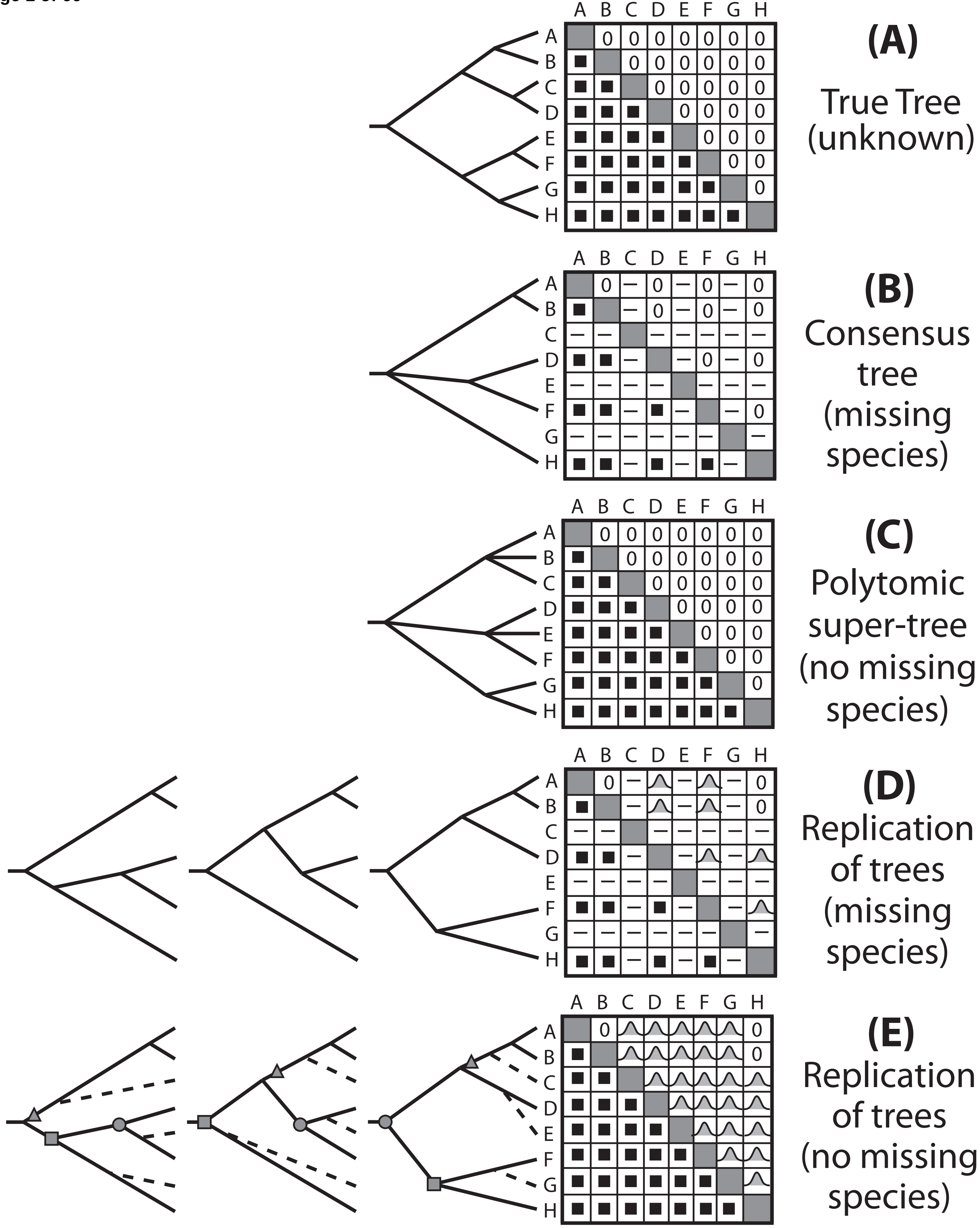
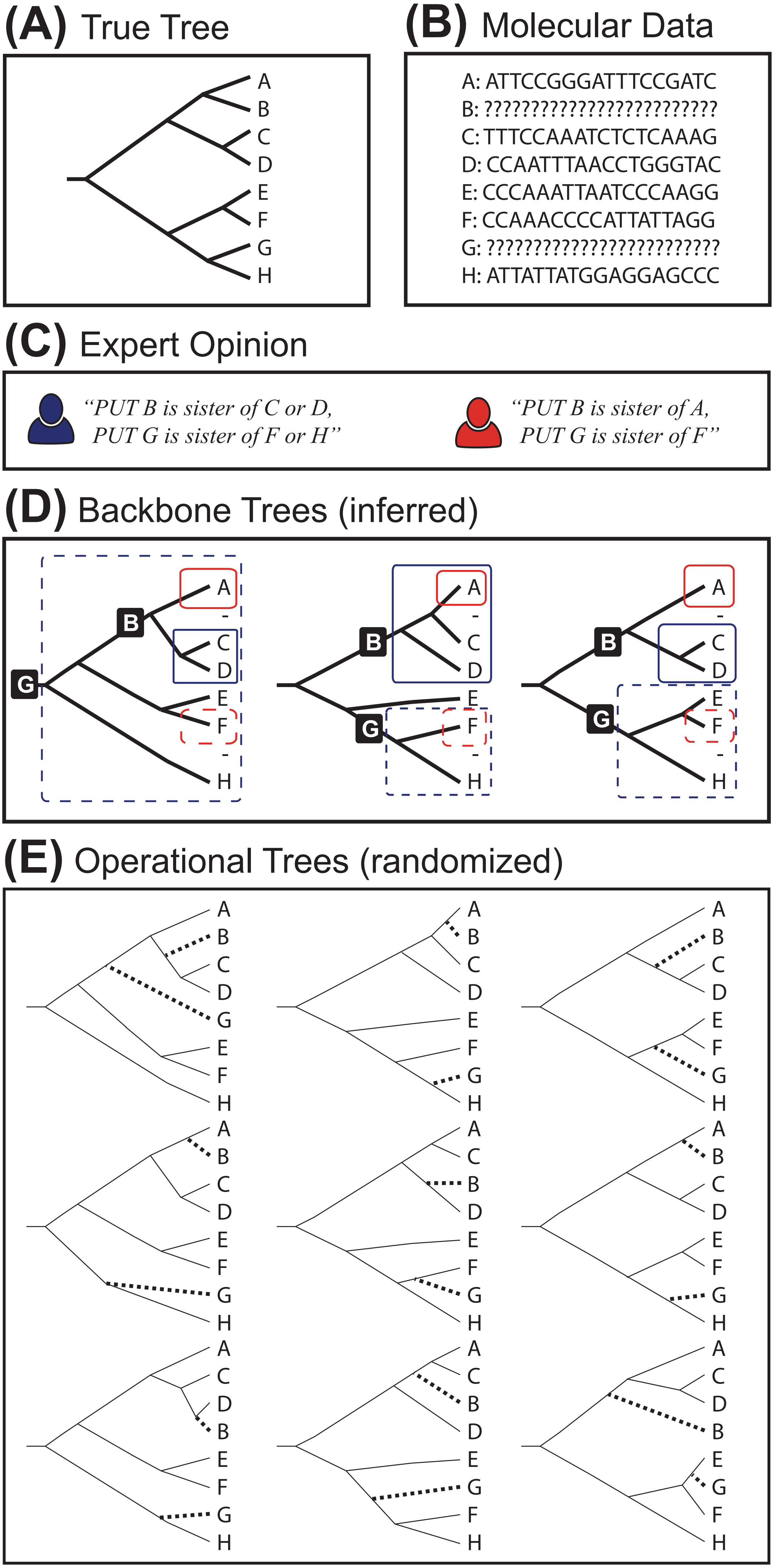
EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Phylogenetic Uncertainty Revisited: Implications for Ecological Analyses

Journal:	*Evolution*
Manuscript ID:	14-0738.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Rangel, Thiago; Universidade Federal de Goiás, Departamento de Ecologia Colwell, Robert; University of Connecticut, Department of Ecology Graves, Gary; Smithsonian Institution, Department of Vertebrate Zoology Fucikova, Karolina; University of Connecticut, Department of Ecology Rahbek, Carsten; Center for Macroecology, Evolution and Climate, University of Copenhagen, Department of Biology Diniz-Filho, José Alexandre; Universidade Federal de Goiás, Departamento de Ecologia
Keywords:	Biogeography, uncertainty quantification, sensitivity analysis, hummingbirds, phylogeny

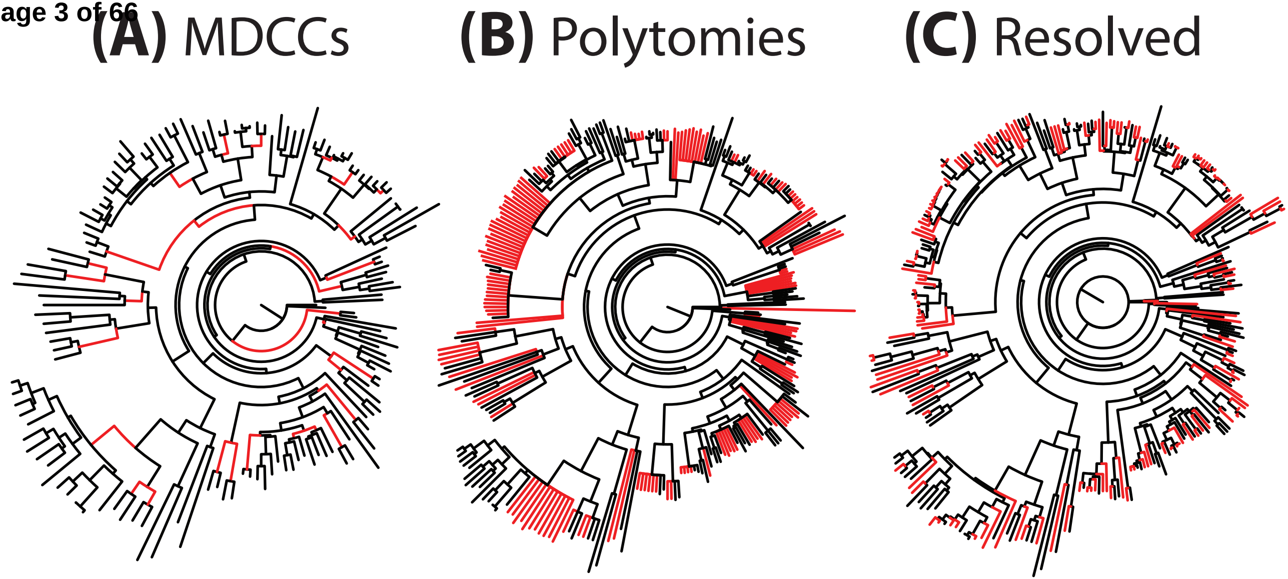

(A)
True Tree
(unknown)
(B)
Consensus
tree
(missing
species)
(C)
Polytomic
super-tree
(no missing
species)
(D)
Replication
of trees
(missing
species)
(E)
Replication
of trees
(no missing
species)
A B C D E F G H

