Syst. Biol. 67(6):1091–1109, 2018

DOI:10.1093/sysbio/syy031
Advance Access publication April 25, 2018

Rethinking phylogenetic comparative methods

Josef C. Uyeda[1,*], Rosana Zenil-Ferguson[2,3], and Matthew W. Pennell[4]

[1] *Department of Biological Sciences, Virginia Polytechnic Institute and State University, 926 West Campus Drive, Blacksburg, VA 24061 USA;*
[2] *Department of Biological Sciences, University of Idaho, 875 Perimeter Drive, Moscow, ID 83844 USA;*
[3] *Department of Ecology, Evolution and Behavior, University of Minnesota, 1479 Gortner Avenue, St. Paul, MN 55108 USA; and*
[4] *Department of Zoology and Biodiversity Research Centre, University of British Columbia, #4200-6700 University Blvd., Vancouver, BC V6T 1Z4, Canada*
Correspondence to be sent to: Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA;
E-mail: juyeda@vt.edu.

Received: 21 November 2017; reviews returned: 17 March 2018; accepted: 17 April 2018
Associate Editor: Nicholas Matzke

Abstract.—As a result of the process of descent with modification, closely related species tend to be similar to one another in a myriad different ways. In statistical terms, this means that traits measured on one species will not be independent of traits measured on others. Since their introduction in the 1980s, phylogenetic comparative methods (PCMs) have been framed as a solution to this problem. In this article, we argue that this way of thinking about PCMs is deeply misleading. Not only has this sowed widespread confusion in the literature about what PCMs are doing but has led us to develop methods that are susceptible to the very thing we sought to build defenses against—unreplicated evolutionary events. Through three Case Studies, we demonstrate that the susceptibility to singular events is indeed a recurring problem in comparative biology that links several seemingly unrelated controversies. In each Case Study, we propose a potential solution to the problem. While the details of our proposed solutions differ, they share a common theme: unifying hypothesis testing with data-driven approaches (which we term "phylogenetic natural history") to disentangle the impact of singular evolutionary events from that of the factors we are investigating. More broadly, we argue that our field has, at times, been sloppy when weighing evidence in support of causal hypotheses. We suggest that one way to refine our inferences is to re-imagine phylogenies as probabilistic graphical models; adopting this way of thinking will help clarify precisely what we are testing and what evidence supports our claims. [Causality; graphical models; macroevolution; phylogenetic natural history]

Every so often, evolution comes up with something totally new and unexpected, a so-crazy-it-just-might-work set of adaptations that is the stuff of nature documentaries. Many biologists likely have a favorite example of a lineage that has evolved something spectacular such as devilishly horned lizards that squirt blood from their eye sockets or marine sloths that grazed ancient seabeds.

As macroevolutionary researchers, it is hard to know what to do with these types of events (Vermeij 2006). Their singular and unreplicated nature seems incompatible with models that we typically use to describe change over time, such as Brownian motion (BM; Felsenstein 1973) or the Mk model (Pagel 1994; Lewis 2001). Such models presume continuity, whereas one-off events, such as the evolution of novel nutritive function in exocrine glands leading to mammalian milk, have no clear precedent in history. The evolution of such traits may set in motion a cascade of changes across an organism, such that descendant lineages may look very different in many ways from their more distant relatives. Or alternatively, a suite of traits may just happen to change at the same time. In either case, it is these sorts of idiosyncratic and unreplicated events that we often think of when we think of the need to consider phylogeny in analyses of comparative data. And this is not an abstract concern; a wide breadth of macroevolutionary data suggest that abrupt shifts and discontinuities have been a major feature of life on Earth (Uyeda et al. 2011, 2017; Landis and Schraiber 2017; Jablonski 2017). But as recent controversies in phylogenetic comparative biology have highlighted, our current methods (reviewed in O'Meara 2012; Pennell and Harmon 2013; Garamszegi 2014) are not designed to deal with such dynamics.

For example, Maddison and FitzJohn (2015) recently demonstrated that common statistical tests (e.g., Maddison 1990; Pagel 1994) for the evolutionary correlation of discrete characters are prone to reporting a significant association even when the pattern is driven by a single (or, very few) independent transition(s) from one character state to another. Maddison and FitzJohn (2015) referred to such scenarios as cases of "phylogenetic pseudoreplication" (see also Read and Nee 1995; Nee et al. 1996).

We will argue that this unresolved challenge permeates not just tests for discrete character correlations, but nearly every method of finding associations in comparative methods (Fig. 1). For example, Rabosky and Goldberg (2015) show that applying trait-dependent diversification models (e.g., BiSSE, Maddison et al. 2007) to real-world phylogenies, which are usually not shaped like trees resulting from simulations of a birth–death stochastic process (Mooers and Heard 1997), often leads to support

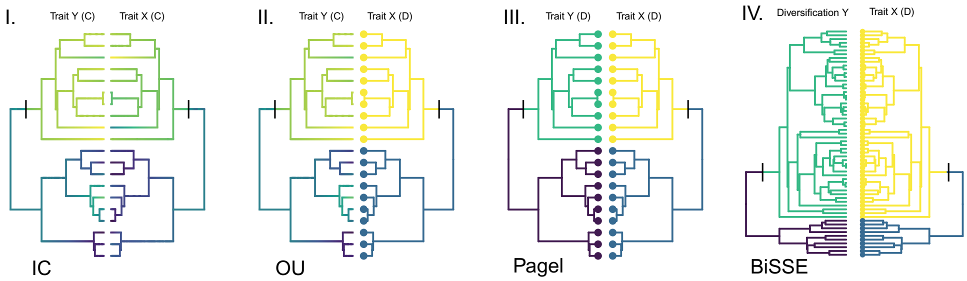

FIGURE 1. Singular, unreplicated events (vertical dashes) can generate apparently significant associations across several types of comparative analyses. Case Studies I–III are indicated in panels I–III, and though we do not consider diversification models such as BiSSE in our examples, they are similarly affected (panel IV). In each case, we map (in some cases, arbitrarily) the dependent variable (Y) on the phylogeny on the left and the predictor trait on the same phylogeny to the right (X), and indicate whether the trait is a continuous trait (C), a discrete trait (D) or a diversification rate. Colors on the branches indicate the state of the character on the phylogeny—either continuous trait value, discrete character state, or diversification rate regime. Panels I and III correspond to variations of "Felsenstein's worst-case scenario" and "Darwin's scenario," respectively. We also suggest a common method used to analyze such associations: IC = Independent Contrasts (Felsenstein 1985); OU = Ornstein–Uhlenbeck models (Butler and King 2004); Pagel = Pagel's correlation test (Pagel 1994).

for trait-dependent diversification models regardless of whether traits are actually affecting speciation and extinction. The work of Beaulieu and O'Meara (Beaulieu et al. 2013; Beaulieu and O'Meara 2014, 2016) has illuminated important underlying reasons behind Rabosky and Goldberg's findings: the failure to consider alternative models in which the "background" diversification rate changes across the tree (i.e., there is a shift in diversification regimes unrelated to the trait being considered). To address this shortcoming, Beaulieu et al. (2013) borrowed an idea from molecular phylogenetics (Galtier 2001; Penny et al. 2001), and developed a Hidden States Model (HSM) for describing the evolution of a binary character along a phylogeny. In their HSM, the transition rates between character states depend on the 'hidden' state of another, unobserved, trait also evolving along the tree (also see Price 1997; Felsenstein 2011 both of whom explored a related model). Applying the same principle to trait-dependent diversification models, they showed how models that include background heterogeneity in diversification rates provide a fairer comparison to the hypothesis of genuine state-dependent diversification (Beaulieu and O'Meara 2016). Rather than considering a biologically unrealistic constant-rate null hypothesis, Beaulieu and colleagues built models that allowed traits and diversification to vary in biologically plausible ways (also see Zenil-Ferguson and Pennell 2017 on this point).

We think that the solution proposed by Beaulieu and O'Meara (2016)—accounting for background shifts in evolutionary regimes unrelated to the focal trait association—is general and applies across comparative biology. In this article, we develop this argument through a series of three Case Studies, depicted in panels I–III of Figure 1. We will show in each Case Study that rare evolutionary events may deceive our methods and distort our interpretations. For each study, we will then sketch out possible solutions for making causal inferences from comparative data. These solutions differ in their modeling details and methods of inference, but they share a core idea.

More specifically, all three Case Studies revolve around the problem of how to discover plausible histories of singular events, or transitions in evolutionary regimes—a practice we call "phylogenetic natural history"—and how to disentangle the impact of these events from that of the hypothesized effects we are investigating. In our examples, we highlight scenarios where a single change in the background evolutionary dynamics can lead to apparent associations between factors of interest, as we find such cases particularly illuminating. But the problems (and potential solutions) we identify apply just as well to situations where the background evolutionary dynamics change more frequently.

By working through the Case Studies, we arrive at two general recommendations for how to move phylogenetic comparative methods (PCMs) forward. First, we advocate for unifying hypothesis-testing and data-driven approaches. Rather than being alternative methods of investigating macroevolutionary processes and patterns, they are complementary, and in our view, essential, to one another. Second, we propose that comparative biologists need to be more careful about how we draw causal inference from phylogenetic data. One particularly elegant solution is to render comparative analyses as graphical models. These graphical models can help clarify exactly what causal statements we are making and what the limits of these inferences are.

CASE STUDY I: FELSENSTEIN'S WORST-CASE SCENARIO

More than anything else, it was the famous series of figures depicting the "worst-case scenario" (Figs. 5, 6, and 7 in the original; our Fig. 2) from Felsenstein's iconic 1985 article "Phylogenies and the comparative method" that awakened biologists to the need for

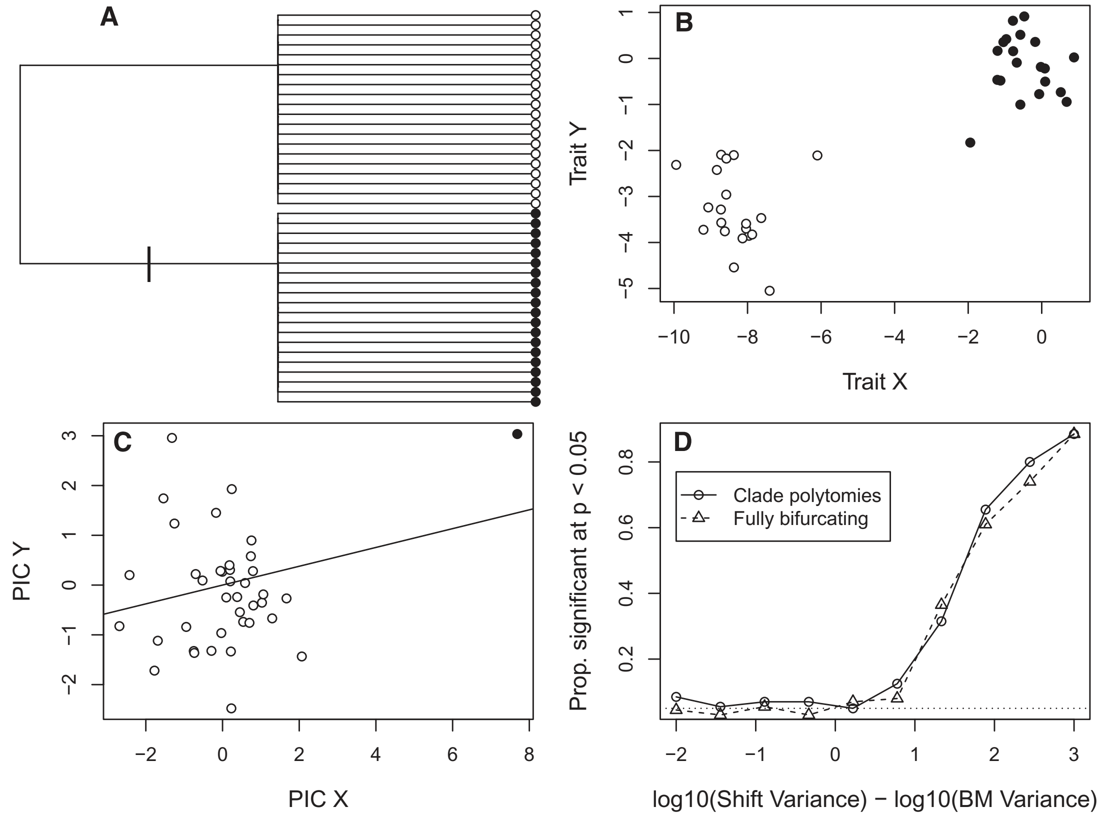

FIGURE 2. Felsenstein's worst-case scenario (Felsenstein 1985) illustrates a problem quite like that identified by Maddison and FitzJohn. Here we modify Felsenstein's original generating process from simple BM, to A) BM with a single burst occurring on the stem branch of one of the two clades (indicated by vertical dash). B) The distribution of trait values produces a figure very similar to Felsenstein's original scenario, but results in C) a single contrast (black) that is not well-described by the estimated BM process, and thereby generates a significant regression of PIC Y and PIC X (dotted line) despite both X and Y in the shift and BM distributions being uncorrelated. D) As the ratio of the shift variance to the BM variance increases, the proportion of contrast regressions that return a significant result increases dramatically (each point represents 200 simulations for a fixed phylogeny, with both the BM process and the random draw from the shift distribution being uncorrelated with equal variance for both traits). While IC corrects for singular events consistent with BM, it does not correct for the more general phenomenon of dramatic singular events driving significant results in comparative analyses. Note that the nonindependence of species is not the issue.

