognized subspecies *P. aurisetus* subsp. *aurilanatus* (JFE). The only exception was for the node separating the population ODA from the remaining *P. a.* subsp. *supthutianus* populations (BOV and EDB), which was validated (PP=1) only in the 'specific' but not (PP=0.95) in the 'diffuse' prior set. To be conservative, we decided to collapse this node, considering all populations once recognized as *P. a.* subsp. *supthutianus* as a single unit. Therefore, BPP resulted in four delimited species, that coincided with the 'splitter' hypothesis, except for the OTUs composed by populations once recognized as *P. a.* subsp. *werdermannianus* and *P. a.* subsp. *densilanatus*, which were collapsed (Fig. 3). When we analyzed our BPP results with the *gdi* index, all estimates were below 0.7 and only three (*P. aurisetus* subsp. *aurilanatus* - JFE; *P. a.* subsp. *aurisetus* - COC; and *P. a.* subsp. *supthutianus* - ODA, BOV and EDB) had values between 0.2 and 0.7 (Fig. S2). Therefore, to be conservative, we decided to collapse all populations in a single species for this hypothesis.

**Delimitation Hypotheses Comparison**

After 250 epochs, our CNN showed accuracies of 96.78% and 92.71% for the training and validation sets, respectively. Some degree of overfitting was observed when we plotted the accuracy of the training and validation sets throughout the epochs. However, this effect decreased as the number of evaluated simulations increased (Fig. S3). Our cross-validation procedure, using a test set of simulations not evaluated during training, also showed that
increasing the number of simulations resulted in accuracy improvement for all models (Fig. 4). It was also possible to observe that CNN presented a higher proportion of simulations that were correctly predicted in relation to their generating model than the ABC results, with all models being correctly assigned with more than 85% accuracy when 10,000 simulations per model were used (Fig. 4 and S4). The percentages of correct predictions with the ABC approach were all higher than 70% when 10,000 simulations per model were used, except for the models with more delimited species that showed ~50% accuracy (Fig. 4). These two types of models with lower accuracy in the ABC analysis ('splitter' and 'BPP_noGDI') showed the most confusion with each other (Fig. S4). When the empirical data were submitted to the trained CNN, the ‘splitter’ hypothesis was selected with a softmax probability of 99.8% (Table 2). Our CNN approach showed consistent results, selecting the same delimitation scenario even when the order of SNPs was shuffled (Supplementary File 1). The ABC approach showed similar results, also recovering the 'splitter' hypothesis (PP=0.994). These results are in agreement with what is observed when we plot the simulated and empirical genotypes with a PCA (Fig. S5), showing that the empirical data are located inside the cloud of simulations from the ‘splitter’ model.

The *Drosophila* dataset also pointed to a higher accuracy of CNN over ABC. The cross-validation tests indicate a correct assignment of CNN simulations just above 78% and 81% in the *D. melanogaster-D simulans* and *D. melanogaster-D. sechellia* pairs, respectively (Fig. S4). The accuracy of ABC was lower, reaching values slightly higher than 70% and 65% for the same species pairs (Fig. S4). When the empirical data were used, both methods supported the hypothesis of two species for the *D. melanogaster-D simulans* pair, showing a probability of 99.2% in CNN and 98.2% for ABC. In *D. melanogaster-D. sechellia*, the CNN pointed to two species with a probability of 94.9%, while the ABC suggested a single species with 89.4% PP (Table 2).
DISCUSSION

Deep learning in coalescent-based species delimitation

The advantages of the simulation approaches adopted here (CNN and ABC) over explicit statistical methods are that they allow testing models with complex demographic scenarios, such as migration events or size fluctuations. In addition to this use, these approaches allow to easily apply different assignments of samples into OTUs (but see Leaché et al. 2014) and even compare hypotheses relying on distinct topologies. Our comparison of these two approximate approaches suggests that the deep learning approach (CNN) showed a higher capacity to distinguish among the simulated demographic scenarios, outperforming the ABC procedure (Fig. 4 and S4). It is important to highlight that CNN achieved this performance with longer, though not impeditive, running times (~21 minutes in a Nvidia K80 GPU with an Intel Xeon 2.30GHz CPU) than ABC (~1 minute in an Intel Core i7 2.9GHz CPU). The results from the Drosophila dataset also supported a higher accuracy of CNN compared to ABC (Fig. S4). Moreover, when the empirical data was evaluated, only CNN recovered two species for the allopatric pair D. melanogaster-D. sechellia (Table 2). These results validated our CNN approach, as they are in agreement with our expectations based on Campillo et al. (2020).