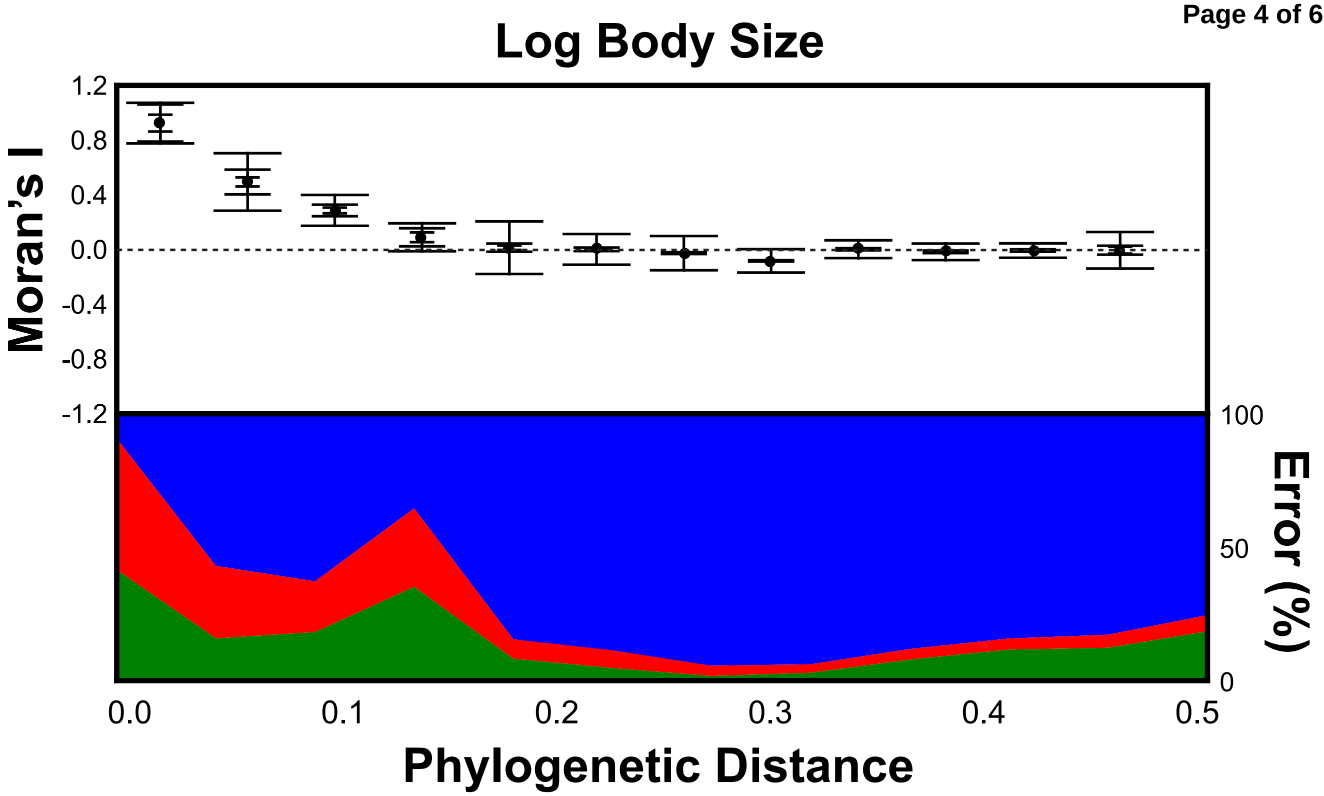


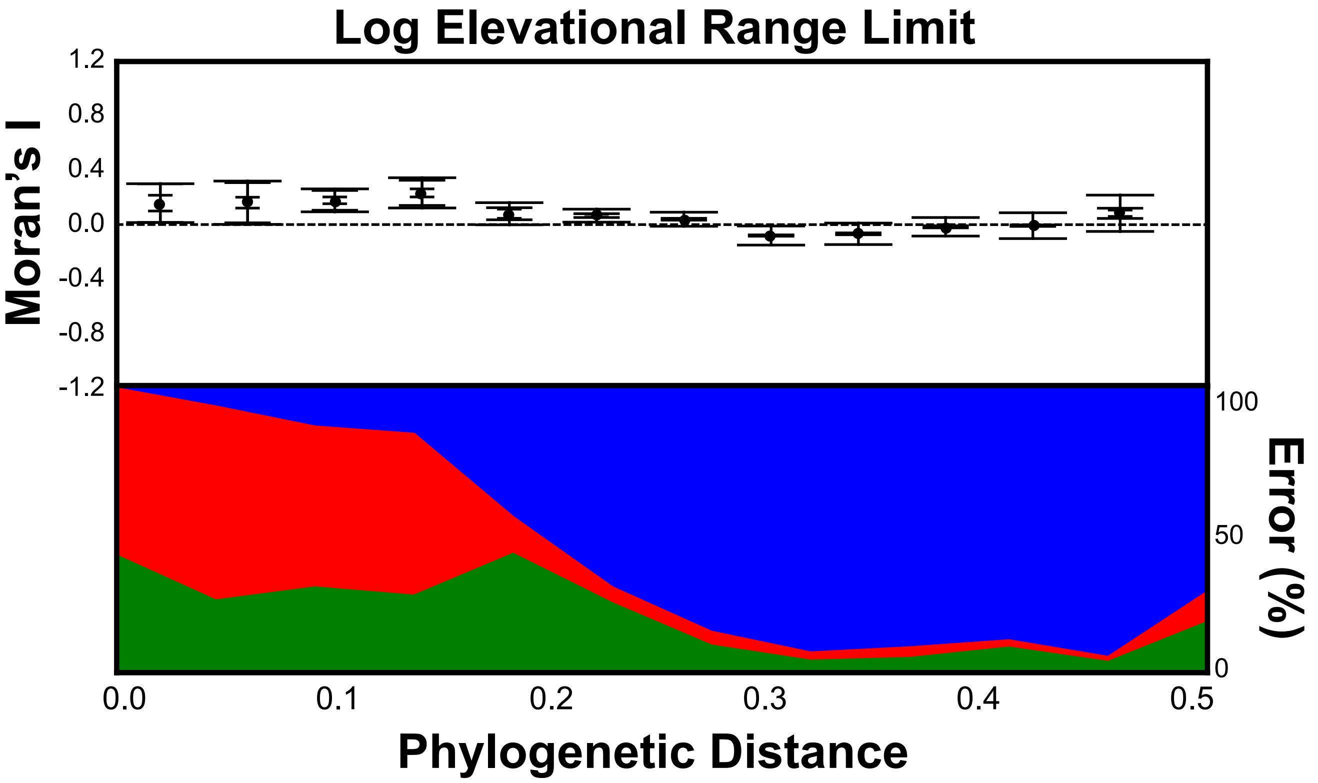
A: ATTCCGGGATTTCCGATC
B: ????????????????????????
C: TTTCCAAATCTCTCAAAG
D: CCAATTTAACCTGGGTAC
E: CCCAAATTAATCCCAAGG
F: CCAAACCCCATTATTAGG
G: ????????????????????????
H: ATTATTATGGAGGAGCCC

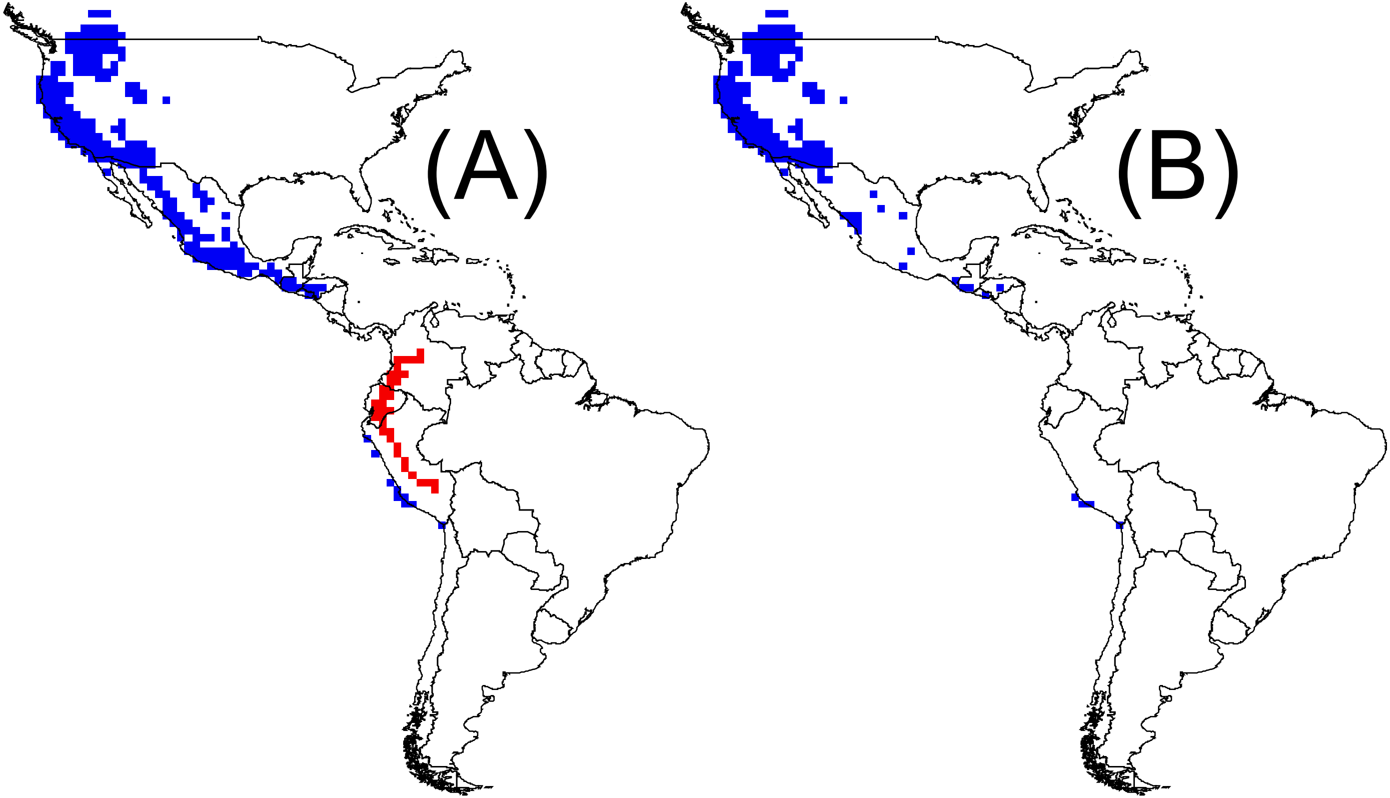
"PUT B is sister of C or D,
PUT G is sister of F or H"

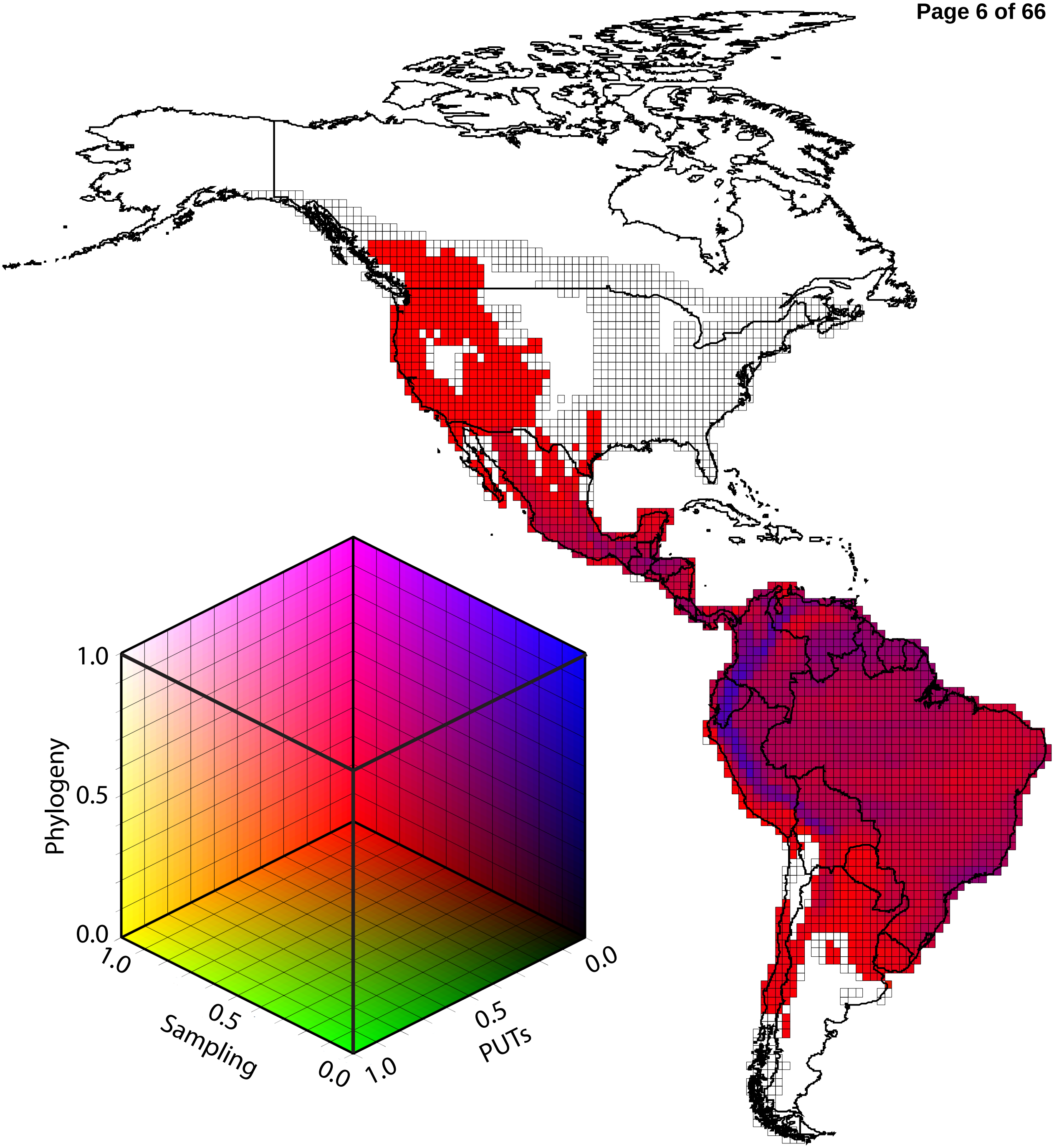
"PUT B is sister of A,
PUT G is sister of F"


(A) MDCCs
(B) Polytomies
(C) Resolved


Log Body Size
Moran's I
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
100
50
0
Error (%)
0.0
0.1
0.2
0.3
0.4
0.5
Phylogenetic Distance


Log Elevational Range Limit
Moran's I
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
100
50
0
Error (%)
0.0
0.1
0.2
0.3
0.4
0.5
Phylogenetic Distance


(A)
(B)


1.0
0.5
0.0
Phylogeny
1.0
0.5
Sampling
0.0
1.0
0.5
PUTs
0.0

For: Evolution

Correspondence to:
Thiago Fernando Rangel
Departamento de Ecologia
Universidade Federal de Goiás
Cx.P: 131, 74001-970
Goiânia, Goiás, Brasil
thiago.rangel@ufg.br

Article:

Phylogenetic Uncertainty Revisited: Implications for Ecological Analyses

Thiago F. Rangel [1]
Robert K. Colwell [1,2,3]
Gary R. Graves [4,5]
Karolina Fučíková [2]
Carsten Rahbek [4]
José Alexandre F. Diniz-Filho [1]

1: Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brasil. TFR: thiago.rangel@ufg.br. JAFD-F: diniz@ufg.br.

2: Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Rd., Unit 3043, 06269-3043, Storrs, CT, United States. RKC: robert.colwell@uconn.edu. KF: karolina.fucikova@uconn.edu.

3. University of Colorado Museum of Natural History, Boulder, Colorado 80309, USA

4: Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, United States. gravesg@si.edu

5: Center for Macroecology, Evolution and Climate, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100, Copenhagen O, Denmark. crahbek@bio.ku.dk

Key words: uncertainty quantification, sensitivity analysis, hummingbirds
Running title: Phylogenetic Uncertainty in Evolutionary Inference
Items to be included in the Expanded Online Edition:
Online Appendix: References used in the compilation of maximum elevational range limit of the 304 species of hummingbirds

Counts: 5,953 Words, 0 Tables, 6 Figures

Abstract

Ecologists and biogeographers usually rely on a single phylogenetic tree to study evolutionary processes that affect macroecological patterns. This approach ignores the fact that each phylogenetic tree is a hypothesis about the evolutionary history of a clade, and cannot be directly observed in nature. Also, trees often leave out many extant species, or include missing species as polytomies because of a lack of information on the relationship among taxa. Still, researchers usually do not quantify the effects of phylogenetic uncertainty in ecological analyses. We propose here a novel analytical strategy to maximizes the use of incomplete phylogenetic information, while simultaneously accounting for several sources of phylogenetic uncertainty that may distort statistical inferences about evolutionary processes. We illustrate the approach using a clade-wide analysis of the hummingbirds, evaluating how different sources of uncertainty affect several phylogenetic comparative analyses of trait evolution and biogeographic patterns. Although no statistical approximation can fully substitute for a complete and robust phylogeny, the method we describe and illustrate enables researchers to broaden the number of clades for which studies informed by evolutionary relationships are possible, while allowing the estimation and control of statistical error that arises from phylogenetic uncertainty. Software tools to carry out the necessary computations are offered.

Introduction

Ecological and biogeographical studies often involve hundreds of species, representing both recent and ancient lineages, and both locally endemic and cosmopolitan species. Although the geographic distribution of most groups of terrestrial vertebrates is increasingly well known, species sampling for phylogenetic analysis is rarely complete for larger clades. Even when a comprehensive phylogeny is available for ecological analyses, key sources of phylogenetic uncertainty are seldom taken into account. We begin this paper by outlining the sources of uncertainty in studies that rely on phylogenetic hypotheses to infer evolutionary processes, and how these sources have been considered or ignored in previous studies. With this motivation, we propose here a novel analytical strategy to quantify and account for key sources of phylogenetic uncertainty in any study that uses phylogenetic input data. As an example of the application of our methodology, we implement it to investigate patterns of trait evolution and phylogenetic assemblages in hummingbirds. We contend that any study that infers evolutionary hypotheses should aim to account for all sources of phylogenetic uncertainty. Finally, we offer freely available software tools to help researchers account for phylogenetic uncertainty in their own research.

