TOOLS AND TECHNIQUES

ES-sim-GLM, a Multiple Regression Trait-Dependent Diversification Approach

Matthew O. Moreira[1] · Carlos Fonseca[1,2] · Danny Rojas[3]

Abstract

Identifying the role of quantitative variables on speciation rates is among the main purposes of trait-dependent diversification methods. *ES-sim*, a recent simulation-based approach that relies on Pearson's correlations, allows testing trait-dependent diversification for single regression models. Here, we modified this approach to include generalized linear models and two independent variables. To examine the effects of multiple traits on speciation we modified *ES-sim* and integrated generalized linear models instead of Pearson's correlations. We named the new approach as *ES-sim*-GLM. We further evaluated how this modified method performs in both single and multiple regression modelling. For this, we analyzed the relationship of speciation rates with geographic range size and snout-to-vent length in 216 species from the family Liolaemidae, a South American radiation of Andean lizards. Based on simulations, *ES-sim*-GLM for single regression models shows high power, low false discovery rates and is robust to incomplete taxon sampling. *ES-sim*-GLM for multiple regression models shows lower power but also low false-discovery rates. Both remained computationally efficient. Using Liolaemidae data, we found that larger species but with smaller species geographic range sizes were associated with higher speciation rates. To the best of our knowledge, no study as addressed these relationships in this clade. Our results provide new insights on macroevolutionary methods that should be relevant to all organisms and facilitate future studies that aim to understand diversification patterns across the Tree of Life.

Keywords Body size · *ES-sim*-GLM · Geographic range size · Liolaemidae · Macroevolution · Speciation

Introduction

Species' diversification is tightly connected with individual traits, including body size (Gittleman & Purvis, 1998), dispersal ability (Claramunt et al., 2012), ecological specialization (Futuyma & Moreno, 1988) and latitudinal range (Cardillo, 1999; reviewed in Harvey & Rabosky,

2018). To examine these relationships, powerful methods have been developed over the years. The quantitative state speciation and extinction model (QuaSSE; FitzJohn, 2010), one of the most popular methods, has been widely used to detect the relationship between diversification and quantitative traits. Although partitioning the phylogeny to account for shifts in speciation rate can provide reliable results (e.g. Rojas et al., 2018), other models in QuaSSE have been associated with high false discovery rates (Fitz-John, 2010; Harvey & Rabosky, 2018; Machac, 2014). In addition, methods like HiSSE account for the limitations of other state-dependent speciation methods (Beaulieu & O'Meara, 2016). While the latter can account for unmeasured factors impacting diversification, there are some alternatives that are less computationally demanding, including tip-rate correlation tests. Such methods usually involve the use of a speciation rate metric, followed by a statistical test for the correlation between speciation rates and a trait. This can be done, for example, using phylogenetic generalized least squared models (PGLS; e.g. Freckleton et al.,

✉ Matthew O. Moreira
matthew.moreira@ua.pt

✉ Danny Rojas
danny.rojas@javerianacali.edu.co

[1] CESAM - Centre for Environmental and Marine Studies, Department of Biology, University of Aveiro, 3810-193 Aveiro, Portugal

[2] ForestWISE - Collaborative Laboratory for Integrated Forest & Fire Management, Quinta de Prados, 5001-801 Vila Real, Portugal

[3] Department of Natural Sciences and Mathematics, Pontificia Universidad Javeriana Cali, Cali, Colombia

2008; Gomes et al., 2016; Jetz et al., 2012). More recently, Harvey and Rabosky (2018) introduced a simulation-based approach (*ES-sim*) that shows low false discovery rates and high power to detect trait-dependent diversification. Overall, these tools help to disentangle possible relationships between traits and species diversification.

The method introduced by Harvey and Rabosky (2018) allows to examine trait-dependent diversification at the species-level and outperforms QuaSSE and PGLS-based approaches in detecting true relationships between speciation rates and species traits. Specifically, *ES-sim* shows lower false discovery rates when compared to QuaSSE, as well as higher power when compared to PGLS-based approaches (Harvey & Rabosky, 2018). First, *ES-sim* focus on the inverse of the equal-splits metric (hereafter, λ; Redding & Mooers, 2006). The metric λ is accurate for estimating speciation rates (but not diversification or extinction rates; Title & Rabosky, 2019). Afterwards, the function simulates under Brownian Motion a null distribution of traits based on the input-trait *n* times. Finally, the method uses Pearson's correlations to assess the significance of the relationship between speciation rates (λ) and the trait. This is done by comparing the true correlation coefficient (i.e. λ ~ input-trait) with the distribution of correlation coefficients obtained for each simulated trait (i.e. λ ~ *n* simulated-trait). Overall, *ES-sim* is accurate (i.e. it has low false discovery rates) under different diversification scenarios and has high power (> 0.9) for moderate (*n* = 250) and large trees (*n* = 1250; Harvey & Rabosky, 2018). Notwithstanding, *ES-sim* is currently restricted to a single model that assess the relationship between speciation rates and a single trait. Although the method has the potential to incorporate multiple independent traits and different functional relationships between speciation rate and traits, these implementations have not been developed yet.

In this paper we modify *ES-sim* to include two independent variables by using generalized linear models. We name this new approach *ES-sim*-GLM and show that it is robust under incomplete taxon sampling and multiple scenarios of trait-dependent diversification, including multiple regression modelling. We apply *ES-sim*-GLM to analyse the relationship between speciation, species geographic range size and body size in Liolaemidae. This family of lizards (322 species described so far; Uetz et al., 2020) offers a useful system for studying the role of niche differentiation on diversification. The family is a well-studied adaptive radiation in South America (e.g. Esquerré et al., 2019; Pincheira-Donoso et al., 2015) and one of the main examples to study how the Andes uplift have shaped diversification (Esquerré et al., 2019). We found that species geographic range size and body size metrics are associated with speciation rates in Liolaemidae

and demonstrate how *ES-sim*-GLM can account for the coupled effect of two traits on speciation.