Regarding the P. aurisetus dataset, both CNN and ABC approaches selected the 'splitter' model without migration as the most likely scenario with high PP (Table 2). The main difference between CNN and ABC functioning is the ability of the former to take information directly from SNP matrices without using SuSt, which is required by the latter. Moreover, while ABC relies on a rejection step that discards most of the simulations and leaves only a small proportion that are more similar to the empirical data (Csilléry et al. 2012), the CNN uses information from the whole set of simulations to learn how to
distinguish the concurrent scenarios (Schrider and Kern 2018). The inferior performance of
the ABC analyses could also be related to the choice of SuSt with overly high dimensionality
(Robert et al. 2011; but see Kousathanas et al. 2016) or with little information to distinguish
among the tested models. To reduce dimensionality, we adopted the suggestion of Perez et al.
(2016a), which consists of the transformation of the raw SuSt with a PCA and the retention of
only the axis necessary to ensure at least 99% of the variance. Other strategies usually applied
for this purpose (reviewed in Prangle 2015) include partial least squares regression
(Wegmann et al. 2009), linear regression (Fearnhead and Prangle 2012) or boosting

Although our CNN results were very encouraging, we note that care must be taken in
interpreting them, as deep learning techniques can suffer from overfitting (Nguyen et al.
2015; Ponti et al. 2017). To evaluate overfitting, we observed the history plot of the accuracy
throughout the epochs (Fig. S3) and adopted a cross-validation approach based on a test set,
which consisted of simulated data that were not evaluated during training. The results pointed
to high accuracy (Fig. 4 and Fig. S4), which is in agreement with the similarity of our
empirical and simulated data (Fig. S5). Another recent approach that could be coupled with
our model comparison strategy is based on unsupervised machine learning (Derkarabetian et
al. 2019). This strategy showed promising results and was particularly robust in
distinguishing population structure from species-level divergence.

Delimiting Species in *P. aurisetus*

*Pilosocereus aurisetus* populations diverged very recently in the last one million years
(Perez et al. 2016a). Delimiting species with recent diversification has been considered
challenging, and the main difficulty arises from the short time for species to accumulate
diagnostic characteristics that allow taxonomic distinction (Salicini et al. 2011). A similar
effect is observed in genetic data and is commonly referred to as incomplete lineage sorting (ILS). ILS appears when ancestral genetic variants persist as polymorphisms across successive diversification events. This phenomenon has been indicated as the primary cause of discordance between gene trees and species trees in Cactaceae (Copetti et al. 2017).

Another issue that imposes additional challenges for species delimitation in *P. aurisetus* is its sky-island distribution, leading to high population divergence (Bonatelli et al. 2014). Topography-driven isolation in species of the Brazilian rupestrian highland habitats (“campos rupestres”), such as *P. aurisetus*, has been considered an important factor driving speciation and high microendemism in these landscapes (Vasconcelos et al. 2020; Zappi et al. 2017).

The major concern in this kind of system is that MSC may fail to discriminate between genetic structure associated with population isolation and species boundaries (Jackson et al. 2017; Sukumaran and Knowles 2017; Leaché et al. 2018). Although gdi appears as a possible solution for the overspliting tendency of MSC methods (Jackson et al. 2017), mainly in highly structured systems, it is still not clear if it could lead to an opposite effect - over lumping differentiated species. Another possibility for overcoming this bias is to incorporate *a priori* knowledge to determine the structure associated with population-level processes and related to the species level. An ideal scenario would incorporate data from a plethora of classes, such as morphological, ecological, and genetic data, to recover more robust species boundaries (Pinheiro et al. 2018). Since no single method can discriminate between these processes, the results from species delimitation using multispecies coalescent models (MSC) in sky-island systems should be treated as species hypotheses that need further confirmation from multiple sources (Sukumaran and Knowles, 2017).

