

## Short Communication

## Does phylogeny have an influence on the date of first description? A comparative study of the world's fishes

Brandon Beukes<sup>a</sup>, Peter Witbooi<sup>b</sup>, Mark J. Gibbons<sup>a,\*</sup><sup>a</sup> Department of Biodiversity and Conservation Biology, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa<sup>b</sup> Department of Mathematics and Applied Mathematics, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa

## ARTICLE INFO

## Article history:

Received 20 March 2020

Received in revised form 24 April 2020

Accepted 24 April 2020

## Keywords:

Biodiversity

Discovery

Estuarine

Freshwater

Marine

Phylogenetic contrasts

## ABSTRACT

The process of species description is not random, and understanding the factors that influence when a species is first described (the date of first description, DoFD) allows us to target environments and/or species' traits to increase our knowledge of diversity. Such studies typically correlate species traits (e.g. maximum size, occupational depth) against DoFD, forgetting that species are not statistically independent of each other, owing to the inheritance of shared characteristics. A recent study of extant fishes by Costello et al. (2015) identified depth and geographic range size as the most important (of many) predictors of the DoFD, implying that newly described species will likely occupy restricted areas and occur deep in the water column. However, these authors failed to accommodate for "identity by descent" in their analyses. We correct that oversight here, and conclude that while the adjustments strengthen the associations between the different predictors and the DoFD, the overall affects are minimal and they do not materially change Costello et al.'s (2015) conclusions. This is briefly discussed.

© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

New species are constantly being described, and a number of factors are now known to influence species discovery and date of first description (DoFD) (Blackburn and Gaston, 1995). These factors are identified using simple regression techniques (Gaston, 1991; Krasnov et al. 2005; Reed and Boback, 2002), and once known they allow us to predict where new species may be found (Randhawa et al. 2015).

Costello et al. (2015) recently conducted a comprehensive study of the factors influencing the DoFD of the global fish fauna. Their results suggest that new species will occupy small geographic ranges, but that they should not necessarily be smaller than previously described species (Costello et al. 2015). While the Costello et al. (2015) paper is insightful, it did not consider the influence of phylogeny on relationships (Blackburn and Gaston, 1995).

Species are descended in a hierarchical manner from their common ancestors, and cannot be considered as independent data points in comparative analyses. Blackburn and Gaston (1995), who were the first to incorporate phylogeny into studies such as this one, explained the problem in their study of South American oscine passerine birds. They suggested that if, by comparison with other taxa, all tanagers were small but had a bright plumage and active behaviour, and that conspicuity increases the likelihood of discovery, then the relationship between DoFD and body size becomes confounded by the in-your-

\* Corresponding author.

E-mail address: [mgibbons@uwc.ac.za](mailto:mgibbons@uwc.ac.za) (M.J. Gibbons).

face nature of tanagers and not size, *per se*. “Small body size and conspicuous plumage may have evolved only once, and then been retained during the radiation of the tanager tribe. But in cross-species comparisons, these traits would be falsely represented as having evolved multiple times” (Blackburn and Gaston, 1995, pp9). Because closely related species may share certain morphological, distributional and biological characteristics through a common ancestor, conventional parametric methods using species units are not suitable, and it is necessary to correct for this accordingly (Blackburn and Gaston, 1995; Grafen, 1989; Harvey and Pagel, 1991; Krasnov et al. 2005; Purvis and Rambaut, 1995).

Here, we re-examine the findings presented by Costello et al. (2015), using their data, in order to correct for “identity by descent”, which allows for an unambiguous interpretation of pattern.

## 2. Method and materials

The raw data used here have been derived from, and are shown in, Costello et al. (2015), though FishBase (Froese and Pauly, 2014) provided the original dataset. FishBase is an online database that contains information about all known species of fish, including (not exhaustively) commercial importance, size, depth of occupation, taxonomy, habitat and distribution. Of these data, we have used vertical depth range (m), minimum depth of occupation (m), mean latitude occupied (degrees), and both latitudinal and longitudinal range (degrees) as well as maximum length (mm) as predictors of the DoFD. All the predictors examined here were also used in the study of Costello et al. (2015).

In order to get around the problem of independent data points for any character (e.g. size, date of first description, geographic range etc), we need to adjust for the “influence” of inheritance on that character in the species, through the taxonomic hierarchy. Here we assume the taxonomy of Nelson et al. (2016) (to Class level) to be a proxy for phylogeny. Although Felsenstein (1985) first came up with a method of comparing pairs of taxa at successive nodes of a branching phylogeny, we have used the methods employed by Pagel and Harvey (1989), Harvey and Pagel (1991) and Grafen (1989), because the exact phylogeny of fishes remains unknown. The method constructs “a single value (a contrast) for each variable within each taxon, which represents the magnitude and direction of the change in the variable within the taxon” (Blackburn and Gaston, 1995; pp10). The contrast is essentially calculated successively as the product of differences between each character at one taxonomic level or node and the average for the higher taxonomic level (species, genus; genus, family; etc.), weighted by nodal diversity.