Phylogenetic uncertainty may arise from two distinct sources (FitzJohn et al. 2009, Diniz-Filho et al. 2013): (1) weak, missing, or conflicting empirical support for hypothesized relationships among species in a given clade (e.g. tree topology, branch length estimation and absolute time calibration), which can be expressed in the form of multiple alternative topologies (e.g., resulting from different analysis methods or different data types), polytomic clades, or low branch support values; and (2) incomplete and unrepresentative sampling of known species. Most ecological studies implicitly assume absolute knowledge of phylogenetic history by simply ignoring the uncertainty caused by the lack of empirical support for phylogenetic trees or issues related to branch length estimation (Fig. 1). When contrasting phylogenetic hypotheses are available, ecologists usually rely on a consensus tree to "average" phylogenetic information, thus failing to take account of variation among trees (an expression of overall phylogenetic uncertainty). In addition, although ecologists acknowledge that most polytomies are expressions of ambiguous or missing empirical data, highly polytomic trees are nonetheless often used in ecological analyses without proper quantification of phylogenetic uncertainty and without conducting the necessary sensitivity analyses.

Ecologists have dealt with incomplete phylogenies in several ways. The first approach is to focus only on clades for which relatively complete phylogenies are available, but this strategy restricts ecological studies to a very small number of groups (Pagel 1999) and can undermine assemblage-level or macroecological studies (Webb and Donoghue 2005). A favored method is to assemble supertrees from smaller, overlapping trees and to fill gaps in phylogenies by placing unsampled species in trees according to their taxonomic classification (Bininda-Emonds 2004, Webb and Donoghue 2005, Hernandez and Vrba 2005, Davies et al. 2004, Ranwez et al. 2007). Nevertheless, even for the best-studied taxa, such as mammals and birds, supertrees that span all species are relatively recent achievements (e.g. Bininda-Emonds et al. 1999, Beck et al. 2006). Unfortunately, supertrees are usually highly polytomic (not fully resolved). Recent approaches based on building megatrees (Roquet et al. 2012) or complex addition of species on backbone trees (Jetz et al. 2012) are still in their infancy. Phylomatic software (Webb and Donoghue 2005), for instance, is a popular tool for assemblage-level ecological studies that constructs customized trees of virtually any size. Although Phylomatic allows additional phylogenetic representations of the clade when compared to the backbone phylogeny, terminal branches are usually based on taxonomic relationships among inserted species and thus are commonly polytomic. Additional drawbacks of supertrees are the absence of information on branch lengths (which require further effort, and assume confidence in the fossil record for time calibration) and, again, the lack of explicit conversion of phylogenetic uncertainty into explicit measures of error associated with statistical inference and parameter estimates.

A second and more radical strategy consists of ignoring the species that are absent from the available phylogeny, under the assumption that the species included in the analysis represent an unbiased and representative sample of all species in the clade (FitzJohn et al. 2009). Of course, the full evolutionary history of a clade of substantial age can be described only by a phylogeny with all species (including extinct ones), although this ideal is achievable only in simulated scenarios (Colwell and Rangel 2010). Estimating the degree of bias due to missing species can, in principle, be achieved by replicating the analysis with random sub-samples of species that are present in the phylogeny (rarefaction or "thinning", Davies et al. 2012), but unfortunately this approach is not common in evolutionary or ecological studies.

Several other approaches to deal with phylogenetic uncertainty have been recently developed, increasing awareness that species missing from phylogenetic trees may seriously affect statistical inference of evolutionary processes, even when the missing species are inserted in the tree in the form of polytomies (Davies et al. 2012, Diniz-Filho et al. 2013). For example, under the Bayesian framework of molecular phylogenetic reconstruction, missing species can be inserted by assigning empty sequences to missing species and by constraining their insertion using priors on the tree topology (Huelsenbeck and Rannala 2003). Several recent studies have employed different analytical strategies to account for phylogenetic uncertainty. For example, Isaac et al. (2007) dealt with missing species in their analysis of evolutionary distinctiveness of mammals by allocating missing species among their presumed closest relatives using a model of constant rate of speciation and extinction. Day et al. (2008) studied the diversification rates of cichlid fish radiation of Lake Tanganyika, accounting for the potential effects of missing species on the estimates of the timing and rate of diversification. Kuhn et al. (2011) proposed a method and a computational tool that uses a birth-death model of diversification to resolve polytomies and to define not only tree topology, but also branch lengths in polytomic trees, and Jetz et al. (2012) built sets of phylogenetic trees for the entire avian clade, combining genetic and taxonomic information and adding a simulation of evolutionary diversification to better establish branch lengths and resolve polytomies. Davies et al. (2012) showed that phylogenetic uncertainty can dramatically inflate estimates of phylogenetic signal and proposed a rarefaction-based method to guide inference. After accounting for phylogenetic uncertainty, Batista et al. (2013) showed that potential loss of evolutionary history caused by extinction of threatened Western Hemisphere anurans would not be significantly higher than that of non-threated anurans. These few examples clearly show the importance of dealing with phylogenetic uncertainty, and demonstrate that systematists and comparative biologists have already begun to develop and employ methods to incorporate estimates of uncertainty in the inference of diversification rates (Day et al. 2008, FitzJohn et al. 2009), character evolution (Losos 1994, Huelsenbeck et al. 2000, Housworth and Martins 2001, Huelsenbeck and Rannala 2003, Ives et al. 2007), reconstruction of ancestral states (Ronquist 2004), and tree topology (Felsenstein 1985, Holder and Lewis 2003).

Here we propose a unified analytical approach to estimate and account for multiple sources of phylogenetic uncertainty. Our method consists of partitioning variance among estimated parameters of evolutionary processes, using random sampling from the universe of probable phylogenies. We illustrate the approach using a clade-wide analysis of the hummingbirds, evaluating how different sources of uncertainty affect several phylogenetic comparative analyses of trait evolution and biogeographic patterns.

Accounting for multiple sources of phylogenetic uncertainty

We propose here an empirical approach, based on simulations, that maximizes the use of incomplete phylogenetic information in ecological studies, while simultaneously accounting for several sources of phylogenetic uncertainty that may distort statistical inferences about evolutionary and ecological processes.

Missing species

Here we define a *phylogenetically uncertain taxon* (henceforth a "PUT" for a single taxon, and "PUTs" for multiple taxa) as a taxonomic unit (e.g. population, species, genus) that is recognized as valid and is accepted as belonging to a particular clade, but is missing from the available phylogenetic tree(s) for that clade. The absence from the tree could be due to multiple causes, such as unavailability of molecular and morphological data (Fig. 2b). However, although the data required to formally reconstruct the phylogenetic relationships of a PUT may be missing, evolutionary information about the PUT is rarely completely non-existent, as no authority would dispute that some other taxon in the clade must be the closest relative of the PUT at some level in the phylogeny. In fact, additional available information, such as taxonomic, morphological or behavioral data, may be useful to define a list of the most likely sister species or lineages of any given PUT (Fig. 2c). Therefore, it makes sense to use all acceptable information available to conservatively define what we will call the *most derived consensus clade* (MDCC), i.e., the node that unequivocally contains each PUT.

Thus, an MDCC can be defined as the most recent common ancestor of all unquestioned candidates for being the closest relative of the PUT (Fig. 2d). Of course, the validity of information (e.g. taxonomic, biogeographical, behavioral, morphological, etc) used to define a MDCC can be questioned. However, if two taxonomists, for example, disagree on the proper MDCC for a given PUT, instead of eliminating the PUT from the statistical analysis, thereby ignoring phylogenetic uncertainty, the MDCC should be conservatively redefined as the most recent common ancestor of the two candidate MDCCs for the PUT, each championed by a different taxonomist (Fig. 2). Thus, it is important to stress that the goal of defining a MDCC to a PUT is not to replace the formal methods of phylogenetic reconstruction, but to conservatively use all available information to allow biologists to make inference of ecological and evolutionary processes using all available information, even with incomplete phylogenies.

We propose here a simple process of building a possible phylogeny that includes all extant species. This process relies on the definition of an MDCC for each PUT and the insertion of the PUT in a random position within its MDCC. We begin by randomizing the order in which PUTs are to be added to the tree. Next, within each MDCC, we assign the PUT to a point along one a branch of the clade. The choice of insertion point along a branch can be uniform random if no additional information is available, or it may be guided by a biologically realistic model of diversification (Kuhn et al. 2011, Davies et al. 2012, Jetz et al. 2012). If the baseline tree is ultrametric, the branch length for the inserted PUT is simply the distance from the attachment point to the end of all other tips. If, however, the baseline tree is not ultrametric, the branch length of the PUT can be sampled from a distribution of possible branch length values. Once a PUT has been inserted, its own branch may serve as a potential insertion point for subsequent species assigned from the PUT queue. Our algorithm iterates until each PUT has been sequentially added to the appropriate MDCC, producing a complete phylogeny of known species (Martins et al. 2013, Batista et al. 2013).

Polytomies

If all polytomies in a tree are the consequence of phylogenetic uncertainty (i.e. "soft" polytomies), then it is necessary to ensure that such uncertainty is quantified and accounted for in statistical analyses that use such a tree (Lewis et al. 2005). To explore the space of all possible dichotomous trees one must resolve the polytomies, assuming no additional information about the evolutionary history of the clade (Batista et al. 2013), or employing a model of diversification (Martins 1996, Housworth and Martins 2001, Kuhn et al. 2011).

Our approach consists in producing fully dichotomous trees by resolving polytomies stochastically. For each node in the tree with three or more branches, we choose two branches at random and reassign them to the same node. Each remaining branch from the original polytomy is then inserted sequentially, in random order, within the clade constructed from the former polytomy, at a randomly chosen position along the length of the existing branches. Dichotomous nodes in the original phylogeny are not changed, thereby preserving the phylogenetic information in the original baseline tree. This process guarantees that the resulting tree is fully resolved and that the phylogenetic uncertainty arising from polytomies is also taken into account when multiple randomized trees are compared (Batista et al. 2013). As with sampling multiple phylogenetic trees (below) or randomizing the position of the PUTs in the phylogeny, not every possible tree topology is examined. However, given a large enough number of replicates, our stochastic procedure ensures that the parameter space will be explored, thereby accounting for phylogenetic uncertainty.

Multiple phylogenetic trees

Modern methods of phylogenetic reconstruction and inference are based on searching the universe of possible phylogenetic trees. The search is guided by empirical data among possible trees, which are judged by their ability to predict the observed data, given a model of molecular evolution (Holder and Lewis 2003). However, because of scarce, ambiguous or missing empirical data, searches are often unable to unambiguously rank all trees within an entire group of equally likely trees, therefore yielding a large set of possible trees. Moreover, especially analyses of large, genome-scale data sets may yield multiple yet well-supported topologies depending on the analytical approach or on selected subsets of data (e.g., Xi et al. 2014, examples in Cooper 2014).

To account for the phylogenetic uncertainty that arises from lack of definitive empirical support for a single phylogenetic hypothesis, one must not limit the statistical analysis to a single sampled tree or to a majority-rule consensus tree that "averages" a larger set of possible trees (Fig. 1). Sampling or averaging among possible trees can potentially improve a point-estimate of a parameter, but does not account for statistical error in evolutionary inference, thereby masking phylogenetic uncertainty associated with tree reconstruction. In practice, variation in the results of ecological analyses that use different phylogenetic trees is an expression of uncertainty in phylogenetic reconstruction. Thus, it is necessary to replicate these analyses over a large number of possible phylogenetic trees, and subsequently study the variation in parameter estimates that arises as a consequence of differences among trees (Fig. 2).

Uncertainty and sensitivity analysis

Computer simulations have long been used to estimate the magnitude of statistical error and to quantify the probability of a given outcome when a system is not fully known (Doubilet et al. 1998, Saltelli et al. 2008). Usually run in tandem with uncertainty quantification, a sensitivity analysis allows one to determine the degree to which the sources of uncertainty in model inputs are responsible for the uncertainty in the model output (Pannell 1997). Coupling uncertainty and sensitivity analysis offers us the opportunity to quantify the robustness of models in the presence of phylogenetic uncertainty, account for uncertainty in evolutionary inference, and enhance communication of the magnitude of statistical error in the study.

Martins (1996) proposed a way to carry out phylogenetic comparative studies when the phylogenetic relationships among species are unknown. In her method, a large sample of trees is generated by randomly resolving a fully polytomic tree ("star phylogeny"), using models of phenotypic evolution and diversification rates (so that uncertainty in branch lengths is also implicitly incorporated). Species traits are then analyzed on each of the possible trees. Although the mean results of such analyses will converge to a non-phylogenetic analysis (Abouheif 1998), the approach of Martins (1996) represents a landmark in phylogenetic inference under uncertainty present in an unresolved phylogeny (see also Housworth and Martins 2001). The mean of the squared standard error of the calculated evolutionary statistic (e.g. the correlation between two species' traits), known as V_s, estimates the true variance of the statistic, which is due to sample variance. The variance of the evolutionary statistic calculated among the randomly generated trees, known as V_p, estimates the variance due to phylogenetic uncertainty. Thus, inferences based on parameter estimation must account for both sources of error, which can be done by computing confidence intervals or *P*-values based on the sum of V_p and V_s.

We expand Martins' (1996) strategy for incorporating phylogenetic uncertainty in parameter estimation further to accommodate multiple sources of phylogenetic uncertainty. We treat each source of uncertainty as a factor in an experimental design, while parameter estimates are treated as the response variable under study. Thus, we can ask how different phylogenetic trees, alternative resolutions of polytomies, and/or probable configurations of PUTs would affect parameter estimates. We partition the amount of variance in a parameter estimate arising from each source of phylogenetic uncertainty with an Analysis of Variance (ANOVA), which not only isolates variance among sources but also calculate the magnitude of standard error in the parameter estimate (a separate Martins' (1996) V_p for each source of uncertainty).

We view the sources of phylogenetic uncertainty we have outlined here as a hierarchy to be employed in the design of an experiment used to assess uncertainty and sensitivity. If multiple empirical phylogenies are available, then this source of uncertainty can be regarded as the highest level of uncertainty in the analysis. Because each empirical phylogeny is unique, its polytomies must be treated individually. Thus, polytomies can be regarded as the second level of phylogenetic uncertainty, immediately below the multiple empirical phylogenies. Of course, for each available empirical phylogeny there is a virtually unlimited number of alternative ways to resolve polytomies randomly or by following some diversification model. Finally, for each combination of original phylogeny and polytomy resolution, PUTs can be inserted according to the procedure described above. Thus, insertion of PUTs is the lowest level in the hierarchy of sources of phylogenetic uncertainty. Because sources of uncertainty are nested within a higher level of classification, and groups representing subordinate levels are randomly chosen, a nested (hierarchic) analysis of variance is ideal for the analysis of sensitivity and uncertainty.

Application example: trait evolution and phylogenetic assemblage patterns in hummingbirds

We illustrate our method by applying it to ecological hypotheses for the hummingbirds (Trochilidae), a large, monophyletic clade (~330 species) with a rich natural history literature and ongoing phylogenetic, morphological, behavioral, and ecological research. McGuire et al. (2007) published a multilocus molecular phylogeny for the hummingbirds that includes only 146 species, but encompasses all the higher-level trochilid diversity (73 of the approximately 104 recognized genera, Schuchmann 1999). Although a more complete hummingbird phylogeny is now available (McGuire et al. 2014), we use the earlier phylogeny(McGuire et al. 2007) to demonstrate our methods because it is typical of current level of phylogenetic knowledge for many comparably diverse taxa.

