Molecular Biology and Evolution, 2025, **42**, 1–16
https://doi.org/10.1093/molbev/msaf032
Advance access publication 11 February 2025
Discoveries

MBE OXFORD

A Tale of Too Many Trees: A Conundrum for Phylogenetic Regression

Richard Adams [1,2,*] **Jenniffer Roa Lozano,**[1,2] **Mataya Duncan,**[1,2] **Jack Green,**[1,2] **Raquel Assis** [3,4,†] **Michael DeGiorgio** [3,†]

[1]Department of Entomology and Plant Pathology, University of Arkansas, Fayetteville, AR, USA
[2]Center for Agricultural Data Analytics, University of Arkansas, Fayetteville, AR, USA
[3]Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA
[4]Institute for Human Health and Disease Intervention, Florida Atlantic University, Boca Raton, FL, USA

***Corresponding author:** E-mail: adamsrh@uark.edu.
[†]These authors contributed equally to this work.
Associate editor: Bui Minh

Abstract

Just exactly which tree(s) should we assume when testing evolutionary hypotheses? This question has plagued comparative biologists for decades. Though all phylogenetic comparative methods require input trees, we seldom know with certainty whether even a perfectly estimated tree (if this is possible in practice) is appropriate for our studied traits. Yet, we also know that phylogenetic conflict is ubiquitous in modern comparative biology, and we are still learning about its dangers when testing evolutionary hypotheses. Here, we investigate the consequences of tree-trait mismatch for phylogenetic regression in the presence of gene tree–species tree conflict. Our simulation experiments reveal excessively high false positive rates for mismatched models with both small and large trees, simple and complex traits, and known and estimated phylogenies. In some cases, we find evidence of a directionality of error: assuming a species tree for traits that evolved according to a gene tree sometimes fares worse than the opposite. We also explored the impacts of tree choice using an expansive, cross-species gene expression dataset as an arguably "best-case" scenario in which one may have a better chance of matching tree with trait. Offering a potential path forward, we found promise in the application of a robust estimator as a potential, albeit imperfect, solution to some issues raised by tree mismatch. Collectively, our results emphasize the importance of careful study design for comparative methods, highlighting the need to fully appreciate the role of accurate and thoughtful phylogenetic modeling.

Keywords: comparative biology, continuous traits, Brownian motion, phylogeny

Introduction

It is a tale nearly as old as time: you measure a set of traits across a sample of organisms, and you seek to gain new insights by testing for statistical relationships between two or more of your studied traits. Both scientists and philosophers alike have strived to understand biology for centuries by employing this strategy. The diversity of traits that can be studied span any variable that can be reliability measured, from those at the molecular (e.g. cell size, cell morphology, and gene expression; Gu 2016; Dunn et al. 2018; Chen et al. 2023) up to the organismal (e.g. body size, head morphology, and behavior; Al-Kahtani et al. 2004; Ross et al. 2004; Kamilar and Cooper 2013) level; we are typically only limited by our own curiosity, time, and funding perhaps. For this study, we focus on quantitative traits of varying architectures. For example, does body size predict brain size? Does the expression of one gene predict the expression of another? Does propagule size predict invasiveness? The possibilities seem endless, and many classical approaches to linear regression appear well-suited to these questions. Because we are living in the 21st century, we also know that phylogeny must be addressed if we seek reasonable and rigorous answers (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Pagel 1997, 1999; Rohlf 2001). What remains less clear, however, is which tree should be considered.

With the advent of phylogenetic comparative methods (PCMs), biologists are now painfully aware of the need to address phylogeny because related species and their traits covary according to shared ancestry—that is, they are not statistically independent (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Pagel 1997, 1999; Rohlf 2001). If ignored, then among-species covariance can lead us astray (Felsenstein 1985; Maddison and FitzJohn 2015; Uyeda et al. 2018; Gardner and Organ 2021). By accounting for the effect of shared ancestry, phylogenetic regression has become an icon of modern comparative biology, inspiring a wave of ecological and evolutionary progress in its wake. In the years since, its principles have been debated, refined, supplemented, and expanded to target diverse questions, hypotheses, and data types (Harvey and Pagel 1991; Hansen 1997; Sanford et al. 2002; Blomberg et al. 2003; Felenstein 2004; O'Meara et al. 2006; Revell et al. 2008; Beaulieu et al. 2012; Adams 2013; Pennell and Harmon 2013; Pennell et al. 2014; Maddison and FitzJohn 2015; Uyeda et al. 2018). Few studies in evolutionary biology are now published without at least a reference to PCMs and their foundations in phylogenetic regression.

Received: September 3, 2024. **Revised:** December 20, 2024. **Accepted:** January 21, 2025

Of course, a fundamental assumption of phylogenetic regression and PCMs generally is that the required input tree, and therefore the among-species covariance structure, is known (e.g. Felsenstein 1985; Martins and Garland 1991; Gittleman and Luh 1992; Miles and Dunham 1993; Boettiger et al. 2012; Cressler et al. 2015; Harmon 2019; Brahmantio et al. (2022); Schraiber et al. 2024). In practice, this assumption is often difficult if not impossible to confirm (Schluter 1995). Rarely does one expect that their assumed tree is indeed the true tree, or even the best tree possible for a given trait. Estimation error in the assumed phylogeny is likely to be an issue; errors tend to beget errors, such that errors in the assumed topology and branch lengths may propagate errors in downstream evolutionary inferences that assume error-free trees (Diaz-Uriarte and Garland 1996, 1998; Symonds 2002; Stone (2001); Mendes et al. 2018). Thus, if an assumed phylogeny is unreliable, then inferences of trait evolution may also be suspect (Harvey and Pagel 1991; Symonds 2002). Bayesian PCMs that fit models to posterior probability distributions of trees or coestimate phylogenetic character-evolution parameters hold promise for incorporating estimation uncertainty into the process (e.g. Villemereuil et al. 2012; Fuentes-G et al. 2020; Bastide et al. 2021; Zhang et al. 2021).

Yet, mismatch between an assumed and true phylogeny can occur for reasons besides just estimation error. Perhaps we are simply looking at the wrong tree. That is, we impose a tree for phylogenetic regression that is completely unrelated to our studied trait, its architecture, or its evolutionary past. Put plainly: how confident are we in our ability to accurately (or at least adequately) match our trait to its true phylogenetic history? We know that variation in phylogenetic history is ubiquitous, arising naturally from speciation, diversification, and evolution (Maddison 1997; Nichols 2001; Degnan and Rosenberg 2009; Kutschera et al. 2014). It is also well understood that traits vary considerably in their architectures with respect to the precise numbers, complexities, and genomic identities of encoding loci, each of which may reflect their own genealogical history. Indeed, gene trees often differ wildly from one another and from the overall species tree as a result of incomplete lineage sorting (ILS; Maddison 1997; Nichols 2001; Degnan and Rosenberg 2009; Hobolth et al. 2011), introgression (Yu et al. 2011; Leaché et al. 2014; Solís-Lemus et al. 2016; Tian and Kubatko 2016; Long and Kubatko 2018), ancestral structure (Slatkin and Pollack 2008; DeGiorgio and Rosenberg 2016; Koch and DeGiorgio 2020), and natural selection (Adams et al. 2018; Borges et al. 2020; He et al. 2020; Wascher and Kubatko 2023). Of these processes, ILS is arguably the most infamous (Maddison 1997; Kubatko and Degnan 2007; Edwards 2009; Liu et al. 2015). One particularly concerning consequence of ILS is hemiplasy (Avise and Robinson 2008), which results from forcing trait data to the wrong tree, which can generate false patterns of homoplasy-like evolution (Avise and Robinson 2008) and mislead PCMs (Hahn and Nakhleh 2016; Mendes and Hahn 2016; Guerrero and Hahn 2018; Mendes et al. 2016, 2018, 2019; Hibbins et al. 2019). Specifically, hemiplasy may artifactually increase the number of independent branches where two different traits match each other rather than the true evolutionary history.

