Journal of
Molluscan Studies

The Malacological Society of London

Journal of Molluscan Studies (2020) **00**: 1–7. doi:10.1093/mollus/eyz031

Geometric morphometrics and machine learning challenge currently accepted species limits of the land snail *Placostylus* (Pulmonata: Bothriembryontidae) on the Isle of Pines, New Caledonia

Mathieu Quenu[1], Steven A. Trewick[1], Fabrice Brescia[2] and Mary Morgan-Richards[1]

[1] *Ecology, College of Science, Massey University, Private Bag 11-222, 4442 Palmerston North, New Zealand; and*
[2] *Institut Agronomique Neo-Caledonien (IAC), Port-Laguerre, Paita, BP73 98890 Noumea, New Caledonia*

Correspondence: M. Quenu; e-mail: mathieuquenu@hotmail.fr

(Received 2 April 2019; editorial decision 28 August 2019)

ABSTRACT

Size and shape variations of shells can be used to identify natural phenotypic clusters and thus delimit snail species. Here, we apply both supervised and unsupervised machine learning algorithms to a geometric morphometric dataset to investigate size and shape variations of the shells of the endemic land snail *Placostylus* from New Caledonia. We sampled eight populations of *Placostylus* from the Isle of Pines, where two species of this genus reportedly coexist. We used neural network analysis as a supervised learning algorithm and Gaussian mixture models as an unsupervised learning algorithm. Using a training dataset of individuals assigned to species using nuclear markers, we found that supervised learning algorithms could not unambiguously classify all individuals of our expanded dataset using shell size and shape. Unsupervised learning showed that the optimal division of our data consisted of three phenotypic clusters. Two of these clusters correspond to the established species *Placostylus fibratus* and *P. porphyrostomus*, while the third cluster was intermediate in both shape and size. Most of the individuals that were not clearly classified using supervised learning were classified to this intermediate phenotype by unsupervised learning, and most of these individuals came from previously unsampled populations. These results may indicate the presence of persistent putative-hybrid populations of *Placostylus* in the Isle of Pines.

INTRODUCTION

Species are defined using many different criteria, in order to classify individuals into distinctive groups (Darwin, 1859; Mayr, 1942; Mahner, 1993; Mallet, 1995; De Queiroz, 2007). However, the continuity of evolution means that probably no single species concept can effectively serve all circumstances. Instead, it has been argued that practical and useful definitions (operational criteria) should be prioritized when applying empirical data to the issue of species delimitation (Dubois, 2011; Vaux, Trewick & Morgan-Richards, 2016). The definition of species as morphological or genotypic clusters of individuals (Mallet, 1995) allows for the presence of individuals that are morphologically or genetically intermediate between two species, as long as these individuals are relatively uncommon. In this view, species are groups of individuals that interact and reproduce mostly together, and morphological and genotypic clustering is a consequence of the correlation of their traits. This definition accommodates evolutionary flux and so allows speciation, gene flow and selection to be studied without the circularity of defining a species by reproductive isolation. In this framework, we need to be able to delimit groups of individuals using character data and, if intermediate individuals between clusters exist, be able to determine whether they are strongly underrepresented (Mallet, 1995). Robust analytical tools are necessary to both identify groups and detect intermediates. Recent advances in machine learning using supervised and unsupervised learning approaches allow identification of specimens with and without *a priori* labels and so provide less subjective ways of delimitating species' limits.

In parallel with advances in machine learning, since the 1980s, there have been developments in the analysis of morphological variation (reflecting the development of a formal theory of shape and geometric morphometric methods) (Kendall, 1986; Mitteroecker & Gunz, 2009). Landmark-based geometric morphometrics use the relationships between a set of landmarks to study the shape of organisms, rather than a set of traditional distance measurements (Bookstein, 1991). In this approach, shape differences between individuals are defined as any variation found among a set of homologous landmarks after they have been scaled, rotated and translated to the same criteria, a process called generalized

Procrustes analysis. Two types of landmarks are the most commonly used. Fixed landmarks sample homologous features in different individuals, whereas semi-landmarks can be used to sample the same positions on any surface located between two fixed landmarks (Zelditch, Swiderski & Sheets, 2004). Statistical analyses can then quantify the differences in shape between individuals by considering variation in a set of aligned landmarks, typically after steps of ordination and dimension reduction (Zelditch *et al*., 2004). Geometric methods are more sensitive to shape variation than traditional morphometry (Rohlf & Marcus, 1993; Maderbacher *et al*., 2008) and have been usefully applied to the study of shell shape variation of molluscs (Carvajal-Rodríguez, Conde-Padín & Rolán-Alvarez, 2005; Cruz, Pante & Rohlf, 2012; Gustafson *et al*., 2014; Dillon & Jacquemin, 2015; Gustafson & Bolek, 2016; Vaux *et al*., 2017, 2018; Rao *et al*., 2018; Verhaegen *et al*., 2018). In the present study, we combine the use of geometric morphometry, with supervised and unsupervised learning algorithms, to test whether the current species-level taxonomy of the land snail *Placostylus* on the Isle of Pines, New Caledonia, corresponds with patterns of shell morphological variation.

Placostylus is a genus of large terrestrial pulmonate snails that is native to the islands of the western Pacific Ocean (Breure, Groenenberg & Schilthuizen, 2010). In New Caledonia, six species are recognized according to current taxonomy (Neubert, Chérel-mora & Bouchet, 2009), but a combination of genotypic and geometric data indicate that the two most abundant nominal species-level taxa *Placostylus fibratus* (Martyn, 1789) and *P. porphyrostomus* (Pfeiffer, 1851) may contain additional undescribed species on the main island, Grande Terre (Dowle *et al*., 2015). These two species are also locally common on the Isle of Pines, a small island (152 km^2) south of Grande Terre, where *P. fibratus* is still harvested for food (Brescia, 2011).