Replicating the phylogenetic reconstruction methodology of McGuire et al. (2007), we sampled 25,000 trees from the posterior distribution (hereafter called "backbone" trees), after discarding a burn-in of 5,000 steps. We relied upon the taxonomic classification of Schuchmann (1999) to designate the most likely MDCC for each PUT (Fig. 3).

From the total of about 330 described species of hummingbirds, we compiled geographical distributions and ecologically important morphological data for 304 species. Species endemic to islands (the West Indies and the Juan Fernández Archipelago) were not included in the analysis. We compiled estimates of average body mass (intersexual mean) for each of the 304 species from Dunning (2007) and Schuchmann (1999). Although large intersexual and geographical variation in hummingbird body size is well documented (Colwell 2000), the averages used here are useful for broad taxonomic analyses. Based on the literature, we also recorded the maximum known elevational range limits for each species (see Appendix A for references). Body mass (as a measure of body size) and maximum elevational range limit were log-transformed to conform to a normal distribution.

We extracted distributional data for 304 species from an updated version (16 July 2010) of the comprehensive database for all land and fresh-water birds known to have breeding populations in the Western Hemisphere, complied by Rahbek and Graves (2000, 2001), and mapped the geographical range of each species on a gridded map at a resolution of 1º x 1º (latitude-longitude). These maps represent a conservative extent-of-occurrence estimate of the breeding range based on museum specimens, published sight records, and spatial distribution of habitats based on documented records for South America.

Uncertainty quantification

Variability in input data is a fundamental source of uncertainty. To estimate the degree of phylogenetic uncertainty in the hummingbird data due to phylogeny reconstruction and missing species we randomly sampled 100 trees from the posterior (here after called "backbone" trees), and for each tree we replicated the insertion of PUTs 100 times, thereby generating a total of 10,000 fully resolved phylogenetic trees. Next, we calculated the tree-to-tree pairwise distance matrix using the weighted Robinson-Foulds (wRF) metric, which measures the minimum number of internal branches that must be collapsed or expanded to make two trees identical. The weighted version of Robinson-Foulds distance accounts for differences in branch lengths between trees (Steel and Penny 1993). Distance matrices of phylogenetic trees have been used in phylogenetic reconstruction to calculate consensus trees (Swofford 1991), produce supertrees (Bansal et al. 2010), and visualize multi-dimensional space of possible trees (Hillis et al. 2005).

The average wRF pair-wise distance between the 10,000 trees is 2.7423, with a standard deviation of 0.136. We used a multivariate analysis of variance (PERMANOVA, Anderson 2001) to partition variance among phylogenetic trees through the pair-wise wRF distance matrix. The estimated components of variance indicated that 17.02% of total variance among trees is attributable to phylogenetic reconstruction (i.e. differences among backbone trees), whereas 82.97% of total variance is due to phylogenetic uncertainty of missing species (PUTs).

Phylogenetic autocorrelation in species traits

We used Moran's I correlograms to estimate the magnitude of phylogenetic autocorrelation in body size and elevational range limits among hummingbird species (see Gittleman and Kot 1990; Diniz-Filho 2001; Pavoine and Ricota 2012). Moran's I is larger when species within a given phylogenetic distance interval have similar trait values and smaller when species within the same distance interval have very different trait values. To account for uncertainty in phylogenetic reconstruction and missing species, the calculation of correlograms was repeated 1,000,000 times (i.e., replicating the insertion of PUTs 1,000 times in each of the 1,000 randomly sampled backbone phylogenies). We used a nested ANOVA to partition sources of uncertainty for each of the 12 distance classes in the correlogram. Applying our method, the average Moran's I within a distance class is the parameter estimate, whereas additive confidence intervals are calculated around the estimate for each source of uncertainty.

Results of these analyses showed that phylogenetic autocorrelation in body size and maximum elevational range limit both decline with phylogenetic distance, although at different rates (Fig. 4, top panel in each plot). Closely related species tend to have similar body sizes but less similar elevational range limits. Accounting for phylogenetic uncertainty widens confidence intervals by at least a factor of two, regardless of trait (Fig. 4, confidence intervals in upper panel; see the caption for details). For all distance classes, standard confidence intervals due to sampling error are always narrower than confidence intervals due to phylogenetic uncertainty. In addition, the magnitude of error attributed to each source of uncertainty is not constant over phylogenetic distance (Fig. 4, bottom panels). The relative proportion of error caused by missing species tends to be higher in short distance classes, as taxonomic information has been used to assign PUTs to relatively derived positions in the phylogeny (MDCCs). In contrast, large phylogenetic confidence intervals at intermediate and long distance classes arise from uncertainty in the empirical phylogeny located at the base of the coquettes and brilliants clades, which together form the Andean clade (McGuire et al. 2007).

Phylogenetic signal and evolutionary models in species traits

We used Blomberg et al.'s (2003) *K* statistic (hereafter *K*) to estimate the magnitude of deviation from Brownian motion in body size and elevational range limits. *K* measures the degree of similarity in species' traits in relation to the similarity expected under a Brownian motion model of phenotypic evolution, given a phylogenetic hypothesis. *K* is based on the ratio between two measures of distance: MSE_0, the squared distance between trait values and the phylogenetically corrected mean trait value, and MSE, the squared distance between trait values estimated from a variance-covariance matrix derived from a phylogenetic hypothesis. Thus, large values of $\mathrm{MSE}_0/\mathrm{MSE}$ indicate a strong phylogenetic signal. To allow comparisons between different traits and trees, observed $\mathrm{MSE}_0/\mathrm{MSE}$ is standardized by expected $\mathrm{MSE}_0/\mathrm{MSE}$, assuming that the trait evolves under a Brownian motion model of phenotypic evolution. *K* values less than one indicate that species are less similar for a given trait than expected under Brownian motion evolution, whereas *K* values greater than one indicate that species are more similar for a given trait than expected under Brownian motion evolution.

To account for uncertainty in phylogenetic reconstruction and missing species, we again calculated 10,000 possible *K* values, replicating the insertion of PUTs 100 times in each of the 100 randomly sampled phylogenies. To estimate the accuracy (standard error) of *K* within each tree we used jackknife permutation for each combination of PUT insertion and sampled phylogeny (Efron and Tibshirani 1994). Finally, we used a nested ANOVA to partition error in estimated *K* among different sources of uncertainty.

Our results revealed a surprisingly low phylogenetic signal in body size among hummingbirds ($\overline{K}$ = 0.0597); thus body size evolution can hardly be explained by a simple model of Brownian evolution. However, the standard error of *K* due to sample size (s_{KSamp}=0.0011) is relatively small (2.23%) compared to the error arising from phylogenetic uncertainty due to missing species ($s_{KSamp+PUT}$=0.0494), as this source of uncertainty represents 97.1% of the total error in *K*. Finally, uncertainty due to phylogenetic reconstruction represents only 0.67% of total error of the evolutionary inference ($s_{KSamp+PUT+Phy}$=0.0497).

The relative importance of sources of uncertainty shift discordantly when maximum elevational range limit is considered, although this trait also has very little phylogenetic signal ($\overline{K}$ = 0.0221). Standard error of *K* due to sampling size contributes 59.7% of total error (s_{KSamp} = 0.0179), whereas error due to missing species represents 40.1% of total error ($s_{KSamp+PUT}$ = 0.0300), and error due to phylogenetic reconstruction represents only 0.2% of total error ($s_{KSamp+PUT+Phy}$=0.03001).

To test the hypothesis that estimated *K* values are not significantly different from random expectation, we calculated 100 *Ks* for each combination of PUT insertion and sampled phylogeny, randomizing trait values among species (Blomberg et al. 2003, Revell et al. 2008). We employed the same nested ANOVA design to estimate standard error of the null distribution due to each source of uncertainty. Finally, we used Welch's *t*-test to evaluate the hypothesis that estimated *Ks* do not differ significantly from the null expectation:

$$t = \frac{\overline{K}_{trait} - \overline{K}_{H0}}{\sqrt{\frac{s^2_{K_{trait}}}{n_{K_{trait}}} + \frac{s^2_{K_{H0}}}{n_{K_{H0}}}}}$$

Estimated *K* for hummingbird body mass is significantly larger than expected under the null expectation ($\overline{K}_{H0}$= 0.0237), and accounting for phylogenetic uncertainty does not affect the hypothesis test (t_{Samp} = 2589.92, $t_{Samp+PUT}$ = 62.5, $t_{Samp+PUT+Phy}$ = 76.72, all *P*-values < 0.001). Conversely, phylogenetic signal in maximum elevational range limit is significantly smaller than expected under the null hypothesis ($\overline{K}_{H0}$= 0.0333). Accounting for phylogenetic uncertainty also does not change the inference of statistical significance in hummingbird maximum elevational range limit (t_{Samp} = -275.36, $t_{Samp+PUT}$ = -34.74, $t_{Samp+PUT+Phy}$ = -44.47, all *P*-values < 0.001).

These results indicate that, although body size evolution deviates significantly from Brownian motion, it carries a small phylogenetic signal. Thus, evolutionary constraints and niche conservatism can be invoked for body size (for instance, evolution under an O-U process – see Hansen et al. 2008), whereas elevational range has less phylogenetic signal than expected by chance.

Phylogenetic community structure at the macroecological scale

Phylogenetic species variability (PSV, Helmus et al. 2007) of an assemblage is maximized (PSV = 1) when the assemblage is composed of the least related species in a clade, and minimized (PSV = 0) when the most related species coexist in an assemblage. We calculated PSV for the 1979 assemblages in the Western Hemisphere based on mapped species ranges with at least two hummingbird species, and tested statistical significance of each PSV value using 300 permutations of species identities. However, because phylogenetic distance between hummingbird species is not known with certainty, we generated 300 possible phylogenies with different PUT insertions, and replicated the calculation of PSV for each of the 1979 assemblages using the each of the 300 randomly selected phylogenetic trees.

Some hummingbird assemblages are composed of species with non-random phylogenetic relationships (Fig. 5). However, identifying any significant departure from randomness requires accounting for all sources of uncertainty. Standard null models to test phylogenetic structure of communities (e.g. Graham et al. 2009), which randomize species identities while preserving species richness, are designed to account only for non-random "sampling" from the phylogeny. For hummingbird assemblages across the Western Hemisphere, an analysis with such a null model would lead to the inference of significant phylogenetic dispersion along the middle and upper Andes (contrary to Graham et al. 2009), and significant phylogenetic clustering across the Pacific coast of Central and North America (Fig 5A). However, uncertainty caused by missing species and phylogenetic reconstruction may seriously affect pattern detectability. Because of substantial uncertainty in the phylogenetic relationship of the two Andean clades (Coquettes and Brilliants), when phylogenetic uncertainty is taken into account no significant phylogenetic dispersion in Andean assemblages is detected. Notice that this result is concordant with the lack of phylogenetic signal and lack of phylogenetic autocorrelation for elevational range at the species level, as we previously discussed. Moreover, once phylogenetic uncertainly is taken into account, many assemblages in Central America can no longer be considered phylogenetically clustered (Fig. 5B).

Because species are neither randomly distributed in the phylogeny nor in geographic space, the magnitude of error in the analysis of phylogenetic structure of assemblages tends to be strongly spatially autocorrelated. In addition, sampling effect is usually higher in species-poor assemblages. On the other hand, assemblages with a higher proportion of species that are members of clades characterized by phylogenetic uncertainty in deep (basal) nodes are subject to a higher proportion of error due to phylogenetic reconstruction. Figure 6 depicts the relative contribution of each source of error in analysis the analysis of phylogenetic structure of hummingbird assemblages. Because of phylogenetic uncertainty in the reconstruction of the relationship between the two, basal, Andean clades (Coquetes and Brilliants), statistical error is substantially higher than the other two sources of error for Andean assemblages. Conversely, statistical error in North American assemblages is mostly due to sampling, as species richness is very low. Uncertainty due to missing species (PUTs) is not particularly pronounced in any assemblage, as no assemblage is composed by more than 17% PUTs.

Concluding Remarks

Phylogenetic uncertainty is not evenly distributed across time and space. As a consequence, the effects of phylogenetic uncertainty in the statistical analysis of phylogenetic data cannot be estimated without a thorough sensitivity analysis on a case-by-case basis. To illustrate the new methods we propose to account for phylogenetic uncertainty, in this paper we used phylogenetic hypotheses derived from molecular data to analyze the hummingbird clade, a well-studied taxonomic group widely accepted as monophyletic. Partition of variance among sources of uncertainty reveals that variance among simulated phylogenies is caused primarily by missing species, in this case, even using the best information available to assign phylogenetically uncertain taxa (PUTs). Conversely, variance among backbone phylogenies, which are estimated through molecular data, is substantially smaller, indicating that efforts to improve the knowledge of evolutionary history of hummingbirds (and other clades that share similar patterns of uncertainty) should be concentrated on gathering molecular data for additional species (e.g. McGuire et al. 2014).

Because phylogenetic uncertainty is expected to be relatively more concentrated within some clades of a phylogeny than within others, statistical analyses of species assemblages or temporal segments of the phylogeny will be affected differently. Thus, statistical analyses of different traits for the same group of species, using the same phylogenetic information, may be differently affected by phylogenetic uncertainty, both in intensity and direction of bias. Of course, comparative analyses among multiple taxonomic groups, based on independently built phylogenetic hypotheses, requires additional caution, as the heterogeneity of variance among groups may seriously distort results in unpredictable directions.

Because the effects of phylogenetic uncertainty in ecological and evolutionary analyses are not subject to generalization, quantifying and accounting for phylogenetic uncertainty through sensitivity analysis is required in all ecological studies. Modern techniques of sensitivity analyses typically involve the application of Monte Carlo methods. Although general simulation methods, such as the one used in this study, could be modified to account for phylogenetic uncertainty in most evolutionary and ecological analyses, the framework for uncertainty quantification and sensitivity analysis should be tailored to the purpose of the study (e.g. estimate of diversification rates, community phylogenetics, comparative analysis of species traits).

Finally, the approach proposed here can be used to quantify the full spectrum of components of phylogenetic uncertainty, guiding sampling strategies for future studies and allowing more reliable interpretations of the relative magnitude of historical and phylogenetic components of biodiversity patterns.

Software Tools

We provide two software toolkits to enable the application of the analytical strategy proposed here. The first software is SUNPLIN (Martins et al. 2013; https://sourceforge.net/projects/sunplin), which is capable of generating randomized phylogenies after the insertion PUTs into backbone trees, with MDCCs assigned. The generated trees can then be applied in any analysis that requires phylogenies as input data. SUNPLIN can be used as an online web service (http://wsmartins.net/sunplin/), as a library that connects through APIs to any compiled software, or directly integrated into R (http://www.ecoevol.ufg.br/pam).

