Uncertain Lines: Analysing phenotype data of Apistogramma species

ABSTRACT: There is great variation within the genus Apistogramma, which occurs in many habitats across South America. This study combined ecological, phenotypic, and lineage data to determine whether ecology or shared history best explain variation in the genus. Multiple Correspondence Analysis (MCA) showed that lineage drives clustering of species. This study highlights the need for comprehensive molecular data and thoughtful sampling to resolve the relationship between ecology, phylogeny, and phenotype in Apistogramma.

INTRODUCTION

Apistogramma is a genus of cichlid found in habitats across South America. Species live in a broad range of conditions and vary in their patterns (Fig. 1). For this project I explored the ecological factors that might explain the variation in *Apistogramma* phenotype and examined ecological and phenotypic traits for approximately 100 species in the genus (Römer 2006). I believe that the visual conditions where the fish live are responsible for the phenotypes of the species that evolve, and I predicted that analysis would cluster species by phenotype best when explained by water type, as this encapsulates numerous ecological factors and determines the visual environment for this highly communicative species.

METHODS: DATA DESCRIPTION

The data come from two tables (Römer 2006), as well as a paper by Tougard *et al* (2017). The first table is categorical ecological data for 102 species of *Apistogramma*, including water body type, water type, habitat type, and river system. The second table is a compilation of 51 phenotype characters of the same *Apistogramma* species. The study by Tougard *et al.* used molecular data from 30 species of *Apistogramma* to construct a phylogenetic tree, resulting in four main clades, or lineages. I compiled the species identified in the paper and assigned lineage to species that are suspected to be within the same species complex.

I removed rows for species that were synonyms as they were not independent observations (Froese & Pauly, 2024). After corrections, the character data had 94 species (rows) and 50 variables, the ecological data had 94 species and 14 variables, and the lineage data had 33 species and one variable (see Supplementary Material). Once combined and filtered for missing data (Wickham *et al* 2019), any variable with a single value was removed as it could not inform the analyses. This process removed all species occurring in the Orinoco and Other river systems and remaining occur in the Amazon river system. The final data set covered 33 species and 55 categorical variables.

METHODS: DATA ANALYSIS

The data were all categorical in nature and so avenues for analysis were limited but prudent given the potential expense associated with gathering quantitative data for the entire genus. I conducted a Multiple Correspondence Analysis (MCA), a method for analysing large sets of categorical variables to discover patterns in low-dimensional space ((Le *et al* 2008; Statistical Tools for High-throughput Data Analysis 2017; Kassambara & Mundt 2020, Wiki Contributors).

This clustered like species together and identified the variables that best explain the shape of the data (Rdocumentation(a)). I plotted the species within the new dimensions of the MCA and coloured iterations by water type as well as lineage. I extracted results for the variable categories to determine which contributed most to the new dimensions (Rdocumentation(c)). I created a new data frame of species and their respective loadings as continuous data from dimension 1 and dimension 2. I created a standard Euclidean distance matrix from the dimension 1 loadings (Rdocumentation(b)) and visualized the strength of relationships between the species in a qgraph (Eskamp *et al* 2012). While Tougard *et al* used maximum likelihood and Bayesian inference to create their phylogeny, I did not have molecular data required for 1:1 comparison. I chose between two common methods available for building trees with distance-based data: neighbour-joining tree estimation (Paradis & Schliep 2019), and unweighted pair group method with arithmetic mean (UPGMA) (Schliep 2011). I compared the species lineages on the trees to see whether the addition of ecological data created a tree in concordance with one based on either molecular or morphological data alone.