tree-thinking and started a revolution in modern comparative biology. The idea is simple: as a result of shared ancestry, measurements taken on one species will not be independent from those collected on another and especially so, if the two species are closely related. This nonindependence can create apparent correlations between traits that, are in truth, evolving independently. To illustrate the effect of nonindependence of characters, Felsenstein generated a scenario in which two clades are separated by long branches (our Fig. 2). He then evolved traits according to a BM process along the phylogeny; he recovered a significant regression slope using Ordinary Least Squares (OLS) despite there being no evolutionary covariance between the traits.

While other researchers had hit upon similar notions throughout the early 1980s (e.g., Clutton-Brock and Harvey 1980; Mace et al. 1981; Ridley 1983; Stearns 1983; Cheverud et al. 1985), none of these had the pervasive impact that Felsenstein's presentation did (see e.g., Losos 2011 who reproduces the figures and the accompanying reasoning in his presidential address for the American Society of Naturalists). The problem is just so obvious—data from different clades clustering in different parts of the bivariate plot—all you have to do is look. And while of course his proposed solution, "independent contrasts" (IC), was widely adopted, we suspect it is the clarity with which Felsenstein articulated the problem that has kept his article a hallmark of biological education and a testament to the importance of tree-thinking, even as his method has largely been superseded by the least squares (Grafen 1989) (which is identical to IC if BM is used to model the covariance of errors: Rohlf 2001; Blomberg et al. 2012) and mixed model (Lynch 1991; Housworth et al. 2004; Hadfield and Nakagawa 2010) approaches.

However, an important part of this story is often missed: Felsenstein also noted that the problem of nonindependence does not occur if "characters respond essentially instantaneously to natural selection in the current environment, so that phylogenetic inertia is essentially absent" (p. 6). Despite this comment, a common misunderstanding of his argument is that the problem inherent in a nonphylogenetic regression of phylogenetically structured data is that species are not independent. In fact, independence of data is not an assumption of standard (nonphylogenetic) linear regression at all. Rather, standard linear regression assumes that the *errors* of the fitted model are independent and identically distributed (i.i.d.). As a result, many applications of a "phylogenetic correction" seem to be missing the point (Revell 2010; Hansen and Bartoszek 2012): if all of the phylogenetic signal in a data set is present in the predictor trait and the errors are i.i.d., then there is no need for any phylogenetic correction (Rohlf 2001, 2006). (However, phylogenetic analyses are nearly always needed to determine this condition in the first place.)

We suggest that what made Felsenstein's *prima facie* argument so compelling was that it appealed to biologists' intuition that many large clades of organisms are just different in many potentially idiosyncratic ways (Vermeij 2006). If the apparent association between traits found in a nonphylogenetic regression analysis is simply a result of these idiosyncratic differences between clades, then we would be inferring a relationship from unreplicated data (Nee et al. 1996), irrespective of the purely statistical consideration of whether errors are i.i.d.

Here, we revisit Felsenstein's worst-case scenario in order to demonstrate that IC and Phylogenetic Generalized Least Squares (PGLS) do not completely address the problem that we tend to think they do—these methods are still susceptible to singular evolutionary events. To demonstrate this, we add a slight twist to Felsenstein's original example. First, we used a phylogeny with two clades, each of which is internally unresolved, similar to that of the 1985 article. We emphasize that the only phylogenetic structure is that stemming from the deepest split. We then simulated two traits under independent BM processes, each with an evolutionary rate (σ^2) of 1. However, at some point on a stem branch of one of the two clades we introduce a singular evolutionary "event"—i.e., a dramatic shift in a lineage's phenotype—drawn from a multivariate normal distribution with uncorrelated divergences and equal variances that are a scalar multiple of σ^2. The resulting distribution of the data suggests a situation very similar to Felsenstein's worst-case scenario—and what we suspect is the type of problem envisioned by most biologists when they warn their students of the dangers of ignoring phylogeny.

One would hope that our tools for "correcting for phylogeny" would recognize that the apparently strong relationship between the two traits in our example was driven by only a single contrast. However, this is not the case. That single contrast results in a very high-leverage statistical outlier that drives significance as the size of the shift increases (Fig. 2). We can repeat the same exercise with more phylogenetically structured data (where the two clades of interest are fully bifurcating following a Yule process) and obtain identical results (Fig. 2, see Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.p8066hd). This is disconcerting since our intuition suggests that we do not have compelling evidence for a causal relationship between these two traits (i.e., there is very little reason for us to believe from this correlation alone that one trait is an adaptation to the other).

How can we formulate a better set of models that can account for what our intuition tells us is a dangerous situation for causal inference? We can do so by including another phylogenetically plausible model: that trait correlations result from a single random shift, drawn from a different distribution than the one used to model trait evolution across the rest of the branches.

Let us consider a situation quite distinct from Felsenstein's multivariate BM (mvBM) scenario. Here traits do not evolve by mvBM, but rather undergo a shift at a single point (perhaps an ancient dispersal event where one clade invaded a new environment). In such a scenario, we only need to consider the phylogeny in as much as a given species exists on either side of the event in question. We can then erect two statistical models: a linear regression model and a singular event model.

Linear regression model:

$$\begin{aligned} Y &= \beta_X X + \beta_0 + \epsilon; \\ X &= \psi(\theta) \end{aligned} \tag{1}$$

where β_X and β_0 are the slope and intercept to the regression of Y on X, ϵ is a vector containing i.i.d. random variables that describe the errors, and the predictor X is generated by some stochastic process $\psi(\theta)$ on the phylogeny (e.g., a random variable describing a single burst in X on the stem branch of one of the two clades with parameters θ). Thus, under the laws of conditional probability, the bivariate probability of X and Y under the linear model conditional on parameters $\theta_{LM} = (\beta_X, \beta_0, \sigma)$ is:

$$\begin{aligned} P(X,Y|\theta_{LM},\theta_\psi) = &P(Y|X,\beta_X,\beta_0,\sigma)P(X|\theta_\psi) \\ &\times P(\beta_X)P(\beta_0)P(\sigma)P(\theta_\psi) \end{aligned} \tag{2}$$

where θ_ψ are the parameters of the process for X on the phylogeny, and σ^2 is the variance associated to ϵ. This equation is derived from the assumed path of causation between X and Y, since the likelihood function of trait X, denoted by $P(X|\theta_\psi)$, is independent of Y, while the likelihood function of Y, denoted by $P(Y|X,\beta_X,\beta_0,\sigma)$ depends on X. The remaining terms in the probability statement are interpreted as prior distributions for the parameters in a Bayesian inferential framework.

Alternatively, X and Y may not be related to one another at all. Rather, they may be the products of singular random evolutionary events denoted by $E1$, and $E2$, that happened to occur on the branch separating two clades.

Singular events model:

$$\begin{aligned} Y &= \beta_Y I_{E1} + \beta_{Y0} + \epsilon_Y \\ X &= \beta_X I_{E2} + \beta_{X0} + \epsilon_X, \end{aligned} \tag{3}$$

where the variables I_{E1} and I_{E2} are indicator random variables that take the value of 1 if an observation is from a lineage that experienced a phylogenetic event or shift, and a value of 0 otherwise. Furthermore, β_{Y0} and β_{X0} are the parameters that describe the trait means had they not experienced the singular evolutionary event in question and each linear model in equation (3) has errors with variances σ_Y^2 and σ_X^2, respectively.

For the singular event model (with parameters $\Lambda_{SE} = (\beta_Y, \beta_{Y0}, \beta_X, \beta_{X0}, \sigma_Y, \sigma_X)$) the bivariate probability becomes:

$$\begin{aligned} &P(X,Y|\Lambda_{SE}) \\ &\quad = P(\beta_Y)P(\beta_{Y0})P(\beta_X)P(\beta_{X0})P(\sigma_X)P(\sigma_Y) \\ &\qquad \times P(N_{E1}=1)P(N_{E2}=1)P(L_{E1}|N_{E1})P(L_{E2}|N_{E2}) \\ &\qquad \times P(Y|L_{E1},\beta_Y,\beta_{Y0},\sigma_Y)P(X|L_{E2},\beta_X,\beta_{X0},\sigma_X), \quad (4) \end{aligned}$$

where $P(N_{E1}=1)$ and $P(N_{E2}=1)$ are the probabilities of observing a single shift on the phylogeny, and $P(L_{E1}|N_{E1})$ and $P(L_{E2}|N_{E2})$ are the probabilities of observing these singular shifts in locations L_{E1} and L_{E2}, respectively.

The linear regression and singular event models lead to potentially very different distributions of trait data at the tips. For example, under the singular event model, the distribution of Y is conditionally independent of X after accounting for $L_{E1}, \beta_Y, \beta_{Y0}$—a testable empirical prediction that will often result in these two models being easily distinguishable with model selection. But failing to consider the singular event model as a possibility is a problem: even for the simple case of two continuous traits, we have shown how easily data simulated under the singular event model can result in highly significant regressions for OLS, PGLS, and IC regressions, regardless if the errors are simulated as independent or phylogenetically correlated with respect to the model and phylogeny. We also note that estimating a λ transformation for the errors (Pagel 1999; Freckleton et al. 2002) will not rescue the analysis; the estimated value of λ will lie between 0 and 1 and we have found both these more extreme cases (OLS and IC, respectively) to be susceptible.

One might argue that the situation we describe is the violation of the assumption of a BM model of evolution—and this would, of course, be correct (see also Maddison and FitzJohn 2015). Indeed, for decades it has been common practice (but unfortunately, not universally so) to test whether contrasts are i.i.d. after conducting an analysis using IC (Garland et al. 1992; Purvis and Rambaut 1995; Slater and Pennell 2013; Pennell et al. 2015) and many researchers have followed Jones and Purvis (1997) in dropping outlying contrasts from regressions. Felsenstein recognized this particular vulnerability in his method and correctly predicted that the underlying model was an "obvious point for future development" (p. 14). While today we have a much wider range of comparative models to choose from including some that allow for adaptive shifts, most continuous trait models are Gaussian (e.g., Pagel 1999; Blomberg et al. 2003; Butler and King 2004; O'Meara et al. 2006; Eastman et al. 2011; Beaulieu et al. 2012; Uyeda and Harmon 2014) and do not accommodate abrupt, discontinuous shifts in phenotypes. It is only recently that alternative classes of models have been considered (Landis et al. 2012; Elliot and Mooers 2014; Schraiber and Landis 2015; Blomberg 2017; Boucher et al. 2017; Duchen et al. 2017). Whether or not these other types of models can sufficiently account for rare, singular events will be examined in the next section.