The second software toolkit, PAM (*Phylogenetic Analysis in Macroecology*, http://www.ecoevol.ufg.br/pam), is a compiled computational platform for inference of ecological and evolutionary processes in a spatially explicit context. In PAM, users can not only generate replicates of phylogenetic trees to be used in other software applications, but can also run several statistical analyses commonly used in biodiversity analysis, while estimating and accounting for multiple sources of uncertainty using the analytical framework proposed here. PAM is a work in progress and will be continuously expanded in the future.

Acknowledgements

Several concepts in this paper were inspired by the phylogenetics course (EEB5349), taught by Dr. Paul Lewis, at the University of Connecticut. TFR is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, 564718/2010-6). RKC was supported by the National Science Foundation (USA, DEB-0639979 and DBI-0851245) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). GRG thanks the Alexander Wetmore Fund of the Smithsonian Institution, the Smoketree Trust, the Danish National Research Foundation, and the Center for Macroecology, Evolution and Climate for support. CR acknowledges the Danish National Research Foundation for support to the Center for Macroecology, Evolution and Climate. KF was supported by the National Science Foundation (USA, DEB-1036448, awarded to Louise A. Lewis and Paul O. Lewis). JAFD-F has been continuously supported by CAPES and CNPq through several research grants.

References

Abouheif, E. 1998. Random trees and the comparative method: a cautionary tale. Evolution 52:1197-1204.

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

Bansal, M. S., J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca. 2010. Robinson-Foulds Supertrees. Algorithms for Molecular Biology 5:18.

Batista, M. C. G, S. F. Gouveia, D. L. Silvano and T. F. Rangel. 2013. Spatially explicit analyses highlight idiosyncrasies: species extinctions and the loss of evolutionary history. Diversity and Distributions (*in press*).

Beck, R. M. D., O. R. P. Bininda-Emonds, M. Cardillo, F.-G., R. Liu and A. Purvis. 2006. A higher-level MRP supertree of placental mammals. Evolutionary Biology 6:93-107.

Bininda-Emonds, O. R., J. L. Gittleman and A. Purvis. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biological Reviews 74: 143-175.

Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. Trends in Ecology and Evolution 19:315-322.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are move labile. Evolution 57:717-745.

Colwell, R. K. 2000. Rensch's Rule crosses the line: Convergent allometry of sexual size dimorphism in hummingbirds and flower mites. American Naturalist 156:495-510.

Colwell, R. K. and T. F. Rangel. 2010. Modelling quaternary range shifts and richness on tropical elevational gradients. Phylosofical Transactions of the Royal Society, B 365: 3695-3707.

Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends in Plant Science 19: 576-582.

Davies, T. J., N. J. B. Kraft, N. Salamin and E. M. Wolkovich. 2012. Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. Ecology 93:242-247

Davies, T.J., T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis and V. Savolainen. 2004 Darwin's abominable mystery: insights from a supertree of the angiosperms. Proceedings of the National Academy of Sciences 107, 1904–1909.

Day, J. D., J. A. Cotton and T. G. Barraclough. 2008. Tempo and Mode of Diversification of Lake Tanganyika Cichlid Fishes. PLoS One 3:e1730

Diniz-Filho, J. A. F. 2001. Phylogenetic autocorrelation under distinct evolutionary processes. Evolution 55:1104-1109.

Diniz-Filho, J. A. F., R. D. Loyola, P. Raia, A. O. Mooers and L. M. Bini. 2014. Darwinian shortfalls in biodiversity conservation. Trends in Ecology and Evolution, *in press*.

Doubilet P., C.B. Begg, M. C. Weinstein, P. Braun, B. J. McNeil. 1985. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. Medical Decision Making 5:157-177.

Dunning, J. B. 2007. CRC handbook of avian body masses. 2nd ed. CRC press.

Efron, B. and R. J. Tibshirani. 1994. An introduction to the bootstrap. Chapman & Hall.

Felsensetein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

FitzJohn, R., W. Maddison, and S. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology 58:595–611.

Gittleman, J. L. and M. Kot. 1990. Adaptation: Statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39:227-241.

Graham, C. H., J. L. Parra, C. Rahbek, and J. A. McGuire. 2009. Phylogenetic structure in tropical hummingbird communities. Proceedings of the National Academy of Sciences of the United States of America 106:19673-19678.

Helmus, M. R., T. J. Bland, C. K. Williams, and A. R. Ives. 2007. Phylogenetic measures of biodiversity. American Naturalist 169:E68-E83.

Hernandez, F.M. and E. S. Vrba. 2005. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. Biological Reviews 80: 269–302.

Hillis, D. M., T. Heath and K. St. John. 2005. Analysis and Visualization of Tree Space. Systematic Biology 54:1-12.

Housworth, E. A. and E. P. Martins. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. Systematic Biology 50:628-639.

Holder, M. and P. O. Lewis. 2003. Phylogeny Estimation: Traditional and Bayesian Approaches. Nature Reviews Genetics 4:275-284.

Huelsenbeck, J. P. and B. Rannala. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57:1237-1247.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accommodating Phylogenetic Uncertainty in Evolutionary Studies. Science 288:2349-2350.

Ives, A. R., P. E. Midford and T. Garland Jr. 2007. Within-species variation and measuremente Error in Phylogenetic Comparative Methods. Systematic Biology 56:252-270.

Isaac, N. J. B., S. T. Turvey, B. Collen, C. Waterman, and J. E. M. Baillie. 2007. Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. PLoS ONE 2:e296.

Kuhn, T. S., A. O. Mooers, and G. H. Thomas. 2011. A simple polytomy resolver for dated phylogenis. Methods in Ecology and Evolution 2:427-436.

Lewis, P. O., M. T. Holder and K. E. Holsinger. 2005. Polytomies and Bayesian Phylogenetic Inference. Systematic Biology 54:241-253.

Losos, J. B. 1994. An approach to the Analysis of Comparative Data When a Phylogeny is Unavailable or Incomplete. Systematic Biology 43:117-123.

Martins, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50:12-22.

Martins, W.S., W. C. Carmo, H. J. Longo, T. C. Rosa and T. F. Rangel. 2013. SUNPLIN: Simulation with Uncertainty for Phylogenetic Investigations. BMC BioInformatics 14:324-335.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen Jr. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Systematic Biology 56:837-856.

McGuire, J. A., C. C. Witt, J. Remsen Jr, A. Corl, D. L. Rabosky, D. L. Altshuler, and R. Dudley. 2014. Molecular phylogenetics and the diversification of hummingbirds. Current Biology 24:910-916.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann and A. O. Mooers. 2013. The global diversity of birds in space and time. Nature 491:444-448.

Pagel, M. D. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Pannell, D.J. 1997. Sensitivity analysis of normative economic models: Theoretical framework and practical strategies. Agricultural Economics 16: 139-152.

Pavoine, S. and C. Ricota. 2012. Testing for phylogenetic signal in biological traits: the ubiquity of cross-product statistics. Evolution, 67:828-840.

Rahbek, C. and G. R. Graves. 2000. Detection of macro-ecological patterns in South American hummingbirds is affected by spatial scale. Proceedings of the Royal Society of London, B 267:2259-2265.

Rahbek, C. and G. R. Graves. 2001. Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences of the USA 98:4534-4539.

Ranwez, V., V. Berry, A. Criscuolo, P. H. Fabre, S. Guillemot, C. Scornavacca and E. J. P. Douzery. 2007. PhySIC: a veto supertree method with desirable properties. Systematic Biology 56:798–817.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Systematic Biology 57:591-601.

Ronquist, F. 2004. Bayesian Inference of Character Evolution. Trends in Ecology and Evolution 19:475-481.

Roquet, C., W. Thuiller and S. Lavergne. 2012. Building megaphylogenies for macroecology: taking up the challenge. Ecography 36:13-26.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. Global Sensitivity Analysis: The Primer. John Wiley & Sons.

Schuchmann, K.L. 1999. Family Trochilidae. Pages 468-680 in J. del Hoyo, A. Elliott, and J. Sargatal, editors. Handbook of the Birds of the World. Vol. 5. Barn-owls to Hummingbirds. Lynx Edicions, Barcelona.

Steel, M. A. and D. Penny. 1993. Distributions of tree comparison metrics - some new results. Systematic Biology 42:126-141.

Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295-333 in M. M. Miyamoto and J. Cracraft, editors. Phylogenetic analysis of DNA sequences. Oxford Univ. Press, New York.

Xi Z., L. Liu, J.S. Rest and C.S. Davis. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. (*in press*).

Webb, C. O. and M. J. Donoghue. 2005. Phylomatic: tree retrieval for applied phylogenetics. Molecular Ecology Notes 5:181-183.

Figure Legends:

Figure 1: A conceptual example of the effect of different sources of uncertainty in phylogenetic trees (left) on the estimation of phylogenetic relationship between species (right, represented as matrices). In each matrix, cells in the lower-left half-matrix represent the existence of phylogenetic information about the relationship between a pair of species, because both species are present in the tree. Conversely, a dash represents the absence of such phylogenetic information, as one or both of the species is missing from the tree. In the upper-right half-matrix a normal distribution represents the variance in the estimated phylogenetic relationship, while zeros represent certainty. (A) Hypothetical, unknown true tree, without missing species or uncertainty in species relationships. (B) Consensus tree with missing species. Relationships are assumed to be known with certainty. (C) Polytomic super-tree. Insertion of missing species in polytomies generates a complete tree, and relationships are assumed to be known with certainty, although the tree differs from the true tree (A). (D) Replication in the use of phylogenetic trees incorporates the uncertainty in the relationship between species, but missing species are ignored. (E) Missing species are inserted in multiple phylogenies, accounting for uncertainty in phylogenetic reconstruction and lack of a complete phylogeny.

Figure 2: Schematic representation of a workflow using the analytical strategy, proposed here, to account for phylogenetic uncertainty. (A) The true, but unknown, tree for a clade of 8 taxa. (B) Molecular data are available for only six taxa; for taxa B and G no molecular data are available. (C) Two experts, based on their knowledge (e.g. taxonomy, behavior, morphology, geographic distribution, etc.), suggest possible sister taxa of the two taxa that lack molecular data. (D) Using the available molecular data and phylogenetic reconstruction methods, three backbone phylogenies are proposed. The variation between backbone phylogenies arises from uncertainty in the process of

phylogenetic reconstruction. Among the backbone phylogenies, the taxa B and G are considered *phylogenetically uncertain taxa* (PUT), because no molecular data are available for them, and therefore they are missing from the backbone phylogenies. Using the information provided by the experts the *most derived consensus clade* (MDCC) is found for each PUT. For PUT B, the MDCC is the clade that necessarily includes taxa C, D, and A (indicated by a bold B in each backbone phylogeny), whereas for PUT G the MDCC must include taxa F and H (indicated by a bold G in each backbone phylogeny). (E) Statistical analyses that use phylogenetic trees as input data should be replicated using samples of operational trees. In each of the operational trees, the PUTs B and G were randomly inserted within their respective MDCCs. The insertion of each PUT was replicated three times (columns) for each backbone tree. The variation among operational trees that use the same backbone tree (each column of the operational trees) arises from variation in placement of the missing taxa (as the backbone tree is not changed by the randomization process), and is caused by uncertainty in the phylogenetic relationships of taxa B and G (both PUTs).

Figure 3: Hummingbird phylogenies. (A) One backbone phylogeny from McGuire et al. (2007), with red internal branches indicating the *Most Derived Consensus Clades* (MDCCs) used in this analysis. (B) Red taxa (polytomies) indicate PUTs inserted in the original phylogeny at the base of their MDCCs. (C) Red taxa indicate PUTs inserted to yield a fully-resolved phylogeny, randomizing their position within the respective MDCC.

Figure 4: Moran's I correlogram (top panels) for two hummingbird traits (body size and maximum elevational range limit). The inner 95% confidence interval (C.I.) around each estimate indicates variance due to sampling error, the intermediate 95% C.I. the variance due to missing species, and the outer 95% C.I. the variance due to phylogenetic reconstruction. The bottom panels indicate the

relative proportion of statistical error caused by each source of uncertainty across evolutionary time: sampling (green), missing species (PUTs, red) and phylogenetic reconstruction (blue).

Figure 5: Hummingbird assemblages with PSV values significantly different from null expectation. Blue cells indicate significant phylogenetic clustering, whereas red cells indicate significant phylogenetic dispersion. (A) Standard PSV analysis, considering only non-random “sampling” from the phylogeny. (B) Re-analysis of PSV accounting for three sources of error: sampling, missing species, and phylogenetic reconstruction. Notice absence of significant phylogenetic dispersion and decreased areas of phylogenetic clustering when all sources of uncertainty are accounted for.

Figure 6: Relative statistical error associated with PSV analysis for hummingbird species present in each map cell, partitioned among sources of uncertainty. Each point within the cube has a unique color that represents the relative proportions of uncertainty due to phylogeny, PUTs and sampling error, as shown by the axes. In the map, the prevalence of red cells indicates sampling error as the main source of uncertainty, whereas blue indicates substantial error caused by phylogenetic uncertainty. Notice the absence of green and yellow areas, indicating that phylogenetic uncertainty due to missing species (PUTs) is relatively irrelevant in the phylogenetic structure of assemblages.

For: Evolution

Correspondence to:
Thiago Fernando Rangel
Departamento de Ecologia
Universidade Federal de Goiás
Cx.P: 131, 74001-970
Goiânia, Goiás, Brasil
thiago.rangel@ufg.br

Article:

Phylogenetic Uncertainty Revisited: Implications for Ecological Analyses

Thiago F. Rangel [1]
Robert K. Colwell [1,2,3]
Gary R. Graves [4,5]
Karolina Fučíková [2]
Carsten Rahbek [4]
José Alexandre F. Diniz-Filho [1]

1: Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brasil. TFR: thiago.rangel@ufg.br. JAFD-F: diniz@ufg.br.

2: Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Rd., Unit 3043, 06269-3043, Storrs, CT, United States. RKC: robert.colwell@uconn.edu. KF: karolina.fucikova@uconn.edu.

3. University of Colorado Museum of Natural History, Boulder, Colorado 80309, USA

4: Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, United States. gravesg@si.edu

5: Center for Macroecology, Evolution and Climate, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100, Copenhagen O, Denmark. crahbek@bio.ku.dk

Key words: uncertainty quantification, sensitivity analysis, hummingbirds
Running title: Phylogenetic Uncertainty in Evolutionary Inference
Items to be included in the Expanded Online Edition:
Online Appendix: References used in the compilation of maximum elevational range limit of the 304 species of hummingbirds

Counts: 5,953 Words, 0 Tables, 6 Figures

Abstract

Ecologists and biogeographers usually rely on a single phylogenetic tree to study evolutionary processes that affect macroecological patterns. This approach ignores the fact that each phylogenetic tree is a hypothesis about the evolutionary history of a clade, and cannot be directly observed in nature. Also, trees often leave out many extant species, or include missing species as polytomies because of a lack of information on the relationship among taxa. Still, researchers usually do not quantify the effects of phylogenetic uncertainty in ecological analyses. We propose here a novel analytical strategy to maximizes the use of incomplete phylogenetic information, while simultaneously accounting for several sources of phylogenetic uncertainty that may distort statistical inferences about evolutionary processes. We illustrate the approach using a clade-wide analysis of the hummingbirds, evaluating how different sources of uncertainty affect several phylogenetic comparative analyses of trait evolution and biogeographic patterns. Although no statistical approximation can fully substitute for a complete and robust phylogeny, the method we describe and illustrate enables researchers to broaden the number of clades for which studies informed by evolutionary relationships are possible, while allowing the estimation and control of statistical error that arises from phylogenetic uncertainty. Software tools to carry out the necessary computations are offered.