Methods

ES-sim-GLM

The original *ES-sim* function (Harvey & Rabosky, 2018) allows testing for trait-dependent diversification for one variable only. To test two variables simultaneously we modified the *ES-sim* test to incorporate generalized linear models (*ES-sim*-GLM). Specifically, the main objective of *ES-sim*-GLM is to test for the coupled effect of two independent variables on speciation. We assessed the performance of this modification in terms of power (section 'Power Analyses') and false-discovery rates (section 'False Discovery Rates'). Afterwards, and using *ES-sim*-GLM, we tested single and multiple-variable regression models of λ against species geographic range size and body size of Liolaemidae (section 'Statistical Analyses').

Similar to the original function (i.e. *ES-sim*; Harvey & Rabosky, 2018), *ES-sim*-GLM uses the ln-transformed inverse of the equal-splits metric (Jetz et al., 2012; Redding & Mooers, 2006). This metric of speciation is estimated using the full root-to-tip topology of the phylogeny, adding more weight to the recent branching patterns (Harvey & Rabosky, 2018). The metric provides a value of speciation rate for every tip of the phylogenetic tree. Afterwards, *ES-sim*-GLM fits a Brownian motion model to obtain diffusion rates and root state estimates for the trait under analysis. Then, we used this information to simulate the new set of trait values under a Brownian motion model of evolution. We used 1000 simulations to create the null-distribution of simulated trait values in the *ES-sim* and *ES-sim*-GLM functions. To account for two variables, we repeated these steps to simulate a new and independent null-distribution of trait values for each trait added to the function. The *P* value is calculated as the proportion of coefficient estimates (i.e. intercept and slope) obtained from the null-distribution of trait values that are higher than or lower than the coefficient estimates obtained when using the trait of interest. We consider parameter effects significant if they had a $P < 0.05$. To compare the relative fit of models we used Akaike weights (AIC$_W$; Burnham & Anderson, 2002). We estimated r^2 values for the *ES-sim*-GLM models as 1 minus the ratio between the residual deviance and the null deviance (Guisan & Zimmermann, 2000). We also calculated the running time for each function in single and multiple regression models for different tree sizes (i.e. 50, 100, 250, 1250 tips; Fig. S1).

Power Analyses

We assessed the performance of *ES-sim*-GLM single regression modelling for trees with different numbers of species, and compared its performance to *ES-sim* (as in Harvey & Rabosky, 2018). First, we simulated four hypothetical datasets with state-dependent diversification using the diversitree v0.9-11 package in R (FitzJohn, 2012). The datasets included states of a continuous trait for the tips of the trees. Each dataset included 100 trees with 50, 100, 250 and 1250 tips. To simulate state-dependent diversification, the linear function relating speciation rate to trait values had a slope of 0.004 and extinction rates were kept constant at 0 (i.e. no extinction), the default values used in Harvey and Rabosky (2018). We simulated the trait values using a diffusion rate of trait change of 0.06, given that *ES-sim* performs better at intermediate values of this parameter (based on comparing values of 0.00006, 0.0006, 0.006, 0.06, 0.6, 6, and 60; Harvey & Rabosky, 2018). To assess power, we ran *ES-sim*-GLM using the simulated datasets and calculated the proportion of datasets for which the function correctly identified state-dependent diversification. The higher the proportion of datasets for which *ES-sim*-GLM correctly identifies trait-dependent diversification, the higher the power of the test.

We then tested how the intensity of the relationship between speciation rate and the trait affected the ability of *ES-sim*-GLM to correctly identify trait-dependent diversification. We also did this for the original *ES-sim* method. We simulated two new datasets using two different values of the slope (0.04 and 0.4; for the linear function relating speciation rate to trait values) for each tree size and performed the power analyses. We selected 0.04 and 0.4 to represent different orders of magnitude relative to the original value (0.004).

Finally, we tested how *ES-sim* and *ES-sim*-GLM performed with incomplete taxon sampling. We randomly pruned the 100 simulated trees with 1250 tips obtained using 0.004 for the slope relating speciation and the trait to obtain three new sets of trees, one with 50 tips, other with 100 tips and the last with 250 tips. We did this 10 times to account for stochastic effects of random pruning. For example, if in the process of random pruning we keep only species with similar values of the trait, most likely we will obtain no relationship between speciation and the trait when in fact there is trait-dependent diversification. Therefore, in this case we would obtain a low power for the function but due to the data and not the function itself. By replicating this pruning process, we assure that we get distinct sets of pruned trees and address this possible bias. This resulted in 10 replicates of a new set of 100 pruned trees (i.e. pruned from the original tree size of 1250 tips) with 50 tips, 10 replicates of a new set of 100 pruned trees with 100 tips and 10 replicates of a new set of 100 pruned trees with 250 tips. We then repeated the power analyses.

To our knowledge, there is no straightforward way to simulate datasets with state-dependent diversification when speciation depends on more than one trait simultaneously. We believe that such simulation would require a dedicated paper. Here we addressed this issue in three steps. First, we simulated a new set of 100 trees with trait-dependent diversification for each tree size (i.e. 50, 100, 250, 1250 tips) as described above, using the diversitree package in R. As for the single regression case, we get a phylogeny and a trait (here we will refer to it as trait A) that evolved with speciation following a linear function relating speciation rate to trait values with different slopes (i.e. 0.004, 0.04 and 0.4). Second, we simulated a new trait (i.e. trait B) for each phylogeny using the mvMORPH v1.1.3 package in R (Clavel et al., 2015). Third, we kept this set (i.e. phylogeny, trait A and trait B) if, and only if, speciation is correlated with trait B. We tested for the relationship between speciation rates and trait B using the *ES-sim* function given that its low false-discovery rates have been demonstrated before for single regression models (Harvey & Rabosky, 2018). In this manner, we obtained a simulated dataset for different tree sizes in which both traits A and B are correlated with speciation. Then, we repeated the power analyses using *ES-sim*-GLM for multiple regression models and calculated the proportion of datasets for which the function correctly identified state-dependent diversification for both traits.