Nevertheless, our primary point here is to suggest that the phenomenon that made Felsenstein's argument so intuitive is not the violation of i.i.d. errors but rather the biologically intuitive realization that unreplicated differences colocalized on a single branch provide only weak evidence of a causal relationship between traits. Furthermore, models that actually describe such scenarios—like our "singular events" model—are rarely considered in comparative analyses. Admittedly, fitting such models to biologically realistic cases more complex than Felsenstein's scenario will require estimating the location and number of events and we therefore view our "singular events" model as primarily an illustrative alternative solution to Felsenstein's thought experiment. Nevertheless, the example illustrates that the phylogeny imposes a challenge to the inference of meaningful associations between traits not because it renders errors nonindependent, but because the structure of the phylogeny allows for ancient, potentially unknowable causal factors (which may be few or even singular) to drive widespread associations between traits. Evaluating the validity of these associations as evidence for a meaningful relationship, even in the case of continuous traits, is precisely the unresolved challenge identified by Maddison and FitzJohn (2015) in the case of discrete character correlations (as we will further elaborate in Case Study III).

Case Study II: Adaptive Hypotheses and Singular Shifts

As stated above, the IC method is based on the BM model of trait evolution. While this model is useful (and has often been used) for testing for adaptation, it is inconsistent with how we think of the *process of adapting* to an optimal state (Lande 1976; Hansen 1997; Hansen and Orzack 2005; Hansen et al. 2008; Hansen and Bartoszek 2012). Hansen's introduction of the Ornstein–Uhlenbeck (OU) process to comparative biology and the suite of methods built on his approach have been the only real attempts to actually try and capture the basic dynamics of adaptive trait evolution on phylogenies.

Multioptima OU models have been widely used to test for the presence of shifts in evolutionary regimes (i.e., parts of the phylogeny with their own optima, or less commonly, their own strength of selection parameters). Tests of adaptive evolution come in two flavors: those with an *a priori* hypothesis (or hypotheses) regarding which lineages belong to which distinct regimes based on ancestral state reconstruction of explanatory factors (Butler and King 2004; Beaulieu et al. 2012) and those where the locations of regime changes are themselves estimated along with the parameters of the OU process

(Ingram and Mahler 2013; Uyeda and Harmon 2014; Khabbazian et al. 2016).

These two types of approaches represent two different philosophies of data analysis that follow a schism that cuts through comparative methods. For example, there are two major ways to investigate the dynamics of lineage diversification: test specific hypotheses about the drivers of diversification rate shifts (e.g., the 'SSE' family of models Maddison et al. 2007; FitzJohn 2012) or search for the most-supported number and configuration of shifts (Alfaro et al. 2009; Stadler 2011; Rabosky 2014). The former (hypothesis testing) seeks to understand the causes of evolutionary shifts, while the latter (data driven) is a descriptive and exploratory approach to understanding evolutionary patterns. As we alluded to above, we refer to these data-driven approaches as "phylogenetic natural history" due to their similarity to the practice of natural history observations in nature but projected backwards through phylogenetic space and time (Maddison and FitzJohn 2015).

Of course, the types of inferences we can make will be limited by our choice of approach. For example, it may be tempting to use exploratory approaches such as *BAMM* (Rabosky 2014) or *bayou* (Uyeda and Harmon 2014) to search a vast range of model space to find a particularly well-supported statistical hypothesis, observe the shifts identified, and then come up with post hoc explanations for why that particular configuration fits an adaptive story that the researcher can suddenly construct with great precision. (Comparative biologists are of course not unique in succumbing to such temptations; see e.g., Pavlidis et al. 2012). In fact, discovering the location of well-supported shifts on the phylogeny does not say anything about causation; it is merely a descriptive technique to find major features of the data where there is evidence that the parameters governing the dynamics of trait evolution have shifted on the phylogeny. It is nonetheless useful—and we argue essential—that a researcher know where these shifts occur. The reasons for this are covered in Case Study I: these major shifts are likely to drown out any biological signal in a data set if they are unaccounted for by our hypothesis-driven models. But even beyond these statistical considerations, observing which lineages and clades differ in evolutionary tempo and mode is as vital to good macroevolutionary inference as traditional natural history is to biology more generally—such knowledge and familiarity with the organisms is essential to generating "empirically-justifiable" synthetic theories of evolution (Futuyma 1998). For these reasons, we argue that hypothesis-driven and phylogenetic natural history approaches are complementary: we must pit our particular causal hypotheses against a descriptive "stuff-happens" model built on idiosyncratic singular evolutionary events.

To illustrate how we might go about uniting these two modes of inference to disentangle the support for causal models of evolution from that attributable to singular events, we reanalyze a data set introduced by Scales et al. (2009) on lizard muscle fiber proportions (hereafter, the 'Scales' data set). (An expanded data set was reanalyzed by Scales and Butler (2016) with slightly modified hypotheses. However, the original 2009 article serves as a clearer example with which to illustrate our perspective; we do not delve into differences between the two.)

Scales et al. (2009) are interested in the composition of muscle fiber types in squamate lizards, and whether these muscle fibers evolve adaptively in response to the changing behavior and ecology of the organisms. They propose three primary adaptive hypotheses for the drivers of fast glycolytic (FG) muscle fiber proportions: i) foraging mode behavior (FM; e.g., sit-and-wait vs. active foraging vs. mixed); ii) predator escape behavior (PE; e.g., active flight vs. crypsis vs. mixed); and iii) a combined hypothesis of foraging mode and predator escape (FMPE) that assigns a unique regime to every combination of FM and PE represented in the data set. For each hypothesis, they reconstruct a likely phylogenetic history of these behavioral modes on the phylogeny by conducting ancestral state reconstructions (Fig. 3). After fitting the multioptimum OU models to the muscle fiber data, they find strong support for the PE hypothesis, which is 13.0 Akaike's Information Criterion (AICc) units better than the next closest model (FMPE). Such a finding appears quite reasonable under the "Life-Dinner Principle" (Dawkins and Krebs 1979), which suggests that escaping a predator may have a far more direct effect on fitness than obtaining a food item (Scales et al. 2009).

However, AIC provides only relative support for a model given a set of alternatives (see Pennell et al. 2015 for more on this point in the context of comparative methods). An examination of the particular configuration of shifts in the three hypotheses may give pause to researchers familiar with squamates. For example, some may want to quibble with the suggestion that the "sit-and-wait" foraging behavior of *Phrynosoma* species, which are often ant-eating specialists that leisurely lap up passing insects, should be grouped with the "sit-and-wait" tactics of species such as *Gambelia wislizenii*, a voracious carnivore that frequently subdues and consumes other lizards close to their own size. Looking at the reconstructions, it is also apparent that the PE hypothesis is the simplest model that allows a shift on the branch leading to *Phrynosoma*, a group that any herpetologist would identify as "weird" for a multitude of reasons (indeed, these are the eyeball-socket-blood-squirters alluded to in the introduction). The question then arises: is the signal in the data set for the PE hypothesis driven entirely by the singular evolution of different muscle fiber composition in *Phrynosoma* lizards? If so, then any number of causal factors that differ between *Phrynosoma* and other lizards could be equally as likely as PE—including FM with a slight reclassification of character states! We want to emphasize that we are not criticizing any of the particular choices the researchers involved in this study made. Rather, we argue that such quandaries are the inexorable result

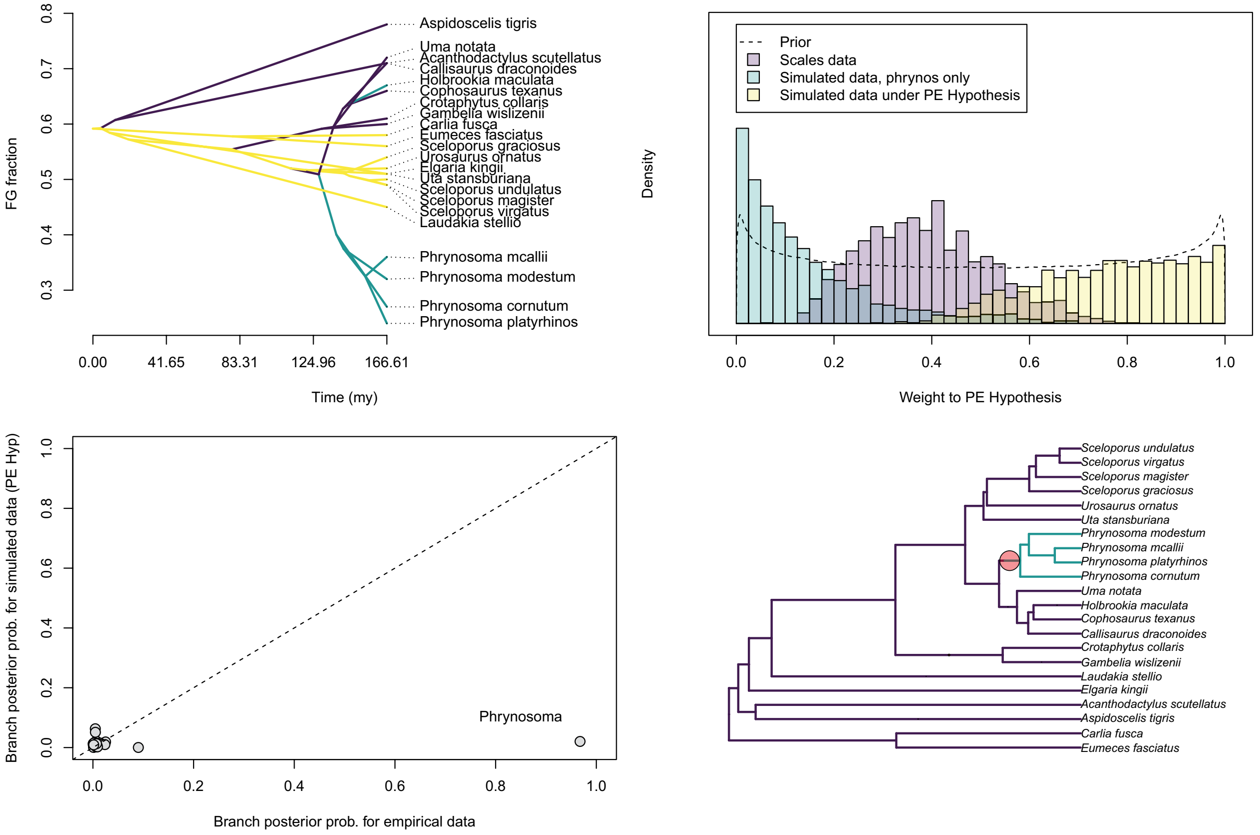

FIGURE 3. A reanalysis of the Scales et al. (2009) data set of fast glycolytic muscle fiber fraction across 22 squamate lizards. A) A traitgram depicting the distribution of the data and the reconstructed regimes for the best-fitting Predator-Escape (PE) hypothesis (blue = cryptic, yellow = active flight, purple = mixed). B) Posterior distributions of weights estimated for the PE hypothesis when mixed with a RJMCMC analysis for the original empirical data (purple), data simulated under the best-fitting estimated parameters for a *Phrynosoma*-only shift model (blue), and a data set simulated under the best-fitting estimated parameters for the full PE model (yellow). Notice that the empirical data set has intermediate weights. C) Posterior probabilities for all branches of the phylogeny estimated for the original empirical data (X-axis) and the simulated data set under the PE hypothesis (dashed line is the 1 to 1 line). D) We estimate a high posterior probability on a regime shift in the genus *Phrynosoma* from the empirical data only (red circle), indicating that while the PE hypothesis explains some patterns in the data, it does not fully explain the shift present in the behaviorally and ecologically unique genus *Phrynosoma*.

whenever the primary signal in the data is due to a singular historical event.

To explore the impact of the distinctiveness of simply being a *Phrynosoma* lizard, we developed a novel Bayesian model by building on the R package *bayou* (Uyeda and Harmon 2014). To do so, we consider the macroevolutionary optimum of a particular species to be a weighted average of past regimes, as is typical in all OU models with discrete shifts in regimes (Butler and King 2004; Beaulieu et al. 2012), but in our case, this weighted average is itself a weighted average of two differing configurations of the locations of adaptive shifts (often referred to as "regime paintings"). One configuration assumes that shifts in the optima have occurred where a discrete character, hypothesized to shape the evolutionary dynamics of the continuous character, is reconstructed to have shifted. The other configuration is estimated directly from the data using *bayou*'s reversible-jump Markov Chain Monte Carlo (RJMCMC) algorithm.

$$E[Y_i] = w(\Psi_{PE}(\alpha)\theta_{PE}) + (1-w)(\Psi_{RJ}(\alpha)\theta_{RJ}). \quad (5)$$

This equation describes the expected value of a trait for species i, Y_i as a weighted average between the expected trait value under the PE hypothesis and the expected trait value under the reversible-jump estimate of regime shift configurations. The vectors θ_{PE} and θ_{RJ} are the values of the trait optima for the N_{PE} and N_{RJ} adaptive regimes, while Ψ_{PE} and Ψ_{RJ} correspond to the standard OU weight matrices that average over the history of adaptive regimes experienced by species i over the course of their evolution, with older regimes being discounted proportional to the OU parameter α (for a full description of how these weight matrices are derived, see Hansen 1997; Butler and King 2004).