Introduction

Ecological and biogeographical studies often involve hundreds of species, representing both recent and ancient lineages, and both locally endemic and cosmopolitan species. Although the geographic distribution of most groups of terrestrial vertebrates is increasingly well known, species sampling for phylogenetic analysis is rarely complete for larger clades. Even when a comprehensive phylogeny is available for ecological analyses, key sources of phylogenetic uncertainty are seldom taken into account. We begin this paper by outlining the sources of uncertainty in studies that rely on phylogenetic hypotheses to infer evolutionary processes, and how these sources have been considered or ignored in previous studies. With this motivation, we propose here a novel analytical strategy to quantify and account for key sources of phylogenetic uncertainty in any study that uses phylogenetic input data. As an example of the application of our methodology, we implement it to investigate patterns of trait evolution and phylogenetic assemblages in hummingbirds. We contend that any study that infers evolutionary hypotheses should aim to account for all sources of phylogenetic uncertainty. Finally, we offer freely available software tools to help researchers account for phylogenetic uncertainty in their own research.

Phylogenetic uncertainty may arise from two distinct sources (FitzJohn et al. 2009, Diniz-Filho et al. 2013): (1) weak, missing, or conflicting empirical support for hypothesized relationships among species in a given clade (e.g. tree topology, branch length estimation and absolute time calibration), which can be expressed in the form of multiple alternative topologies (e.g., resulting from different analysis methods or different data types), polytomic clades, or low branch support values; and (2) incomplete and unrepresentative sampling of known species. Most ecological studies implicitly assume absolute knowledge of phylogenetic history by simply ignoring the uncertainty caused by the lack of empirical support for phylogenetic trees or issues related to branch length estimation (Fig. 1). When contrasting phylogenetic hypotheses are available, ecologists usually rely on a consensus tree to "average" phylogenetic information, thus failing to take account of variation among trees (an expression of overall phylogenetic uncertainty). In addition, although ecologists acknowledge that most polytomies are expressions of ambiguous or missing empirical data, highly polytomic trees are nonetheless often used in ecological analyses without proper quantification of phylogenetic uncertainty and without conducting the necessary sensitivity analyses.

Ecologists have dealt with incomplete phylogenies in several ways. The first approach is to focus only on clades for which relatively complete phylogenies are available, but this strategy restricts ecological studies to a very small number of groups (Pagel 1999) and can undermine assemblage-level or macroecological studies (Webb and Donoghue 2005). A favored method is to assemble supertrees from smaller, overlapping trees and to fill gaps in phylogenies by placing unsampled species in trees according to their taxonomic classification (Bininda-Emonds 2004, Webb and Donoghue 2005, Hernandez and Vrba 2005, Davies et al. 2004, Ranwez et al. 2007). Nevertheless, even for the best-studied taxa, such as mammals and birds, supertrees that span all species are relatively recent achievements (e.g. Bininda-Emonds et al. 1999, Beck et al. 2006). Unfortunately, supertrees are usually highly polytomic (not fully resolved). Recent approaches based on building megatrees (Roquet et al. 2012) or complex addition of species on backbone trees (Jetz et al. 2012) are still in their infancy. Phylomatic software (Webb and Donoghue 2005), for instance, is a popular tool for assemblage-level ecological studies that constructs customized trees of virtually any size. Although Phylomatic allows additional phylogenetic representations of the clade when compared to the backbone phylogeny, terminal branches are usually based on taxonomic relationships among inserted species and thus are commonly polytomic. Additional drawbacks of supertrees are the absence of information on branch lengths (which require further effort, and assume confidence in the fossil record for time calibration) and, again, the lack of explicit conversion of phylogenetic uncertainty into explicit measures of error associated with statistical inference and parameter estimates.

A second and more radical strategy consists of ignoring the species that are absent from the available phylogeny, under the assumption that the species included in the analysis represent an unbiased and representative sample of all species in the clade (FitzJohn et al. 2009). Of course, the full evolutionary history of a clade of substantial age can be described only by a phylogeny with all species (including extinct ones), although this ideal is achievable only in simulated scenarios (Colwell and Rangel 2010). Estimating the degree of bias due to missing species can, in principle, be achieved by replicating the analysis with random sub-samples of species that are present in the phylogeny (rarefaction or "thinning", Davies et al. 2012), but unfortunately this approach is not common in evolutionary or ecological studies.

Several other approaches to deal with phylogenetic uncertainty have been recently developed, increasing awareness that species missing from phylogenetic trees may seriously affect statistical inference of evolutionary processes, even when the missing species are inserted in the tree in the form of polytomies (Davies et al. 2012, Diniz-Filho et al. 2013). For example, under the Bayesian framework of molecular phylogenetic reconstruction, missing species can be inserted by assigning empty sequences to missing species and by constraining their insertion using priors on the tree topology (Huelsenbeck and Rannala 2003). Several recent studies have employed different analytical strategies to account for phylogenetic uncertainty. For example, Isaac et al. (2007) dealt with missing species in their analysis of evolutionary distinctiveness of mammals by allocating missing species among their presumed closest relatives using a model of constant rate of speciation and extinction. Day et al. (2008) studied the diversification rates of cichlid fish radiation of Lake Tanganyika, accounting for the potential effects of missing species on the estimates of the timing and rate of diversification. Kuhn et al. (2011) proposed a method and a computational tool that uses a birth-death model of diversification to resolve polytomies and to define not only tree topology, but also branch lengths in polytomic trees, and Jetz et al. (2012) built sets of phylogenetic trees for the entire avian clade, combining genetic and taxonomic information and adding a simulation of evolutionary diversification to better establish branch lengths and resolve polytomies. Davies et al. (2012) showed that phylogenetic uncertainty can dramatically inflate estimates of phylogenetic signal and proposed a rarefaction-based method to guide inference. After accounting for phylogenetic uncertainty, Batista et al. (2013) showed that potential loss of evolutionary history caused by extinction of threatened Western Hemisphere anurans would not be significantly higher than that of non-threated anurans. These few examples clearly show the importance of dealing with phylogenetic uncertainty, and demonstrate that systematists and comparative biologists have already begun to develop and employ methods to incorporate estimates of uncertainty in the inference of diversification rates (Day et al. 2008, FitzJohn et al. 2009), character evolution (Losos 1994, Huelsenbeck et al. 2000, Housworth and Martins 2001, Huelsenbeck and Rannala 2003, Ives et al. 2007), reconstruction of ancestral states (Ronquist 2004), and tree topology (Felsenstein 1985, Holder and Lewis 2003).

Here we propose a unified analytical approach to estimate and account for multiple sources of phylogenetic uncertainty. Our method consists of partitioning variance among estimated parameters of evolutionary processes, using random sampling from the universe of probable phylogenies. We illustrate the approach using a clade-wide analysis of the hummingbirds, evaluating how different sources of uncertainty affect several phylogenetic comparative analyses of trait evolution and biogeographic patterns.

Accounting for multiple sources of phylogenetic uncertainty

We propose here an empirical approach, based on simulations, that maximizes the use of incomplete phylogenetic information in ecological studies, while simultaneously accounting for several sources of phylogenetic uncertainty that may distort statistical inferences about evolutionary and ecological processes.

Missing species

Here we define a *phylogenetically uncertain taxon* (henceforth a "PUT" for a single taxon, and "PUTs" for multiple taxa) as a taxonomic unit (e.g. population, species, genus) that is recognized as valid and is accepted as belonging to a particular clade, but is missing from the available phylogenetic tree(s) for that clade. The absence from the tree could be due to multiple causes, such as unavailability of molecular and morphological data (Fig. 2b). However, although the data required to formally reconstruct the phylogenetic relationships of a PUT may be missing, evolutionary information about the PUT is rarely completely non-existent, as no authority would dispute that some other taxon in the clade must be the closest relative of the PUT at some level in the phylogeny. In fact, additional available information, such as taxonomic, morphological or behavioral data, may be useful to define a list of the most likely sister species or lineages of any given PUT (Fig. 2c). Therefore, it makes sense to use all acceptable information available to conservatively define what we will call the *most derived consensus clade* (MDCC), i.e., the node that unequivocally contains each PUT.

Thus, an MDCC can be defined as the most recent common ancestor of all unquestioned candidates for being the closest relative of the PUT (Fig. 2d). Of course, the validity of information (e.g. taxonomic, biogeographical, behavioral, morphological, etc) used to define a MDCC can be questioned. However, if two taxonomists, for example, disagree on the proper MDCC for a given PUT, instead of eliminating the PUT from the statistical analysis, thereby ignoring phylogenetic uncertainty, the MDCC should be conservatively redefined as the most recent common ancestor of the two candidate MDCCs for the PUT, each championed by a different taxonomist (Fig. 2). Thus, it is important to stress that the goal of defining a MDCC to a PUT is not to replace the formal methods of phylogenetic reconstruction, but to conservatively use all available information to allow biologists to make inference of ecological and evolutionary processes using all available information, even with incomplete phylogenies.

We propose here a simple process of building a possible phylogeny that includes all extant species. This process relies on the definition of an MDCC for each PUT and the insertion of the PUT in a random position within its MDCC. We begin by randomizing the order in which PUTs are to be added to the tree. Next, within each MDCC, we assign the PUT to a point along one a branch of the clade. The choice of insertion point along a branch can be uniform random if no additional information is available, or it may be guided by a biologically realistic model of diversification (Kuhn et al. 2011, Davies et al. 2012, Jetz et al. 2012). If the baseline tree is ultrametric, the branch length for the inserted PUT is simply the distance from the attachment point to the end of all other tips. If, however, the baseline tree is not ultrametric, the branch length of the PUT can be sampled from a distribution of possible branch length values. Once a PUT has been inserted, its own branch may serve as a potential insertion point for subsequent species assigned from the PUT queue. Our algorithm iterates until each PUT has been sequentially added to the appropriate MDCC, producing a complete phylogeny of known species (Martins et al. 2013, Batista et al. 2013).

Polytomies

If all polytomies in a tree are the consequence of phylogenetic uncertainty (i.e. "soft" polytomies), then it is necessary to ensure that such uncertainty is quantified and accounted for in statistical analyses that use such a tree (Lewis et al. 2005). To explore the space of all possible dichotomous trees one must resolve the polytomies, assuming no additional information about the evolutionary history of the clade (Batista et al. 2013), or employing a model of diversification (Martins 1996, Housworth and Martins 2001, Kuhn et al. 2011).

Our approach consists in producing fully dichotomous trees by resolving polytomies stochastically. For each node in the tree with three or more branches, we choose two branches at random and reassign them to the same node. Each remaining branch from the original polytomy is then inserted sequentially, in random order, within the clade constructed from the former polytomy, at a randomly chosen position along the length of the existing branches. Dichotomous nodes in the original phylogeny are not changed, thereby preserving the phylogenetic information in the original baseline tree. This process guarantees that the resulting tree is fully resolved and that the phylogenetic uncertainty arising from polytomies is also taken into account when multiple randomized trees are compared (Batista et al. 2013). As with sampling multiple phylogenetic trees (below) or randomizing the position of the PUTs in the phylogeny, not every possible tree topology is examined. However, given a large enough number of replicates, our stochastic procedure ensures that the parameter space will be explored, thereby accounting for phylogenetic uncertainty.

Multiple phylogenetic trees

Modern methods of phylogenetic reconstruction and inference are based on searching the universe of possible phylogenetic trees. The search is guided by empirical data among possible trees, which are judged by their ability to predict the observed data, given a model of molecular evolution (Holder and Lewis 2003). However, because of scarce, ambiguous or missing empirical data, searches are often unable to unambiguously rank all trees within an entire group of equally likely trees, therefore yielding a large set of possible trees. Moreover, especially analyses of large, genome-scale data sets may yield multiple yet well-supported topologies depending on the analytical approach or on selected subsets of data (e.g., Xi et al. 2014, examples in Cooper 2014).

To account for the phylogenetic uncertainty that arises from lack of definitive empirical support for a single phylogenetic hypothesis, one must not limit the statistical analysis to a single sampled tree or to a majority-rule consensus tree that "averages" a larger set of possible trees (Fig. 1). Sampling or averaging among possible trees can potentially improve a point-estimate of a parameter, but does not account for statistical error in evolutionary inference, thereby masking phylogenetic uncertainty associated with tree reconstruction. In practice, variation in the results of ecological analyses that use different phylogenetic trees is an expression of uncertainty in phylogenetic reconstruction. Thus, it is necessary to replicate these analyses over a large number of possible phylogenetic trees, and subsequently study the variation in parameter estimates that arises as a consequence of differences among trees (Fig. 2).

Uncertainty and sensitivity analysis

Computer simulations have long been used to estimate the magnitude of statistical error and to quantify the probability of a given outcome when a system is not fully known (Doubilet et al. 1998, Saltelli et al. 2008). Usually run in tandem with uncertainty quantification, a sensitivity analysis allows one to determine the degree to which the sources of uncertainty in model inputs are responsible for the uncertainty in the model output (Pannell 1997). Coupling uncertainty and sensitivity analysis offers us the opportunity to quantify the robustness of models in the presence of phylogenetic uncertainty, account for uncertainty in evolutionary inference, and enhance communication of the magnitude of statistical error in the study.

Martins (1996) proposed a way to carry out phylogenetic comparative studies when the phylogenetic relationships among species are unknown. In her method, a large sample of trees is generated by randomly resolving a fully polytomic tree ("star phylogeny"), using models of phenotypic evolution and diversification rates (so that uncertainty in branch lengths is also implicitly incorporated). Species traits are then analyzed on each of the possible trees. Although the mean results of such analyses will converge to a non-phylogenetic analysis (Abouheif 1998), the approach of Martins (1996) represents a landmark in phylogenetic inference under uncertainty present in an unresolved phylogeny (see also Housworth and Martins 2001). The mean of the squared standard error of the calculated evolutionary statistic (e.g. the correlation between two species' traits), known as V_s, estimates the true variance of the statistic, which is due to sample variance. The variance of the evolutionary statistic calculated among the randomly generated trees, known as V_p, estimates the variance due to phylogenetic uncertainty. Thus, inferences based on parameter estimation must account for both sources of error, which can be done by computing confidence intervals or *P*-values based on the sum of V_p and V_s.