In the presence of phylogenetic conflict, how do we choose a tree or trees for phylogenetic regression? Growing evidence suggests that this decision matters, but it can be difficult to know a priori whether to assume the overall species tree, a particular gene tree, a specific set of gene trees, or even every possible gene tree. To model trait evolution, studies may assume a species tree that has been estimated using coalescent-based (Doña and Johnson 2023) or traditional concatenation (Hensen et al. 2023) approaches, or perhaps a specific gene tree (Al-Kahtani et al. 2004; Ross et al. 2004; Kamilar and Cooper 2013; Gu 2016; Dunn et al. 2018; Chen et al. 2023; Adams et al. 2016). However, making such assumptions may (Dimayacyac et al. 2023) or may not (Hahn and Nakhleh 2016) be the best strategy. Modeling evolution as a function of a particular gene tree may prove beneficial for traits predicted to exhibit a one-to-one correspondence with a single tree, such as the expression of a gene largely regulated by *cis* elements near its encoded locus (Chen et al. 2019 Bastide et al. 2023; Bertram et al. 2023; Dimayacyac et al. 2023). Perhaps such scenarios represent a best case in which we might at least hope to match tree with trait. PCMs have also garnered great interest for modeling functional genomic evolution across cells, tissues, and species (Rohlfs and Nielsen 2015; Chen et al. 2019; Bastide et al. 2023; Bertram et al. 2023; Dimayacyac et al. 2023; Adams et al. 2024). Somewhat surprisingly, a recent study found that modeling gene expression as a function of the overall species tree rather than local gene trees improved model fit (Dimayacyac et al. 2023). Some traits, however, may be subject to more complex architectures encoded by multiple genetic loci, each with their own genealogical history. Taking this idea further, several models assume that all gene trees contribute to a given trait (Mendes et al. 2018; Hibbins et al. 2023). Importantly, a choice of trees is made each and every time a PCM is applied.

Therein lies a conundrum: we must choose a tree, but how can we be certain of which tree to choose? Rarely do we appreciate or even understand the ripple effects of this choice that is profoundly central to comparative biology. This study seeks to gauge our level of concern about tree mismatch that is not only possible but probable. Specifically, we explore the behavior of phylogenetic regression for testing trait associations when the true and assumed trees are mismatched due to the well-known and wide-spread phenomenon of gene tree–species tree conflict (Figs. 1 and 2). We employ a large-scale battery of evolutionary simulations with varying degrees of model mismatch for traits of both simple and complex architectures, and with both known and estimated trees. Given our findings, we then investigate a best-case scenario for matching tree with trait by using an extensive cross-species gene expression dataset sampled from mammals. Through this work, we seek to advance our understanding of the consequences of tree choice for comparative studies, while arguing for the promise of more robust and thoughtful evolutionary modeling.

Methods

Simulations with Known Trees and Simple Architectures

We explored the performance of phylogenetic regression when using matched versus mismatched trees for testing associations between two continuous traits **x** and **y**. We generated trait data using a linear model with phylogenetic signal in both the input predictor trait **x** and the response trait **y**, following the approach of similar studies (Pennell et al. 2014; Mazel et al 2016; Revell 2010; Fig. 1). The familiar linear regression equation for these traits can be written as

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\epsilon},$$

where **y** is an n-dimensional vector containing measurements of the response trait in each of n species, **x** is an n-dimensional vector containing measurements of the input predictor trait in each of n

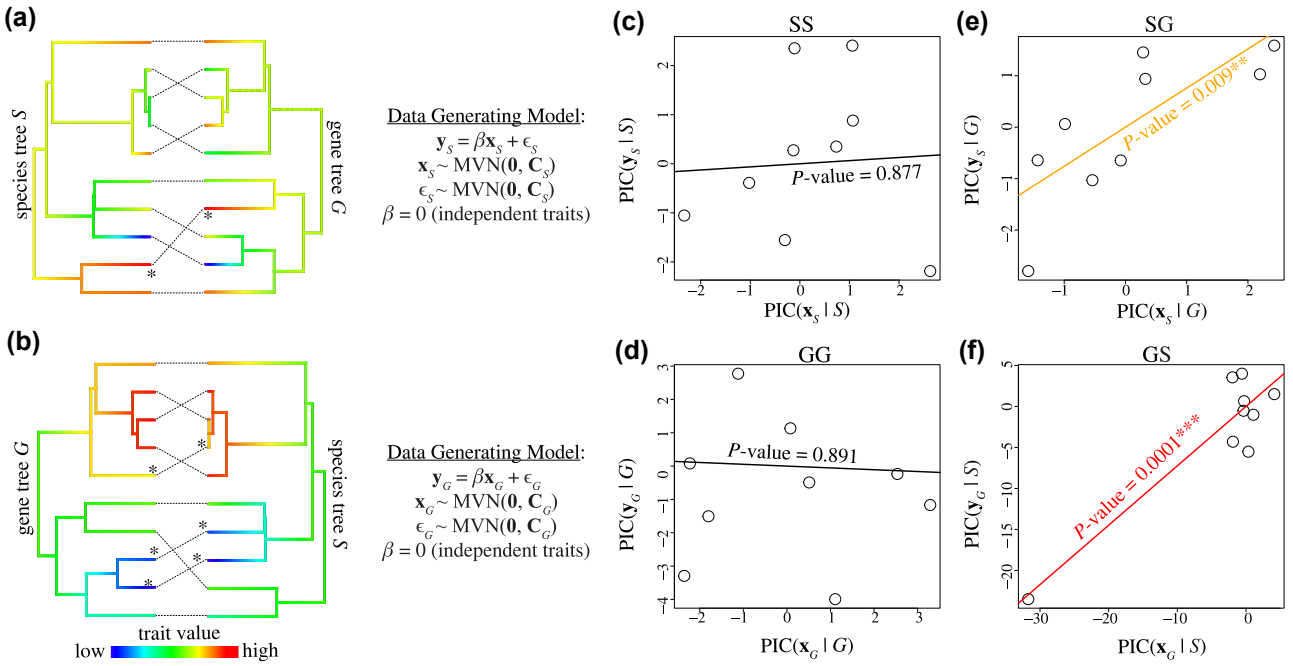

Fig. 1. Illustrating the phylogenetic conundrum. Examples showing species tree and gene tree pairs (a and b) and their associated data generating models (center) for scenarios in which both traits are generated according to the species tree *S* (top row) or the gene tree *G* (bottom row). Thus, the true generating tree is shown on the left in (a) and (b). Branch colors (a and b) illustrate values of the response trait **y** when mapped to the respective tree using the contMap function from phytools. Two examples (random replicates) of matched phylogenetic regression are shown for SS (c) and GG (d), in which the same tree was used for both generating the trait data and computing PICs, and two examples (random replicates) of mismatched regression are shown for SG (e) and GS (f), in which different trees were used for generating the trait data and computing PICs.

species, β is the regression coefficient that measures the relationship between $\mathbf{x}$ and $\mathbf{y}$, and $\boldsymbol{\epsilon}$ is an n-dimensional vector of residuals. Under the null hypothesis (no association between $\mathbf{x}$ and $\mathbf{y}$), $\beta = 0$, whereas the alternative hypothesis states that $\beta \neq 0$. Ordinary least squares assumes that the residuals $\boldsymbol{\epsilon}$ are independent and identically distributed as normal with mean zero and some standard deviation; this assumption is inherently violated with comparative data, in which traits tend to covary among species. Phylogenetic regression relaxes this assumption by considering the variance–covariance structure across a set of n species that is defined by their evolutionary relationships. Phylogenetic independent contrast (PIC) computes a set of $n-1$ contrasts that are statistically independent (at which point the null hypothesis of $\beta = 0$ can be tested), whereas phylogenetic generalized least squares (PGLS) incorporates the phylogenetic variance–covariance structure directly into the model. Both methods provide equivalent estimates of significance levels (Blomberg et al. 2012).