False Discovery Rates

We performed false discovery rates (FDR) analyses for single and multiple regression *ES-sim*-GLM models to avoid estimating false positives. First, we used the 63 datasets simulated in Harvey and Rabosky (2018). Each dataset represents a different scenario without state-dependent diversification (details in Harvey & Rabosky, 2018). The lower the proportion of datasets for which *ES-sim*-GLM incorrectly identifies trait-dependent diversification (i.e. $P < 0.05$), the lower the FDR. Second, to address if *ES-sim*-GLM had low FDR with the data from the study case, we simulated 9 traits without trait-dependence (i.e. neutral) on the Liolaemidae tree (section 'Liolaemidae Data'; Esquerré et al., 2019), following Harvey and Rabosky (2018). We simulated each trait under a different evolutionary scenario: (i) '1cladefixed': Brownian motion with one randomly selected clade (> 10% of the total number of tips) fixed for a single trait value from one of its tips; (ii) '1cladenosignal': Brownian motion overall but with one randomly selected clade lacking phylogenetic signal; (iii) 'BM': Brownian motion; (iv) 'BMjump': Brownian motion with a jump in the mean values in one random clade; (v) 'BMmultirate': Brownian motion with one rate shift in a randomly selected clade; (vi) 'discretetrait': one shift between two discrete trait distributions; (vii) 'nosignal': no phylogenetic signal in the trait; (viii)

'OUstrong': Ornstein–Uhlenbeck process with a single optimum and "strong" pull; (ix) 'OUweak': Ornstein–Uhlenbeck process with a single optimum and "weak" pull (Table S1). Each represents a distinct trait evolution setting, decoupled from diversification rates. We simulated 1000 replicates for each trait on the Liolaemidae tree. We then performed FDR analyses on the *ES-sim*-GLM function.

To test for FDR in multiple regression modelling, we simulated a new set of nine traits without trait-dependence on the Liolaemidae phylogeny. By doing so, we were able to perform the FDR analyses on the *ES-sim*-GLM with pairs of traits that were decoupled from diversification rates. We performed the FDR analyses on the *ES-sim*-GLM function using the pairs of traits and the same evolutionary scenarios that we describe above. The lower the proportion of datasets for which *ES-sim*-GLM incorrectly identifies trait-dependent diversification for both traits under analyses, the lower the FDR. For multiple regression models, we used two significance levels as threshold (i.e. $P < 0.05$ and 0.1).

Liolaemidae Data

We focused on Liolaemidae and obtained phylogenetic data (Supplementary material 3) for 258 species from Esquerré et al. (2019), geographic range data (in kilometre squared) for 217 species from Roll et al. (2017) and snout-to-vent length data (in millimetres) for 293 species from Feldman et al. (2016). After matching all three datasets, we pruned the tree using ape v5.4–1 package in R (Paradis & Schliep, 2019) to keep 216 species with both geographic range data and snout-to-vent length data available. Even though this represent approximately 67% of described liolaemids (numbers from Uetz et al., 2020), simulations showed that both *ES-sim* and *ES-sim*-GLM have high power for trees with sampling fractions of 20% (see below). Therefore, our overall results should not be strongly affected by incomplete taxon sampling. Additionally, the main objective of this study is to demonstrate the use of *ES-sim*-GLM to help explain ecological and evolutionary patterns.

We used the geographic range maps for the 216 Liolaemidae species and calculated the geographic range size for each species. We extracted the overall area that each species occupies using the rgdal v1.5–23 package in R (Bivand et al., 2019). The geographic range size metric is the total area of each species polygon(s). Although this metric can yield overestimations (e.g. if geographic range maps extend outside the actual species range), we consider this is a suitable approach at the macroevolutionary framework of the study and should impact all species under analyses. Regarding the snout-to-vent length, we used this metric as a proxy of body size following standard procedure throughout the field (e.g. Mahler et al., 2010, 2013). We applied the natural logarithm to the geographic range size and body size (Fig. S2).

Speciation Rates and Sensitivity Analyses

We estimated speciation rates for each species of Liolaemidae using the inverse of the equal-splits metric (Jetz et al., 2012; Redding & Mooers, 2006; Supplementary material 2). However, the estimation of λ can be biased by incomplete taxon sampling (Harvey & Rabosky, 2018). To account for this bias, we estimated λ on the most comprehensive tree of Liolaemidae with 258 species (Esquerré et al., 2019). Then we examined the relationship between those λ values and the λ values we had obtained in the pruned tree with just 216 species. The two sets of λ values show a strong significant correlation (Pearson's $r = 0.95$, $P < 0.001$; Fig. S3). Although the λ values for the tree with 258 species are significantly higher than those from the pruned tree with 216 species ($\bar{x} = -1.35$ and $\bar{x} = -1.43$, respectively; $t = -4.67$, df = 215, $P < 0.001$; Fig. S4), the effect size for these differences is small (Cohen's $d = 0.318$). Alongside these marginal differences, the positive correlation between the two sets of λ values (Fig. S3) suggests that we are estimating faster/slower rates for the same lineages using distinct approaches. Although we worked with 216 species due to data availability, we performed all analyses using the λ values that we calculated from the full tree with 258 species (Esquerré et al., 2019). This allowed us to account for incomplete taxon sampling (i.e. increased sample size) in the estimation of λ.

Statistical Analyses

We ran 2 single regression models (i.e. 1 for each metric) and 1 multiple regression models. We checked if both traits (i.e. ln-transformed geographic range size and body size) were not highly correlated to avoid multicollinearity. Both variables are not correlated with each other (Pearson's $r = -0.09$; P value $= 0.177$). The multiple regression model included the geographic range size and body size metrics, but did not consider the interaction between the two. The main goal of this study is to modify *ES-sim* and include multiple regression models, and for that purpose we combined only pairs of traits. We tested only linear models because initial graphical exploration of the data did not indicate non-linear relationships. We tested each model using 1000 simulations.