Analysis of 26 specimens from the Isle of Pines has shown that shell morphology seems sufficient to discriminate the two sympatric *Placostylus* species, with concordance between shell morphology and nuclear genetic clusters being unequivocal (Dowle *et al*., 2015). Shell differences in sympatry indicate that the size and shape of adult *Placostylus* shells are largely controlled by genetic factors, rather than environmental differences. *Placostylus fibratus* is larger and has a wider ventral aperture than *P. porphyrostomus*, which is usually smaller with a thicker lip around a smaller aperture. Although shell morphology was concordant with two genetic clusters in a previous sample from the Isle of Pines, some crucial snail populations on the island were not sampled. Those from the Comwagna district are of particular interest because adult snails are intermediate in size to the two recognized species on the island and thus difficult to identify (Brescia, 2011). Available mt (mitochondrial) DNA sequence data suggest some historical genetic exchange consistent with hybridization between the two species (Dowle *et al*., 2015), and this may explain why these populations are intermediate in size to *P. fibratus* and *P. porphyrostomus*.

Based on geometric morphometric analyses of data on shell size and shell shape data, the aim of this study is to apply supervised and unsupervised learning algorithms to test whether the recognition of two *Placostylus* species on the Isle of Pines is sufficient and to quantify the frequency of potential novel phenotypes.

MATERIAL AND METHODS

Rationale

Supervised learning is a type of algorithm that is 'taught' to create a function that will link a set of features (inputs) to a set of labels (outputs) (Hastie, Tibshirani & Friedman, 2009). This is done by providing example data, called the training dataset, for which features are already associated with some known labels. A training process is then used to find the optimal function that links feature values (shell shape and size) to label values (species identification). Here, we use a class of artificial neural networks called a multilayer perceptron for supervised learning (Ripley, 1994). Multilayer perceptrons are recognized for their predictive efficiency and rely on the use of gradient descent and backpropagation algorithms. Associations of these algorithms with geometric morphometrics can be used to classify individuals into different categories based on shape and size information (Baylac, Villemant & Simbolotti, 2003; Dubey *et al*., 2006; Soda, Slice & Naylor, 2017). We chose to use this algorithm as it requires few assumptions and can be applied to a wide range of data types (Ripley, 1994). A sample of snails where we have both morphometric and genetic information from a previous study was used as a training dataset (Dowle *et al*., 2015). These shells form two separate morphological groups that could also be separated using population genetic tools with nuclear loci; using the genotypic cluster definition, they are regarded as two species. If the current taxonomy of *Placostylus* reflects biological reality and shell morphology can be used to separate individuals into two distinct species, we expect the supervised learning algorithms to classify without ambiguities all (or almost all) individuals in our expanded sample.

We also use Gaussian mixture models (GMMs), a class of unsupervised learning algorithm, to find the optimal number of morphological groups in our new (expanded) dataset. Unsupervised learning is a class of machine learning algorithms, which infer structure in datasets that do not have *a priori* labels (Hastie *et al*., 2009). In the case of GMMs, the dataset is considered to be a mixture of population samples with different Gaussian distributions, and the modelling process divides the data into different classes, each one corresponding to a single Gaussian distribution (Fraley & Raftery, 2006). We use GMMs because this approach can model any kind of data, as long as the data can be separated into different Gaussian components. Different models are created for a different number of hypothetical clusters. Bayesian information criteria (BIC) scores are used to determine which model is the best fit to the data, and in this way, the optimal number of clusters in our sample is estimated. If the current shell-based taxonomy of *Placostylus* in the Isle of Pines is correct, we expect an optimal GMM that divides our dataset into two groups, corresponding to the two currently recognized species.

Sampling strategy

A total of 337 *Placostylus* snails were sampled across eight sites on the Isle of Pines in April 2015 (Fig. 1). Sample sites were arrayed along a set of elevation gradients from the coast to the centre of the island (Fig. 1). All live snails within a 20 m radius of a fixed point were collected by hand, at three locations along each of eight transects. Two-dimensional geometric morphometric data were gathered from the 337 shells sampled in 2015 and from the 26 shells that were collected and analysed previously (Dowle *et al*., 2015). Among the eight population samples examined, five had previously been sampled (Vao, Gadji, Kere, Youaty, Touete), and three were sampled for the first time (Comagna, Wapan, Waatchia) (Fig. 1). Once snails are sexually mature, age-related variation in shell shape and size are considered to be minimal (F. M. Brescia *et al*., 2008). Only shells of adult snails (identified by the thickened lip) were photographed for shape analysis. As these snails are hermaphroditic, it was not necessary to consider sexual dimorphism.

Geometric morphometrics

The set of two-dimensional landmarks used for this study was derived from a set of optimal landmarks and semi-landmarks established in an analysis of New Zealand *Placostylus* (Daly, 2016). A total of 40 landmarks and semi-landmarks were used, consisting of nine fixed landmarks and 31 sliding/semi-landmarks (Fig. 2). Digital images of the ventral surface of each shell were obtained using a Canon EOS 600D with EF100 mm f2.8 USM macro lens after careful positioning of the shell in a bed of sand of contrasting