To simulate trait evolution, we included phylogenetic signal into the linear model by simulating $\mathbf{x}$ and $\boldsymbol{\epsilon}$ according to a multivariate normal (MVN) distribution with mean zero and an $n \times n$ phylogenetic variance–covariance matrix $\mathbf{C}$, which is defined according to a specific species tree or gene tree (Grafen 1989; Martins and Garland 1991; Martins 1996). When denoting the data generating process, we use the subscript S for traits with signals matching a species tree S and the subscript G for traits with signals matching a gene tree G. Therefore, to generate trait data with phylogenetic signal according to a tree $T \varepsilon \{S, G\}$, we used

$$\mathbf{y}_T = \mathbf{x}_T \beta + \boldsymbol{\epsilon}_T$$

$$\mathbf{x}_T \sim \text{MVN}(\mathbf{0}, \mathbf{C}_T)$$

$$\boldsymbol{\epsilon}_T \sim \text{MVN}(\mathbf{0}, \mathbf{C}_T)$$

where both $\mathbf{x}_T$ and $\boldsymbol{\epsilon}_T$ are distributed as MVN with mean n-dimensional vector $\mathbf{0}$ containing all zero elements and phylogenetic variance–covariance $\mathbf{C}_T$ defined according to tree T. Note that when $\beta = 0$, the response trait is simply distributed as $\mathbf{y}_T \sim \text{MVN}(\mathbf{0}, \mathbf{C}_T)$, representing independent Brownian motion evolution for both $\mathbf{y}_T$ and $\mathbf{x}_T$ on the same tree. After generating trait data with these two data generating models (one according to $T = S$ and another according to $T = G$), phylogenetic regression was conducted by computing PICs using either the species tree S or the gene tree G, allowing us to explore scenarios of tree mismatch in which the data generating process and the assumed phylogeny for PICs are different (details provided below). To assess false positive rates under different scenarios, we set the true regression coefficient $\beta = 0$, whereas a nonzero $\beta \neq 0$ was used to evaluate statistical power.

To investigate impacts of tree choice for phylogenetic regression, we compared "matched" regression (Fig. 1c and d), for which the same tree is used to generate and compute PICs, and "mismatched" regression, for which different trees are assumed to generate and compute PICs (Fig. 1e and f). We conducted a multifactorial simulation study to investigate a range of scenarios with increasing probabilities of tree mismatch due to ILS. Our overall simulation protocol can be described in four steps: (1) a species tree S is generated using a diversification process (Yule 1925), (2) a gene tree G is simulated according to the multispecies coalescent with the species tree S obtained from Step 1, (3) traits are simulated using the phylogenetic variance–covariance matrix $\mathbf{C}_S$ from the species tree obtained from Step 1 to obtain $\mathbf{y}_S$ and $\mathbf{x}_S$, or using $\mathbf{C}_G$ from the gene tree from Step 2 to obtain to obtain $\mathbf{y}_G$ and $\boldsymbol{x}_G$, and (4) phylogenetic regression is applied to the simulated trait data (Fig. 1) with either matched or mismatched trees. That is, the opportunity for mismatch occurs in Steps 3 and

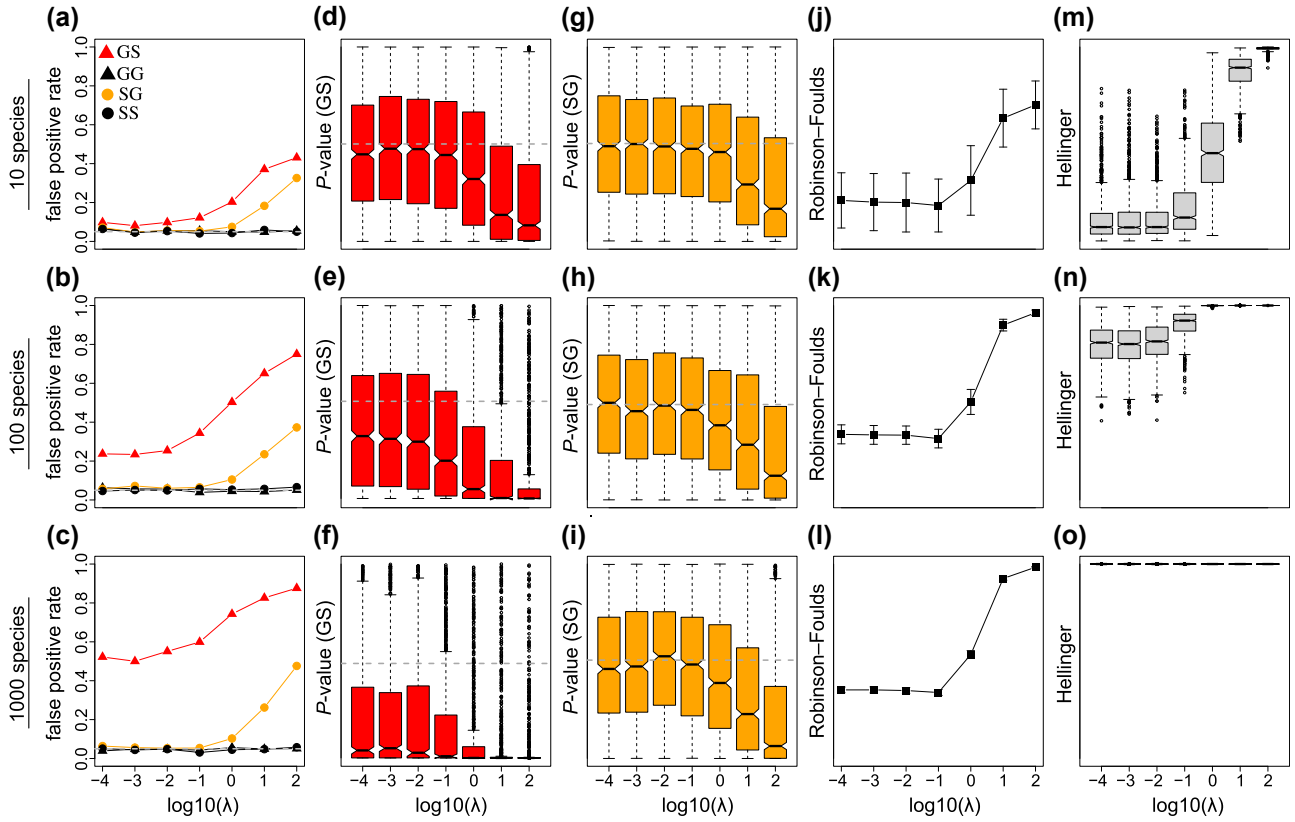

Fig. 2. Tree mismatch exacerbates evidence of false trait associations with phylogenetic regression. Estimates of the false positive rate (a to c), *P*-value distributions for GS (d to f), *P*-value distributions for SG (g to i), means and standard deviations of Robinson–Foulds (j to l), and Hellinger (m to o) between gene trees and species trees from simulations including 10 species (top row), 100 species (middle row), and 1,000 species (bottom row) for birth–death simulations with birth rate λ, death rate $\lambda/2$, and root age of 10 coalescent units. The two traits were statistically independent ($\beta = 0$) for all simulations. Dashed horizontal lines mark the commonly used false positive rate $\alpha = 0.05$ in (a) to (c), median *P*-values taken from matched GG scenarios in (d) to (f), and median *P*-values from matched SS scenarios in (g) to (i). The *y*-axis ranges from 0 to 1 in all panels.