Results

Simulation Analyses

ES-sim-GLM for multiple regressions performed strongly in terms of correctly identifying state-dependent diversification. Except for one of the scenarios with two discrete traits (FDR = 0.054; Fig. 1), we obtained false discovery rates
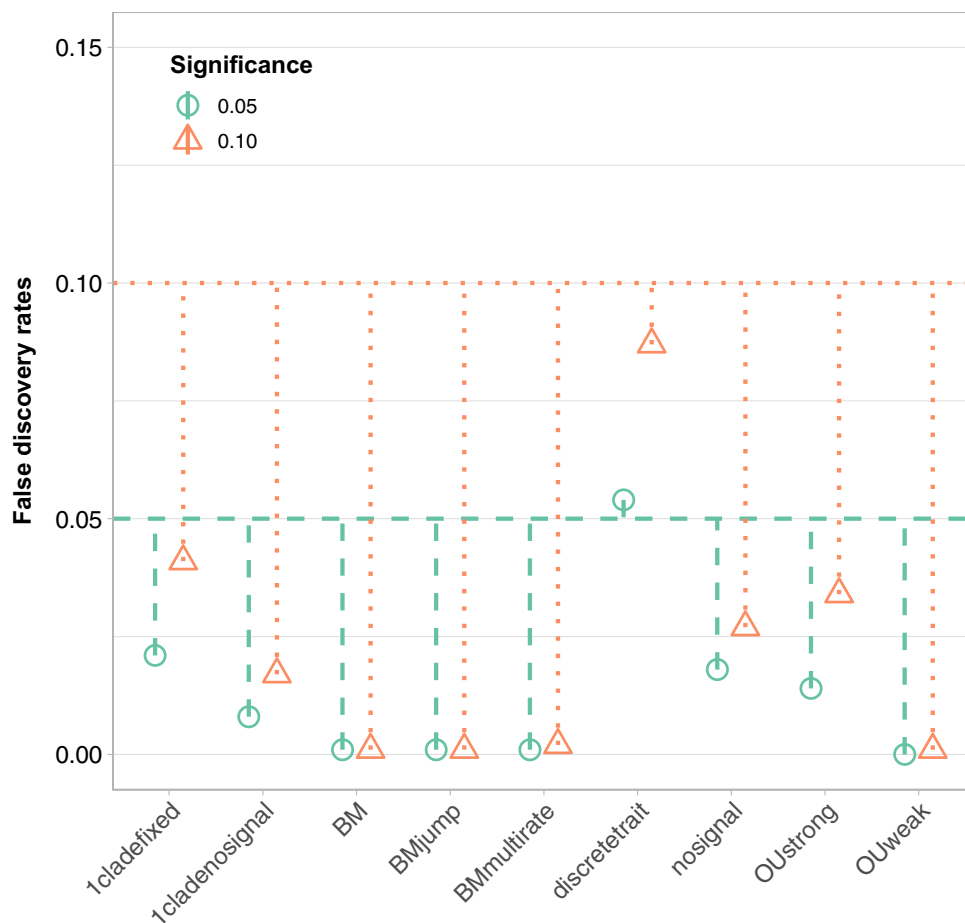
Fig. 1 False discovery rates of the *ES-sim*-GLM function in multiple regression modelling. False discovery rates were calculated as the proportion of datasets for which *ES-sim*-GLM incorrectly identifies trait-dependent diversification for two significance levels (0.05 and 0.1). Each pair of points represents one of the 9 simulated datasets of the liolaemids-based simulations on the Liolaemidae tree (Esquerré et al., 2019). Each scenario refers to a unique simulated diversification scenario without trait-dependence. The 9 datasets represent different trait evolution settings. (i) '1cladefixed': Brownian motion with one randomly selected clade (> 10% of the total number of tips) fixed for a single trait value from one of its tips; (ii) '1cladenosignal': Brownian motion overall but with one randomly selected clade lacking phylogenetic signal; (iii) 'BM': Brownian motion; (iv) 'BMjump': Brownian motion with a jump in the mean values in one random clade; (v) 'BMmultirate': Brownian motion with one rate shift in a randomly selected clade; (vi) 'discretetrait': one shift between two discrete trait distributions; (vii) 'nosignal': no phylogenetic signal in the trait; (viii) 'OUstrong': Ornstein–Uhlenbeck process with a single optimum and "strong" pull; (ix) 'OUweak': Ornstein–Uhlenbeck process with a single optimum and "weak" pull

below 0.021 in all cases when considering a significance level of 0.05 as a threshold. For a 0.1 significance threshold, all false discovery rates were below 0.087 (Fig. 1). Nevertheless, the power of *ES-sim*-GLM for multiple regression models was relatively low. The power peaked for intermediate tree size (i.e. 250 tips) with proportions of true positives ranging between 14 and 29%.

ES-sim-GLM for single regressions performed as well as or better than *ES-sim* for larger samples sizes ($n = 250$ and 1250; Fig. 2). However, when the tree had few tips ($n = 50$) the two functions showed low power, especially for lower slopes of the relationship between speciation and the trait (0.004; Fig. 2). Still, *ES-sim*-GLM was more robust for intermediate and higher slopes of the

relationship between speciation and the trait (0.04 and 0.4; Fig. 2). For trees with 100 tips, *ES-sim*-GLM performed better for higher slopes (0.4) while both functions performed similarly for lower and intermediate slopes (0.004 and 0.04; Fig. 2). When analysing datasets with 50, 100 and 250 tips pruned from the full tree, *ES-sim* and *ES-sim*-GLM performed similarly (Fig. S5). Both functions (*ES-sim* and *ES-sim*-GLM) showed low power when using a subset with 50 tips (4% sampling fraction), with proportions of true positives below 50%. With a subset of 100 tips (8% sampling fraction) proportions of true positives were below 80% for *ES-sim*-GLM and 70% for *ES-sim*. However, when we increased this subset to 250 tips (20% sampling fraction) these values improved