*Taxonomic Implications for P. aurisetus*
Our species delimitation investigation within *P. aurisetus* is discordant from the currently recognized taxonomic diversity of this taxon (Hunt et al. 2006), as the ‘splitter’ hypothesis has been recovered with higher probability with different methods. Convergent evolution as well as the level of plasticity commonly found in cactus species (Guerrero et al. 2018) might be related to the disagreement in the number of species delimited by current taxonomy and the detected genetic lineages, as observed within the genus *Pilosocereus* (Bonatelli et al. 2014; Perez et al. 2016b). Therefore, our results should provide some guidance for taxonomists interested in re-evaluating the taxonomy of *P. aurisetus*. In this context, our results may help to fill gaps in the traditional taxonomy and provide recommendations for a taxonomic treatment that better reflects the diversity of species, leading us to a multidisciplinary approach for better understanding the boundaries within species complexes considering their evolutionary history.

Previous genetic studies have identified well-differentiated lineages within *P. aurisetus* and related species, with monophyletic groups more frequently being associated with the geographic distribution than the proposed taxonomic boundaries (Fig. 2; Bonatelli et al. 2014; Perez et al. 2016b; Calvente et al. 2016; Khan et al. 2018; Lavor et al. 2019). Different taxonomic descriptions have been considered in the past for some of the recovered lineages within *P. aurisetus* that agree with our results. For instance, the populations occurring in the Serra Negra Mountains (ITA and PMN localities), in the center (MEN and INA) and in the northern distribution of the species (EDB and BOV localities) were previously reported as *P. aurisetus* subsp. *densilanatus*, *P. aurisetus* subsp. *werdermannianus*, and *P. aurisetus* subsp. *supthutianus* (Braun and Esteves Pereira, 1995), respectively. These subspecies propositions were based on some level of morphological variation that initially seemed rather distinctive, such as the densely wooly stems observed in the ITA populations (Fig. S1). However, Taylor and Zappi (2004) argued that such differences are not sufficiently important to merit
recognition as an additional subspecies. These authors reasoned that if this level of
morphological difference was to be recognized, the logical extension would be to give similar
status to other populations that show slightly divergent characteristics in terms of size, white-
wooly stems, and glaucousness.

CONCLUSIONS

Although the splits within species could be an artifact of the MSC, we might recognize
previous taxonomic issues that agree with the inferred species in this investigation.

_Pilosocereus aurisetus_ exhibits many biological characteristics that make species delimitation
challenging, such as recent divergence, a sky-island distribution and unstable taxonomic
history. Here, we sought to investigate the species limits in this complex system and, for the
first time, incorporated a deep learning approach to select the best species hypothesis
according to our data and simulations. Finally, we stress that all the species hypotheses
reported here should be considered, as any of the taxonomic species in fact are hypotheses
prone to validation with other methods and sources of information.

SUPPLEMENTARY MATERIAL

All scripts and datasets used are available in GitHub:


FUNDING

This work was supported by grants from the São Paulo Research Foundation (FAPESP)
(2015/06160-5 to EMM, 2012/22943-1 to MFP, and 2012/22857-8 to IASB); the National
Council for Scientific and Technological Development (CNPq) (03940/2019-0 to EMM,
305301/2018-7 to DCZ); and the Coordination for the Improvement of Higher Education Personnel (CAPES) (Finance Code 001 to MRB).

ACKNOWLEDGMENTS

We thank Heidi Utsunomiya and Juliana de Fátima Martinez for laboratory assistance and Gerardus Olsthoorn and Marlon Machado for sampling assistance.

REFERENCES


Ponti M., Ribeiro L., Nazare T., Bui T., Collomosse J. “Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask”. In: 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 17–41. IEEE.


Differences and Historical Demography in South American Arowanas

(Osteoglossiformes, Osteoglossidae, Osteoglossum). Genes 10: 693.


Genetics 182:1207–1218.