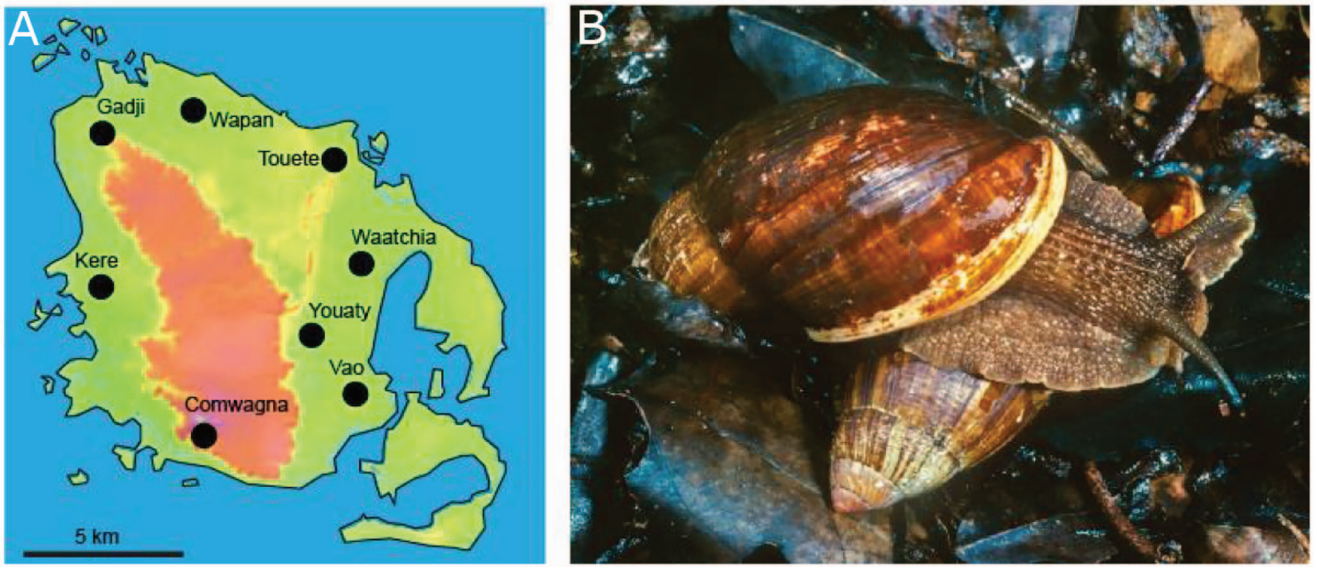

Figure 1. A. Topographic map of the Isle of Pines, New Caledonia, highlighting the different locations where snails were sampled. Colours represent different elevation levels. **B.** Two live *Placostylus fibratus* in their natural environment (image: Rod Morris).

colour. The camera equipment was mounted on a high-precision Kaiser stand to allow reproducible positioning and orientation (Dowle *et al.*, 2015). All images were captured with the lens fixed at the same distance from shell, and shells were positioned by the same person to minimize operator error (Schilthuizen & Haase, 2010). In order to place the semi-landmarks along the curves of the shells, two 'combs' were placed manually using Adobe Photoshop CS6, with their teeth perpendicular to a line from the shell apex to the intersection of the lips and outer shell (Fig. 2). Digitization of landmarks and semi-landmarks was conducted using the programme tpsDig2 v. 1.1 (Rohlf, 2015). Landmark type assignment and Procrustes analysis were then performed using Coordgen v.8.0 (Zelditch *et al*., 2004). Principal component analysis (PCA) was executed on the covariance matrices of the aligned landmark coordinates using MorphoJ (Klingenberg, 2011).

The size of the shells was incorporated into the analysis using the 'centroid size' tool included in the Procrustes analysis in Coordgen8. This size estimation is calculated as the square root of the sum of the arrays coming from the centroid position of a shape to each of the landmarks (Klingenberg, 2016). Both shape information (principal components (PCs)) and the centroid size estimate (from here on referred to as 'shell size') were used as input variables in the supervised and unsupervised learning algorithms.

Errors linked to shell manipulation and digitization were assessed by experimental replication; this involved taking five photographs of the same shell and comparing the variance attributed to the repeat process with the variance of the whole dataset using the *morphol.disparity*() function in the package *geomorph* v. 3.1.2 (Adams, Collyer & Kaliontzopoulou, 2018) in R v.3.6.1 (R Development Core Team & R Core Team, 2017). Overall, the variance linked to manipulation and digitization error corresponded to <0.3% of the variance of the rest of the dataset, which we deemed to be negligible for subsequent analyses.

Supervised learning algorithms

For the training process of the supervised learning algorithm, we used 26 shells that had been genetically assigned to one of two species (our training dataset). These specimens had been unequivocally assigned by Dowle *et al.* (2015) to one of the two recognized *Placostylus* species on the basis of a nuclear DNA dataset comprising 661 SNPs. Neural network models were built using the *neuralnet*() function in the R package *neuralnet* v. 1.44.2 (Fritsch *et al.*, 2016). Input data (shape and size) were normalized and scaled using a minmax function. Neural network models containing two hidden layers of five and three neurons were used (Fig. 3).

We first tested the efficacy of different neural network models using exhaustive cross validation on our training dataset ($n = 26$). For model validation purposes, this dataset was split into different train and test datasets of various proportions, using different numbers of input variables (for split proportions, number of input variables and other details; see Supplementary Material Table S1). The objectives of this process were to estimate whether we had a sufficiently large training dataset and to determine the optimal number of input variables. Efficacy of the models was assessed by looking at the mean proportion of individuals assigned correctly in the test datasets for a large number of replicates. We found that the optimal models used only two PCs and one size of a variable. In these models, the cross validation correctly identified all snails most of the time (mean proportion of correct assignment was between 98.8 and 99.8% for different-sized training datasets; Supplementary Material Table S1).

We then trained neural network models using the whole training dataset and used trained models on our new dataset of 337 shells to assign each snail to one of the two species. We repeated this process for a large number of neural network models (×1,000) to avoid overfitting issues arising from single model interpretation (Zhou, 2009), retaining the average value of the neural networks outputs as assignment score for each individual (ensemble learning).

Unsupervised learning algorithms

For GMMs, we used the R package *mclust* v. 5.4 (Fraley & Raftery, 2006). GMMs were built using shell size and the same number of PCs as in the supervised learning analyses (1 and 2). Different models were built for a range of *a priori* clusters in the dataset, with the optimal model being selected on the basis of BIC scores (Supplementary Material Fig. S1). Results of the modelling process with additional input variables (PC1–5, shell size) were also computed and yielded similar results.