Fig. 2 Performance of two tests of trait-dependent diversification based on simulations with complete phylogenies in single and multiple regression modelling. Methods are compared under different models relating speciation rates and trait values. Intensity refers to the slope used in the simulations of the datasets. The two methods are: *ES-sim* (based on Pearson's correlation) and *ES-sim*-GLM (Generalized Linear Model). We used four different tree sizes: 50, 100, 250, and 1250 species
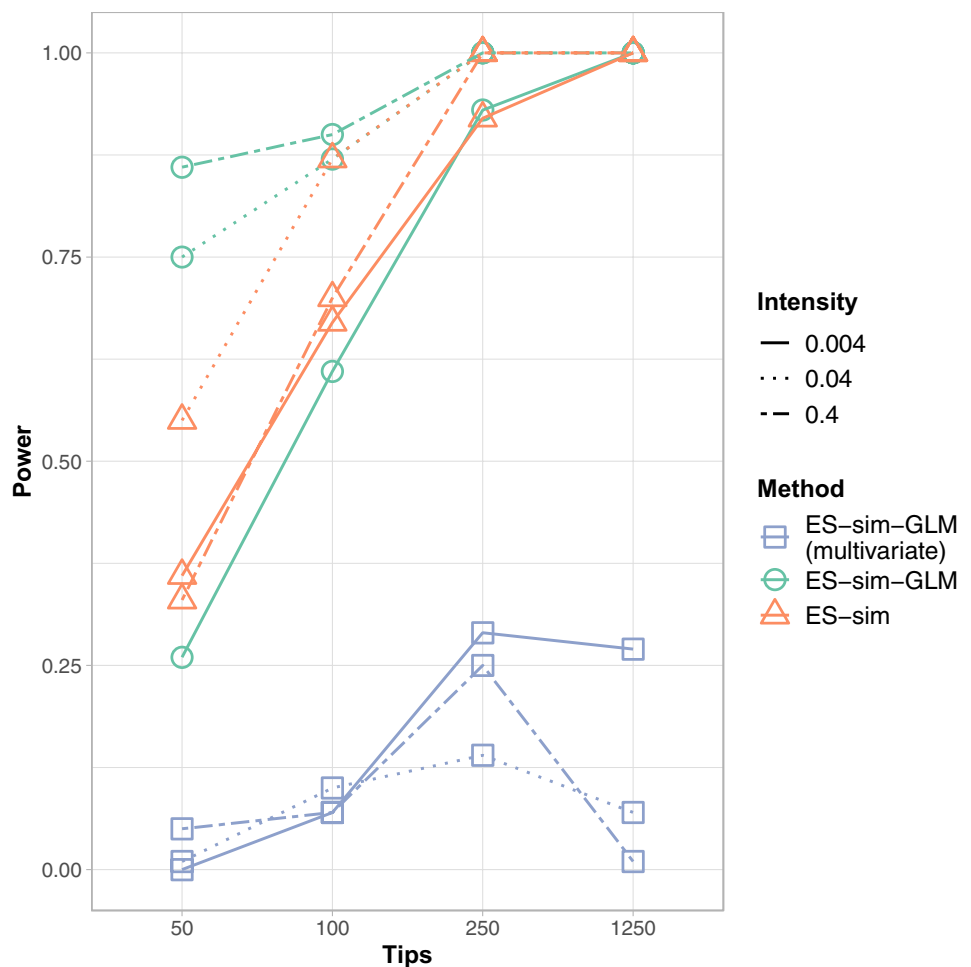
significantly, ranging between 86 and 96% for *ES-sim*, and 91 to 100% for *ES-sim*-GLM (Fig. S5). *ES-sim*-GLM for single regression showed higher false discovery rates than *ES-sim*, with high proportions of false discovery rates in several scenarios without trait-dependence (Fig. S6).

Liolaemidae-Based Simulations

Under different Liolaemidae-based simulations without state-dependent diversification, *ES-sim*-GLM showed false discovery rates below 0.05 for 4 out of the 9 datasets (Fig. S6). These include the following scenarios: 66) 'BM'; 67) 'BMjump'; 68) 'BMmultirate'; and 72) 'OUweak'. False discovery rates were below 0.08 for scenario 65) '1cladenosignal' and 0.1 for 70) 'nosignal'. The highest values of false discovery rates for *ES-sim*-GLM single regression modelling were 0.22 for scenario 69) 'discretetrait', 0.21 for scenario 64) '1cladefixed' and 0.13 for scenario 71) 'OUstrong'.