In our model, the regime painting for our *a priori* hypothesis Ψ_{PE} is fixed, while we estimate the parameters the configuration of shifts for the reversible-jump component, Ψ_{RJ}, as well as the values for the optima θ_{PE} and θ_{RJ}; and standard parameters for the OU model such as α and σ^2 which are assumed constant across the phylogeny. We also estimate the weight parameter w, which determines the degree of support for the PE hypothesis against the reversible-jump regime painting. We place a truncated Poisson prior on the number of shifts for the reversible-jump analysis to be quite low, with a mean of $\lambda = 0.5$ and a maximum of $\lambda = 10$ (meaning that we are placing a prior expectation of 0.5 shifts on the tree). Furthermore, we place a symmetric β-distributed prior on the w parameter with shape parameters of (0.8, 0.8). Additional details on the model-fitting can be found in the Supplementary material available on Dryad.

We then fit this model to three different data sets: i) the original Scales data; ii) data simulated using the Maximum Likelihood estimates for the parameters of the PE model fitted to the Scales data set; and iii) data simulated under the Maximum Likelihood estimates for a "*Phrynosoma*-only" model in which a single shift occurs leading to the genus *Phrynosoma*. We then compared the posterior distribution of the weight parameter w to evaluate the weight of evidence for each hypothesis in each data set.

We find that our approach places intermediate weight on the PE hypothesis for the original Scales data set. When we simulated data under the PE hypothesis, the estimated weight given to the PE hypothesis was likewise high (Fig. 3B). When data were simulated under the *Phrynosoma*-only hypothesis, the weight given to the PE hypothesis was low, as predicted (Fig. 3B). Furthermore, the RJ portion of the model fit to the Scales data set recovers only a single highly supported shift on the stem branch of the *Phrynosoma* lizards (Fig. 3C,D). This suggests that the PE hypothesis has statistically supported explanatory power as its estimated weight is well bounded away from 0. But it does not explain everything. In particular, the PE hypothesis fails to fully explain the shift leading to the *Phrynosoma* lizards (Fig. 3C,D), which are more extreme than they should be considering the other taxa in their regime (there is only one, *Holbrookia maculata*, which does not show such an extreme shift). In summary, the signal for an association between muscle fiber composition and predation escape behavior is generated in part, but not completely, by variation that is specific to the genus *Phrynosoma*. Therefore, the conclusions of Scales et al. (2009) hold up, even when accommodating evolutionary regime shifts unrelated to the factors being considered, although the weight of evidence appears to be larger than it actually is. This more subtle view of muscle fiber evolution conforms quite well with our biological intuition—variation in PE behavior is a good explanation for observed patterns of muscle fiber divergence, but *Phrynosoma* are a unique group with other factors likely influencing their trait evolution beyond PE.

We can conduct the same analysis where we test not the PE hypothesis, but the *Phrynosoma*-only hypothesis against the reversible-jump hypotheses (Fig. 4). In this case, we recover high weights for the *Phrynosoma*-only hypothesis regardless if the model is fit to the Scales data set, or to data simulated under either the *Phrynosoma*-only hypothesis or the PE hypothesis. This is because accounting for the *Phrynosoma* shift is the primary feature of all three data sets (though weights are somewhat higher for data simulated under the *Phrynosoma*-only hypothesis than others). It may appear unsatisfying that such high weights are recovered for the *a priori* hypothesis when a singular event, which is easily reconstructed by the RJMCMC, explains the distribution of the data just as well.

However, the analysis favors the *Phrynosoma*-only hypothesis simply because of the vague priors placed on the number and location of shifts in the reversible-jump analysis. Guessing correctly which of the 42 branches on the phylogeny has a single shift with our hypothesis is rewarded by the analysis (we will return to this issue in Case Study III). In the original Scales data set, there are weakly supported shifts in the clades leading to the sister group of *Phrynosoma* lizards, and the branch leading to *Acanthodactylus scutellatus* and *Aspidoscelis tigris*. Finally, we can combine all three hypothesis simultaneously by placing a Dirichlet prior on the vector $w = [w_{\mathrm{RJ}}, w_{\mathrm{PE}}, w_{Phrynosoma}]$. Doing so recovers strongest support for the *Phrynosoma*-only model, intermediate support for the PE hypothesis, and very little weight on the reversible-jump hypothesis, which has no strongly supported shifts (Fig. 5).

By combining phylogenetic natural history approaches with our *a priori* hypotheses, we show that we can account for singular evolutionary events that are not well-accounted for by our generating model. In the case of the PE hypothesis, we show that it does indeed have explanatory power beyond simply explaining a singular shift in *Phrynosoma* and support the original authors' conclusions. However, the intermediate result likely only occurs because the PE hypothesis places *Phrynosoma* in the same regime as *Holbrookia maculata*, which does not share the extreme shift that is found in *Phrynosoma*. Were this not the case (as in our fitting of the *Phrynosoma*-only hypothesis), it would still require visual inspection of the phylogenetic distribution of traits under the hypothesis in question to determine that a singular evolutionary event is driving support for a particular model. As discussed above, given a large enough tree such *a priori* hypotheses are likely to be strongly supported; if you can predict which one branch out of many will contain a shift then you may be on to something. But given the dangers of ascertainment bias and our biological intuition, we find this interpretation unsatisfying (Maddison and FitzJohn 2015). As with Case Study I, these scenarios ultimately reduce to whether or not coincident unreplicated events are evidence of a causal link between traits (Maddison and FitzJohn 2015), a problem we will address in Case Study III.

Nevertheless, we show the value in combining a hypothesis-testing framework with a natural history approach to identifying patterns of evolution. We show here that allowing for unaccounted shifts can provide a stronger test and more nuanced conclusions regarding the support for a particular predictor driving trait evolution across a phylogeny. Furthermore,

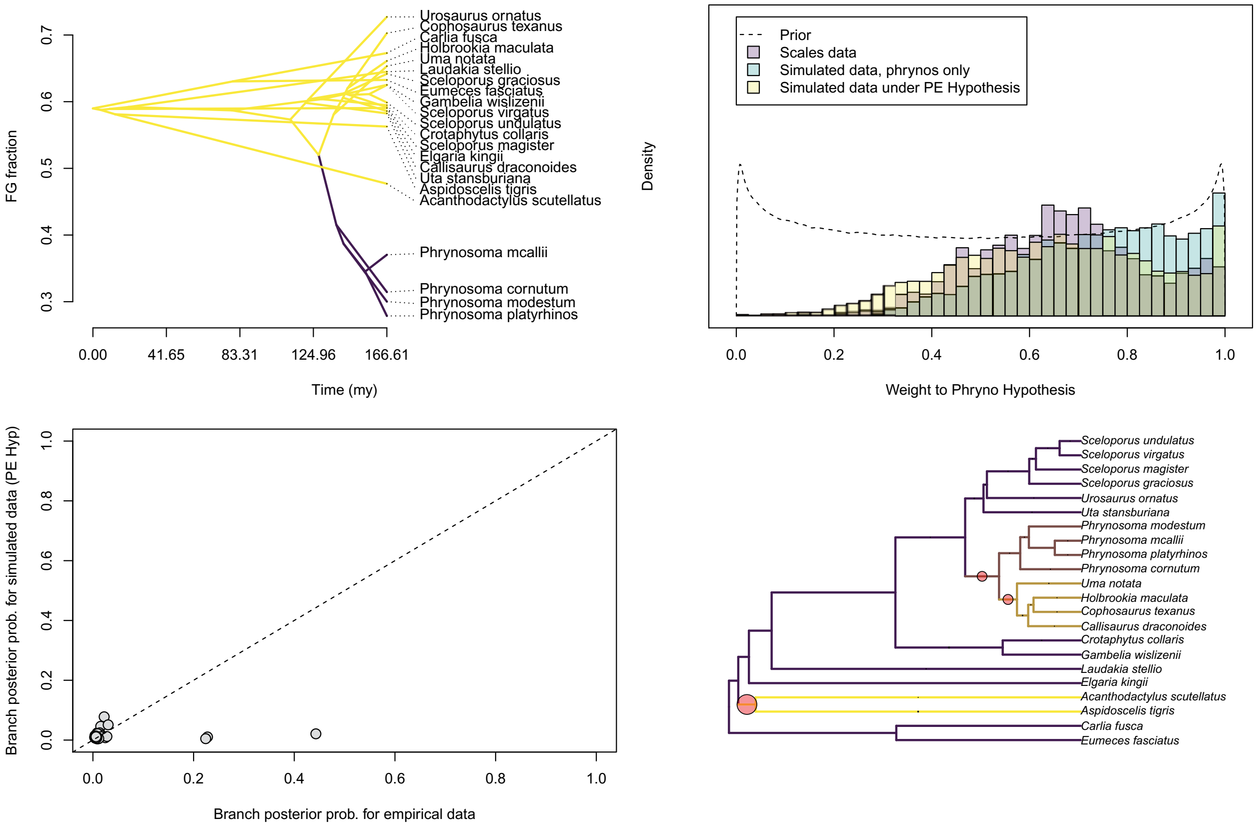

FIGURE 4. A reanalysis of the Scales et al. (2009) data set of fast glycolytic muscle fiber fraction across 22 squamate lizards against the *Phyrnosoma*-only hypothesis. A) A traitgram depicting the distribution of simulated data under the *Phyrnosoma*-only hypothesis (yellow = squamates, purple = *Phyrnosoma*). B) Posterior distributions of weights estimated for the *Phyrnosoma*-only hypothesis when mixed with a RJMCMC analysis for the original empirical data (purple), data simulated under the best-fitting estimated parameters for a *Phrynosoma*-only shift model (blue), and a data set simulated under the best-fitting estimated parameters for the full PE model (yellow). All analysis recover high weights. C) Posterior probabilities for all branches of the phylogeny estimated for the original empirical data (X-axis) and the simulated data set under the PE hypothesis (dotted line is the 1 to 1 line). D) Modest support for two additional shifts are recovered for the empirical data only (red circles).

predictors which provide additional explanatory power (if e.g., regimes are convergent or if predictors vary continuously) will be even more favored over natural history models. Thus, our framework certainly does not automatically reward more complex, freely estimated models. Rather, the great uncertainty in possible models is incorporated as a prior on the arrangement of shifts and is limited in explanatory power, something that researcher-driven biological hypotheses are much more capable of accomplishing.

CASE STUDY III: DARWIN'S SCENARIO AND UNREPLICATED BURSTS

We now turn to a case where both the explanatory variable and the focal trait are discrete characters. Detecting a signal of evolutionary covariation is more difficult in discrete characters, but examining this situation isolates the recurring problem we have identified in the two previous Case Studies—whether or not coincident unreplicated events are evidence of causal links between traits. As we mention above, Maddison and FitzJohn (2015) recently demonstrated that commonly used methods return significant correlations all the time—and in scenarios that seem to defy our statistical intuition. For example, Pagel's (1994) correlation test would find the phylogenetic codistribution of milk production and middle ear bones highly statistically significant even though they both are a defining characteristic of mammals, an inference so obviously dubious that even Darwin (1872) warned against it. This seems to be a clear case of phylogenetic pseudoreplication (Read and Nee 1995; Maddison and FitzJohn 2015). More broadly, Maddison and FitzJohn describe the goal of correlation tests as finding the "weak" conclusion that "the two variables of interest appear to be part of the same adaptive/functional network, causally linked either directly, or indirectly through other variables" (p. 128). They assert that with our current approaches, we cannot even clear this (arguably low) bar. Here, we delve into this idea a bit deeper. What constitutes good evidence of such a relationship and why precisely do phylogenetic tests for correlations provide such apparently unreasonable results?