**FIGURE CAPTIONS**

**FIGURE 1.** A general protocol for species delimitation in *P. aurisetus*, comparing the adopted CNN and ABC approaches. (1) Evaluate delimitation hypotheses based on taxonomic and genetic information with the BPP (Yang and Rannala 2014); (2) simulate genetic data for each delimitation hypothesis, with and without gene flow among lineages (+M models); (3) transform the simulated data for the two approaches, CNN (convert to images, with black pixels representing the major and white pixels representing the minor frequency alleles for each segregating site) and ABC (extract SuSt averaged over loci); (4) train a neural network (CNN) to recognize simulations generated from each model; (5) use independent simulations to evaluate and compare the accuracy of CNN and ABC; (6) predict model probability, with the trained network for the CNN approach or by approximating the posterior (retaining only the 20% most similar simulations) for ABC.

**FIGURE 2.** (A) Geographic distribution of the sampled localities and elevational range in eastern Brazil (codes according to Table 1). Symbols (triangles and circle) represent the currently recognized *P. aurisetus* subspecies, and gray shades represent heterotypic synonmys of *P. aurisetus* subsp. *aurisetus* (in parenthesis). The delimited species are shown with lines colored according to the delimited entities in Fig. 3.

**FIGURE 3.** Species delimitation results are represented as bars at the tips of the topology obtained in BPP. Values in parentheses above each bar present the number of species for each
hypothesis. The numbers at the nodes represent Beast PP/BPP ‘specific’/and ‘diffuse’ prior values. Symbols (triangles and circles) represent both currently recognized and synonymized (in parenthesis) \textit{P. aurisetus} taxa. A representative of each inferred species is shown, next to the limits according to the 'splitter' hypothesis.

\textbf{FIGURE 4.} Cross-validation test to compare the performance of different model comparison methods (CNN - filled lines and ABC - dashed lines) and numbers of simulations. Each point represents the probability of choosing the "correct" model (i.e., the model from which the data were simulated) over incorrect models. Each color represents a different scenario, including (+M) or not migration between the OTUs.

\textbf{APPENDICES}

\textbf{SUPPLEMENTARY FILE 1.} Detailed simulation and model selection methods.

\textbf{SUPPLEMENTARY TABLE S1.} Primer sequence of each marker used in this study for microfluidic amplification. The genomic region for each marker is coded as follows: \texttt{nr = nuclear, \texttt{mt = mitochondrial, and cp = plastidial. Substitution models are provided for loci used in the final analyses.}

SUPPLEMENTARY FIGURE S2. Posterior distribution of genealogical divergence index ($gdi$) values generated by the BPP analysis. Color scheme according to branches from the tree depicted in the upper right corner, with locality names following Fig. 2.

SUPPLEMENTARY FIGURE S3. Accuracy of the CNN along the training epochs with varying numbers of simulations (500 in orange, 1,000 in yellow and 10,000 in green) for the training (filled lines) and validation (dashed lines) sets.

SUPPLEMENTARY FIGURE S4. Confusion matrices showing the percentage of predicted classes for each simulated scenario (model representation for $P. aurisetus$ following Fig. 1) using test data (independent simulations not evaluated in the training step). The results are shown for both methods (CNN and ABC) in the three taxonomic systems analyzed ($P. aurisetus$ and the species $D. melanogaster-D. simulans$ and $D. melanogaster-D. sechellia$).

SUPPLEMENTARY FIGURE S5. Principal coordinate analysis for simulated and empirical data in $P. aurisetus$. Simulations from the 'splitter' model are shown following the same color scheme in Fig. 1.
Table 1. Sampling localities for *P. aurisetus* and their assignment to operational taxonomic units according to three different species delimitation hypotheses (‘splitter’, ‘lumper’, ‘BPP_noGDI’ and ‘BPP_GDI’).