RESULTS

The Multiple Correspondence analysis created five new dimensions for the data, the first explaining 20.7% of the variation, the second explaining 16.9% (Fig. 2). For dimension 1, the most correlated variable was lineage ($r^2 = 0.9304$) (Table 1), whereas the closest ecological variable was whitewater ($r^2 = 0.1707868$). The variables most correlated with dimension 2 were head-body proportion, followed by lineage. Ecological variables did not rank in the top 25 most correlated with dimension 2. Biplots of species grouped by water type do not encompass the data as completely as when grouped by lineage (Fig. 3). The qgraph of the distance matrix displays strong correlations between species in lineage 1 and 3, with weaker relationships in lineages 2 and 4 (Fig. 4). The topology of the UPGMA tree most resembled the tree in Tougard *et al* and was selected for comparison. While lineages 1 and 3 are nearly identical to the layout in the paper, lineages 2 and 4 are mixed (Fig. 5).

DISCUSSION

This analysis showed that lineage, not ecology, was the main contributor to the new dimensions incorporating the data. Limitations of the data may have been significant as all fish were from the Amazon river system, the largest in the world. The weak correlations of lineages 2 and 4, and shifts in the phylogeny, may be due to incomplete sampling across lineages with greater endemism. Little detail was given by Römer's regarding his data collection process. Beyond this uncertainty, there is a great deal of variation not captured by the MCA. Research has suggested that *Apistogramma* evolve a paedomorphic phenotype when co-occurring with sister clade *Geophagus* (Steele 2018). Key ecological variables were missed, notably community structure and predation. This work highlights the pitfalls of sampling only widely distributed species when you need to capture all the variation that exists. A complete molecular database and the adoption of standard best practices when collecting field data would be of immense help in understanding the phenotypes, and resolving the phylogeny, of the genus.

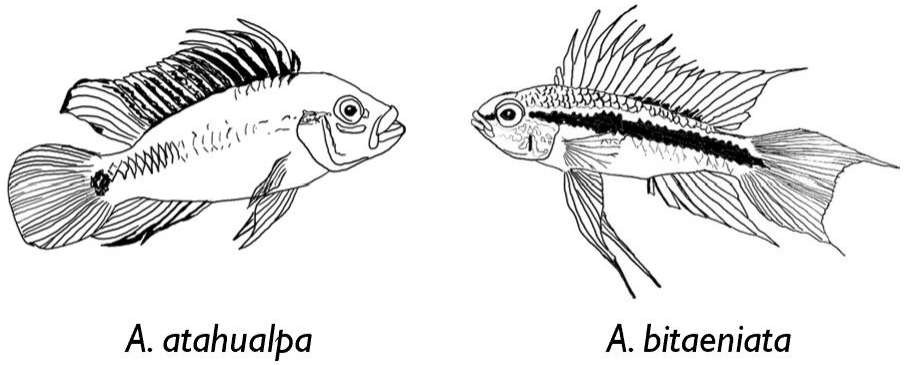

Figure 1: Apistogramma species vary in morphology, colour, and dimorphism, as demonstrated by the two males. Sketches by J. Bullock.