4 when computing PICs using an incorrect tree that is unrelated to the data generating process of the studied traits. When the same tree is used in Steps 3 and 4, the scenario represents matched regression because the same tree used to simulate the traits is also used to compute PICs. Conversely, when different trees are used for Steps 3 and 4, the scenario represents mismatched regression, as the assumed tree is not the tree that generated the data (e.g. a species tree is assumed for traits simulated on a gene tree).

Our simulation approach examined four distinct scenarios: matched gene tree–gene tree (GG), matched species tree–species tree (SS), mismatched gene tree–species tree (GS), and mismatched species tree–gene tree (SG), where the first tree in each pair indicates the tree used to simulate traits, and the second tree is assumed for phylogenetic regression (Fig. 1). For example, GG represents the matched scenario for which the same gene tree G is used to both generate traits and compute their PICs using $\mathbf{y}_G$ and $\boldsymbol{x}_G$, whereas GS is mismatched because a gene tree G is used to generate $\mathbf{y}_G$ and $\boldsymbol{x}_G$, but the species tree S is incorrectly assumed for their PICs. Likewise, both $\mathbf{y}_S$ and $\boldsymbol{x}_S$ traits and their PICs are generated with the same species tree for SS scenarios, whereas SG represents tree mismatch because the traits $\mathbf{y}_S$ and $\mathbf{x}_S$ are both generated via the species tree, but a gene tree is incorrectly assumed. Thus, we evaluated phylogenetic regression with two forms of correctly specified models (GG and SS) and with two forms of incorrectly specified models (GS and SG; Fig. 1).

Throughout our simulations, we varied both the total number of taxa $n \in \{10,\ 100,\ 1,000\}$ and speciation rate $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$ used to simulate the species trees. This strategy allowed us to effectively incorporate variability in the expected amount of phylogenetic conflict due to ILS, as λ is inversely proportional to the expected branch lengths in the species tree. Slow rates ($\lambda = 10^{-4}$) yield long internal branch lengths and lower ILS, whereas fast rates ($\lambda = 10^{2}$) generate short internal branch lengths, exacerbating ILS. We generated species trees under a birth–death model in which the death rate was set to half the speciation rate; we also investigated a simple pure-birth model of diversification (Yule 1925) with zero death rates. We employed the R package TreeSim (Stadler 2011) using the *sim.bd.taxa.age* function with a most recent common ancestor age of either 1, 10, or 100 to generate species trees of varying depths. Therefore, both the number of species and the total tree height were held constant within each set of simulation conditions. The *sim.coaltree.phylo* function in Phybase (Liu and Yu 2010) was used to simulate gene trees from species trees. Trait data were then simulated according to either S or G using the linear models described above for a total of 10^3 replicates for each value of λ and for each of the four scenarios GG, GS, SS, and SG. For each replicate, phylogenetic regression was conducted using PICs computed according to the four scenarios (Fig. 1) using the *pic* function provided in the software package APE (Paradis and Schliep 2019). We evaluated the false positive rates for each scenario by setting the true regression coefficient $\beta = 0$ and quantifying the number of replicates with P-value < 0.05 that incorrectly reject the null hypothesis, whereas four values of nonzero $\beta \in \{0.25,\ 0.50,\ 0.75,\ 1.0\}$ were used to investigate statistical power for correctly rejecting

the null hypothesis when $\beta \neq 0$. To provide context on the degree of tree discordance, we computed Robinson–Foulds distances (Robinson and Foulds 1981) and probabilistic Hellinger distances (Pardo 2005; Adams et al. 2021) between the gene tree and species tree for each replicate. The Robinson–Foulds metric considers only the topological distance between two trees, whereas Hellinger measures the distance between two MVN distributions based on models of trait evolution.

Simulations with Known Trees and Complex Architectures

Our first array of simulations described above applied simple architectures in which traits were generated according to a single species tree or alternative, a single gene tree (Fig. 1). We also conducted a case study that explored more complex architectures in which trait data were generated according to multiple gene trees, which is expected for some continuous traits. For these simulations, we followed the same general protocol as above with the addition of the seastaR approach (Hibbins et al. 2023), by computing a phylogenetic variance–covariance matrix $\mathbf{C}^*_{\mathcal{T}}$ as a weighted mean of the individual gene tree variance–covariance matrices taken from a set $\mathcal{T}$ of t different gene trees. The primary change is that we used $\mathbf{C}^*_{\mathcal{T}}$ instead of $\mathbf{C}_S$ (a single species tree S) or $\mathbf{C}_G$ (a single gene tree G) to generate the traits. More specifically, the traits were encoded by t gene trees, each with equal contribution. We conducted four case study simulations in which the number of gene trees $t \in \{2,\ 5,\ 10,\ 100\}$ varied to represent traits with architectures encoded by 2, 5, 10, or 100 genomic loci and their associated gene trees.

Here, we explored matched scenarios in which the same generating $\mathbf{C}^*_{\mathcal{T}}$ was used to both simulate traits and conduct phylogenetic regression. We also investigated two additional scenarios of mismatched regression: (i) one $\mathbf{C}^*_{\mathcal{T}}$ was used to simulate the traits, and a different $\mathbf{C}^*_{\mathcal{T}'}$ was generated from a separate set $\mathcal{T}'$ of t different gene trees that was incorrectly assumed for phylogenetic regression, and (ii) one $\mathbf{C}^*_{\mathcal{T}}$ was used to simulate the traits, and the species tree $\mathbf{C}_S$ was incorrectly assumed for regression. We refer to these three scenarios as matched gene trees (i.e. same $\mathbf{C}^*_{\mathcal{T}}$ used for both simulation and inference), mismatched gene trees (i.e. different sets of gene trees $\mathbf{C}^*_{\mathcal{T}}$ and $\mathbf{C}^*_{\mathcal{T}'}$ for simulation and inference), and mismatched species tree (i.e. $\mathbf{C}_S$ used for inference instead of the true $\mathbf{C}^*_{\mathcal{T}}$), respectively. We conducted phylogenetic regression using PGLS (Grafen 1989; Martins and Garland 1991; Martins 1996) using the *gls* function in the R package nlme (Pinheiro et al. 2017) because the *pic* function requires strictly bifurcating trees. Importantly, the regression slope estimates and levels of significance are equivalent with PGLS and PIC under Brownian motion (Blomberg et al. 2012). For these analyses, the same birth–death process was used to simulate species tree with a depth of 10 coalescent units and either 10 or 100 species, and we focused on assessing false positive rates when $\beta = 0$ with 10^3 replicates for each value of the birth rate $\lambda \in \{10^{-4}, 10^{-3},\ \ldots,\ 10^2\}$.