Correlates of Speciation Rates in Liolaemidae

Speciation rates (λ) are heterogeneous among lineages (Fig. 3). Geographic range size and body size metrics also show high/low values dispersed throughout the phylogeny (Fig. 3). Speciation rates are best explained by a multiple regression linear model including the geographic range size metric and the body size metric ($AIC_W = 0.97$). Considering these two variables together improved the model in relation to the single regression models involving each variable independently; the geographic range size metric ($AIC_W < 0.001$; $r^2 = 0.05$) and the body size metric ($AIC_W = 0.03$; $r^2 = 0.16$; Table 1). In the multiple regression model, the geographic range size metric shows a significant negative partial slope (-0.05, $P < 0.001$), while the body size metric shows a significant but positive partial slope (1.38, $P < 0.001$). This indicates that smaller geographic range sizes favour higher speciation rates while an increase in species' body size can accelerate speciation. This model explains ∼19% of the variation in λ (Table 1).
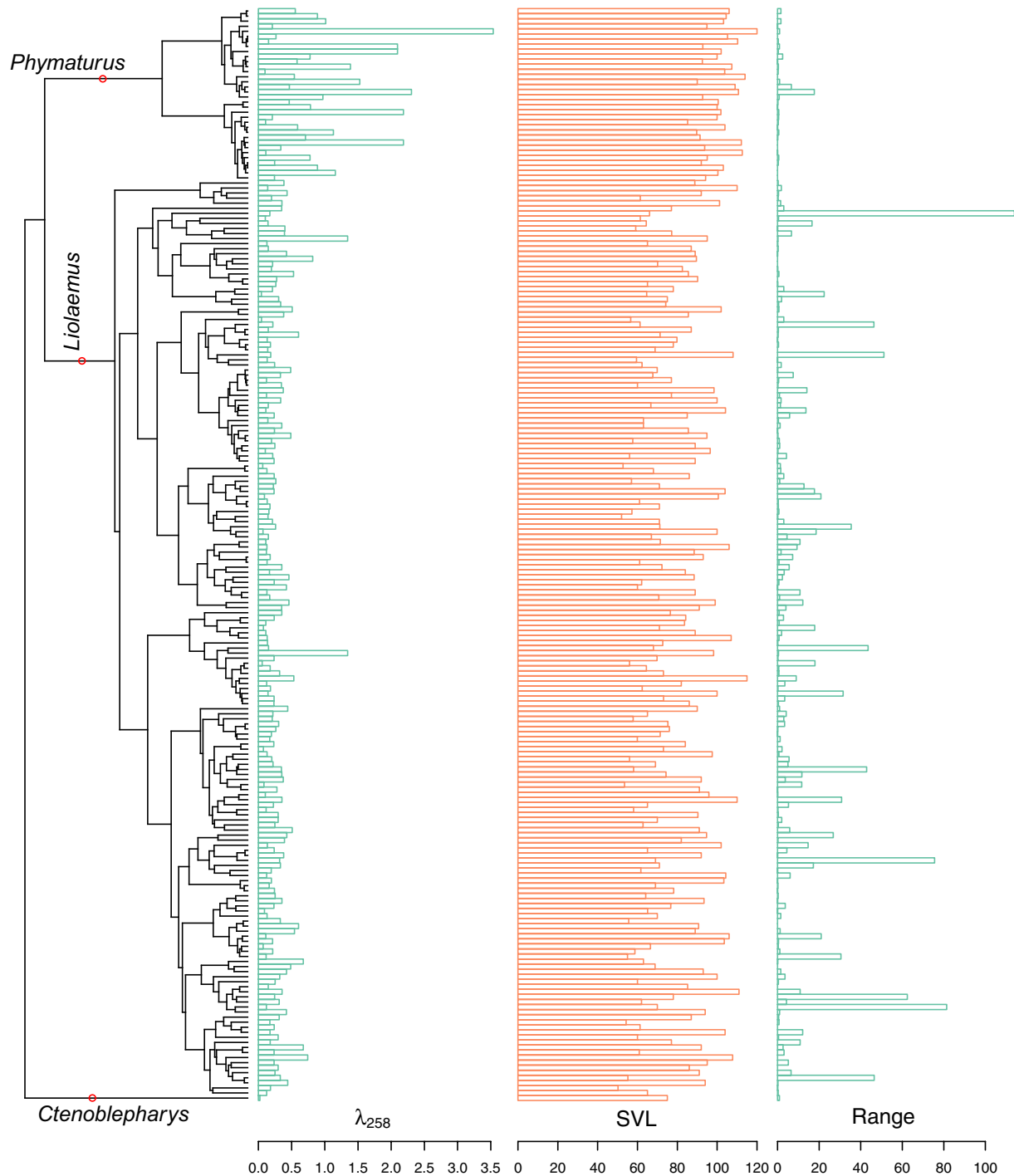
Fig. 3 From left to right: phylogeny of Liolaemidae lizards ($n=216$) with the genera name highlighted for simplicity, inverse of the equal-splits metric of speciation rates obtained from the full phylogeny (λ_{258}), snout-to-vent length metric (SVL) and geographic range size metric (Range). Phylogeny is from Esquerré et al. (2019)

Discussion

In this study we show that *ES-sim*, a simulation-based approach introduced by Harvey and Rabosky (2018), can be modified to test multiple regression modelling in scenarios with state-dependent diversification. We extended the *ES-sim* approach to test simultaneously multiple correlates of species diversification and incorporated a different model

Table 1 *ES-sim*-GLM regression metrics and support for different models testing the relationship between speciation rates and traits in Liolaemidae lizards

λ ~ trait(s)	Intercept	Slope$_1$ (P value)	Slope$_2$ (P value)	r^2	AIC	AIC$_w$
Range	− 1.374 (0.552)	− 0.063 (**< 0.001**)	–	0.048	496.288	< 0.001
SVL	− 7.665 (**< 0.001**)	1.442 (**< 0.001**)	–	0.160	469.255	0.031
Range + SVL	− 7.420 (**< 0.001**)	− 0.053 (**< 0.001**)	1.381 (**< 0.001**)	0.194	462.368	0.969

Models test the relationship between ln-transformed values of the inverse of the equal-splits metric (λ), the ln-transformed geographic range size metric (Range), and the ln-transformed snout-to-vent length metric (SVL). The Akaike Information Criterion (AIC) and AIC weights (AIC$_w$) were used to compare the fit of models. *P* values are based on 1000 simulation replicates. Significant *P* values are boldfaced

(i.e. GLM). The single regression version of the method we highlight in this study—*ES-sim*-GLM—is robust to incomplete taxon sampling, has high power for moderate and large sized trees ($n = 100$, 250 and 1250) and low FDR. More importantly, *ES-sim*-GLM allows inclusion of multiple variables in each model while maintaining low proportions of false positives. However, we advise caution for the possibility of false negatives in multivariate modelling, and dataset-specific simulations (i.e. simulate neutral scenarios of trait evolution for a phylogeny of interest and check for the proportion of false-discovery rates) should be conducted to address any possible bias.