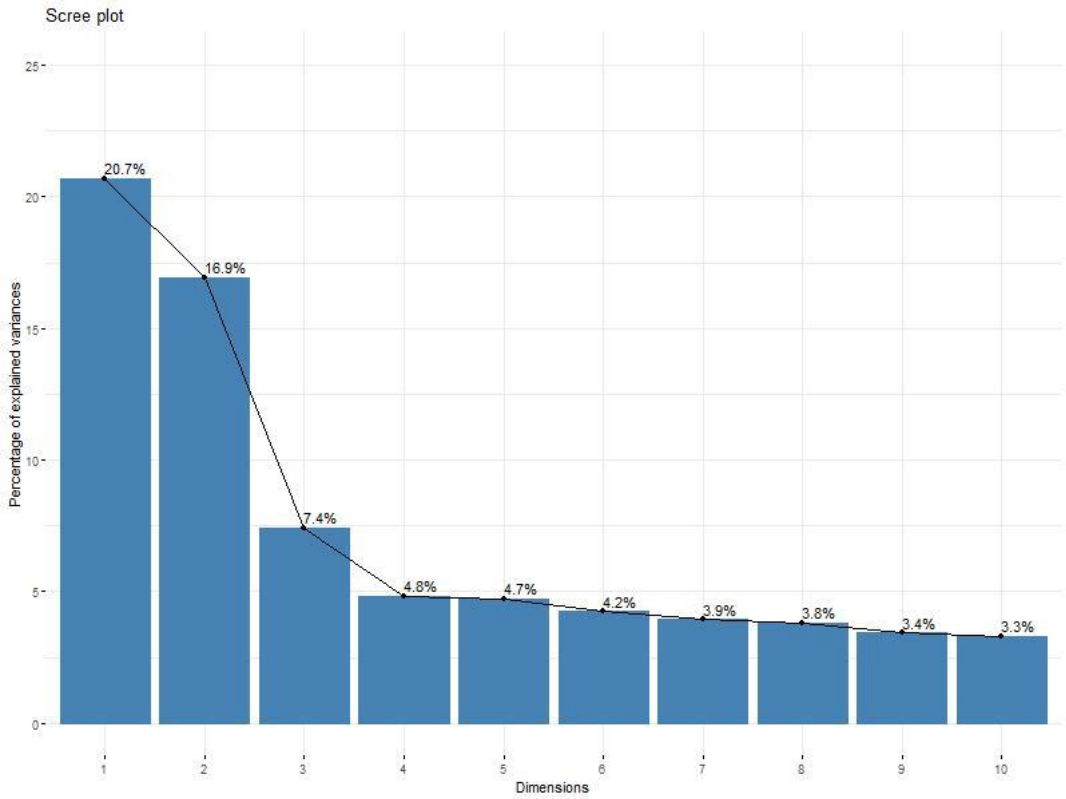

Figure 2: Scree plot showing % variance explained by each dimension of the MCA

TABLE 1: Variables and their r^2 values, which is a measure of the strength of association between the variables and the axis in question, for the top 5 contributors to dimension 1 and 2.
Determined by 1-way ANOVA

Dimension 1	r2	Dimension 2	r2
Lineage	0.9304433	Head-body proportion	0.9139043
Body form	0.7619608	Lineage	0.9136878
Infraorbital pores	0.7619608	Jaw spot	0.7130203
Longitudinal band	0.7172129	Chin spot	0.6938044
Caudal spot	0.7093105	Chest spot	0.6105135