Simulation Case Study: How Does Phylogenetic Estimation Error Influence Mismatch?

Results obtained when using true trees may not hold when instead using estimated trees, which is important for empirical studies. Thus, in addition to simulations that utilized known phylogenies (i.e. those without estimation error), we also conducted a simulation case study that incorporated phylogenetic estimation error. We followed the above simulation protocol but included additional steps for estimating gene trees and species trees. Though a multitude of parameters are likely to influence tree estimation and error, we chose several factors predicted to be important while ensuring computational feasibility. The first three steps of our simulation protocol for this case study are analogous to those of the simulations described above for known trees: (i) simulate a species tree with varying speciation rate λ, (ii) simulate 10 gene trees for each species tree from Step i, and (iii) simulate continuous traits using either the known species tree from Step i (SS and SG scenarios) or a known gene tree from Step ii (GG and GS scenarios). Next, we added components for estimating gene trees and species trees: (iv) simulate 2.5 kb alignments for each of the 10 gene trees using an HKY model with a molecular clock and per-base population-scaled mutation rate $\theta = 0.01$, transition/transversion ratio of 4.6, and base equilibrium frequencies of $f_A = 0.3$, $f_C = 0.2$, $f_G = 0.2$, and $f_T = 0.3$ for nucleotides A, C, G, and T, respectively, (v) estimate gene trees using IQ-TREE2 (Minh et al. 2020), (vi) infer a species tree with the gene tree estimates from Step v using STELLS2 (Pei and Wu 2017), and (vii) conduct phylogenetic regression using either the estimated species tree from the Step vi or the estimated gene tree from Step v for computing PICs. Because our simulations were conducted using a molecular clock, estimated gene trees were midpoint rooted. Therefore, this simulation protocol matches our above simulations with the addition of gene tree estimation (Step v) and species tree inference (Step vi), with phylogenetic regression conducted using these estimated trees instead of the known trees. For this case study, we focused on the impacts of phylogenetic estimation error on false positive rates of phylogenetic regression by simulating two statistically independent traits with $\beta = 0$. Because STELLS2 requires at least two samples per species to estimate external branch lengths, the known gene trees simulated in Step ii and estimated in Step v include two samples per species for STELLS2. However, for both simulating trait data and fitting regression models, we pruned these trees to include only one lineage sampled per species to allow direct comparisons of regression models on gene trees versus species trees with the same numbers of lineages. Because of the computational requirements needed to simulate this multistep experiment (species tree to gene trees to sequence alignments to inferences of each), we generated 10^2 replicates with $n = 10$ species for each $\lambda \in \{10^{-4}, 10^{-3},\ \ldots,\ 10^2\}$.

An Empirical Best-Case Study: Does Tree Choice Impact Gene Expression Phylogenetic Regression?

Our simulations revealed evidence of profound bias with mismatched models for traits with both simple and complex architectures, and when using known and estimated trees (see *Results*). Given these findings, we sought to explore the empirical impacts of tree choice for a best-case scenario in which one might be able to better match tree with trait. PCMs have gained recent promise for providing exciting insights into the origins and evolution of functional genomic traits (e.g. Rohlfs and Nielsen 2015; Chen et al. 2019; Bastide et al. 2023; Bertram et al. 2023; Dimayacyac et al. 2023; Adams et al. 2024). We sought to use gene expression evolution as an example of a best-case scenario because one might predict that it should, at least in theory, be easier to match the expression trait of a given gene to one specific tree—either the species tree or respective gene tree.

We explored the effects of phylogenetic tree specification on tests of trait association using an empirical gene expression dataset from 11 female and male tissues in eight mammals and chicken (Brawand et al. 2011). In particular, we obtained normalized gene expression abundance measurements computed in reads per kilobase of exon model per million mapped reads (RPKM; Mortazavi et al. 2008) from female and male brain (whole brain without cerebellum), female and male cerebellum, female and male heart, female and male kidney, female and male liver, and testis in human (*Homo sapiens*), chimpanzee (*Pan trogodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus abelii*), macaque (*Macaca mulatta*), mouse (*Mus musculus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), and chicken (*Gallus gallus*; Brawand et al. 2011). We focused our comparisons by restricting analyses to the most conservative 5,321 orthologous genes, or those with constitutive exons that aligned across all nine species in the original dataset (Brawand et al. 2011), and computed the median expression level for tissues containing multiple replicates.

To understand the impacts of tree specification on phylogenetic regression, we obtained the estimated species tree from Brawand et al. (2014) and estimated gene trees from nucleotide and amino acid alignments downloaded from the UCSC Genome Browser (Navarro Gonzalez et al. 2021) at http://www.genome.ucsc.edu. Specifically, the UCSC alignments included all protein-coding exons in human (GRCh38/hg38) and 99 vertebrates (Blanchette et al. 2004; Dreszer et al. 2012), from which we extracted those pertaining to 5,267 genes in the nine species considered here. We concatenated the nucleotide and exon alignments for each gene and constructed gene trees by applying PhyML (Guindon et al. 2010) with default parameters to these alignments. The species tree and all gene trees were scaled to unit depth. We investigated the statistical performance of phylogenetic regression in three experimental settings: expression in female brain–male brain, female heart–male heart, and female kidney–male kidney. For each experiment, we conducted PIC regression based on log-transformed RPKM values across the nine species and assessed relationships between tissues via evidence of statistical significance (*P*-values).

Specifically, we evaluated impacts of tree choice when modeling female and male expression evolution across species. For each gene, we conducted phylogenetic regression to test associations between male and female expression in three separate analyses based on phylogenetic regression fit to: (i) the species tree, (ii) the gene tree inferred from nucleotide sequences, and (iii) the gene tree inferred from amino acid sequences. To explore genome-wide patterns and identify interesting case studies, we also computed three distance statistics: d_S, d_N, and d_A representing analyses that assumed the species tree (ST), the nucleotide gene tree (NT), and the amino acid gene tree (AT), respectively. These statistics have the forms

$$d_S = \frac{D(\mathrm{ST}, \mathrm{NT}) + D(\mathrm{ST}, \mathrm{AT}) - D(\mathrm{NT}, \mathrm{AT})}{2}$$

$$d_N = \frac{D(\mathrm{NT}, \mathrm{ST}) + D(\mathrm{NT}, \mathrm{AT}) - D(\mathrm{ST}, \mathrm{AT})}{2}$$

$$d_A = \frac{D(\mathrm{AT}, \mathrm{ST}) + D(\mathrm{AT}, \mathrm{NT}) - D(\mathrm{ST}, \mathrm{NT})}{2}$$

where $D(\text{Tree 1}, \text{Tree 2}) = |\log_{10}(P_{\text{Tree 1}}) - \log_{10}(P_{\text{Tree 2}})|$ represents the magnitude of the difference between the log-transformed *P*-values of a pair of trees. Thus, each distance statistic will evaluate whether the *P*-value for a given analysis tree is substantially different from the *P*-values of the other two trees. These measures are akin to those that have been used for identifying population branches with extreme differences using allele frequency (Shriver et al. 2004; Yi et al. 2010) or expression (Assis 2019; Jiang and Assis 2020) data. We applied these measurements here to identify genes that appear particularly sensitive to tree choice. For example, a large d_A value might reflect scenarios in which regression is strongly significant (P-value $< 10^{-6}$) based on the amino acid tree, but not significant in the nucleotide or species tree-based regression (P-value > 0.05).