<table>
<thead>
<tr>
<th>Taxon/Locality</th>
<th>Code</th>
<th>N</th>
<th>Longitude</th>
<th>Latitude</th>
<th>‘splitter’</th>
<th>‘lumper’</th>
<th>‘BPP_noGDI’</th>
<th>‘BPP_GDI’</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. a. subsp. aurisetus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cocais-MG</td>
<td>COC</td>
<td>4</td>
<td>-43,450000</td>
<td>-19,86667</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Itamarandiba-MG (a)</td>
<td>ITA</td>
<td>2</td>
<td>-42,935972</td>
<td>-18,001944</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pedra Menina-MG (a)</td>
<td>PMN</td>
<td>2</td>
<td>-43,044500</td>
<td>-18,136250</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Olhos D’Água-MG (b)</td>
<td>ODA</td>
<td>3</td>
<td>-43,622361</td>
<td>-17,438833</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Engenheiro Dolabela-MG (b)</td>
<td>EDB</td>
<td>3</td>
<td>-43,793667</td>
<td>-17,451417</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Bocaiúva-MG (b)</td>
<td>BOV</td>
<td>3</td>
<td>-43,755833</td>
<td>-17,545833</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Mendanha-MG (c)</td>
<td>MEN</td>
<td>4</td>
<td>-43,536583</td>
<td>-18,118917</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Inhaí-MG (c)</td>
<td>INA</td>
<td>4</td>
<td>-43,605833</td>
<td>-17,146444</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><em>P. a. subsp. aurilanatus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joaquim Felicio-MG</td>
<td>JFE</td>
<td>5</td>
<td>-44,178333</td>
<td>-17,757833</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) and (c) indicate the samples from the previously synonymized forms *P. a. subsp. densilanatus*, *P. a. subsp. supthutianus* and *P. a. subsp. werdermannianus*, respectively.
Table 2. Model selection results for *P. aurisetus* and the two *Drosophila* species pairs using empirical data. CNN values are softmax probabilities for each scenario, while posterior probabilities are shown for ABC. The scenario with the highest probability for each method is shown in bold.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>CNN</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P. aurisetus</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splitter</td>
<td>0.998</td>
<td>0.994</td>
</tr>
<tr>
<td>Splitter+Migration</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>Lumper</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Lumper+Migration</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BPP+GDI</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BPP_noGDI</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>BPP_noGDI+Migration</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>D. melanogaster_D.simulans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Species</td>
<td>0.008</td>
<td>0.018</td>
</tr>
<tr>
<td>Two Species</td>
<td>0.992</td>
<td>0.982</td>
</tr>
<tr>
<td><strong>D. melanogaster_D. sechellia</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Species</td>
<td>0.051</td>
<td><strong>0.894</strong></td>
</tr>
<tr>
<td>Two Species</td>
<td><strong>0.949</strong></td>
<td>0.106</td>
</tr>
</tbody>
</table>
1) Define taxonomic and BPP (with and without GDI) hypotheses

- Splitter
- Lumper
- BPP+GDI
- BPP_noGDI

5 species
2 species
1 species
4 species

2) Simulate data with $ms$ for each hypothesis, with (+M) and without migration

1. Splitter
2. Splitter+M
3. Lumper
4. Lumper+M
5. BPP+GDI
6. BPP
7. BPP+M

3) Transform to image

- CNN
- ABC

3) Extract SuSt from each sim and for empirical data (averaged over Loci)

- $\pi_S$
- $\theta_H$
- $\pi_W$
- $\pi_B$

4) Train CNN with all data (70,000 sims)

- Empirical dataset

4) Extract SuSt in empirical dataset

5) Evaluate accuracy with test data by building a confusion matrix

- CNN
- ABC

6) Predict model prob. from empirical data with trained CNN from step (2)

- Empirical dataset
- Softmax Probability

6) Approximate the posterior with empirical data

- Discard 80% most dissimilar simulations
- Vector of SuSt
- Parameter $\theta$
- Posterior Probability

---

The copyright holder for this version posted December 24, 2020. doi: bioRxiv preprint
Espinhaço Mountain Range

Southern Espinhaço Mountain Range

Serra do Cabral Mountains

Serra Negra Mountains

P. a. subsp. supthutianus

P. a. subsp. werdermannianus

P. a. subsp. densilanatus

P. a. subsp. aurisets

P. a. subsp. aurilanatus

0 50 100 Km

0 m 250 m 500 m 750 m 1000 m 1250 m 1500 m

CC-BY-NC 4.0 International license

preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in

The copyright holder for this version posted December 24, 2020. ; https://doi.org/10.1101/2020.12.23.424219
doi: bioRxiv preprint
\[ P. a. \text{ subsp. supthutianus} \]
\[ P. a. \text{ subsp. werdermannianus} \]
\[ P. a. \text{ subsp. densilanatus} \]
\[ P. a. \text{ subsp. aurisetus} \]
\[ P. a. \text{ subsp. aurilanatus} \]