We expand Martins' (1996) strategy for incorporating phylogenetic uncertainty in parameter estimation further to accommodate multiple sources of phylogenetic uncertainty. We treat each source of uncertainty as a factor in an experimental design, while parameter estimates are treated as the response variable under study. Thus, we can ask how different phylogenetic trees, alternative resolutions of polytomies, and/or probable configurations of PUTs would affect parameter estimates. We partition the amount of variance in a parameter estimate arising from each source of phylogenetic uncertainty with an Analysis of Variance (ANOVA), which not only isolates variance among sources but also calculate the magnitude of standard error in the parameter estimate (a separate Martins' (1996) V_p for each source of uncertainty).

We view the sources of phylogenetic uncertainty we have outlined here as a hierarchy to be employed in the design of an experiment used to assess uncertainty and sensitivity. If multiple empirical phylogenies are available, then this source of uncertainty can be regarded as the highest level of uncertainty in the analysis. Because each empirical phylogeny is unique, its polytomies must be treated individually. Thus, polytomies can be regarded as the second level of phylogenetic uncertainty, immediately below the multiple empirical phylogenies. Of course, for each available empirical phylogeny there is a virtually unlimited number of alternative ways to resolve polytomies randomly or by following some diversification model. Finally, for each combination of original phylogeny and polytomy resolution, PUTs can be inserted according to the procedure described above. Thus, insertion of PUTs is the lowest level in the hierarchy of sources of phylogenetic uncertainty. Because sources of uncertainty are nested within a higher level of classification, and groups representing subordinate levels are randomly chosen, a nested (hierarchic) analysis of variance is ideal for the analysis of sensitivity and uncertainty.

Application example: trait evolution and phylogenetic assemblage patterns in hummingbirds

We illustrate our method by applying it to ecological hypotheses for the hummingbirds (Trochilidae), a large, monophyletic clade (~330 species) with a rich natural history literature and ongoing phylogenetic, morphological, behavioral, and ecological research. McGuire et al. (2007) published a multilocus molecular phylogeny for the hummingbirds that includes only 146 species, but encompasses all the higher-level trochilid diversity (73 of the approximately 104 recognized genera, Schuchmann 1999). Although a more complete hummingbird phylogeny is now available (McGuire et al. 2014), we use the earlier phylogeny(McGuire et al. 2007) to demonstrate our methods because it is typical of current level of phylogenetic knowledge for many comparably diverse taxa.

Replicating the phylogenetic reconstruction methodology of McGuire et al. (2007), we sampled 25,000 trees from the posterior distribution (hereafter called "backbone" trees), after discarding a burn-in of 5,000 steps. We relied upon the taxonomic classification of Schuchmann (1999) to designate the most likely MDCC for each PUT (Fig. 3).

From the total of about 330 described species of hummingbirds, we compiled geographical distributions and ecologically important morphological data for 304 species. Species endemic to islands (the West Indies and the Juan Fernández Archipelago) were not included in the analysis. We compiled estimates of average body mass (intersexual mean) for each of the 304 species from Dunning (2007) and Schuchmann (1999). Although large intersexual and geographical variation in hummingbird body size is well documented (Colwell 2000), the averages used here are useful for broad taxonomic analyses. Based on the literature, we also recorded the maximum known elevational range limits for each species (see Appendix A for references). Body mass (as a measure of body size) and maximum elevational range limit were log-transformed to conform to a normal distribution.

We extracted distributional data for 304 species from an updated version (16 July 2010) of the comprehensive database for all land and fresh-water birds known to have breeding populations in the Western Hemisphere, complied by Rahbek and Graves (2000, 2001), and mapped the geographical range of each species on a gridded map at a resolution of 1° x 1° (latitude-longitude). These maps represent a conservative extent-of-occurrence estimate of the breeding range based on museum specimens, published sight records, and spatial distribution of habitats based on documented records for South America.

Uncertainty quantification

Variability in input data is a fundamental source of uncertainty. To estimate the degree of phylogenetic uncertainty in the hummingbird data due to phylogeny reconstruction and missing species we randomly sampled 100 trees from the posterior (here after called "backbone" trees), and for each tree we replicated the insertion of PUTs 100 times, thereby generating a total of 10,000 fully resolved phylogenetic trees. Next, we calculated the tree-to-tree pairwise distance matrix using the weighted Robinson-Foulds (wRF) metric, which measures the minimum number of internal branches that must be collapsed or expanded to make two trees identical. The weighted version of Robinson-Foulds distance accounts for differences in branch lengths between trees (Steel and Penny 1993). Distance matrices of phylogenetic trees have been used in phylogenetic reconstruction to calculate consensus trees (Swofford 1991), produce supertrees (Bansal et al. 2010), and visualize multi-dimensional space of possible trees (Hillis et al. 2005).

The average wRF pair-wise distance between the 10,000 trees is 2.7423, with a standard deviation of 0.136. We used a multivariate analysis of variance (PERMANOVA, Anderson 2001) to partition variance among phylogenetic trees through the pair-wise wRF distance matrix. The estimated components of variance indicated that 17.02% of total variance among trees is attributable to phylogenetic reconstruction (i.e. differences among backbone trees), whereas 82.97% of total variance is due to phylogenetic uncertainty of missing species (PUTs).

Phylogenetic autocorrelation in species traits

We used Moran's I correlograms to estimate the magnitude of phylogenetic autocorrelation in body size and elevational range limits among hummingbird species (see Gittleman and Kot 1990; Diniz-Filho 2001; Pavoine and Ricota 2012). Moran's I is larger when species within a given phylogenetic distance interval have similar trait values and smaller when species within the same distance interval have very different trait values. To account for uncertainty in phylogenetic reconstruction and missing species, the calculation of correlograms was repeated 1,000,000 times (i.e., replicating the insertion of PUTs 1,000 times in each of the 1,000 randomly sampled backbone phylogenies). We used a nested ANOVA to partition sources of uncertainty for each of the 12 distance classes in the correlogram. Applying our method, the average Moran's I within a distance class is the parameter estimate, whereas additive confidence intervals are calculated around the estimate for each source of uncertainty.

Results of these analyses showed that phylogenetic autocorrelation in body size and maximum elevational range limit both decline with phylogenetic distance, although at different rates (Fig. 4, top panel in each plot). Closely related species tend to have similar body sizes but less similar elevational range limits. Accounting for phylogenetic uncertainty widens confidence intervals by at least a factor of two, regardless of trait (Fig. 4, confidence intervals in upper panel; see the caption for details). For all distance classes, standard confidence intervals due to sampling error are always narrower than confidence intervals due to phylogenetic uncertainty. In addition, the magnitude of error attributed to each source of uncertainty is not constant over phylogenetic distance (Fig. 4, bottom panels). The relative proportion of error caused by missing species tends to be higher in short distance classes, as taxonomic information has been used to assign PUTs to relatively derived positions in the phylogeny (MDCCs). In contrast, large phylogenetic confidence intervals at intermediate and long distance classes arise from uncertainty in the empirical phylogeny located at the base of the coquettes and brilliants clades, which together form the Andean clade (McGuire et al. 2007).

Phylogenetic signal and evolutionary models in species traits

We used Blomberg et al.'s (2003) *K* statistic (hereafter *K*) to estimate the magnitude of deviation from Brownian motion in body size and elevational range limits. *K* measures the degree of similarity in species' traits in relation to the similarity expected under a Brownian motion model of phenotypic evolution, given a phylogenetic hypothesis. *K* is based on the ratio between two measures of distance: MSE_0, the squared distance between trait values and the phylogenetically corrected mean trait value, and MSE, the squared distance between trait values estimated from a variance-covariance matrix derived from a phylogenetic hypothesis. Thus, large values of MSE_0/MSE indicate a strong phylogenetic signal. To allow comparisons between different traits and trees, observed MSE_0/MSE is standardized by expected MSE_0/MSE, assuming that the trait evolves under a Brownian motion model of phenotypic evolution. *K* values less than one indicate that species are less similar for a given trait than expected under Brownian motion evolution, whereas *K* values greater than one indicate that species are more similar for a given trait than expected under Brownian motion evolution.

To account for uncertainty in phylogenetic reconstruction and missing species, we again calculated 10,000 possible *K* values, replicating the insertion of PUTs 100 times in each of the 100 randomly sampled phylogenies. To estimate the accuracy (standard error) of *K* within each tree we used jackknife permutation for each combination of PUT insertion and sampled phylogeny (Efron and Tibshirani 1994). Finally, we used a nested ANOVA to partition error in estimated *K* among different sources of uncertainty.

Our results revealed a surprisingly low phylogenetic signal in body size among hummingbirds ($\overline{\boldsymbol{K}}$ = 0.0597); thus body size evolution can hardly be explained by a simple model of Brownian evolution. However, the standard error of *K* due to sample size (s_{KSamp}=0.0011) is relatively small (2.23%) compared to the error arising from phylogenetic uncertainty due to missing species ($s_{KSamp+PUT}$=0.0494), as this source of uncertainty represents 97.1% of the total error in *K*. Finally, uncertainty due to phylogenetic reconstruction represents only 0.67% of total error of the evolutionary inference ($s_{KSamp+PUT+Phy}$=0.0497).

The relative importance of sources of uncertainty shift discordantly when maximum elevational range limit is considered, although this trait also has very little phylogenetic signal ($\overline{\boldsymbol{K}}$ = 0.0221). Standard error of *K* due to sampling size contributes 59.7% of total error (s_{KSamp} = 0.0179), whereas error due to missing species represents 40.1% of total error ($s_{KSamp+PUT}$ = 0.0300), and error due to phylogenetic reconstruction represents only 0.2% of total error ($s_{KSamp+PUT+Phy}$=0.03001).

To test the hypothesis that estimated *K* values are not significantly different from random expectation, we calculated 100 *K*s for each combination of PUT insertion and sampled phylogeny, randomizing trait values among species (Blomberg et al. 2003, Revell et al. 2008). We employed the same nested ANOVA design to estimate standard error of the null distribution due to each source of uncertainty. Finally, we used Welch's *t*-test to evaluate the hypothesis that estimated *K*s do not differ significantly from the null expectation:

$$t = \frac{\overline{K}_{trait} - \overline{K}_{H0}}{\sqrt{\frac{s^2_{K_{trait}}}{n_{K_{trait}}} + \frac{s^2_{K_{H0}}}{n_{K_{H0}}}}}$$

Estimated *K* for hummingbird body mass is significantly larger than expected under the null expectation ($\overline{K}_{H0}$= 0.0237), and accounting for phylogenetic uncertainty does not affect the hypothesis test (t_{Samp} = 2589.92, $t_{Samp+PUT}$ = 62.5, $t_{Samp+PUT+Phy}$ = 76.72, all *P*-values < 0.001). Conversely, phylogenetic signal in maximum elevational range limit is significantly smaller than expected under the null hypothesis ($\overline{K}_{H0}$= 0.0333). Accounting for phylogenetic uncertainty also does not change the inference of statistical significance in hummingbird maximum elevational range limit (t_{Samp} = -275.36, $t_{Samp+PUT}$ = -34.74, $t_{Samp+PUT+Phy}$ = -44.47, all *P*-values < 0.001).

These results indicate that, although body size evolution deviates significantly from Brownian motion, it carries a small phylogenetic signal. Thus, evolutionary constraints and niche conservatism can be invoked for body size (for instance, evolution under an O-U process – see Hansen et al. 2008), whereas elevational range has less phylogenetic signal than expected by chance.

Phylogenetic community structure at the macroecological scale

Phylogenetic species variability (PSV, Helmus et al. 2007) of an assemblage is maximized (PSV = 1) when the assemblage is composed of the least related species in a clade, and minimized (PSV = 0) when the most related species coexist in an assemblage. We calculated PSV for the 1979 assemblages in the Western Hemisphere based on mapped species ranges with at least two hummingbird species, and tested statistical significance of each PSV value using 300 permutations of species identities. However, because phylogenetic distance between hummingbird species is not known with certainty, we generated 300 possible phylogenies with different PUT insertions, and replicated the calculation of PSV for each of the 1979 assemblages using the each of the 300 randomly selected phylogenetic trees.

Some hummingbird assemblages are composed of species with non-random phylogenetic relationships (Fig. 5). However, identifying any significant departure from randomness requires accounting for all sources of uncertainty. Standard null models to test phylogenetic structure of communities (e.g. Graham et al. 2009), which randomize species identities while preserving species richness, are designed to account only for non-random “sampling” from the phylogeny. For hummingbird assemblages across the Western Hemisphere, an analysis with such a null model would lead to the inference of significant phylogenetic dispersion along the middle and upper Andes (contrary to Graham et al. 2009), and significant phylogenetic clustering across the Pacific coast of Central and North America (Fig 5A). However, uncertainty caused by missing species and phylogenetic reconstruction may seriously affect pattern detectability. Because of substantial uncertainty in the phylogenetic relationship of the two Andean clades (Coquettes and Brilliants), when phylogenetic uncertainty is taken into account no significant phylogenetic dispersion in Andean assemblages is detected. Notice that this result is concordant with the lack of phylogenetic signal and lack of phylogenetic autocorrelation for elevational range at the species level, as we previously discussed. Moreover, once phylogenetic uncertainly is taken into account, many assemblages in Central America can no longer be considered phylogenetically clustered (Fig. 5B).

Because species are neither randomly distributed in the phylogeny nor in geographic space, the magnitude of error in the analysis of phylogenetic structure of assemblages tends to be strongly spatially autocorrelated. In addition, sampling effect is usually higher in species-poor assemblages. On the other hand, assemblages with a higher proportion of species that are members of clades characterized by phylogenetic uncertainty in deep (basal) nodes are subject to a higher proportion of error due to phylogenetic reconstruction. Figure 6 depicts the relative contribution of each source of error in analysis the analysis of phylogenetic structure of hummingbird assemblages. Because of phylogenetic uncertainty in the reconstruction of the relationship between the two, basal, Andean clades (Coquetes and Brilliants), statistical error is substantially higher than the other two sources of error for Andean assemblages. Conversely, statistical error in North American assemblages is mostly due to sampling, as species richness is very low. Uncertainty due to missing species (PUTs) is not particularly pronounced in any assemblage, as no assemblage is composed by more than 17% PUTs.

Concluding Remarks

Phylogenetic uncertainty is not evenly distributed across time and space. As a consequence, the effects of phylogenetic uncertainty in the statistical analysis of phylogenetic data cannot be estimated without a thorough sensitivity analysis on a case-by-case basis. To illustrate the new methods we propose to account for phylogenetic uncertainty, in this paper we used phylogenetic hypotheses derived from molecular data to analyze the hummingbird clade, a well-studied taxonomic group widely accepted as monophyletic. Partition of variance among sources of uncertainty reveals that variance among simulated phylogenies is caused primarily by missing species, in this case, even using the best information available to assign phylogenetically uncertain taxa (PUTs). Conversely, variance among backbone phylogenies, which are estimated through molecular data, is substantially smaller, indicating that efforts to improve the knowledge of evolutionary history of hummingbirds (and other clades that share similar patterns of uncertainty) should be concentrated on gathering molecular data for additional species (e.g. McGuire et al. 2014).