Investigating a Potential Robust Path Forward

We recently found promise in the application of robust estimators for improving the resistance of phylogenetic regression to evolutionary outliers (Adams et al. 2024). Given these findings, we sought to assess whether a robust estimator may yield comparatively better performance than standard L2-based phylogenetic regression, which minimizes the mean squared error of predictions and is thus sensitive to outliers. To address this question, we employed the robust L1 estimator, which instead minimizes the mean absolute error (Rousseeuw and Yohai 1984), helping to alleviate false positive rates associated with strong outliers by de-emphasizing large residuals. We applied L1-based regression to the same simulation conditions as before with $n \in \{10,\ 100,\ 1,000\}$ species and varying levels of tree mismatch for known (simulated) trees, our case study that included gene tree estimation in addition to tree mismatch with $n = 10$ species, and finally, our empirical case study.

Results

Illustrating the Phylogenetic Conundrum

We chose two simulation replicates to illustrate this phylogenetic conundrum (Fig. 1). For these examples, we simulated a species tree *S* and an associated gene tree *G* given the multispecies coalescent process on S. We then simulated two statistically independent ($\beta = 0$) traits $\mathbf{x}$ and $\mathbf{y}$ using the species tree S (Fig. 1; top row), and separately using the gene tree G (Fig. 1; bottom row). Here, we show two examples of matched models in which the same tree was used to both generate and analyze the trait data (Fig. 1c and d). Likewise, we provide two examples of mismatched regression models in which the traits were generated according to the species tree, but the gene tree was incorrectly assumed for PICs (Fig. 1e), and the alternative scenario in which the traits were generated according to the gene tree, but the species tree was incorrectly assumed (Fig. 1f). *P*-values from matched phylogenetic regression SS (Fig. 1c) and GG (Fig. 1d) were not statistically significant, consistent with the null hypothesis of independence ($\beta = 0$). However, both examples of mismatched phylogenetic regression based on SG (Fig. 1e) and GS (Fig. 1f) were statistically significant, yielding a false positive result due to the wrong tree choice. In these examples, the degree of false significance was higher for GS than for SG, as demonstrated by their *P*-values. Comparing the trait mappings provides some intuition, with evidence of hemiplasy when a trait is mapped to the incorrect tree (Fig. 1a and b).

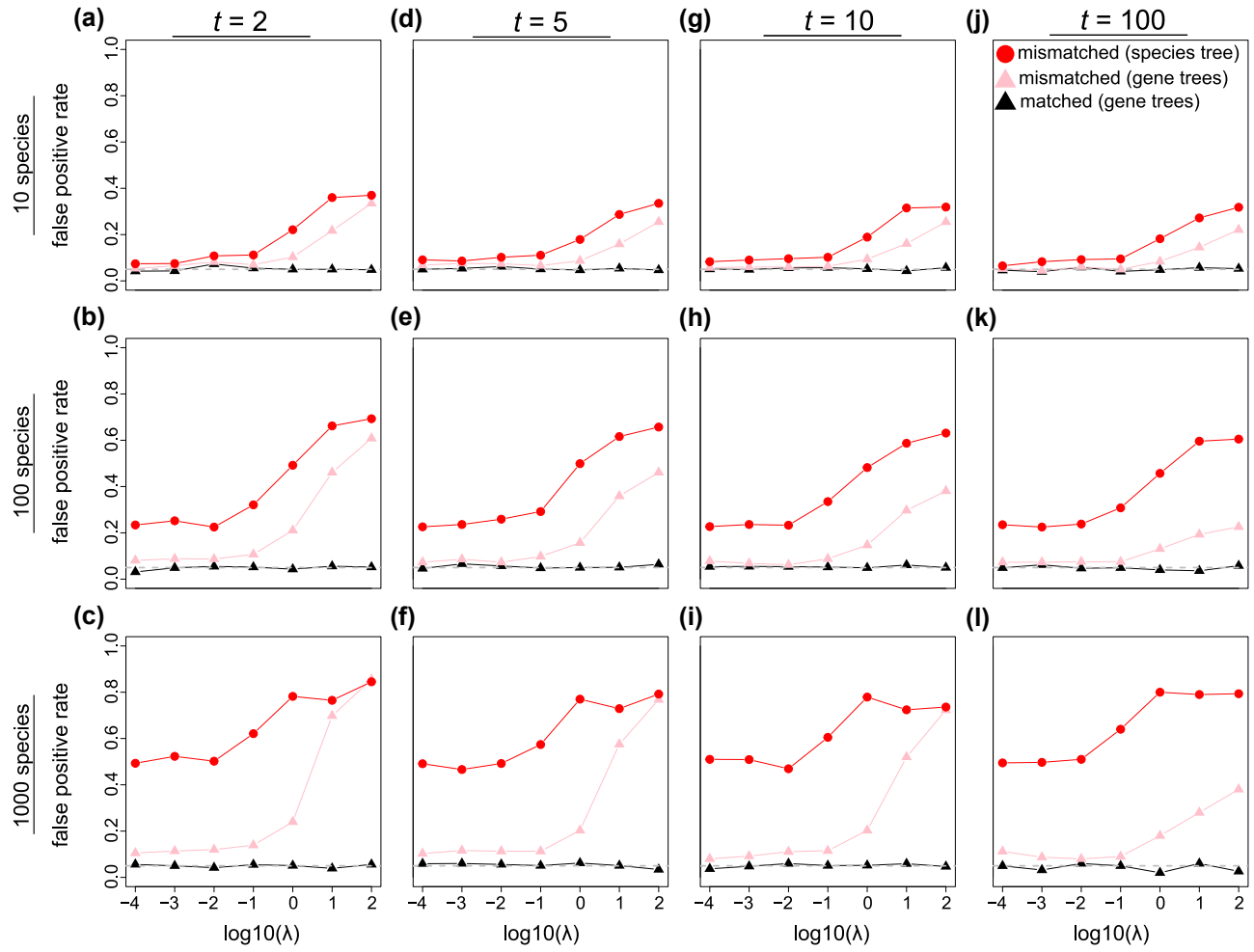

Fig. 3. Tree mismatch misleads phylogenetic regression for traits with more complex architectures. Estimates of the false positive rates from simulations including 10 species (top row), 100 species (middle row), and 1,000 species (bottom row) for birth–death simulations with birth rate λ, death rate $\lambda/2$, and root age of 10 coalescent units for mismatched species tree regression (red lines), mismatched gene tree regression (pink links), and matched gene tree sets (black lines). Results shown for traits encoded by two loci (a to c), five loci (d to f), 10 loci (g to i), and 100 loci (j to l). The two traits were statistically independent ($\beta = 0$) for all simulations. Horizontal dashed lines mark the commonly used false positive rate of $\alpha = 0.05$.