Phylogenies and species trait data can inform about trait-dependent diversification scenarios. But to detect this relationship and avoid false positives, the diversification process must be tightly related to a trait (and usually a single trait; Herrera-Alsina et al., 2019). If we consider a scenario in which diversification depends on not one but two traits simultaneously, it should be even more difficult to detect trait-dependent diversification while avoiding Type I errors. We show that *ES-sim*-GLM under multiple regression modelling (i.e. for two traits) performed well under different scenarios without trait-dependence and was robust to false positives (Fig. 1). In contrast, it performed relatively poorly in terms of power when compared with the single regression version of the method (Fig. 2). Multiple regressions with two independent variables show low power for small ($n = 50$ and 100 tips) and large ($n = 1250$ tips) trees, with its power peaking at intermediate tree sizes ($n = 250$). Importantly, we argue that when *ES-sim*-GLM finds trait-dependent diversification in multiple regression modelling it will almost certainly be true, although *ES-sim*-GLM might not always detect that signal in multiple regression scenarios.

ES-sim-GLM for single regression modelling often performed as well as or better than *ES-sim* in terms of power (Fig. 2 and S5). However, *ES-sim*-GLM showed higher false discovery rates in some scenarios without state-dependent diversification, although these rates were below 0.22 in all cases and below 0.1 in most scenarios (Fig. S6). Therefore, we recommend testing for high rates of false positives to avoid Type I errors when *ES-sim*-GLM is used, especially in the single regression case.

Also, caution should be taken with small sized trees. Our analyses showed low power in *ES-sim*-GLM in identifying trait-dependent diversification in trees with 50 tips. Nevertheless, *ES-sim*-GLM for single regression models showed high power when we considered at least 20% of sampled taxa on our simulations with trait-dependent diversification (e.g. the phylogeny and trait data are available for 250 species out of 1250 extant species; Fig. S5).

To illustrate the use of *ES-sim*-GLM for multiple regressions, we show that higher speciation rates in Liolaemidae are best explained by the total geographic range size of each species and the body size of each species. These two traits together explained 19% of the variation in speciation rates. This can be seen as a small fraction of the variation in speciation. However, even if strong factors such as time could better explain speciation rates (Scholl & Wiens, 2016) or species richness (Li & Wiens, 2019), there is an intrinsic value in understanding what other aspects of species' biology and ecology are related to species' diversification.

On one hand, the geographic range size each species occupies could be relevant for species diversification. For example, large range sizes can help buffer species from extinction, promote range fragmentation and ultimately allopatric speciation (Rosenzweig, 1995). However, if species geographic range sizes are large enough relative to potential barrier sizes, large range sizes could actually act as a buffer for range fragmentation and reduce the likelihood of speciation events (Gaston, 1998). In such a scenario, speciation rates would be higher not for species with the largest distribution but for species with intermediate range sizes. For example, if species are mainly widely distributed relative to the barrier size, the species with the lower values of geographic range size will be more impacted by the potential effect of range fragmentation and subsequent speciation; the relationship between speciation and species geographic range size will be negative. This scenario might be the case for the family Liolaemidae. We found that on average the genus *Phymaturus* shows narrower geographic range sizes and faster speciation rates (Fig. 3). The Andean uplift had a strong role in the diversification of liolaemids (Esquerré et al., 2019). The fast Andean-orogeny promoted allopatric

speciation, but this mechanism could have impacted more strongly species with small to intermediate range sizes.

On the other hand, it is well established that smaller species are more abundant than larger species in nature (e.g. Hutchinson, 1959). One hypothesis suggests that this occurs because larger species are younger and so had less time to diversify (Etienne et al., 2012). In fact, previous results suggest that there is no relationship between body size and diversification in lepidosaurs (Feldman et al., 2016). At a smaller scale, however, ecological opportunity has been shown to have a role in the Andean radiation of the genus *Liolaemus*, with body size negatively correlated with net diversification (i.e. speciation rates minus extinction rates; Pincheira-Donoso et al., 2015). This indicates that evolution towards larger body sizes decelerated speciation rates and/ or accelerated extinction rates. In contrast, we found that the evolution towards larger body sizes is positively related to speciation. This suggests a scenario in which the lack of spatial overlap among diversifying lineages (Pincheira-Donoso et al., 2015)—which reduced the role of competition—favoured speciation towards larger body sizes following ecological opportunity. Notwithstanding, more studies would be needed to disentangle the role of both speciation and extinction for the evolution of body size in the diversification of this clade. Overall, this should further the debate on how diversification operates in the Liolaemidae Andean radiation, specifically regarding the role of morphological evolution in the diversification of this clade.

Conclusions

In summary, we show that the *ES-sim* approach can be modified to include multiple regression models. Overall, these results should facilitate future studies that aim to understand diversification patterns across the Tree of Life. Our analyses using a new approach (*ES-sim*-GLM) support that species geographic range size and body size are significantly related to speciation in Liolaemidae. The simulation analyses in this study address the role of quantitative variables on speciation rate, which has been one of the main purposes of trait-dependent diversification methods (Harvey & Rabosky, 2018).

Declarations

References

Beaulieu, J. M., & O'Meara, B. C. (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology, 65*(4), 583–601. https://doi.org/10.1093/sysbio/syw022

Bivand, R., Keitt, T., & Rowlingson, B. (2019). rgdal: bindings for the "Geospatial" data abstraction library. R package version 1.4–8. https://CRAN.R-project.org/package=rgdal.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer.