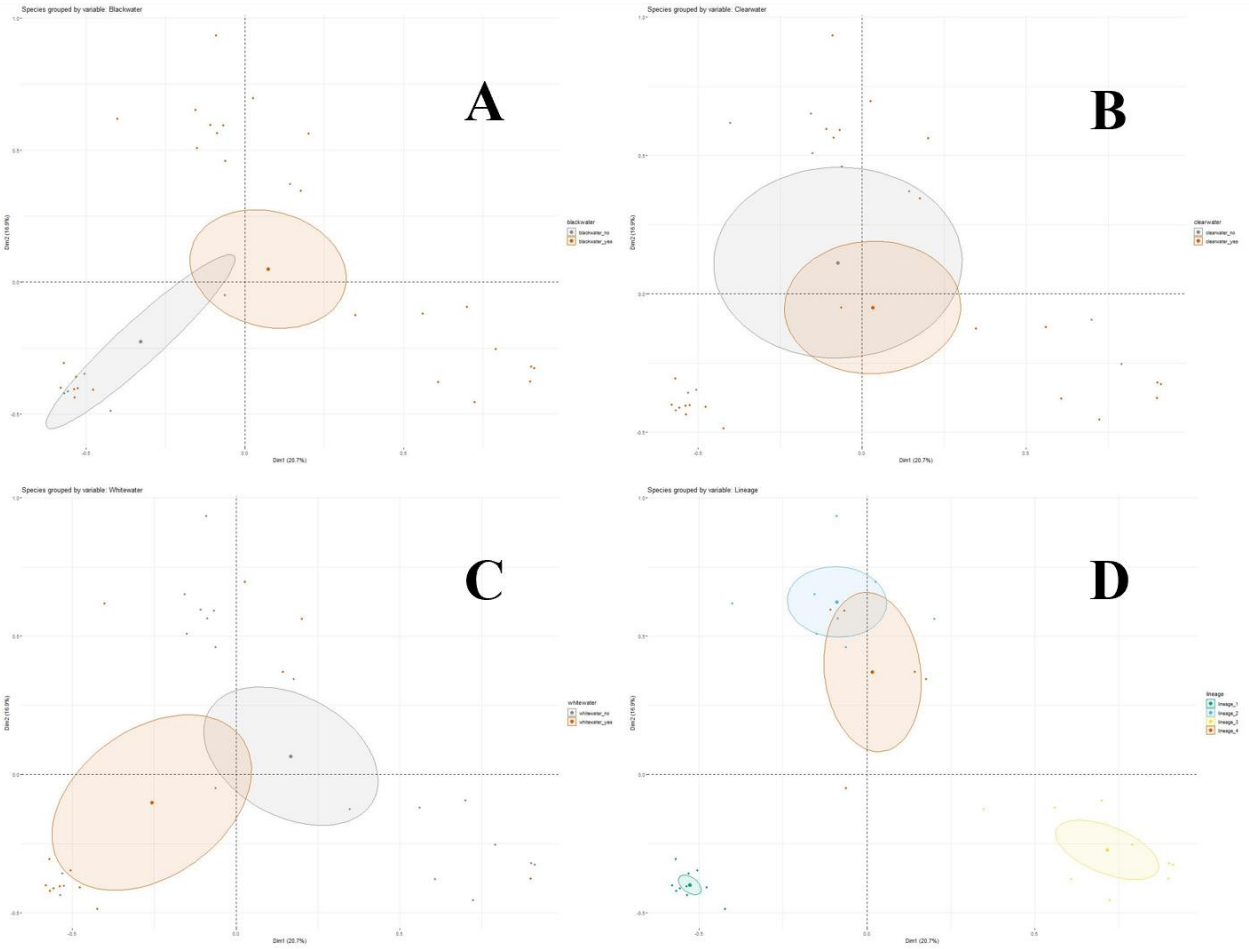

Figure 3: Biplots showing MCA grouping of species by A) blackwater, B) clearwater, C) whitewater, and D) lineage on dimensions 1 and 2. Ellipses denote confidence around category mean.