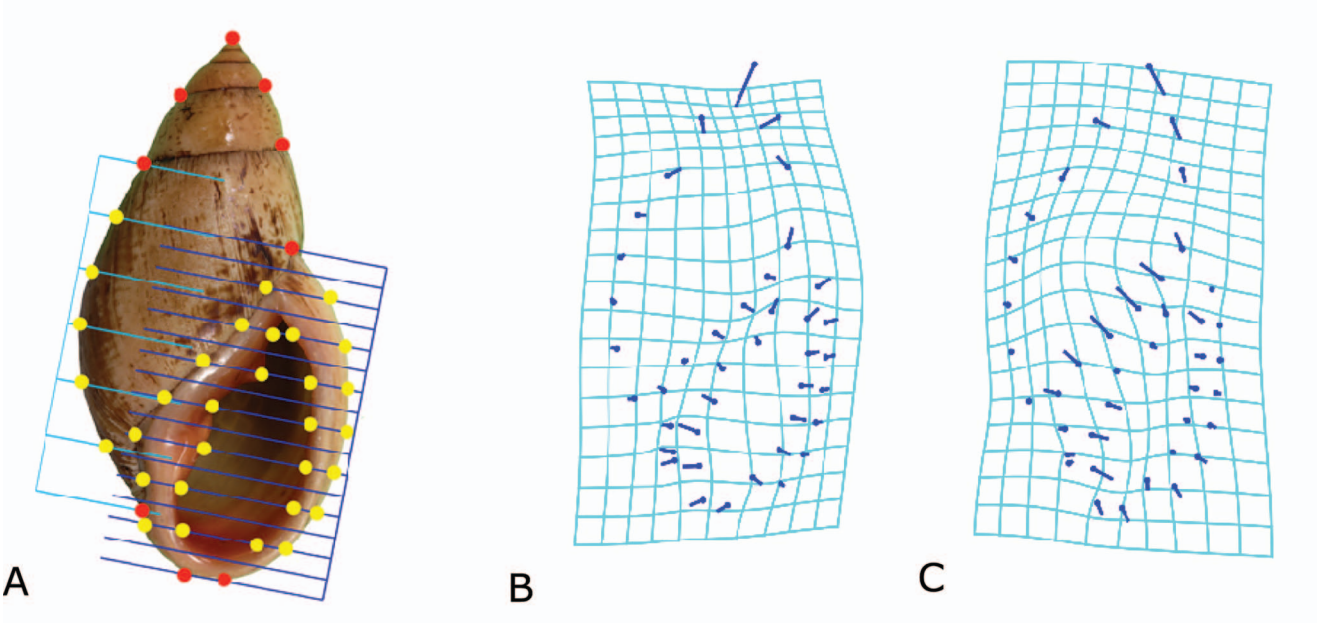

Figure 2. A. Shell of New Caledonian *Placostylus fibratus* with the 40 landmarks used in the geometric morphometry analysis. Digitizing combs are in blue, and the orientation is indicated by the white dashed line. Red and yellow dots indicate permanent and semi-landmarks, respectively. **B, C.** Relative displacement of landmarks using thin-plate splines, showing shell shape variation for PCs 1 (**B**) and 2 (**C**) of the geometric morphometric analysis of the whole dataset. PCs 1 and 2 explained 35.3 and 15.5% of total variance, respectively. Length of lollipop lines is proportional to warping in shape space for each PC.

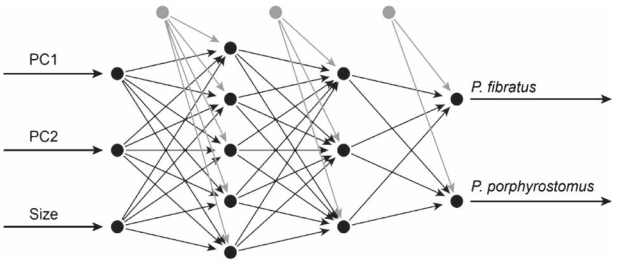

Figure 3. Schematic of a multilayer perceptron used for the supervised learning algorithm examining classification of *Placostylus* land snail shells with the two shape PCs and centroid size. Neural networks contained two hidden layers of five and three artificial neurons, respectively. Here, weight and bias parameters are represented by black and grey arrows.

RESULTS

First and second PCs of variation of the geometric morphometric analysis accounted for 35.3 and 15.5% respectively of the total variance in shell shape. Warp transformation grids indicate that spire height, aperture size and lip thickness varied the most (Fig. 2).

Supervised learning algorithms

Using the supervised learning algorithm, individual specimens were assigned to one or other of the two species, *Placostylus fibratus* or *P. porphyrostomus*, if their assignment probability was > 95%. On the basis of this criterion, 224 individual snails were identified as *P. fibratus* (67%) and 61 as *P. porphyrostomus* (18%); 51 (15%) could not be assigned to either species (Table 1, Figs 2B, 4A). Most of the unassigned individuals came from two populations: Comwagna (28 individuals) and Wapan (15 individuals). The six other population samples contained individuals that could be confidently identified to either *P. fibratus* or *P. porphyrostomus*, with only a few unassigned individuals (Table 1). Youaty was the only population sample with just one snail species (*P. fibratus*).

Unsupervised learning algorithms

The shell size and shape variations had an optimal fit to a GMM with three clusters (elliposidal, equal volume, shape and orientation (EEE) model; Supplementary Material Fig. S1). Assignment probabilities to the three phenotypic clusters were computed for each of the 337 individuals (Table 1, Fig. 4). Two of the phenotypic clusters identified using shell size and shape variables correspond to the recognized species *P. fibratus* and *P. porphyrostomus*. A third cluster was also observed. This consisted mostly of individuals from Comwagna and Wapan, which were unassigned by the supervised learning algorithm (Fig. 4, Supplementary Material Table S1). The snails in the third cluster were intermediate in shell size and shape to *P. fibratus* and *P. porphyrostomus* (Fig. 5).