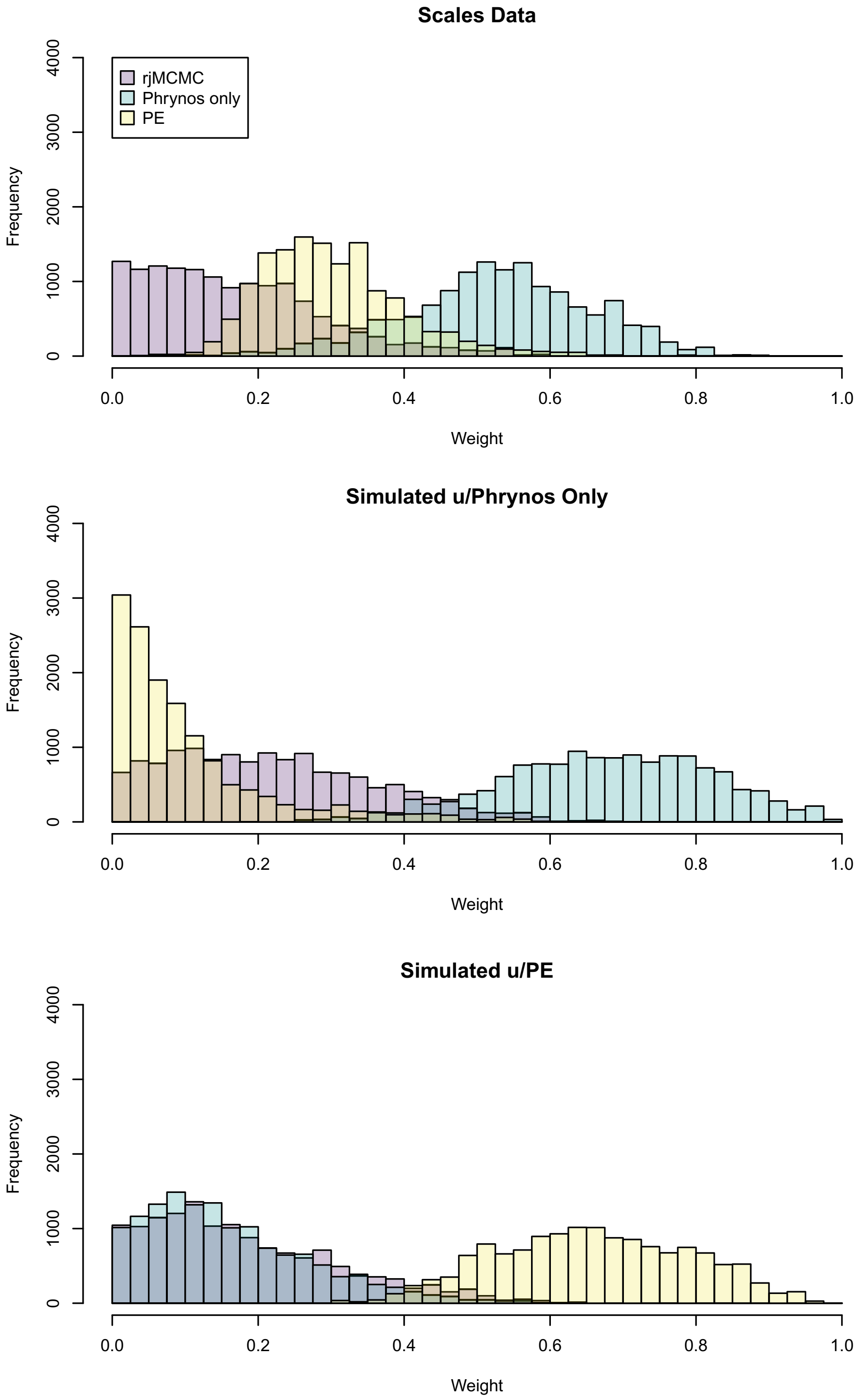

FIGURE 5. A reanalysis of the Scales et al. (2009) data set of fast glycolytic muscle fiber fraction across 22 squamate lizards against with both the *Phrynosoma*-only hypothesis and the PE hypotheses. Weights are depicted for each of the three data sets A) the original Scales data set B) A data set simulated under the *Phrynosoma*-only model C) A data set simulated under the PE hypothesis. In B and C, the correct model receives highest support with neither of the alternatives being well-supported. In the original Scales data set, the *Phrynosoma*-only hypothesis receives the most weight (indicating a singular shift best explains the patterns observed in the data), while an intermediate weight is given to the PE hypothesis (which explains a good amount of the remaining variation). In no analysis did the reversible-jump portion recover support for any additional shifts.

Maddison and FitzJohn highlight two hypothetical situations, that they refer to as "Darwin's scenario" and an "unreplicated burst." They argue that these scenarios provide little evidence for an adaptive/functional relationship between two traits because the patterns of codistribution only reflect singular evolutionary events (Fig. 1). In Darwin's scenario, two traits are coextensive on the phylogeny, meaning that in every lineage where one trait is in the derived character state, the other trait is as well. As an example, consider the aforementioned phylogenetic distribution of middle ear bones and milk production in animals; all mammals (and only mammals) have middle ear bones and produce milk. These traits (depending on how they are defined) have only appeared once on the tree of life and both occurred on the same branch (the stem branch of mammals). The unreplicated burst scenario is identical to Darwin's scenario except that rather than a single transition occurring in both traits, there is a single transition in the state of one trait (e.g., the gain of middle ear bones) and a sudden shift in the transition *rates* in another trait (e.g., the rates by which external testes are gained and lost across mammals). Note that these scenarios do not differ qualitatively from Felsenstein's worst-case scenario nor the *Phrynosoma*-only model scenario from Case Studies I and II (Fig. 1). In all three scenarios, something novel and interesting happened on a single branch and the distribution of traits at the tips of the phylogeny reflects this.

In their article, Maddison and FitzJohn (2015) simulated comparative data and reported a preponderance of significant results using Pagel's correlation test (1994) and Maddison's (1990) concentrated changes test. In order to hone our intuition of the problems they present, we dig a bit deeper and investigate the mathematical reason that Pagel's discrete correlation test (1994) returns a significant result in Darwin's scenario. [We should note here that Brookfield (1993) conducted a similar analysis that was more-or-less completely overlooked.] To make the problem tractable, we assume that the traits were selected for study without first looking at their phylogenetic distribution, a condition that we (as well as Maddison and FitzJohn 2015) suspect is rarely met in practice (more on this below).

Again, under Darwin's scenario, there is a single concurrent origin of two traits leading to perfect codistribution across the phylogeny for all taxa stemming from branch L (a condition we define mathematically as event A). Under the independent model, both traits X and Y have to switch from 0 to 1 on the same branch L once. For these traits, we can make the assumption that the likelihood of evolving these traits at all is quite small. In other words, replaying the tape of life, under Markovian assumptions, will likely lead to many worlds where milk and middle ear bones don't exist at all. However, we do not study traits that don't exist. Thus, for traits such as these, we can expect that there is likely to be only one origin of the trait on the phylogeny. Therefore, the probability of the independent model denoted as $P(M_{\text{ind}}|A)$ given the assumptions above is

$$\begin{aligned} P(M_{\text{ind}}|A) &= P(N_x(t)|N_x(T)\geq 1)P(N_y(t)|N_y(T)\geq 1) \\ &= (t/T)^2, \end{aligned} \tag{6}$$

where t is the branch length of branch L containing both shifts (Karlin and Taylor 1981) and N_x and N_y are the stochastic processes that denote the number of shifts of trait X and Y at time t, respectively (see Supplementary Material available on Dryad for exact derivation of the probability the independent model). Since this is the only way Darwin's scenario can occur on a branch for M_{ind}, the probability in Eq. (6) is equivalent to the likelihood value $L(M_{\text{ind}})$ evaluated at the maximum likelihood estimate of M_{ind}.

In contrast, for the completely dependent model M_{dep}, it is enough to follow what happens in a single trait since the second will just simply change along. The probability of the dependent model $P(M_{\text{dep}}|A)$ under Darwin's scenario is

$$P(M_{\text{dep}}|A) = (t/T) \tag{7}$$

(see full derivation of this probability in the Supplementary Material available on Dryad). The likelihood value at the maximum likelihood estimate of this model turns out to be just the probability from Eq. (7), that is $L(M_{\text{dep}}) = t/T$. Therefore, the test statistic D used in the likelihood ratio test for Pagel's discrete correlation test comparing models M_{ind} and M_{dep} is simply proportional to the ratio of the length of the branch where the shift occurred to the total length of the tree:

$$\begin{aligned} D &= 2(lnL(M_{\text{dep}}) - lnL(M_{\text{ind}})) \\ &= 2(ln(t) - ln(T)) - 4(ln(t) - ln(T)) \\ &= 2(ln(T) - ln(t)). \end{aligned} \tag{8}$$

In other words, the results of the analysis are predetermined. Under Darwin's scenario, including additional taxa in the analysis will increase the support for the dependent model simply as a consequence of increasing the total length of the tree (i.e., because T increases and the difference between $ln(T)$ and $ln(t)$ will get larger as long as the additional sampled taxa do not break Darwin's scenario).

The assumptions used to derive this result differ very slightly from those used in available software; however, we can use simulation to test the validity of our result and to demonstrate that this is the mathematical reason that Pagel's test returns a significant result. Using the R package *diversitree* (FitzJohn 2012), we simulated a set of 20 taxon trees where both traits underwent a irreversible transition on a single, randomly chosen, internal branch. We then fit a Pagel model with constrained (M_{dep}) and unconstrained (M_{ind}) transition rates. We also constrained the root state in both traits to 0, rates of losses of both the traits to 0, and gain rates in the dependent model following the gain of the other trait

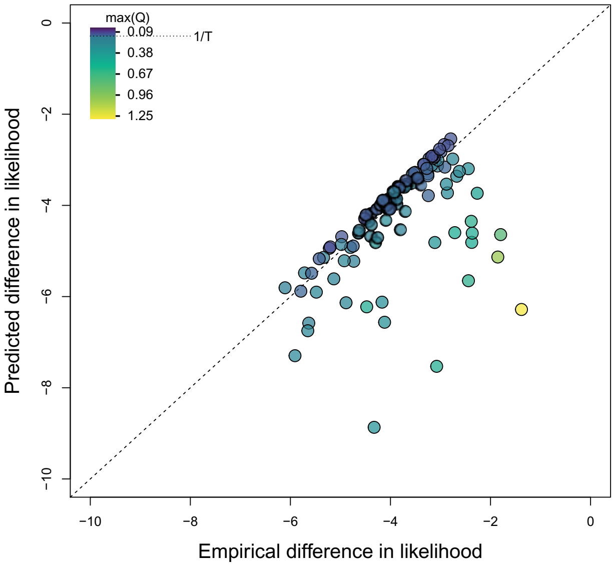

FIGURE 6. Darwin's scenario–the singular origin of two coextensive traits on the phylogeny–represents a boundary case to finding the correlation between discrete characters. Pagel's correlation test for Darwin's scenario can essentially be reduced to the difference in probability between choosing the same branch twice vs. choosing the branch only once. We demonstrate that here, showing our predicted differences in log likelihood between the independent and dependent trait models (y-axis) against the empirical estimates of the difference in log likelihood between models for simulated Darwin's scenarios on different phylogenies. Dotted line indicates equality. Points falling off the line represent slight violations of the assumptions we used to derive our prediction. Particularly, we assume that the rates of gain of the traits are so low that only one shift is ever observed. The shading of the points indicates cases where this assumption is violated, as outlying points with max(Q) values much greater than $1/T$ (where only 1 shift is expected) are much more likely to fall off the predicted line.

to be extremely high. Plotting the empirically estimated differences in the MLEs against the predictions making the simplifying assumptions above reveals a strong modal correlation between them (Fig. 6). Differences likely reflect the fact that we have not explicitly made the assumption that $P(N_x(t)=1)=P(N_y(t)=1)\approx 1$ when we fit the model with *diversitree*. Furthermore, we compare here only fully dependent and independent models. This can be seen when calculating the probability of one switch in each trait $P(N_x(t)=1,N(t)_y=1)$. In the fully dependent case that simply becomes $P(N_x(t)=1)$, in the independent case it becomes $P(N_x(t)=1)P(N_y(t)=1)$ but in the correlated case it becomes $P(N_y(t)=1|N_x(t)=1)P(N(t)_x=1)\neq 1$ affecting the likelihood ratio test based on estimations of the correlation (see Supplementary Material available on Dryad). However, such intermediate cases will only introduce slight differences and may not be distinguishable from the fully dependent case under Darwin's Scenario (though they will be important in more intermediate cases, see Supplementary Material available on Dryad).