Prior to the calculation of the contrasts, all data were  $\log_{10}(x+1)$  transformed. Contrasts on each predictor were pooled across the phylogeny and separately regressed against contrast on year of first description, forced through the origin. The datasets for these variables were not consistent in size, owing to variable gaps in knowledge. In order to explore synergistic responses, a multiple regression analysis was implemented in STATISTICA v5.5, with contrasts on all predictors regressed against contrast on year of first description.

## 3. Results

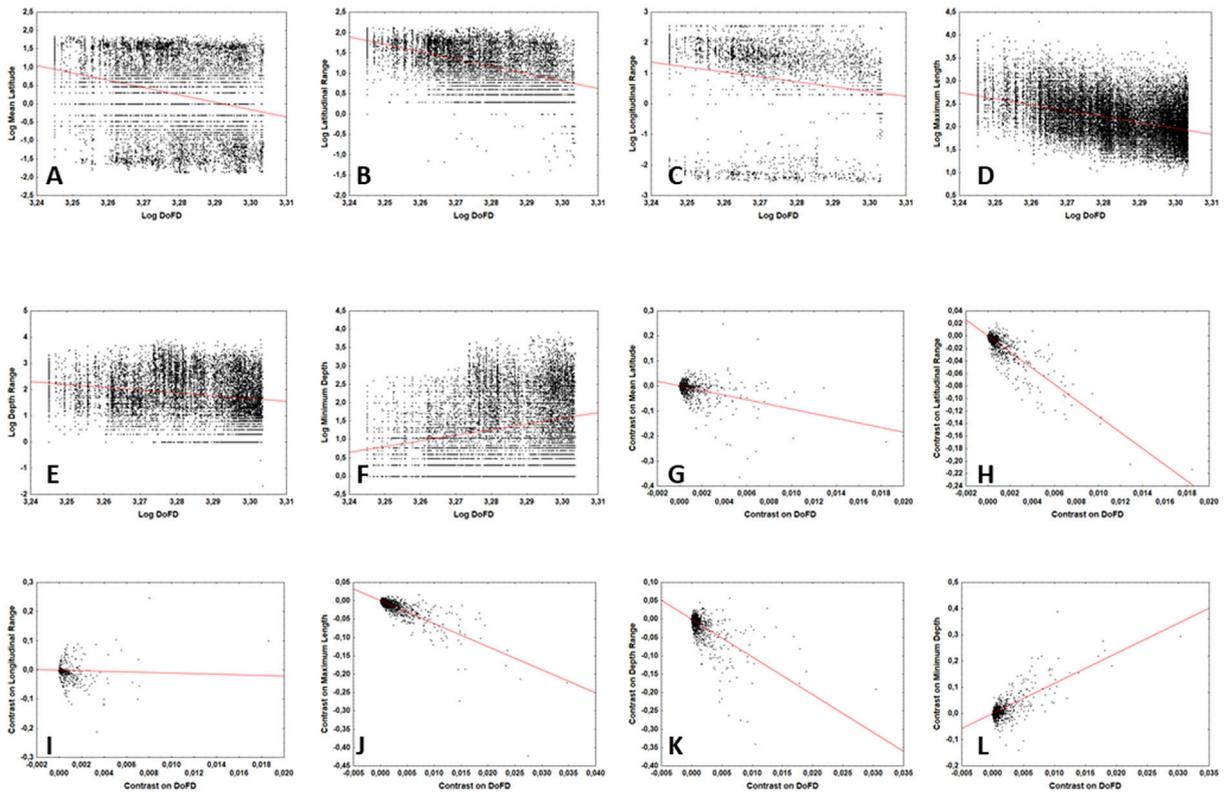
Correlation coefficients between the different  $\log_{10}(x+1)$  predictors and DoFD are presented in Table 1, as too are those of the independent contrasts. Comparative scatterplots are presented in Fig. 1. It can be seen from Table 1 that all relationships between independent contrasts are significant, though the strength of the relationships vary. In summary, there was a negative relationship between DoFD and mean latitude and latitudinal range, maximum length and depth range, implying that more recently described species are from the Southern Hemisphere and have restricted distributions, that they are small and occupy narrow depth ranges. Positive relationships were observed only between DoFD and minimum depth and longitudinal range (albeit weakly), suggesting that recently described species occur at deep depths and that they occur across a relatively wide longitudinal range.