Cardillo, M. (1999). Latitude and rates of diversification in birds and butterflies. *Proceedings of the Royal Society B: Biological Sciences, 266*(1425), 1221–1225. https://doi.org/10.1098/rspb.1999.0766

Claramunt, S., Derryberry, E. P., Remsen, J. V., & Brumfield, R. T. (2012). High dispersal ability inhibits speciation in a continental radiation of passerine birds. *Proceedings of the Royal Society B: Biological Sciences, 279*(1733), 1567–1574. https://doi.org/10.1098/rspb.2011.1922

Clavel, J., Escarguel, G., & Merceron, G. (2015). mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution, 6*(11), 1311–1319. https://doi.org/10.1111/2041-210X.12420

Esquerré, D., Brennan, I. G., Catullo, R. A., Torres-Pérez, F., & Keogh, J. S. (2019). How mountains shape biodiversity: The role of the Andes in biogeography, diversification, and reproductive biology in South America's most species-rich lizard radiation (Squamata: Liolaemidae). *Evolution, 73*(2), 214–230. https://doi.org/10.1111/evo.13657

Etienne, R. S., de Visser, S. N., Janzen, T., Olsen, J. L., Olff, H., & Rosindell, J. (2012). Can clade age alone explain the relationship between body size and diversity? *Interface Focus, 2*(2), 170–179. https://doi.org/10.1098/rsfs.2011.0075

Feldman, A., Sabath, N., Pyron, R. A., Mayrose, I., & Meiri, S. (2016). Body sizes and diversification rates of lizards, snakes, amphisbaenians and the tuatara. *Global Ecology and Biogeography, 25*(2), 187–197. https://doi.org/10.1111/geb.12398

FitzJohn, R. G. (2010). Quantitative traits and diversification. *Systematic Biology, 59*(6), 619–633. https://doi.org/10.1093/sysbio/syq053

FitzJohn, R. G. (2012). Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution, 3*(6), 1084–1092. https://doi.org/10.1111/j.2041-210X.2012.00234.x

Freckleton, R. P., Phillimore, A. B., & Pagel, M. (2008). Relating traits to diversification: A simple test. *The American Naturalist, 172*(1), 102–115. https://doi.org/10.1086/588076

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. *Annual Review of Ecology and Systematics, 19*(1), 207–233. https://doi.org/10.1146/annurev.es.19.110188.001231

Gaston, K. J. (1998). Species-range size distributions: Products of speciation, extinction and transformation. *Philosophical Transactions of the Royal Society b: Biological Sciences, 353*(1366), 219–230. https://doi.org/10.1098/rstb.1998.0204

Gittleman, J. L., & Purvis, A. (1998). Body size and species–richness in carnivores and primates. *Proceedings of the Royal Society B: Biological Sciences, 265*(1391), 113–119. https://doi.org/10.1098/rspb.1998.0271

Gomes, A. C. R., Sorenson, M. D., & Cardoso, G. C. (2016). Speciation is associated with changing ornamentation rather than stronger sexual selection. *Evolution, 70*(12), 2823–2838. https://doi.org/10.1111/evo.13088

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Harvey, M. G., & Rabosky, D. L. (2018). Continuous traits and speciation rates: Alternatives to state-dependent diversification models. *Methods in Ecology and Evolution, 9*(4), 984–993. https://doi.org/10.1111/2041-210X.12949

Herrera-Alsina, L., van Els, P., & Etienne, R. S. (2019). Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. *Systematic Biology, 68*(2), 317–328. https://doi.org/10.1093/sysbio/syy057

Hutchinson, G. E. (1959). Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist, 93*(870), 145–159. https://doi.org/10.1086/282070

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature, 491*(7424), 444–448. https://doi.org/10.1038/nature11631

Li, H., & Wiens, J. J. (2019b). Time explains regional richness patterns within clades more often than diversification rates or area. *The American Naturalist*. https://doi.org/10.1086/702253

Machac, A. (2014). Detecting trait-dependent diversification under diversification slowdowns. *Evolutionary Biology, 41*(2), 201–211. https://doi.org/10.1007/s11692-013-9258-z

Mahler, D. L., Ingram, T., Revell, L. J., & Losos, J. B. (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science, 341*(6143), 292–295. https://doi.org/10.1126/science.1232392

Mahler, D. L., Revell, L. J., Glor, R. E., & Losos, J. B. (2010). Ecological opportunity and the rate of morphological evolution in the diversification of greater Antillean anoles. *Evolution, 64*(9), 2731–2745. https://doi.org/10.1111/j.1558-5646.2010.01026.x

Paradis, E., & Schliep, K. (2019c). ape 5.4–1: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pincheira-Donoso, D., Harvey, L. P., & Ruta, M. (2015). What defines an adaptive radiation? Macroevolutionary diversification dynamics of an exceptionally species-rich continental lizard radiation. *BMC Evolutionary Biology, 15*(1), 153. https://doi.org/10.1186/s12862-015-0435-9

Redding, D. W., & Mooers, A. Ø. (2006). Incorporating evolutionary measures into conservation prioritization. *Conservation Biology, 20*(6), 1670–1678. https://doi.org/10.1111/j.1523-1739.2006.00555.x

Rojas, D., Ramos Pereira, M. J., Fonseca, C., & Dávalos, L. M. (2018). Eating down the food chain: Generalism is not an evolutionary dead end for herbivores. *Ecology Letters, 21*(3), 402–410. https://doi.org/10.1111/ele.12911

Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A. M., Bernard, R., et al. (2017). The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nature Ecology & Evolution, 1*(11), 1677–1682. https://doi.org/10.1038/s41559-017-0332-2

Rosenzweig, M. L. (1995). *Species diversity in space and time. Cambridge University Press, Cambridge*. Cambridge: Cambridge University Press.

Scholl, J. P., & Wiens, J. J. (2016). Diversification rates and species richness across the Tree of Life. *Proceedings of the Royal Society b: Biological Sciences, 283*(1838), 20161334. https://doi.org/10.1098/rspb.2016.1334

Title, P. O., & Rabosky, D. L. (2019). Tip rates, phylogenies and diversification: What are we estimating, and how good are the estimates? *Methods in Ecology and Evolution, 10*(6), 821–834. https://doi.org/10.1111/2041-210X.13153

Uetz, P., Freed, P., & Hošek, J. (2020). The reptile database. http://www.reptile-database.org. Accessed 22 Oct 2020