Maddison and FitzJohn (2015) hinted that the coincident occurrence of single events could be a way of measuring the evidence for a correlation, but did not work out the details as we have done here. The key to understanding the above results is to recall Gould and Eldredge's famous dictum (1977) that "stasis is data." The remarkable coincidence is not just that the two characters happened to evolve on the same branch but that they were never subsequently gained or lost throughout the rest of the tree. For even a modestly sized tree, this coincidence is so unlikely that the alternative hypothesis of correlated evolution is preferred over the null. It is therefore not completely unreasonable that Pagel's test tells us that these traits have evolved in an entirely correlated fashion.

However, one key consideration should make us suspect of this line of reasoning. As Maddison and FitzJohn (2015) point out, the traits we use in comparative analyses are not chosen independently with respect to their phylogenetic distribution (as we assumed in our analysis). Rather, researchers' prior ideas about how traits map unto trees likely inform which traits they choose to test for correlated evolution. For example, it is common practice among systematists to search for defining and diagnostic characteristics for named clades; these traits are of especial interest and are likely the same sorts of traits that are researchers might include in comparative analysis, thereby greatly increasing the likelihood of finding traits with independent, unrelated origins that align with Darwin's scenario. We agree with Maddison and FitzJohn (2015) that this type of ascertainment bias is likely prevalent in empirical studies, even if it is usually more subtle than testing for a correlation between milk and middle ear bones. However, the presence of ascertainment bias does not mean it is not worth attempting to discover the source of the signal. Understanding the exact mathematical reasons why Pagel's test infers a significant correlation in a given case provides a clear boundary condition that can help develop quantitative corrections for ascertainment bias. Furthermore, the issues of ascertainment bias are likely to rapidly dissipate as we move away from the boundary case of Darwin's scenario. As a result, extending our analytical approach to more complicated scenarios will likely provide an even more meaningful estimate of the weight of evidence supporting a hypothesis of correlation.

The Structure of a Solution

We have shown in the three Case Studies that many PCMs, including those that form the bedrock of our field, are susceptible to being misled by singular evolutionary events. This fundamental problem has sown doubts about the suitability and reliability of many methods in comparative biology (e.g., Losos 2011), even if it was not obvious that these issues were connected. But again, the fact that apparently different issues share a common root makes us hopeful that there can be a common solution.

As we illustrate through our Case Studies, we think that accounting for the possibility of idiosyncratic evolutionary events will be an essential step towards such a solution. However, we will need to think hard about how best to model such events. In Case Study II, we present one solution to the problem that involves explicitly accounting for the possibility of unaccounted adaptive shifts using Bayesian Mixture modeling. We believe this approach has a great deal of promise as it provides simultaneous identification of biologically interesting shifts and the explanatory power of a particular hypothesis.

However, we do not claim that such an approach is the only solution or that it solves the problem completely. Indeed, we find that in all three Case Studies, the uniting philosophy is to consider models that account for background shifts in evolutionary regime, rather than strict adherence to a particular methodology. For example, we highlighted in the introduction that we think HSMs (following Beaulieu et al. 2013; Beaulieu and O'Meara 2016) are a potentially powerful, and widely applicable solution, even though we did not consider these in detail here.

And there are still other potential solutions which we have not even mentioned yet. In our own work (Uyeda et al. 2017), we have used a strategy similar to the Bayesian Mixture Modeling presented in Case Study II, but instead of modeling the trait dynamics as a joint function of our hypothesized factors and background changes (represented by the RJMCMC component), we did the analyses in a two-step process: first, we used *bayou* (Uyeda and Harmon 2014) to locate shifts points on the phylogeny, then used Bayes Factors to determine if predictors could "explain away" shifts found through exploratory analyses. For PGLS and other linear modeling approaches, modeling the errors using fat-tailed distributions (Blomberg et al. 2012; Landis et al. 2012; Elliot and Mooers 2014; Duchen et al. 2017) may mitigate the impact of singular evolutionary events on the estimation of the slope (also see Slater and Pennell 2013, for an alternative approach using robust regression). Furthermore, we also think that rigorous examination of goodness-of-fit and model adequacy following any comparative analysis is critical for finding unforeseen singular events driving signal in the data set (Garland et al. 1992; Boettiger et al. 2012; Slater and Pennell 2013; Pennell et al. 2015). Which of these solutions (including those that were included in our Case Studies and those that were not) will be the most profitable to pursue will probably differ depending on the question, data set, and application—we anticipate that there will not be a one-size-fits-all solution—but we do think that any compelling solution will involve a unification of phylogenetic natural history and hypothesis-testing approaches.

But we want to take this a step further. While it is useful to account for phylogenetic events in our statistical models, a greater goal of comparative biology should be explain why these events exist in the first place. We return to Maddison and FitzJohn's (2015) "weak" goal of finding whether or not "two variables of interest appear to be part of the same adaptive/functional network, causally linked either directly, or indirectly through other variables." We ultimately disagree with them that this constitutes a weak conclusion; the challenges of making these inferences from any comparative data set are significant. Furthermore, we find the often repeated axiom "correlation does not mean causation" to be unhelpful. While it is accurate in the strict sense, some patterns of correlations will *at least be consistent* with a given model of causation whereas others will not be. And it is clear from reading the macroevolutionary literature, biologists do not shy away from forming causal statements from correlative data regardless. While some might reasonably wish us to simply highlight these claims as violating statistical principles, we think it is worthwhile to take seriously the question: "What would it take to infer causation from comparative data?" And even if we are to conclude that all the evidence for a hypothesized causal relationship stems from one or a few evolutionary events, is this finding biologically meaningful?

Phylogenies are Graphical Models of Causation

One way to gain a foothold on the problem of causation is to build, communicate, and analyze PCMs in a graphical modeling framework—a perspective that has recently been advocated by Höhna et al. (2014, 2016). Graphical models that depict hypothesized causal links between variables make explicit key underlying assumptions that may otherwise remain obscured; indeed, the precise assumptions of PCMs were hotly debated in the early days of their development (McNab 1988; Harvey et al. 1995; Westoby et al. 1995a,b; Nee et al. 1996; Westoby 2007) and remain poorly understood to this day (Hansen and Orzack 2005; Hansen and Bartoszek 2012). As examples of how using graphical models force us to be more clear in our reasoning, consider the graphs in Figure 7. We depict three

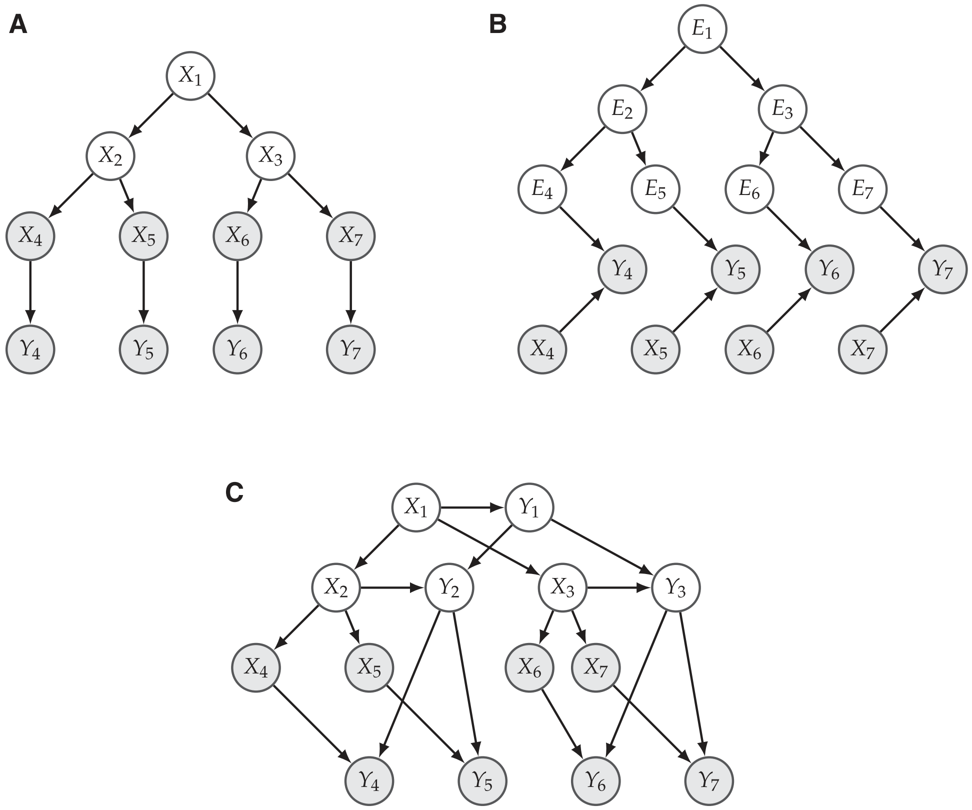

FIGURE 7. Graphical models of alternative causal relationships between a predictor (X) and a trait of interest (Y). Note that each node has independent, uncorrelated error as an input, but these have not been shown for clarity. A) X follows the phylogeny with observed states (gray) and unobserved ancestral states (white) and is a cause of trait Y. However, the phylogeny and pattern of evolution of X are irrelevant, and this graph can be modeled with methods such as OLS regression. B) The trait Y has unobserved causes (E_i) that follow the phylogeny (gray) that can be modeled using, for example, BM. The trait X is a cause of Y. This graph can be modeled using methods such as PGLS and PIC. C) The trait Y evolves on the phylogeny and is affected by trait X all throughout its history. Thus, the history of both X and Y must be modeled (e.g., BM of X and Ornstein–Uhlenbeck for Y). This graph can be modeled using methods such as SLOUCH.

different models of causation that have phylogenetic effects that each require alternative methods of analysis to estimate the effect of trait X on trait Y. In our example, a four species phylogeny provides possible pathways for causal effects, but variables may have entirely nonphylogenetic causes or may be blocked from ancestral causes by observed measurements, rendering the phylogeny irrelevant (e.g., Fig. 7A). Edges connect nodes and indicate the direction of causality, where the nature of phylogenies allows us to assume that ancestors are causes of descendants, and not vice versa. This asymmetry results in a what is known as a probabilistic Bayesian Network (a type of directed acyclic graph, or DAG) that predicts a specific set of conditional probabilities among the data.

Depending on the Bayesian network structure, the appropriate method of analysis can range from a nonphylogenetic regression (Fig. 7A), to commonly used comparative methods such as PGLS (Fig. 7B), to methods that require modeling both the evolutionary history of interaction of both trait X and trait Y (Fig. 7C) (Hansen 1997; Butler and King 2004; Hansen et al. 2008; Revell 2010; Hansen and Bartoszek 2012). We emphasize that this implies that the use of phylogeny in interspecific comparisons is an *assumption* that depends on the precise question being asked and the hypothesized causal network. It is often assumed and asserted that PCMs are simply a more rigorous version of standard regression. This is simply not true.