Because phylogenetic uncertainty is expected to be relatively more concentrated within some clades of a phylogeny than within others, statistical analyses of species assemblages or temporal segments of the phylogeny will be affected differently. Thus, statistical analyses of different traits for the same group of species, using the same phylogenetic information, may be differently affected by phylogenetic uncertainty, both in intensity and direction of bias. Of course, comparative analyses among multiple taxonomic groups, based on independently built phylogenetic hypotheses, requires additional caution, as the heterogeneity of variance among groups may seriously distort results in unpredictable directions.

Because the effects of phylogenetic uncertainty in ecological and evolutionary analyses are not subject to generalization, quantifying and accounting for phylogenetic uncertainty through sensitivity analysis is required in all ecological studies. Modern techniques of sensitivity analyses typically involve the application of Monte Carlo methods. Although general simulation methods, such as the one used in this study, could be modified to account for phylogenetic uncertainty in most evolutionary and ecological analyses, the framework for uncertainty quantification and sensitivity analysis should be tailored to the purpose of the study (e.g. estimate of diversification rates, community phylogenetics, comparative analysis of species traits).

Finally, the approach proposed here can be used to quantify the full spectrum of components of phylogenetic uncertainty, guiding sampling strategies for future studies and allowing more reliable interpretations of the relative magnitude of historical and phylogenetic components of biodiversity patterns.

Software Tools

We provide two software toolkits to enable the application of the analytical strategy proposed here. The first software is SUNPLIN (Martins et al. 2013; https://sourceforge.net/projects/sunplin), which is capable of generating randomized phylogenies after the insertion PUTs into backbone trees, with MDCCs assigned. The generated trees can then be applied in any analysis that requires phylogenies as input data. SUNPLIN can be used as an online web service (http://wsmartins.net/sunplin/), as a library that connects through APIs to any compiled software, or directly integrated into R (http://www.ecoevol.ufg.br/pam).

The second software toolkit, PAM (*Phylogenetic Analysis in Macroecology*, http://www.ecoevol.ufg.br/pam), is a compiled computational platform for inference of ecological and evolutionary processes in a spatially explicit context. In PAM, users can not only generate replicates of phylogenetic trees to be used in other software applications, but can also run several statistical analyses commonly used in biodiversity analysis, while estimating and accounting for multiple sources of uncertainty using the analytical framework proposed here. PAM is a work in progress and will be continuously expanded in the future.

Acknowledgements

Several concepts in this paper were inspired by the phylogenetics course (EEB5349), taught by Dr. Paul Lewis, at the University of Connecticut. TFR is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, 564718/2010-6). RKC was supported by the National Science Foundation (USA, DEB-0639979 and DBI-0851245) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). GRG thanks the Alexander Wetmore Fund of the Smithsonian Institution, the Smoketree Trust, the Danish National Research Foundation, and the Center for Macroecology, Evolution and Climate for support. CR acknowledges the Danish National Research Foundation for support to the Center for Macroecology, Evolution and Climate. KF was supported by the National Science Foundation (USA, DEB-1036448, awarded to Louise A. Lewis and Paul O. Lewis). JAFD-F has been continuously supported by CAPES and CNPq through several research grants.

References

Abouheif, E. 1998. Random trees and the comparative method: a cautionary tale. Evolution 52:1197-1204.

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

Bansal, M. S., J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca. 2010. Robinson-Foulds Supertrees. Algorithms for Molecular Biology 5:18.

Batista, M. C. G, S. F. Gouveia, D. L. Silvano and T. F. Rangel. 2013. Spatially explicit analyses highlight idiosyncrasies: species extinctions and the loss of evolutionary history. Diversity and Distributions (*in press*).

Beck, R. M. D., O. R. P. Bininda-Emonds, M. Cardillo, F.-G., R. Liu and A. Purvis. 2006. A higher-level MRP supertree of placental mammals. Evolutionary Biology 6:93-107.

Bininda-Emonds, O. R., J. L. Gittleman and A. Purvis. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biological Reviews 74: 143-175.

Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. Trends in Ecology and Evolution 19:315-322.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are move labile. Evolution 57:717-745.

Colwell, R. K. 2000. Rensch's Rule crosses the line: Convergent allometry of sexual size dimorphism in hummingbirds and flower mites. American Naturalist 156:495-510.

Colwell, R. K. and T. F. Rangel. 2010. Modelling quaternary range shifts and richness on tropical elevational gradients. Phylosofical Transactions of the Royal Society, B 365: 3695-3707.

Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends in Plant Science 19: 576-582.

Davies, T. J., N. J. B. Kraft, N. Salamin and E. M. Wolkovich. 2012. Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. Ecology 93:242-247

Davies, T.J., T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis and V. Savolainen. 2004 Darwin’s abominable mystery: insights from a supertree of the angiosperms. Proceedings of the National Academy of Sciences 107, 1904–1909.

Day, J. D., J. A. Cotton and T. G. Barraclough. 2008. Tempo and Mode of Diversification of Lake Tanganyika Cichlid Fishes. PLoS One 3:e1730

Diniz-Filho, J. A. F. 2001. Phylogenetic autocorrelation under distinct evolutionary processes. Evolution 55:1104-1109.

Diniz-Filho, J. A. F., R. D. Loyola, P. Raia, A. O. Mooers and L. M. Bini. 2014. Darwinian shortfalls in biodiversity conservation. Trends in Ecology and Evolution, *in press*.

Doubilet P., C.B. Begg, M. C. Weinstein, P. Braun, B. J. McNeil. 1985. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. Medical Decision Making 5:157-177.

Dunning, J. B. 2007. CRC handbook of avian body masses. 2nd ed. CRC press.

Efron, B. and R. J. Tibshirani. 1994. An introduction to the bootstrap. Chapman & Hall.

Felsensetein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

FitzJohn, R., W. Maddison, and S. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology 58:595–611.

Gittleman, J. L. and M. Kot. 1990. Adaptation: Statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39:227-241.

Graham, C. H., J. L. Parra, C. Rahbek, and J. A. McGuire. 2009. Phylogenetic structure in tropical hummingbird communities. Proceedings of the National Academy of Sciences of the United States of America 106:19673-19678.

Helmus, M. R., T. J. Bland, C. K. Williams, and A. R. Ives. 2007. Phylogenetic measures of biodiversity. American Naturalist 169:E68-E83.

Hernandez, F.M. and E. S. Vrba. 2005. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. Biological Reviews 80: 269–302.

Hillis, D. M., T. Heath and K. St. John. 2005. Analysis and Visualization of Tree Space. Systematic Biology 54:1-12.

Housworth, E. A. and E. P. Martins. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. Systematic Biology 50:628-639.

Holder, M. and P. O. Lewis. 2003. Phylogeny Estimation: Traditional and Bayesian Approaches. Nature Reviews Genetics 4:275-284.

Huelsenbeck, J. P. and B. Rannala. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57:1237-1247.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accommodating Phylogenetic Uncertainty in Evolutionary Studies. Science 288:2349-2350.

Ives, A. R., P. E. Midford and T. Garland Jr. 2007. Within-species variation and measuremente Error in Phylogenetic Comparative Methods. Systematic Biology 56:252-270.

Isaac, N. J. B., S. T. Turvey, B. Collen, C. Waterman, and J. E. M. Baillie. 2007. Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. PLoS ONE 2:e296.

Kuhn, T. S., A. O. Mooers, and G. H. Thomas. 2011. A simple polytomy resolver for dated phylogenis. Methods in Ecology and Evolution 2:427-436.

Lewis, P. O., M. T. Holder and K. E. Holsinger. 2005. Polytomies and Bayesian Phylogenetic Inference. Systematic Biology 54:241-253.

Losos, J. B. 1994. An approach to the Analysis of Comparative Data When a Phylogeny is Unavailable or Incomplete. Systematic Biology 43:117-123.

Martins, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50:12-22.

Martins, W.S., W. C. Carmo, H. J. Longo, T. C. Rosa and T. F. Rangel. 2013. SUNPLIN: Simulation with Uncertainty for Phylogenetic Investigations. BMC BioInformatics 14:324-335.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen Jr. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Systematic Biology 56:837-856.

McGuire, J. A., C. C. Witt, J. Remsen Jr, A. Corl, D. L. Rabosky, D. L. Altshuler, and R. Dudley. 2014. Molecular phylogenetics and the diversification of hummingbirds. Current Biology 24:910-916.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann and A. O. Mooers. 2013. The global diversity of birds in space and time. Nature 491:444-448.

Pagel, M. D. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Pannell, D.J. 1997. Sensitivity analysis of normative economic models: Theoretical framework and practical strategies. Agricultural Economics 16: 139-152.

Pavoine, S. and C. Ricota. 2012. Testing for phylogenetic signal in biological traits: the ubiquity of cross-product statistics. Evolution, 67:828-840.

Rahbek, C. and G. R. Graves. 2000. Detection of macro-ecological patterns in South American hummingbirds is affected by spatial scale. Proceedings of the Royal Society of London, B 267:2259-2265.

Rahbek, C. and G. R. Graves. 2001. Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences of the USA 98:4534-4539.

Ranwez, V., V. Berry, A. Criscuolo, P. H. Fabre, S. Guillemot, C. Scornavacca and E. J. P. Douzery. 2007. PhySIC: a veto supertree method with desirable properties. Systematic Biology 56:798–817.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Systematic Biology 57:591-601.

Ronquist, F. 2004. Bayesian Inference of Character Evolution. Trends in Ecology and Evolution 19:475-481.

Roquet, C., W. Thuiller and S. Lavergne. 2012. Building megaphylogenies for macroecology: taking up the challenge. Ecography 36:13-26.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. Global Sensitivity Analysis: The Primer. John Wiley & Sons.

Schuchmann, K.L. 1999. Family Trochilidae. Pages 468-680 in J. del Hoyo, A. Elliott, and J. Sargatal, editors. Handbook of the Birds of the World. Vol. 5. Barn-owls to Hummingbirds. Lynx Edicions, Barcelona.

Steel, M. A. and D. Penny. 1993. Distributions of tree comparison metrics - some new results. Systematic Biology 42:126-141.

Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295-333 in M. M. Miyamoto and J. Cracraft, editors. Phylogenetic analysis of DNA sequences. Oxford Univ. Press, New York.

Xi Z., L. Liu, J.S. Rest and C.S. Davis. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. (*in press*).

Webb, C. O. and M. J. Donoghue. 2005. Phylomatic: tree retrieval for applied phylogenetics. Molecular Ecology Notes 5:181-183.

Figure Legends:

Figure 1: A conceptual example of the effect of different sources of uncertainty in phylogenetic trees (left) on the estimation of phylogenetic relationship between species (right, represented as matrices). In each matrix, cells in the lower-left half-matrix represent the existence of phylogenetic information about the relationship between a pair of species, because both species are present in the tree. Conversely, a dash represents the absence of such phylogenetic information, as one or both of the species is missing from the tree. In the upper-right half-matrix a normal distribution represents the variance in the estimated phylogenetic relationship, while zeros represent certainty. (A) Hypothetical, unknown true tree, without missing species or uncertainty in species relationships. (B) Consensus tree with missing species. Relationships are assumed to be known with certainty. (C) Polytomic super-tree. Insertion of missing species in polytomies generates a complete tree, and relationships are assumed to be known with certainty, although the tree differs from the true tree (A). (D) Replication in the use of phylogenetic trees incorporates the uncertainty in the relationship between species, but missing species are ignored. (E) Missing species are inserted in multiple phylogenies, accounting for uncertainty in phylogenetic reconstruction and lack of a complete phylogeny.

Figure 2: Schematic representation of a workflow using the analytical strategy, proposed here, to account for phylogenetic uncertainty. (A) The true, but unknown, tree for a clade of 8 taxa. (B) Molecular data are available for only six taxa; for taxa B and G no molecular data are available. (C) Two experts, based on their knowledge (e.g. taxonomy, behavior, morphology, geographic distribution, etc.), suggest possible sister taxa of the two taxa that lack molecular data. (D) Using the available molecular data and phylogenetic reconstruction methods, three backbone phylogenies are proposed. The variation between backbone phylogenies arises from uncertainty in the process of

phylogenetic reconstruction. Among the backbone phylogenies, the taxa B and G are considered *phylogenetically uncertain taxa* (PUT), because no molecular data are available for them, and therefore they are missing from the backbone phylogenies. Using the information provided by the experts the *most derived consensus clade* (MDCC) is found for each PUT. For PUT B, the MDCC is the clade that necessarily includes taxa C, D, and A (indicated by a bold B in each backbone phylogeny), whereas for PUT G the MDCC must include taxa F and H (indicated by a bold G in each backbone phylogeny). (E) Statistical analyses that use phylogenetic trees as input data should be replicated using samples of operational trees. In each of the operational trees, the PUTs B and G were randomly inserted within their respective MDCCs. The insertion of each PUT was replicated three times (columns) for each backbone tree. The variation among operational trees that use the same backbone tree (each column of the operational trees) arises from variation in placement of the missing taxa (as the backbone tree is not changed by the randomization process), and is caused by uncertainty in the phylogenetic relationships of taxa B and G (both PUTs).

Figure 3: Hummingbird phylogenies. (A) One backbone phylogeny from McGuire et al. (2007), with red internal branches indicating the *Most Derived Consensus Clades* (MDCCs) used in this analysis. (B) Red taxa (polytomies) indicate PUTs inserted in the original phylogeny at the base of their MDCCs. (C) Red taxa indicate PUTs inserted to yield a fully-resolved phylogeny, randomizing their position within the respective MDCC.

Figure 4: Moran's I correlogram (top panels) for two hummingbird traits (body size and maximum elevational range limit). The inner 95% confidence interval (C.I.) around each estimate indicates variance due to sampling error, the intermediate 95% C.I. the variance due to missing species, and the outer 95% C.I. the variance due to phylogenetic reconstruction. The bottom panels indicate the

relative proportion of statistical error caused by each source of uncertainty across evolutionary time: sampling (green), missing species (PUTs, red) and phylogenetic reconstruction (blue).

Figure 5: Hummingbird assemblages with PSV values significantly different from null expectation. Blue cells indicate significant phylogenetic clustering, whereas red cells indicate significant phylogenetic dispersion. (A) Standard PSV analysis, considering only non-random “sampling” from the phylogeny. (B) Re-analysis of PSV accounting for three sources of error: sampling, missing species, and phylogenetic reconstruction. Notice absence of significant phylogenetic dispersion and decreased areas of phylogenetic clustering when all sources of uncertainty are accounted for.

Figure 6: Relative statistical error associated with PSV analysis for hummingbird species present in each map cell, partitioned among sources of uncertainty. Each point within the cube has a unique color that represents the relative proportions of uncertainty due to phylogeny, PUTs and sampling error, as shown by the axes. In the map, the prevalence of red cells indicates sampling error as the main source of uncertainty, whereas blue indicates substantial error caused by phylogenetic uncertainty. Notice the absence of green and yellow areas, indicating that phylogenetic uncertainty due to missing species (PUTs) is relatively irrelevant in the phylogenetic structure of assemblages.