Using GMMs, 226 (67%) individuals were assigned to the *P. fibratus* cluster, 56 (17%) to the *P. porphyrostomus* cluster and 54 (16%) to the intermediate group. The assignment of individuals was similar to the result produced by the supervised learning algorithm, with specimen assignment probabilities <0.95 (from supervised learning) placed in the third cluster (Fig. 4B). Only for 11 of the 337 snails (3.27%) did the classification differ between the two methods: the three individuals not assigned by supervised learning were assigned by GMM to *P. fibratus* (two individuals) and *P. porphyrostomus* (one individual), whereas the eight shells classified as *P. porphyrostomus* by the supervised learning algorithm were classified as either *P. fibratus* or the intermediate type by GMM. The geographic source of the snails that contribute to the three morphological clusters identified by GMM suggests that the intermediate shell phenotype has a limited distribution. The Youaty population sample consisted only of *P. fibratus*, but the Touete, Waatchia, Gadji, Kere and Vao population samples contained both *P. fibratus* and *P. porphyrostomus*. The Wapan population sample contained examples of all three clusters, while the Comwagna population sample consisted mostly of individuals assigned to the intermediate type, with a few individuals of *P. porphyrostomus* (Table 1).

Table 1. Numbers of *Placostylus* in the three phenotype clusters (species) as assigned by supervised and unsupervised analysis of shell size and shape for eight population samples from the Isle of Pines, New Caledonia.

Population	Comwagna	Gadji	Kere	Touete	Vao	Waatchia	Wapan	Youaty	Total
Supervised learning classification									
P. fibratus	0	29	31	26	52	38	23	25	224
P. porphyrostomus	8	9	7	7	16	9	5	0	61
Not assigned	28	1	2	1	3	1	15	0	51
Unsupervised learning classification									
P. fibratus	0	29	32	26	53	38	23	25	226
P. porphyrostomus	4	9	7	7	15	9	5	0	56
Intermediate phenotype	32	1	1	1	3	1	15	0	54

Individuals were accepted as assigned to a cluster only if their assignment score was >0.95.

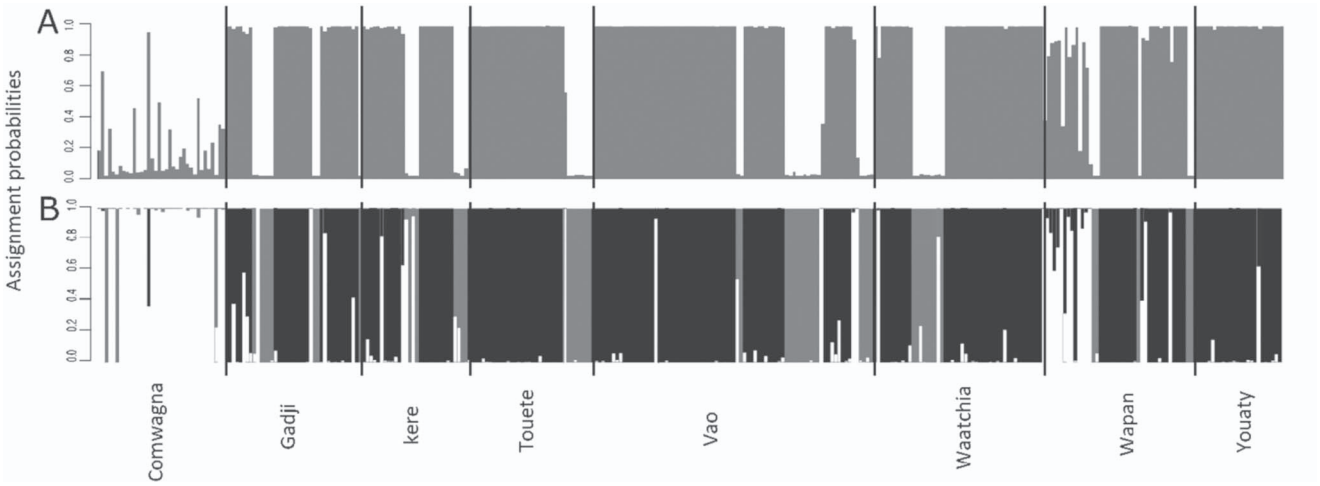

Figure 4. Assignment probabilities calculated for shell shape and size variation in eight population samples of *Placostylus* from the Isle of Pines. **A.** Neural network supervised learning analysis for two phenotypes (the two recognized species). Colour coding: grey, *P. fibratus*; white, *P. porphyrostomus*. **B.** An unsupervised analysis based on an optimal Gaussian mixture models that found three phenotypes (the two recognized species and the intermediate phenotype). Colour coding: grey, *P. porphyrostomus*; black, *P. fibratus*; white, intermediate phenotype.

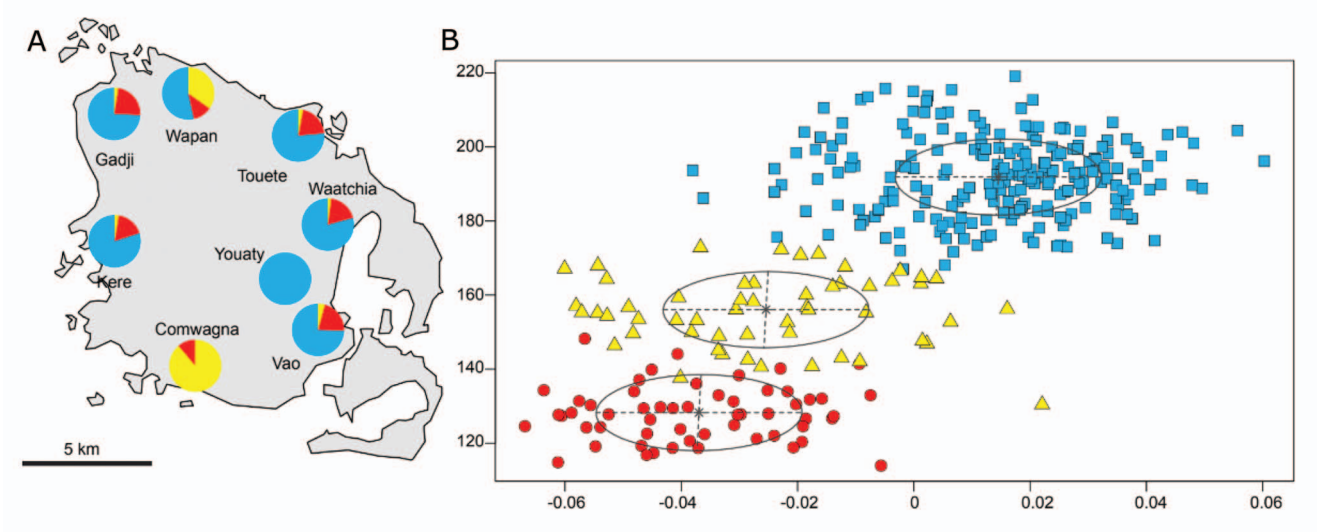

Figure 5. A. Relative frequency of three phenotypes of *Placostylus* in eight population samples from the Isle of Pines, New Caledonia, as indicated by the optimal GMM, using the first two PCs of variation in shell shape (50.8% of total shape variation) and shell size. **B.** Gaussian mixture model–based classification of *Placostylus* in relation to variation of shell size and shape along PC 1 (35.3%). Colour coding: red, *P. porphyrostomus*; blue, *P. fibratus*; yellow, intermediate phenotype.