In cases where phylogeny does matter, we must specify the generating model for unobserved states in our causal graphs. For example, it is common to assume a BM model for residual variation in PGLS or that ancestral states are reconstructed using stochastic character mapping in OU modeling of adaptation. However, BM and other continuous Gaussian or Markov processes are only a few of the many types of processes that may generate change on a phylogeny. We have shown that discontinuous processes and singular events are poorly handled in our current framework and lead to much confusion about what exactly, our statistical methods are allowing us to infer from comparative data. Such models can be similarly illustrated using graphical models (Fig. 8). By making our models explicit, we see that the phylogeny is best thought of as a pathway for past factors to causally influence the present-day distribution of observed states. These "singular-event" models are alternatives to the more continuous models

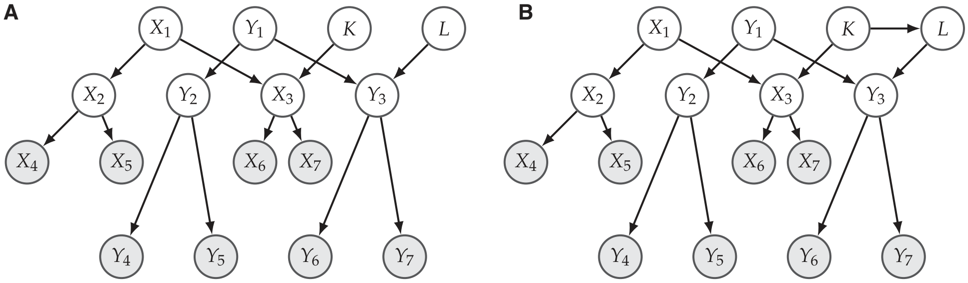

FIGURE 8. Graphical models of Darwin's Scenario between a predictor (X) and a trait of interest (Y). Note that each node has independent, uncorrelated error as an input, but these have not been shown for clarity. A) Singular event model. Here two independent factors cause a change on ancestral states X_3 and Y_3 (K and L respectively). However, they are independent events and coincidentally occur at the same point on the phylogeny. B) Similar to the previous model, but K and L are causally linked. Thus, whenever K occurs, it probabilistically causes L which causes a shift in Y. If only one event occurs however, this model is only distinguishable from graph (D) proportional to the probability that events K and L occur on the same branch (see Case Study III).

we typically examine. Furthermore, representing our models as graphs, we are poised to take advantage of the sophisticated approaches for causal reasoning (e.g., Pearl 1995, 2009; Sugihara et al. 2012; Shipley 2016) that subsume familiar tools such as path analysis, structural equation modeling, and graphical modeling into a more generally applicable structural theory of causation. These tools have been embraced by fields such as computer science, epidemiology and the social sciences, but largely ignored by comparative biologists [rare exceptions are the recent introductions of phylogenetic path analysis by Hardenberg and Gonzalez-Voyer (2013) and the application of causal models to make inferences from paleontological time-series by Reitan and Liow (2017)].

One clear case where such graphical modeling would improve inference are cases where considering phylogeny reverses the sign of the relationship between two variables. This is precisely what Nee et al. (1991) found looking at the relationship between body size and abundance in British birds; depending on how they aggregated the data (means of species, means of genera, means of tribes, etc.) the direction of correlation flipped back and forth. This reversal in the sign of the relationship between two variables X and Y when conditioning on a third Z is a general, and widely studied, statistical phenomenon known as "Simpson's paradox" (Blyth 1972). Nee et al. (1991, 1996) hold up their findings of the British bird study to be emblematic; in their view, the presence of Simpson's paradox in their data clearly implies that phylogeny is key to making sense of interspecific data.

However, as Pearl (2014) has convincingly demonstrated, Simpson's paradox is not really paradoxical at all when considered from the standpoint of Bayesian Networks. In fact, Pearl shows that the appropriate way to analyze the data depends crucially on what one assumes is causing what. To understand how causal inference resolves Simpson's Paradox, we now present a rather artificial, but nevertheless illustrative example (Pearl 2009). Consider three traits: body size (B), abundance (N) and migratory behavior (M) in birds. Given the Bayesian Networks presented in Figure 9, we have two possible hypotheses for the causal relationships between the traits. We further consider the possibility that we do not have adequate data on M, and thus only B and N are observed. Our goal is to estimate the causal effect of B on N. In Figure 9A, body size influences whether or not species become migratory, and both migratory status and body size influence species abundance (but in opposite directions). Furthermore, under this scenario, both body size and migratory status will have phylogenetic signal. We can evolve traits along the phylogeny depicted in Figure 9C and obtain a bivariate plot that looks like Figure 9D. Under the alternative Bayesian Network, migratory behavior still has a positive effect on species abundance, but also increases body size, which in turn causes decreases species abundance. These two causal structures are observationally equivalent—meaning that any distribution simulated under one can be replicated under the alternative causal structure. Therefore, both networks can produce data sets with phylogenetic signal in both body size and migratory behavior, and both can produce a data set with the distribution in Figure 9D (see Supplementary Material available on Dryad for additional details on generating Fig. 9).

How then should we analyze the data if we want to understand the effect of body size on species abundance? If we assume that body size influences migratory behavior, then increasing body size (e.g., if natural selection leads a species to become larger) will increase the probability of that species becoming migratory—and the two opposing effects will result in relatively little change in species abundance. Therefore, we should perform OLS regression to estimate the net causal effect of increasing body size. We also note that all the phylogenetic signal is coming from the evolution of body size, which becomes irrelevant once we observe body size, and thus we do not need to perform PGLS. In contrast, if migratory behavior causes changes in body size, then selecting for an increase in body size will not result in a lineage changing their migratory status at all. Therefore, we are assured that increasing body size will likewise always decrease species abundance. Consequently, we should perform PGLS to account for

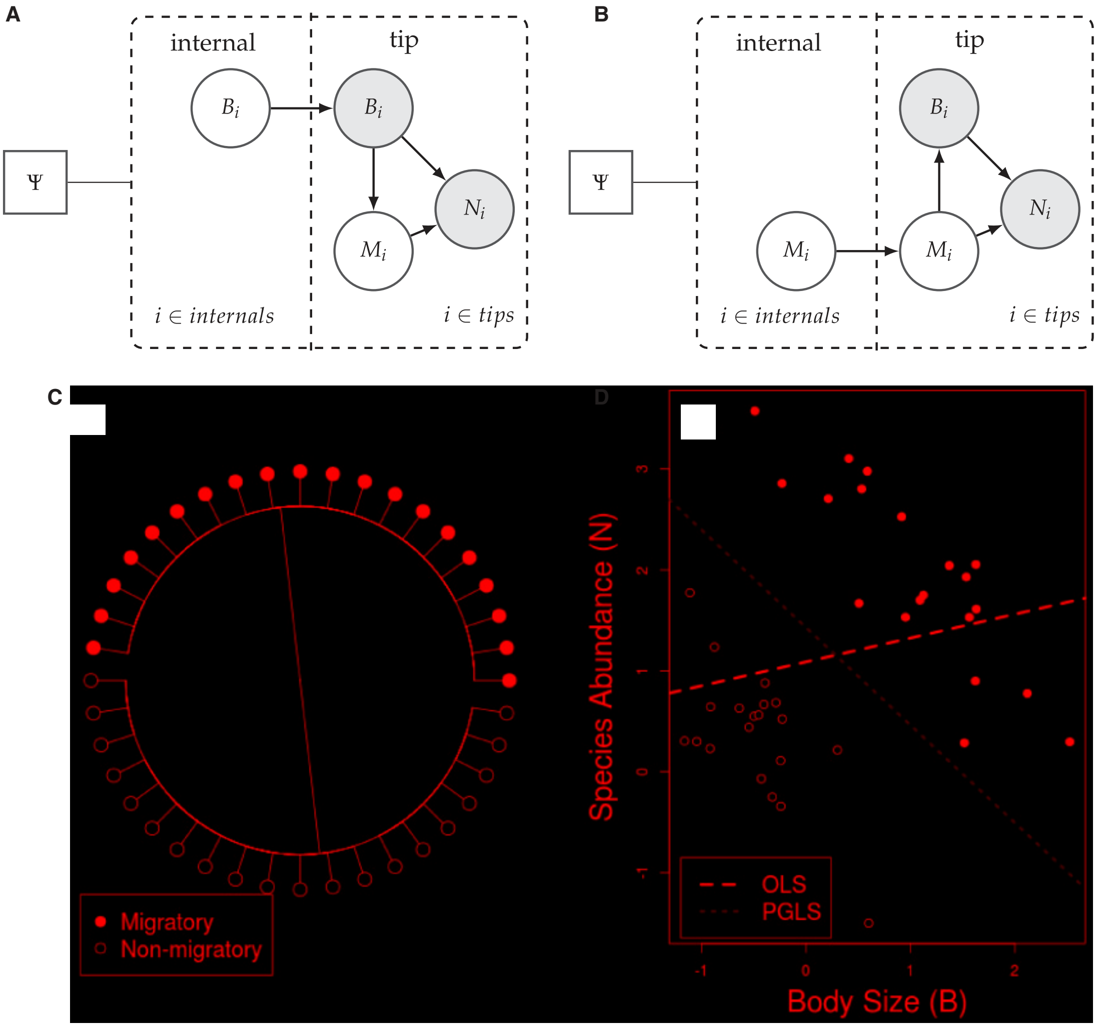

FIGURE 9. Simpson's paradox in PCMs. Panels (A) and (B) depict two alternative Bayesian Networks. In (A), body size is a cause of both species abundance and migratory behavior, and trait B evolves on the phylogeny Ψ. We represent evolution on the phylogeny using a "tree plate" (see Höhna et al. 2014; dashed box) where (unobserved) node states can influence tip states. In (B), body size still affects species abundance, but migratory behavior itself is a cause of both body size and species abundance, but the phylogenetic effect is present in migratory behavior (in this case, we simulated with a Brownian threshold model). (C) A phylogeny similar to that of Darwin's scenario used to simulate the data set (D), with migratory species (black) and nonmigratory (white) taxa. The data in (D) can be generated by either causal structure. However, to estimate the effect of *B* in both networks, one must use different analytical approaches. To estimate the net effect of *B* on *N* in network (A), the appropriate method of analysis is OLS regression (black line). This is because increasing body size will simultaneously decrease species abundance and increase migratory behavior, which itself increases abundance, leading to a net slight increase in abundance. However, under network (B) the correct method is PGLS (gray line) as increasing body size will have no effect on migratory behavior, and unaccounted phylogenetic residual error is present in the observed data. Here, increasing body size will only have a direct effect of decreasing species abundance, which is reflected in the estimate of the slope. The resolution of Simpson's paradox rests entirely on causal assumptions; which are immediately apparent from graphical models but difficult to express with standard mathematical formulae.

the phylogenetic signal in the residual variation imposed by (unobserved) migratory status.

By working through the logic of comparative analyses using graphical models we have come to essentially the same line of reasoning of Westoby et al. (1995a,b), who, in the early days of PCMs, challenged the growing consensus that phylogeny needed to be included in any interspecific comparison—a consensus which has only gotten stronger as the years passed by (also see McNab 2003 for a related critique). Westoby and colleagues were concerned that including phylogeny in interspecific comparisons necessarily favored some causal explanations over others. At the time, their critique was dismissed as innumerate hogwash (Harvey et al. 1995; Nee et al. 1996) and this evaluation has largely stuck. However, from our example of bird size and abundance, it is apparent that Westoby et al. were right all along: PCMs are a powerful tools for drawing inferences from interspecific data but they necessarily imply some types of causal structures and negate others. It is too much to ask of our methods to decide what questions we ought to ask. As Westoby et al. (1995a) put it: “No statistical procedure can substitute for thinking about alternative evolutionary scenarios and their plausibility” (p. 534).

Concluding Remarks: Are Our Models Valid Tests of Our Causal Hypotheses?

By explicitly including phylogeny into our graphical models of causation, we are forced to reckon with the scope of the inference problem and the ability of our data to be informative. While most of the statistical assumptions of methods are often well-known (e.g., for linear models, we assume that errors have equal variance and are normally distributed, etc.), Gelman and Hill (2006) argue that there is a more fundamental assumption—validity of data—that is almost always implicit and often overlooked:

> “Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied.” (Gelman and Hill 2006)

We believe that far less discussion in comparative methods has been focused on the issue of statistical validity of the data collected to the research questions being posed by a given study. This is in large part because comparative data and the phylogeny that underly it are largely beyond the control of the researcher, but careful consideration of the data is required to understand what research questions can be reasonably answered. For example, we find that most comparative research questions have a poorly defined scope of inference: it is unclear to what population a model or inference should generalize to. If we ask “are milk and middle ear bones correlated?”, we must also specify “in what organisms?”. Since no organisms other than mammals have the particular traits we define as “milk” and “middle ear bones,” we actually do not need statistics at all to determine whether these traits are correlated—we have sampled nearly the entire population relevant to the question! In nature, they are perfectly collinear. If we wish to expand our scope of inference to hypothetical organisms that evolve milk and/or middle ear bones we are free to do so. However, we have collected a very poor data sample for such a question. It is not the fault of the statistical method to demonstrate that a poorly designed experiment does not represent its scope of inference, rather it is our job as researchers and statisticians to ask whether or not such a relationship addresses our biological question and whether the sample of data collected is valid for the question being asked.