Impacts of Tree Mismatch on False Positive Rates of Phylogenetic Regression with Simple Architectures

Across our simulations, we found evidence of strong biases with incorrectly mismatched phylogenetic regression (GS and SG) compared to correctly matched regression (GG and SS; Fig. 2, supplementary figs. S1 and S2, Supplementary Material online). Specifically, false positive rates for GS and SG (red and orange) were higher than those for matched GG and SS (black), which yielded acceptable false positive rates of ~5% across all simulations (Fig. 2a to c). Thus, assuming the incorrect tree tended to mislead phylogenetic regression to reject the null hypothesis when the two traits were statistically independent ($\beta = 0$). The impact of phylogenetic mismatch was exacerbated with more species (Fig. 2a to c; top to bottom), shorter tree depths (Fig. 2, supplementary figs. S1 and S2, Supplementary Material online), and as the expected amount of ILS increased: false positive rate increased with speciation rate for both GS and SG (Fig. 2a to c; left to right in each panel). Specifically comparing the two mismatched scenarios (GS vs. SG) revealed evidence of higher false positive rates for GS (red) than for SG (orange) across our simulations (Fig. 2, supplementary figs. S1 and S2, Supplementary Material online). That is, performance was worse when incorrectly assuming the species tree for traits generated from a gene tree (GS) than the reverse situation in which an incorrect gene tree was assumed for traits generated from a species tree (SG). The severity of false positive rate inflation was influenced by the overall depth of the species tree, with shorter tree depths exacerbating false positive rates comparatively (supplementary fig. S1, Supplementary Material online vs. Fig. 2 vs. supplementary fig. S2, Supplementary Material online). Our results were similar when simulations under pure-birth and birth–death models (comparing Fig. 2 and supplementary fig. S3, Supplementary Material online). These findings were consistent with the overall distributions of *P*-values (Fig. 2d to i), and these impacts reflected topological (Fig. 2j to l) and probabilistic (Fig. 2m to o) distances between gene trees and species trees.

Complex Architectures and Mismatched Phylogenetic Regression

Our simulation study with more complex architectures for traits encoded by 2, 5, 10, or 100 loci continued to mirror these results (Fig. 3). We found unacceptably high false positive rates for mismatched models across all tree sizes (10, 100, or 1,000 species; rows in Fig. 3) and architectures (2, 5, 10, or 100 loci; columns

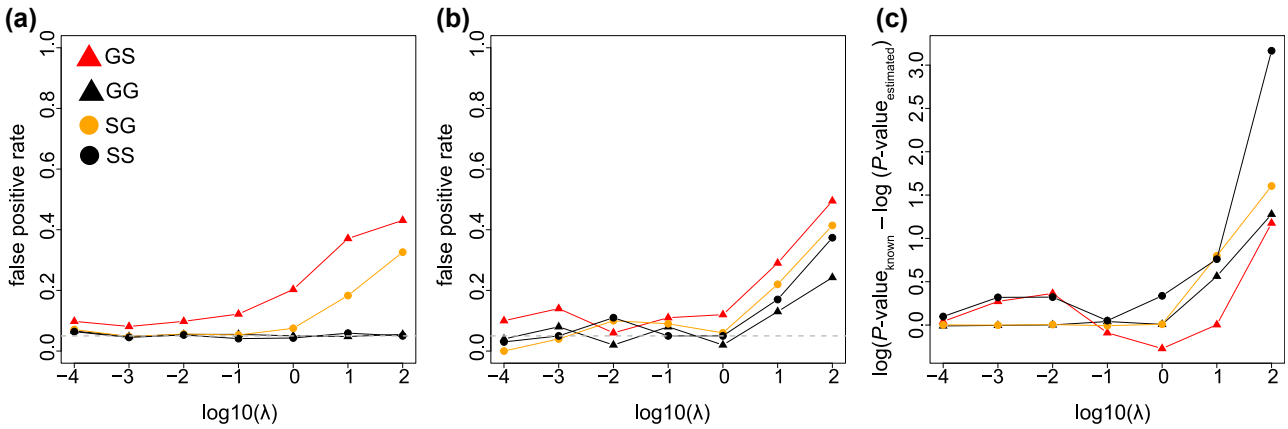

Fig. 4. Case studying the impacts of both tree mismatch and tree estimation error on phylogenetic regression. Depicted are false positive rates of the two mismatched scenarios (GS and SG) and the two matched scenarios (GG and SS) when regression was performed with known trees (a) and estimated trees (b) for $n = 10$ species. Difference between log-scaled *P*-values obtained with known and estimated trees (c).

in Fig. 3) that were amplified with higher amounts of ILS (left to right on *x*-axes in Fig. 3). As with single tree regression (Fig. 2), increasing the sample size (i.e. increasing the number of species) only made the situation worse. Moreover, both scenarios of tree mismatch (i.e. red and pink lines in Fig. 3) tended to produce high false positive rates compared to the appropriate false positive rate for correctly matched gene tree sets (black lines; Fig. 3). However, incorrectly assuming a species tree tended to generate higher false positive rates than assuming an incorrect gene tree set (red vs. pink; Fig. 3). Increasing the architecture complexity (i.e. the number of loci encoding a trait; left to right columns in Fig. 3) yielded slight improvements, though false positive rates remained substantially higher than the typical $\alpha = 0.05$ cutoff for many scenarios. With large trees (100 or 1,000 tips), even the smallest birth rates still exhibited remarkably high false positive rates. For example, false positive rates were estimated at ~50% for 1,000-tip trees with a birth rate of $\lambda = 10^{-3}$ (Fig. 3c).

Simulation Case Study: Phylogenetic Estimation Error and Tree Mismatch Together

Regardless of whether known (Fig. 4a) or estimated (Fig. 4b) trees are assumed, mismatched regression amplified false positive rates. Perhaps expectedly, we found higher false positive rates when using estimated versus known phylogenies in many cases. Estimation error tended to increase false positive rates for matched GG and SS phylogenetic regression (black lines; Fig. 4). The effects of estimation error on these matched scenarios were still less pronounced than those on mismatched GS and SG (red and orange lines; Fig. 4), however. Increasing the speciation rate tended to exacerbate false positive rates for all GG, SS, GS, and SG scenarios with estimated trees (Fig. 4b), whereas known matched regression scenarios (GG and SS) were unaffected (Fig. 4a). Comparing differences between log-scaled *P*-values of known and estimated trees further highlighted these findings (Fig. 4c), with the largest differences between known and estimated analyses observed in the matched SS, followed by SG, GG, and GS. This result likely reflects the higher relative false positive rates for SS when using estimated versus known trees (black lines; Fig. 4b), whereas known matched analyses demonstrate acceptable false positive rates of 0.05 (black lines; Fig. 4a). In these scenarios, phylogenetic regression with GS and SG were strongly influenced by tree mismatch with known trees and estimated trees (red and orange lines in Fig. 4a and b).

Tree Mismatch and Statistical Power of Phylogenetic Regression

Next, we evaluated the potential for phylogenetic mismatch to influence the power of regression to detect trait associations when $\beta > 0$. When compared with false positive rates, the effects of mismatched trees on power appear to be less dramatic and fluctuate depending on the value of β and number of species (Fig. 5). In many cases, however, we found evidence that mismatched regression can decrease power. This finding is perhaps most apparent in our simulations with and 100 species (Fig. 5e), as well as with and 1,000 species (Fig. 5c), in which mismatched GS scenarios demonstrated comparatively lower power (red lines; Fig. 5). Mismatched SG scenarios also exhibited lower power than matched regression in some examples (orange vs. black; Fig. 5). However, impacts were less apparent for species trees that were smaller ($n = 10$; top row of Fig. 5) and deeper (Fig. 5 vs. supplementary fig. S4, Supplementary Material online vs. supplementary fig. S5, Supplementary Material online).