DISCUSSION

Shell morphology, separable into shape and size components through geometric morphometrics, is controlled by a suite of heritable characters that can be influenced by environmental factors. Aquatic gastropods have provided evidence that phenotypic plasticity of shell size and shape is widespread (Bourdeau *et al.*, 2015). The New Zealand land snail *Placostylus ambagiosus* shows size variation that is, in part, due to environmental factors (Parrish, Stringer & Sherley, 2014). However, previous work on the two New Caledonian *Placostylus* species *P. fibratus* and *P. porphyrostomus* has shown that they are genetically distinct, despite coexisting at many locations on the Isle of Pines (Brescia *et al.*, 2008; Dowle *et al.*, 2015). Distinct shell traits of sympatric species are unlikely to be due to phenotypic plasticity, as the snails experience the same environment.

When applied to our new dataset, supervised learning algorithms failed to unambiguously assign all snails into one or other of the two species that are recognized on the basis of traditional shell-based taxonomy. Most of the specimens with low assignment probabilities originate from two newly sampled locations on the Isle of Pines. The presence of many unassigned snails suggests that the classification into two morphological groups (species) is too simplistic in the context of our expanded dataset. At the same time, unsupervised learning algorithms indicate the presence of a third group, and this result is consistent with our finding that supervised learning failed to classify some individuals. This third group is intermediate in both shape and size to the two recognized species (Fig. 5).

The snails from the Comwagna district, which constitute the majority of intermediate-type individuals, have never been harvested for commercial purposes due to their smaller size. To date, these individuals have been considered to be a subgroup of *P. fibratus* (Brescia *et al.*, 2008). Our analysis suggests that they are distinct and could potentially represent an unusual population of hybrid individuals. While available data on mt DNA haplotype distributions suggest historical hybridization between *P. fibratus* and *P. porphyrostomus*, the presence of morphologically and genetically distinct sympatric populations does not indicate ongoing hybridization (Dowle *et al.*, 2015). However, here, our morphological data suggest hybridization could go beyond simple mitochondrial introgression and could be the origin of an exclusively hybrid population at Comwagna and rare hybrids at other localities.

Hybridization can be viewed as a gradient that extends from rare introgression through stable hybrid zones to the formation of exclusively hybrid populations (Nieto Feliner *et al.*, 2017). Introgression can generate gene flow between distinct species, but parental lineages can remain distinct when hybrids normally backcross to one of the parental forms (Anderson, 1953; Folk *et al.*, 2018). Observations of five populations of *Placostylus* from the Isle of Pines (Gadji, Kere, Touete, Waatchia, Vao) seem to be consistent with the rare introgression hypothesis suggested by mt DNA haplotype data; this is because of the presence of just two distinctive parental phenotypic clusters and a small number (1.4–2.9%) of individuals with intermediate shell phenotype in these five populations (Fig. 3). Given the absence of intermediates at localities where *P. fibratus* and *P. porphyrostomus* are sympatric, we could apply the genotypic cluster definition (Mallet, 1995) and recognize two species. On the other hand, the numerical dominance of the intermediate phenotype in the population samples from Wapan and Comwagna suggests that if hybridization explains this phenotype, then hybridization has been locally common and has led to the loss of distinct parental phenotypic clusters at these sites.

Environmental conditions instead of interspecific hybridization may potentially explain the prevalence of individuals with an intermediate phenotype. Of the eight locations where snails were sampled, Comwagna is situated on red ferralitic (=laterite) soils, whereas all the others are on uplifted coral reef terrain (Lagarde & Ouetcho, 2016). Experiments have shown a positive correlation between calcium availability and shell growth in some terrestrial gastropods (Beeby & Richmond, 2007). Phenotypic plasticity is likely to account for a part of the morphological diversity observed for New Caledonian *Placostylus* (Brescia *et al.*, 2008; Dowle *et al.*, 2015). Lower calcium availability on ferralitic soils is, therefore, a plausible explanation for the intermediate phenotype at Comwagna, but this is challenged by the presence of that phenotype on calcium-rich substrate at Wapan. Although hybridization and recent gene flow between *P. fibratus* and *P. porphyrostomus* seems to be the most likely explanation for the presence of the intermediate phenotype, at this stage we cannot rule out the possibility that the intermediate phenotype represents a novel lineage distinct from *P. fibratus* and *P. porphyrostomus,* or the possibility of an explanation based on environmental factors.

Here, we used two approaches to classify samples: one uses a training dataset, the other does not. The first method seeks to classify individuals within a framework of knowledge, while the second explores a dataset to find optimal divisions. Both methods are subject to overfitting issues (Adams, Rohlf & Slice, 2004). Supervised learning can misclassify individuals that are not inside the range of variance of the training data, and unsupervised learning can under- or overestimate the number of clusters in a dataset. We suggest that low support values for individual classification from supervised learning can be biologically interesting. While low support assignment can arise from technicalities such as a small training dataset or mislabelling, here, the concordance of supervised learning results with unsupervised learning classification suggests our models were not lacking power. Our results, therefore, support the use of both methods, rather than either of them individually. Overall, our work joins the growing number of studies that have demonstrated that the association of geometric morphometrics, and machine learning can be useful in addressing biological questions (Dubey *et al.*, 2006; Bocxlaer & Schultheiß, 2010; Mapp *et al.*, 2017; Nattier *et al.*, 2017; Soda *et al.*, 2017; Fang *et al.*, 2018).