A multiple regression analysis was performed on the independent contrasts of the predictors and date of first description. The results of the multiple regression are presented in Table 2. The results of the regression indicated that the model explained 74% (adjusted  $R^2$ ) of the variance and that the model was a significant predictor of the DoFD;  $F(6, 961) = 459.89$ ,  $P < 0.000$ . As seen in Table 2, each of the variables contributes significantly to the model, though latitudinal range contributed the most significantly, relationships between each predictor and the DoFD are as described for the univariate regressions (Table 1).

**Table 1**

Correlation coefficients between the different  $\log_{10}(x+1)$  predictors and date of first description for both the independent contrasts and raw data.

Predictor	N	Raw		Contrasts	
		r	p	r	p
Minimum depth	17 726	0.394	0.0000	0.735	0.0000
Long range	25 505	-0.130	0.0000	-0.048	0.0000
Latitude mean	20 094	-0.276	0.0000	-0.432	0.0000
Latitude range	20 552	-0.419	0.0000	-0.873	0.0000
Depth range	18 105	0.055	0.0000	-0.674	0.0000
Length max	4756	-0.315	0.0000	-0.831	0.0000



**Fig. 1.** The relationship between the  $[\log_{10}]$  date of first description (DoFD) of global fish species and  $[\log_{10}]$  selected predictor variables (from Costello et al., 2015) (A–F). The relationships between the contrast on the DoFD and the contrast on predictor variables are represented from G – L. For example, a negative correlation between maximum length infers that an increase in date of first description of fish species will be correlated to a decrease in the maximum length of those species described. Statistics generated for these relationships can be found in Table 1.

#### 4. Discussion

Our results show strong evidence for a relationship between the DoFD and all of the selected predictor variables independently and in combination. Latitudinal range is the most significant predictor: a finding in common with that of Costello et al. (2015) and most of the other studies mentioned previously. Mean latitude was also an important predictor, and together these results imply that recently described species occupy narrow latitudinal ranges within the Southern Hemisphere. This is perhaps not unexpected as most historical centres of taxonomy were centred in London, Paris, Berlin and various museums in North America (Funk, 2018), and mounting expeditions to the Southern Hemisphere would have been very costly. It is also a truism that recently described species will have small geographic ranges because there has been little time to properly analyse their ranges.

Maximum length was also found to be an important predictor of the DoFD, which is contradictory to the findings of Costello et al. (2015). Despite this, Costello et al. (2015) did acknowledge that since the early 1970s the rate of description for small fish has surpassed that of larger fish (Costello et al. 2015). Rhandhawa et al. (2015) also found that body size was a strong predictor of the DoFD for sharks and a number of other authors have noted its similar influence in Australian beetles, North American butterflies and fleas (Allsopp, 1997; Gaston et al. 1995; Krasnov et al. 2005). Conversely, other studies have found body size to be a poor, if nevertheless significant, predictor for passerine birds, reptiles and branchiopods (Adamowicz and Purvis, 2005; Blackburn and Gaston, 1995; Reed and Boback, 2002). Large organisms are more likely to be detected and caught than small ones, unless animals are collected in a size-independent manner (Gibbons et al. 2005). Moreover, in recent times, the technology to examine small organisms and recognise unique identity has improved (Valdemarsen, 2001).

The results indicate that recently described species of fish are likely to be deep living and occupy narrow depth ranges. Neither of these relationships are unexpected, in large part due to the increase in deep sea exploration with major technological advancements in research tools. The former result is in agreement with the suggestion of (Gibbons et al., 2005), and both concur with the results of Costello et al. (2015).

The results of this study agree, in general, with those obtained by Costello et al. (2015). However, there are some important differences. Firstly, the nature of the relationships between the date of first description and the various predictors are stronger using phylogenetic contrasts (Table 1; Table 4 of Costello et al., 2015). In addition, although latitudinal range and maximum

**Table 2**

Results of the multiple regression analysis of the variables on date of first description.  $R = 0.861$ ,  $R^2 = 0.741$  Adjusted  $R^2 = 0.739$ ;  $F(6,961) = 457.89$   
 $p < 0.00001$  Std. Error of estimate: 0.00024.

	Beta	Std.Err. of Beta	B	Std.Err. of B	t (961)	p-level
Intercept			0.0000	0.0000	4.1043	0.0000
Depth range	0.0738	0.0219	0.0063	0.0019	3.3728	0.0008
Latitude mean	-0.2133	0.0202	-0.0067	0.0006	-10.5772	0.0000
Maximum length	-0.2057	0.0199	-0.0364	0.0035	-10.3207	0.0000
Longitude range	0.0821	0.0181	0.0025	0.0005	4.5383	0.0000
Minimum depth	0.2584	0.0182	0.0187	0.0013	14.2129	0.0000
Latitude range	-0.5381	0.0241	-0.0438	0.0020	-22.3393	0.0000

length are the first and second most useful predictors in both univariate analyses, we have identified minimum depth as being third most useful, whilst [Costello et al. \(2015\)](#) recognised mid-latitude. These relationships vary slightly when considered together in the multiple regression, but the conclusions of both studies are similar.