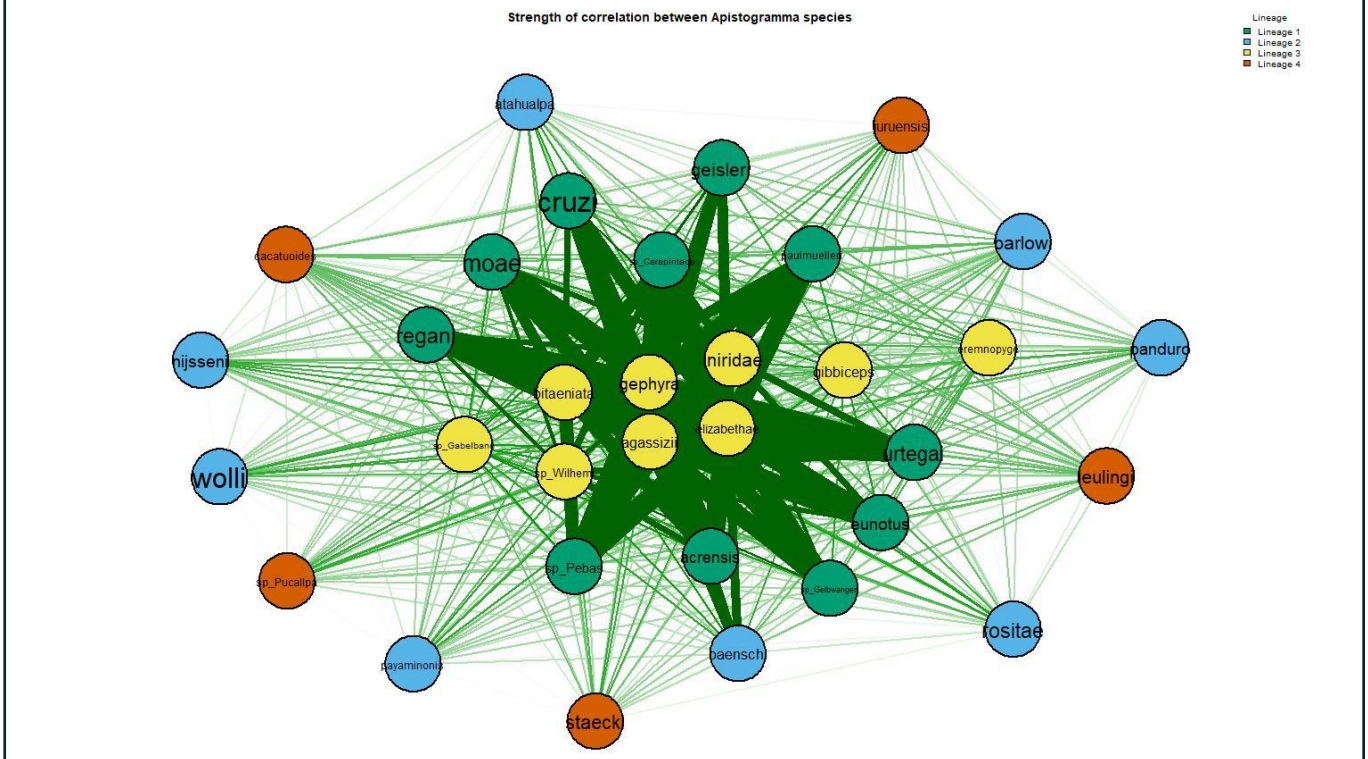

FIGURE 4: Qgraph representing relationships between nodes of the distance matrix created from dimension 1 loadings, species coloured by lineage. The thickness of the lines represents the strength of correlation between species; stronger correlation will show a thicker line,

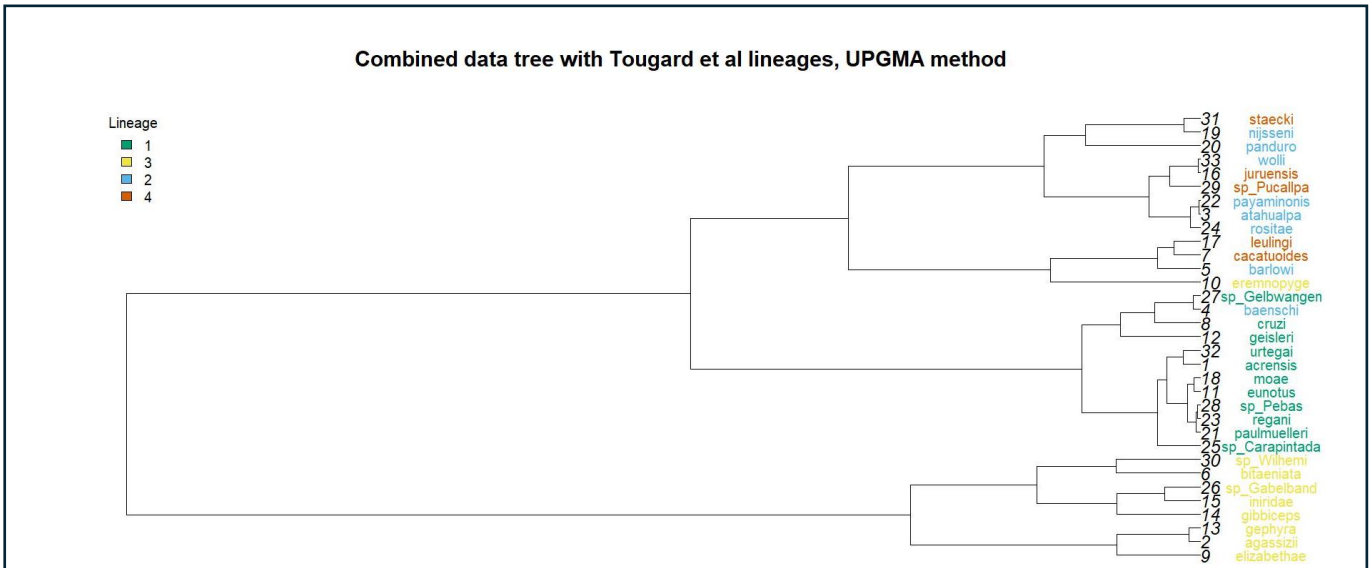

FIGURE 5: Phylogenetic tree constructed using UPGMA method from the MCA dimension 1 loadings distance matrix. Species coloured by lineage from Tougard *et al* 2017, disparity in grouping shows shifts in position.