ACKNOWLEDGEMENTS

We thank the New Caledonian people, the Institut Agronomique Neo-Caledonien and members of the Phoenix group for Evolutionary Ecology and Genetics (evolves.massey.ac.nz). This research was supported by the Institute Agronomique Neo-Caledonien and a Massey University PhD scholarship (awarded to MQ).

REFERENCES

ADAMS, D., COLLYER, M. & KALIONTZOPOULOU, A. 2018. *Geomorph: software for geometric morphometric analyses, version 3.1.2*. Available at: https://CRAN.R-project.org/package=geomorph. Accessed 1 September 2019.

ADAMS, D.C., ROHLF, F.J. & SLICE, D.E. 2004. Geometric morphometrics: ten years of progress following the 'revolution. *Italian Journal of Zoology*, **71**: 5–16.

ANDERSON, E. 1953. Introgressive hybridization. *Biological Reviews*, **28**: 280–307.

BAYLAC, M., VILLEMANT, C. & SIMBOLOTTI, G. 2003. Comining geometric morphometrics with pattern recognition for the investigation of species complexes. *Biological Journal of the Linnean Society*, **80**: 89–98.

BEEBY, A. & RICHMOND, L. 2007. Differential growth rates and calcium-allocation strategies in the garden snail *Cantareus aspersus*. *Journal of Molluscan Studies*, **73**: 105–112.

BOCXLAER, B.V. & SCHULTHEIß, R. 2010. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine-learning approaches to biological discrimination. *Paleobiology*, **36**: 497–515.

BOOKSTEIN, F.L. 1991. *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, Cambridge.

BOURDEAU, P.E., BUTLIN, R.K., BRÖNMARK, C., EDGELL, T.C., HOVERMAN, J.T. & HOLLANDER, J. 2015. What can aquatic gastropods tell us about phenotypic plasticity? A review and meta-analysis. *Heredity*, **115**: 312–321.

BRESCIA, F. 2011. *Ecology and population trends in New Caledonian* Placostylus *snails (Mollusca: Gastropoda: Bulimulidae)*. PhD thesis. Massey University, Palmerston North, New Zealand.

BRESCIA, F.M., PÖLLABAUER, C.M., POTTER, R.A. & ROBERTSON, A.W. 2008. A review of the ecology and conservation of *Placostylus* (Mollusca: Gastropoda: Bulimulidae) in New Caledonia. *Molluscan Research*, **28**: 111–122.

BREURE, A.S.H., GROENENBERG, D.S.J. & SCHILTHUIZEN, M. 2010. New insights in the phylogenetic relationswithin the *Orthalicoidea* (*Gastropoda*, *Stylommatophora*) based on 28S sequence data. *Basteria*, **74**: 25–32.

CARVAJAL-RODRÍGUEZ, A., CONDE-PADÍN, P. & ROLÁN-ALVAREZ, E. 2005. Decomposing shell form into size and shape by geometric morphometric methods in two sympatric ecotypes of *Littorina saxatilis*. *Journal of Molluscan Studies*, **71**: 313–318.

CRUZ, R.A.L., PANTE, M.J.R. & ROHLF, F.J. 2012. Geometric morphometric analysis of shell shape variation in *Conus* (*Gastropoda*: *Conidae*). *Zoological Journal of the Linnean Society*, **165**: 296–310.

DALY, E. 2016. *Fine scale population structure through space and time*. PhD thesis. Massey University, Palmerston North, New Zealand.

DARWIN, C. 1859. *On the origin of species*. John Murray, London.

DE QUEIROZ, K. 2007. Species concepts and species delimitation. *Systematic Biology*, **56**: 879–886.

DILLON, R.T. & JACQUEMIN, S.J. 2015. The heritability of shell morphometrics in the freshwater pulmonate gastropod Physa. *PLoS One*, **10**: 1–13.

DOWLE, E.J., MORGAN-RICHARDS, M., BRESCIA, F. & TREWICK, S.A. 2015. Correlation between shell phenotype and local environment suggests a role for natural selection in the evolution of *Placostylus* snails. *Molecular Ecology*, **24**: 4205–4221.

DUBEY, B.P., BHAGWAT, S.G., SHOUCHE, S.P. & SAINIS, J.K. 2006. Potential of artificial neural networks in varietal identification using morphometry of wheat grains. *Biosystems Engineering*, **95**: 61–67.

DUBOIS, A. 2011. Species and "strange species" in zoology: do we need a "unified concept of species"? *Comptes Rendus Palevol*, **10**: 77–94.

FANG, Z., FAN, J., CHEN, X. & CHEN, Y. 2018. Beak identification of four dominant octopus species in the East China Sea based on traditional measurements and geometric morphometrics. *Fisheries Science*, **84**: 975–985.

FOLK, R.A., SOLTIS, P.S., SOLTIS, D.E. & GURALNICK, R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany*, **105**: 364–375.

FRALEY, C. & RAFTERY, A.E. 2006. *MCLUST version 3: an R Package for normal mixture modeling and model-based clustering*. Available at: https://cran.r-project.org/src/contrib/Archive/mclust.

FRITSCH, S., GUENTHER, F., SULING, M. & M. MUELLER, S. 2016. *Neuralnet: training of neural networks, version 1.44.2*. Available at: https://CRAN.R-project.org/package=neuralnet. Accessed 1 September 2019.

GUSTAFSON, K.D. & BOLEK, M.G. 2016. Effects of trematode parasitism on the shell morphology of snails from flow and nonflow environments. *Journal of Morphology*, **277**: 316–325.

GUSTAFSON, K.D., KENSINGER, B.J., BOLEK, M.G. & LUTTBEG, B. 2014. Distinct snail (*Physa*) morphotypes from different habitats converge in shell shape and size under common garden conditions. *Evolutionary Ecology Research*, **16**: 77–89.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York.