Our data suggest the absence of a significant phylogenetic effect on the overall relationship between the DoFD and the selected predictor variables. In other words, the general conclusions of [Costello et al. \(2015\)](#) are upheld. Is this realistic? Undoubtedly, controlling for phylogeny is a more robust and statistically rigorous method of analysis when it is suspected that there may be a fundamentally phylogenetic bias. Hence, studies of (e.g.) cranium and body size in primates need to take cognisance of shared identity ([Collen et al. 2004](#)). However, when the questions asked reflect the human endeavour (as DoFD), it could be argued that phylogeny is relatively immaterial. Certainly, in the studies published to date, correcting for phylogeny in studies such as these serve to strengthen and not weaken the patterns made, and conclusions drawn, from the analysis of raw data ([Blackburn and Gaston, 1995](#); [Krasnov et al. 2005](#)).

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

We would like to thank The National Research Foundation and the University of the Western Cape for financial and logistic support. We are grateful to the comments provided by the reviewer and editor, which have helped to clarify the text. Brandon Beukes is especially grateful to the NRF's Phuhliisa programme for bursary assistance.

### References

- Adamowicz, S.J., Purvis, A., 2005. How many branchiopod crustacean species are there? Quantifying the components of underestimation. *Global Ecol. Biogeogr.* 14, 455–468.
- Allsopp, P., 1997. Probability of describing an Australian scarab beetle: (r) Influence of body size and distribution. *J. Biogeogr.* 24, 717–724.
- Blackburn, T.M., Gaston, K.J., 1995. What determines the probability of discovering a species? A study of South American oscine passerine birds. *J. Biogeogr.* 7–14.
- Collen, B., Purvis, A., Gittleman, J.L., 2004. Biological correlates of description date in carnivores and primates. *Global Ecol. Biogeogr.* 13, 459–467.
- Costello, M.J., Lane, M., Wilson, S., Houlding, B., 2015. Factors influencing when species are first named and estimating global species richness. *Global Ecol. Conserv.* 4, 243–254.
- Felsenstein, J., 1985. Phylogenies and the comparative method. *Am. Nat.* 125, 1–15.
- Froese, R., Pauly, D., 2014. FishBase. World Wide Web electronic publication version (8/2014).
- Funk, V.A., 2018. Collections-based science in the 21st century. *J. Systemat. Evol.* 56, 175–193.
- Gaston, K.J., 1991. Body size and probability of description: the beetle fauna of Britain. *Ecol. Entomol.* 16, 505–508.
- Gaston, K.J., Blackburn, T.M., Loder, N., 1995. Which species are described first? The case of North American butterflies. *Biodivers. Conserv.* 4, 119–127.
- Gibbons, M.J., Richardson, A.J., Angel, M., Buecher, E., Esnal, G., Fernandez Alamo, M.A., Gibson, R., Itoh, H., Pugh, P., Boettger-Schnack, R., 2005. What determines the likelihood of species discovery in marine holozooplankton: is size, range or depth important? *Oikos* 109, 567–576.
- Grafen, A., 1989. The phylogenetic regression. *Phil. Trans. Roy. Soc. B* 326, 119–156.
- Harvey, P.H., Pagel, M.D., 1991. *A Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.
- Krasnov, B.R., Shenbrot, G.I., Mouillot, D., Khokhlova, I.S., Poulin, R., 2005. What are the factors determining the probability of discovering a flea species (Siphonaptera)? *Parasitol. Res.* 97, 228–237.
- Nelson, J.S., Grande, T.C., Wilson, M.V.H., 2016. *Fishes of the World*. Wiley, New Jersey.
- Pagel, M.D., Harvey, P.H., 1989. Comparative methods for examining adaptation depend on evolutionary models. *Folia Primatol.* 53, 203–220.
- Purvis, A., Rambaut, A., 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Bioinformatics* 11, 247–251.
- Randhawa, H.S., Poulin, R., Krkošek, M., 2015. Increasing rate of species discovery in sharks coincides with sharp population declines: implications for biodiversity. *Ecography* 38, 96–107.
- Reed, R.N., Boback, S.M., 2002. Does body size predict dates of species description among North American and Australian reptiles and amphibians? *Global Ecol. Biogeogr.* 11, 41–47.
- Valdemarsen, J.W., 2001. Technological trends in capture fisheries. *Ocean Coast Manag.* 44, 635–651.