SUPPLEMENTARY MATERIAL

Apisto_character_data.csv
Apisto_ecological_data.csv
Apisto_lineage_data.csv

Apistogramma character data description.docx
Apistogramma ecological data description.docx
Apistogramma lineage data description.docx

EEB313 Project Code.Rmd

REFERENCES

Chang, W. A colorblind-friendly palette. http://www.cookbook r.com/Graphs/Colors_(ggplot2)/#a-colorblind-friendly-palette. Creative Commons Attribution-Share Alike 3.0 Unported License.

Epskamp, S, Cramer, A, Waldorp, L, Schmittmann, V, Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. Journal of Statistical Software, 48(4), 1-18. URL http://www.jstatsoft.org/v48/i04/.

Froese, R. and Pauly, D. [Editors]. (2024). FishBase. www.fishbase.org, version (06/2024).

Jombart, Thibaut. Introduction to phylogenetics using R. Imperial College, London. https://adegenet.r-forge.r-project.org/files/MSc-intro-phylo.1.1.pdf

Kassambara A, Mundt F. (2020). _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7, <https://CRAN.R-project.org/package=factoextra>.

Le, S, Josse, J, Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

Linck, Ethan. Quick and dirty tree building in R. February 26, 2016. https://www.molecularecologist.com/2016/02/26/quick-and-dirty-tree-building-in-r/

Paradis E, Schliep K. (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.” _Bioinformatics_, *35*, 526-528.

Römer U. (2006). Cichlid Atlas Volume 2. 1st Ed. Melle, Germany: Mergus.
Table 2: Habitat preferences and degree of specialisation of *Apistogramma*-species, pp. 194-195
Basic data for clusteranalysis of relationships within Apistogramma-species, pp. 1290-1297

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rdocumentation(a). MCA: Multiple Correspondence Analysis (MCA). https://www.rdocumentation.org/packages/FactoMineR/versions/2.9/topics/MCA

Rdocumentation(b). dist: Distance Matrix Computation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/dist

Rdocumentation(c). fviz_ellipses: Draw confidence ellipses around the categories. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_ellipses

Statistical Tools for High-throughput Data Analysis. MCA - Multiple Correspondence Analysis in R: Essentials. September 24, 2017. https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/

Steele, S. (2018). Chapter 3: Extreme body size reduction, morphological modifications, and allometry in Neotropical cichlid fishes (Cichliformes:Cichlidae:Cichlinae) in *Body size evolution and diversity of fishes using the neotropical cichlids (cichlinae) as a model system* [Thesis]. Available from Dissertations & Theses @ University of Toronto; ProQuest Dissertations & Theses Global

Schliep K. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593

Tougard C, García Dávila CR, Römer U, Duponchelle F, Cerqueira F, Paradis E, Guinand B, Angulo Chávez C, Salas V, Quérouil S, Sirvas S, Renno JF. (2017). Tempo and rates of diversification in the South American cichlid genus Apistogramma (Teleostei: Perciformes: Cichlidae). PLoS One. Sep 5;12(9):e0182618.

The Comprehensive R Archive Network. Confidence Ellipse. https://cran.r-project.org/web/packages/ConfidenceEllipse/vignettes/confidence-ellipse.html

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). Welcome to the tidyverse. Journal of Open Source Software, *4*(43), 1686.

Wikipedia contributors, "Multiple correspondence analysis," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Multiple_correspondence_analysis&oldid=1252487684 (accessed December 10, 2024).