In this article, we have tried to synthesize a wide variety of statistical and philosophical concepts to lay out a roadmap for where we think comparative biology should go. We certainly do not have all the answers. Of the paths we have explored, there are many details that need to be worked out, and we fully anticipate that there are many alternative paths that we have not even considered. However, we argue that if we are going to make substantial progress in using phylogenetic data to test evolutionary hypotheses, we will need to reckon more seriously with the idiosyncratic nature of evolutionary history, and to more clearly articulate precisely what we want to test and whether our models and data are suitable for the task.

Code Availability

Data and code needed to reproduce all analyses in this manuscript are available at https://github.com/uyedaj/pnh-ms/.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.p8066hd.

Funding

This work was supported by a NSERC Discovery Grant to M.W.P.; NSF Grants to Luke Harmon (DEB-1208912 to J.C.U. and RZF and DBI-1661516 to J.C.U.).

Acknowledgments

We thank Luke Harmon, Dolph Schluter, Mark Westoby, Daniel Caetano, Eliot Miller, Ben Freeman, Florent Mazel, Joel McGlothlin, Martha Muñoz, Barbara Neto-Bradley, Francisco Henao Diaz, Mauro Sugawara, Emma Goldberg, Michael Landis, Simone Blomberg, and Nicholas Matzke for their critical feedback on these ideas and this manuscript. JCU would like to specially thank the insightful knowledge and teaching gleaned from conversations over the years with Thomas Hansen that inspired the bulk of this manuscript (though he holds no culpability for the contents and opinions therein).

References

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. USA 106:13410–13414.

Beaulieu J.M., O’Meara, B.C. 2014. Hidden Markov models for studying the evolution of binary morphological characters. In: Modern phylogenetic comparative methods and their application in evolutionary biology. Springer. p. 395–408.

Beaulieu, J.M., O’Meara, B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.

Beaulieu, J.M., Jhwueng, D.-C., Boettiger, C., O’Meara, B.C. 2012. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Beaulieu, J.M., O'Meara, B.C., Donoghue, M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst. Biol. 62:725–737.

Blomberg, S.P. 2017. Beyond Brownian motion and the Ornstein–Uhlenbeck process: stochastic diffusion models for the evolution of quantitative characters. bioRxiv, doi:10.1101/067363.

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

Blomberg, S.P., Lefevre, J.G., Wells, J.A., Waterhouse, M. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol. 61:382–391.

Blyth, C.R. 1972. On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. 67:364–366.

Boettiger, C., Coop, G., Ralph, P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Boucher, F.C., Démery, V., Conti, E., Harmon, L.J., Uyeda, J. 2017. A general model for estimating macroevolutionary landscapes. Syst. Biol. 67:304–319.

Brookfield, J. 1993. Haldane's rule is significant. Evolution 47:1885–1888.

Butler, M.A., King, A.A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Cheverud, J.M., Dow, M.M., Leutenegger, W. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution 39:1335–1351.

Clutton-Brock, T.H., Harvey, P.H. 1980. Primates, brains and ecology. J. Zool. 190:309–323.

Darwin, C.R. 1872. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 2nd ed. London: John Murray.

Dawkins, R., Krebs, J.R. 1979. Arms races between and within species. Proc. R. Soc. Lond. B 205:489–511.

Duchen, P., Leuenberger, C., Szilágyi, S.M., Harmon, L., Eastman, J., Schweizer, M., Wegmann, D. 2017. Inference of evolutionary jumps in large phylogenies using levy processes. Syst. Biol. 66:950–963.

Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L., Harmon, L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Elliot, M.G., Mooers, A.Ø. 2014. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC Evol. Biol. 14:226.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genetics 25:471.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Felsenstein, J. 2011. A comparative method for both discrete and continuous characters using the threshold model. Am. Nat. 179:145–156.

FitzJohn, R.G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3:1084–1092.

Freckleton, R.P., Harvey, P.H., Pagel, M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160:712–726.

Futuyma, D.J. 1998. Wherefore and whither the naturalist? Am. Nat. 151:1–6.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.

Garamszegi, L.Z. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice. Heidelberg: Springer.

Garland, T., Harvey, P.H., Ives, A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Gelman, A., Hill, J. 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Gould, S.J., Eldredge, N. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

Grafen, A. 1989. The phylogenetic regression. Philos.Trans. R. Soc. Lond. B 326:119–157.

Hadfield, J., Nakagawa, S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J. Evol. Biol. 23:494–508.

Hansen, T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T.F., Bartoszek, K. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst. Biol. 61:413–425.

Hansen, T.F., Orzack, S.H. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59:2063–2072.

Hansen, T.F., Pienaar, J., Orzack, S.H. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977.

Harvey, P.H., Read, A.F., Nee, S. 1995. Why ecologists need to be phylogenetically challenged. J. Ecol. 83:535–536.

Höhna, S., Heath, T.A., Boussau, B., Landis, M.J., Ronquist, F., Huelsenbeck, J.P. 2014. Probabilistic graphical model representation in phylogenetics. Syst. Biol. 63:753–771.

Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., Huelsenbeck, J.P., Ronquist, F. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

Housworth, E.A., Martins, E.P., Lynch, M. 2004. The phylogenetic mixed model. Am. Nat. 163:84–96.

Ingram, T., Mahler, D.L. 2013. Surface: detecting convergent evolution from comparative data by fitting Ornstein–Uhlenbeck models with stepwise Akaike information criterion. Methods Ecol. Evol. 4:416–425.

Jablonski, D. 2017. Approaches to macroevolution: 1. General concepts and origin of variation. Evol. Biol. 44:1–24.

Jones, K., Purvis, A. 1997. An optimum body size for mammals? Comparative evidence from bats. Funct. Ecol. 11:751–756.

Karlin, S., Taylor, H.E. 1981. A second course in stochastic processes. New York: Academic Press.

Khabbazian, M., Kriebel, R., Rohe, K., Ané, C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol. 7:811–824.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Landis, M.J., Schraiber, J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes. Proc. Natl. Acad. Sci. USA 114:13224–13229.

Landis, M.J., Schraiber, J.G., Liang, M. 2012. Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. Syst. Biol. 62:193–204.

Lewis, P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Losos, J.B. 2011. Seeing the forest for the trees: the limitations of phylogenies in comparative biology: (American Society of Naturalists Address). Am. Nat. 177:709–727.

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.

Mace, G.M., Harvey, P.H., Clutton-Brock, T. 1981. Brain size and ecology in small mammals. J. Zool. 193:333–354.

Maddison, W.P. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539–557.

Maddison, W.P., FitzJohn, R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 64:127–136.

Maddison, W.P., Midford, P.E., Otto, S.P. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.

McNab, B.K. 1988. Complications inherent in scaling the basal rate of metabolism in mammals. Q. Rev. Biol. 63:25–54.

McNab, B.K. 2003. Standard energetics of phyllostomid bats: the inadequacies of phylogenetic-contrast analyses. Comp. Biochem. Physiol. A Mol. & Integr. Physiol. 135:357–368.

Mooers, A.O., Heard, S.B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Nee, S., Read, A.F., Greenwood, J.J., Harvey, P.H. 1991. The relationship between abundance and body size in British birds. Nature 351:312–313.

Nee, S., Read, A.F., Harvey, P.H. 1996. Why phylogenies are necessary for comparative analysis. Phylogenies and the comparative method in animal behavior. Oxford: Oxford University Press. p. 399–411.

O'Meara, B.C. 2012. Evolutionary inferences from phylogenies: a review of methods. Annu. Rev. Ecol. Evol. Syst. 43:267–285.

O'Meara, B.C., Ané, C., Sanderson, M.J., Wainwright, P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37–45.

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

Pavlidis, P., Jensen, J.D., Stephan, W., Stamatakis, A. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol. Biol. Evol. 29:3237–3248.

Pearl, J. 1995. Causal diagrams for empirical research. Biometrika 82:669–688.

Pearl, J. 2009. Causality. Cambridge: Cambridge University Press.

Pearl, J. 2014. Comment: understanding Simpson's paradox. Am. Stat. 68:8–13.

Pennell, M.W., Harmon, L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. N. Y. Acad. Sci. 1289:90–105.

Pennell, M.W., FitzJohn, R.G., Cornwell, W.K., Harmon, L.J. 2015. Model adequacy and the macroevolution of angiosperm functional traits. Am. Nat. 186:E33–E50.

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53:711–723.

Price, T. 1997. Correlated evolution and independent contrasts. Philos. Trans. R. Soc. Lond. B 352:519–529.

Purvis, A., Rambaut, A. 1995. Comparative analysis by independent contrasts (caic): an apple macintosh application for analysing comparative data. Bioinformatics 11:247–251.

Rabosky, D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9:e89543.

Rabosky, D.L., Goldberg, E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol. 64:340–355.

Read, A.F., Nee, S. 1995. Inference from binary comparative data. J. Theor. Biol. 173:99–108.

Reitan, T., Liow, L.H. 2017. An unknown phanerozoic driver of brachiopod extinction rates unveiled by multivariate linear stochastic differential equations. Paleobiology 43:537–549.

Revell, L.J. 2010. Phylogenetic signal and linear regression on species data. Methods Ecol. Evol. 1:319–329.

Ridley, M. 1983. The explanation of organic diversity: the comparative method and adaptations for mating. Oxford and New York: Oxford University Press.

Rohlf, F.J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

Rohlf, F.J. 2006. A comment on phylogenetic correction. Evolution 60:1509–1515.

Scales, J.A., Butler, M.A. 2016. Adaptive evolution in locomotor performance: how selective pressures and functional relationships produce diversity. Evolution 70:48–61.

Scales, J.A., King, A.A., Butler, M.A. 2009. Running for your life or running for your dinner: what drives fiber-type evolution in lizard locomotor muscles? Am. Nat. 173:543–553.

Schraiber, J.G., Landis, M.J. 2015. Sensitivity of quantitative traits to mutational effects and number of loci. Theor. Popul. Biol. 102:85–93.

Shipley, B. 2016. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. Cambridge: Cambridge University Press.

Slater, G.J., Pennell, M.W. 2013. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst. Biol. 63:293–308.

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. USA 108:6187–6192.

Stearns, S.C. 1983. The influence of size and phylogeny on patterns of covariation among life-history traits in the mammals. Oikos 41:173–187.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., Munch, S. 2012. Detecting causality in complex ecosystems. Science 338:496–500.

Uyeda, J.C., Harmon, L.J. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Syst. Biol. 63:902–918.

Uyeda, J.C., Hansen, T.F., Arnold, S.J., Pienaar, J. 2011. The million-year wait for macroevolutionary bursts. Proc. Natl. Acad. Sci. USA 108:15908–15913.

Uyeda, J.C., Pennell, M.W., Miller, E.T., Maia, R., McClain, C.R. 2017. The evolution of energetic scaling across the vertebrate tree of life. Am. Nat. 190:185–199.

Vermeij, G.J. 2006. Historical contingency and the purported uniqueness of evolutionary innovations. Proc. Natl. Acad. Sci. USA 103:1804–1809.

von Hardenberg, A. Gonzalez-Voyer, A. 2013. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. Evolution 67:378–387.

Westoby, M. 2007. Generalization in functional plant ecology: the species-sampling problem, plant ecology strategy schemes, and phylogeny. In: Pugnaire F., Valladares F., editors. Functional plant ecology. 2nd. ed. Boca Raton: CRC Press.

Westoby, M., Leishman, M., Lord, J. 1995a. Further remarks on phylogenetic correction. J. Ecol. 83:727–729.

Westoby, M., Leishman, M.R., Lord, J.M. 1995b. On misinterpreting thephylogenetic correction'. J. Ecol. 83:531–534.

Zenil-Ferguson, R. Pennell, M.W. 2017. Digest: trait-dependent diversification and its alternatives. Evolution 71:1732–1734.