KENDALL, D.G. 1986. A survey of the statistical theory of shape. *Statistical Science*, **10**: 354–363.

KLINGENBERG, C.P. 2011. MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*, **11**: 353–357.

KLINGENBERG, C.P. 2016. Size, shape, and form: concepts of allometry in geometric morphometrics. *Development Genes and Evolution*, **226**: 113–137.

LAGARDE, L. & OUETCHO, A. 2016. Horticultural structures on ultramafic soils: the case of Isle of Pines and other parts of southern Grande Terre (New Caledonia). In: *La pratique de l'espace en Océanie: découverte, appropriation et émergence des systèmes sociaux traditionnels* (F. Valentin & G. Molle, eds), pp. 79–90. Société préhistorique Française, Paris.

MADERBACHER, M., BAUER, C., HERLER, J., POSTL, L., MAKASA, L. & STURMBAUER, C. 2008. Assessment of traditional versus geometric morphometrics for discriminating populations of the *Tropheus moorii* species complex (*Teleostei*: *Cichlidae*), a Lake Tanganyika model for allopatric speciation. *Journal of Zoological Systematics and Evolutionary Research*, **46**: 153–161.

MAHNER, M. 1993. What is a species? *Journal for General Philosophy of Science*, **24**: 103–126.

MALLET, J. 1995. A species definition for the modern synthesis. *Trends in Ecology & Evolution*, **10**: 294–299.

MAPP, J., HUNTER, E., KOOIJ, J., Van Der SONGER, S. & FISHER, M. 2017. Otolith shape and size: the importance of age when determining indices for fish-stock separation. *Fisheries Research*, **190**: 43–52.

MAYR, E. 1942. *Systematics and the origin of species*. Columbia University Press, New York.

MITTEROECKER, P. & GUNZ, P. 2009. Advances in geometric morphometrics. *Evolutionary Biology*, **36**: 235–247.

NATTIER, R., PELLENS, R., ROBILLARD, T., JOURDAN, H., LEGENDRE, F., CAESAR, M., NEL, A. & GRANDCOLAS, P. 2017. Updating the phylogenetic dating of New Caledonian biodiversity with a meta-analysis of the available evidence. *Scientific Reports*, **7**: 1–9.

NEUBERT, E., CHÉREL-MORA, C. & BOUCHET, P. 2009. Polytypy, clines, and fragmentation: the bulimes of New Caledonia revisited. *Mémoires du Muséum d'Histoire Naturelle*, **198**: 37–131.

NIETO FELINER, G., ÁLVAREZ, I., FUERTES-AGUILAR, J., HEUERTZ, M., MARQUES, I., MOHARREK, F., PIÑEIRO, R., RIINA, R., ROSSELLÓ, J.A., SOLTIS, P.S. & VILLA-MACHÍO, I. 2017. Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, **118**: 513–516.

PARRISH, G.R., STRINGER, I.A.N. & SHERLEY, G.H. 2014. The biology of *Placostylus ambagiosus* (*Pulmonata*: *Bulimulidae*) in New Zealand: part 1. Behaviour, habitat use, abundance, site fidelity, homing and the dimensions of eggs and snails. *Molluscan Research*, **34**: 139–154.

R CORE TEAM. 2017. R: a language and environment for statistical computing, version 2.6.2. R Foundation for Statistical Computing, Vienna. Available at: http://CRAN.R-project.org/.

RAO, S.R., LIEW, T., YOW, Y. & RATNAYEKE, S. 2018. Cryptic diversity: two morphologically similar species of invasive apple snail in Peninsular Malaysia. *PLoS One*, **13**: e0196582.

RIPLEY, B.D. 1994. Neural network and related methods for classification. *Journal of the Royal Statistical Society, Series B (Methodological)*, **56**: 409–456.

ROHLF, F.J. 2015. The tps series of software. *Hystrix, the Italian Journal of Mammalogy*, **26**: 9–12.

ROHLF, F.J. & MARCUS, L.F. 1993. A revolution in morphometrics. *Trends in Ecology & Evolution*, **8**: 129–132.

SCHILTHUIZEN, M. & HAASE, M. 2010. Disentangling true shape differences and experimenter bias: are dextral and sinistral snail shells exact mirror images? *Journal of Zoology*, **282**: 191–200.

SODA, K.J., SLICE, D.E. & NAYLOR, G.J.P. 2017. Artificial neural networks and geometric morphometric methods as a means for classification: a case-study using teeth from *Carcharhinus* sp. (*Carcharhinidae*). *Journal of Morphology*, **278**: 131–141.

VAUX, F., CRAMPTON, J.S., MARSHALL, B.A., TREWICK, S.A. & MORGAN-RICHARDS, M. 2017. Geometric morphometric analysis reveals that the shells of male and female siphon whelks *Penion chathamensis* are the same size and shape. *Molluscan Research*, **37**: 194–201.

VAUX, F., TREWICK, S.A., CRAMPTON, J.S., MARSHALL, B.A., BEU, A.G., HILLS, S.F.K. & MORGAN-RICHARDS, M. 2018. Evolutionary lineages of marine snails identified using molecular phylogenetics and geometric morphometric analysis of shells. *Molecular Phylogenetics and Evolution*, **127**: 626–637.

VAUX, F., TREWICK, S.A. & MORGAN-RICHARDS, M. 2016. Speciation through the looking-glass. *Biological Journal of the Linnean Society*, **120**: 480–488.

VERHAEGEN, G., MCELROY, K.E., BANKERS, L., NEIMAN, M. & HAASE, M. 2018. Adaptive phenotypic plasticity in a clonal invader. *Ecology and Evolution*, **8**: 4465–4483.

ZELDITCH, M., SWIDERSKI, D. & SHEETS, H.D. 2004. *Geometric morphometrics for biologists: a primer*. Elsevier Academic Press, San Diego.

ZHOU, ZH. 2009. Ensemble learning. In: *Encyclopedia of biometrics*. (S.Z. Li, ed.), pp. 270–273. Springer, Berlin.