Empirical Case Study: Investigating Phylogenetic Mismatch and Gene Expression Data

Most apparent in our exploration of phylogenetic regression using mammalian gene expression data are the differences in inferred significance depending on tree choice (Fig. 6a to c). For instance, in heart tissue, we observed the smallest number of outliers (red points) for the statistic corresponding to analyses using the species tree (Fig. 6a; 38 genes), followed by for analyses using the nucleotide gene tree (Fig. 6b; 107 genes), and finally by for analyses using the amino acid gene tree (Fig. 6c; 643 genes). Thus, using the amino acid gene tree for phylogenetic regression resulted in the identification of many more significantly associated genes in female and male heart tissue than using either the species tree or nucleotide gene tree.

These genome-level explorations also allowed us to identify the largest outlier genes based on their values of d_S, d_N, and d_A and (Fig. 6d to f). The largest outlier based on d_S was *UQCR11* (Fig. 6d), which is involved in the mitochondrial electron transport chain. All three analyses revealed a positive relationship between female and male expression in this gene,

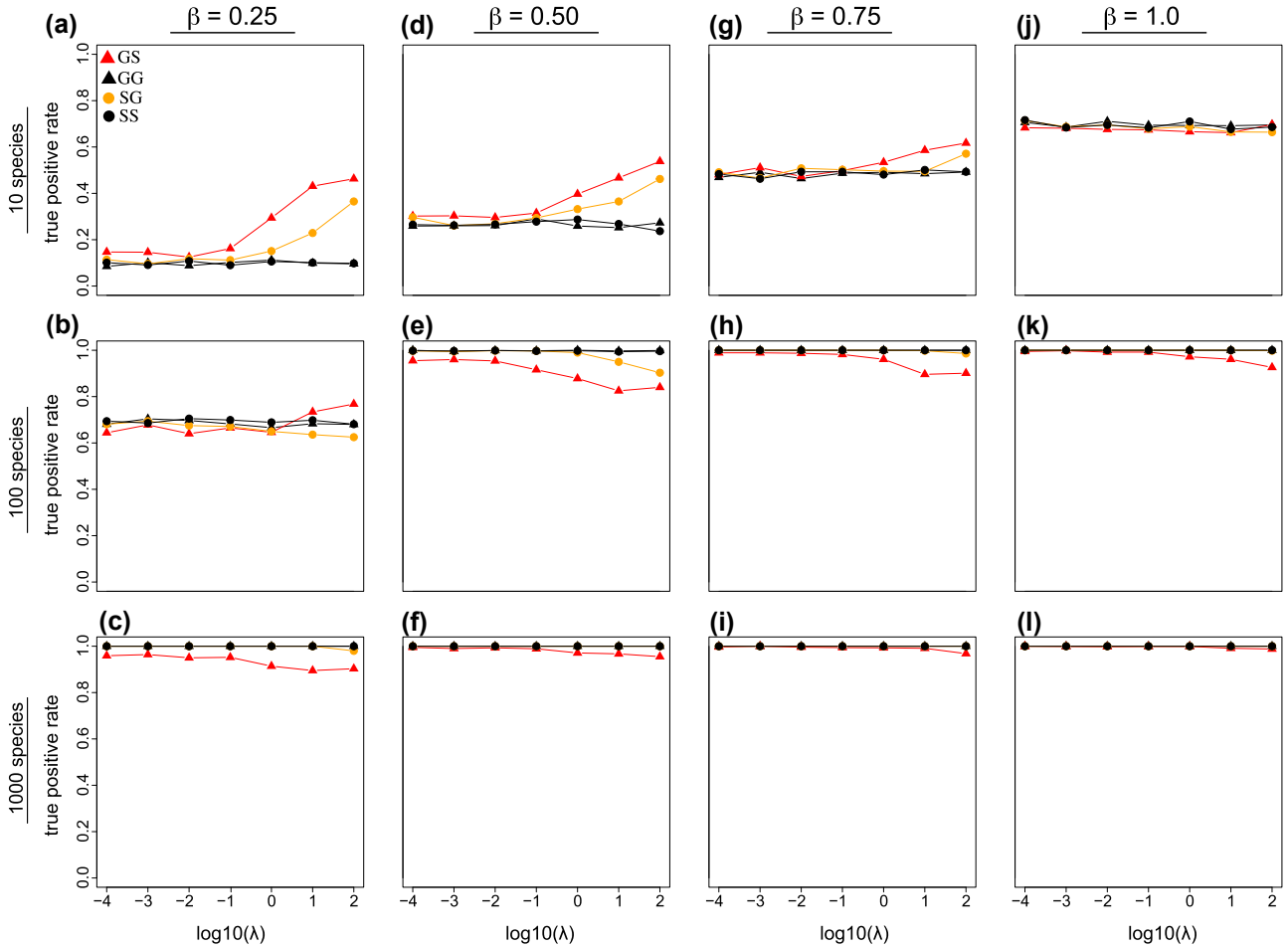

Fig. 5. Tree mismatch influences power to detect true trait associations. Estimates of true positive rates for 10 species (top row), 100 species (middle row), and 1,000 species (bottom row) for birth–death simulations with birth rate λ, death rate $\lambda/2$, and root age of 10 coalescent units. Results are shown for $\beta = 0.25$ (a to c), $\beta = 0.50$ (d to f), $\beta = 0.75$ (g to i), and $\beta = 1.0$ (j to l).

though with much weaker significance when using the species tree than when using either of the two gene trees (Fig. 6d). The largest outlier based on d_N was *RAB14* (Fig. 6e), which is involved in intracellular membrane trafficking. For this gene, the nucleotide tree yielded a highly significant negative relationship between female and male expression, whereas the other two trees did not produce significant results (Fig. 6e). Finally, the largest outlier based on d_A was *TBCC* (Fig. 6f), which is one of four genes involved in the pathway leading to correctly folded beta-tubulin from folding intermediates. For this gene, the amino acid tree regression produced a highly significant positive relationship between female and male expression, whereas the other two analyses did not yield a significant association (Fig. 6f).

Next, we evaluated overlap in statistically significant genes estimated using the three regression strategies (i.e. assuming the species tree, nucleotide tree, or amino acid tree) across tissues. We first considered the fraction of significant (P-value < 0.05) analyses from a tissue-level perspective. For each of three tissues considered (brain, heart, and kidney), we found substantial overlap in the percentages of genes with estimates of significant relationships between female and male expression (Fig. 7). That is, within a given tissue, the fractions of significant genes were similar for regression based on the species tree, amino acid tree, and nucleotide tree. However, consistent with our previous findings in heart (Fig. 7a to c), phylogenetic regression based on the amino acid gene tree yielded the largest percentage of uniquely significant genes for all tissues, with 24%, 23%, and 22% significant for brain, heart, and kidney, respectively (Fig. 7). Given these results, we then computed log-likelihoods of the fitted phylogenetic regression model for the three tissues and the three strategies. All three tissues agree that model fit was highest on average when assuming the species tree, followed by the nucleotide tree, and finally the amino acid tree (Fig. 8). That is, phylogenetic models tend to fit the species tree best and the amino acid tree worst, with the nucleotide tree fit representing an intermediate. Thus, suggesting that excess of uniquely significant genes for phylogenetic regression using the amino acid tree may be due to a poor fit.

Exploring the Potential for Robust Phylogenetic Regression

We found evidence that robust L1-based regression can reduce false positive rates, at least compared to conventional L2-based regression for both known (Fig. 9a to c) and estimated trees (Fig. 9d). In particular, L1-based regression yielded comparatively fewer false positives for GS (solid vs. dashed red lines; Fig. 9) and SG (solid vs. dashed orange lines; Fig. 9) under most conditions of mismatched regression. When considering our analyses of simulations that used estimated

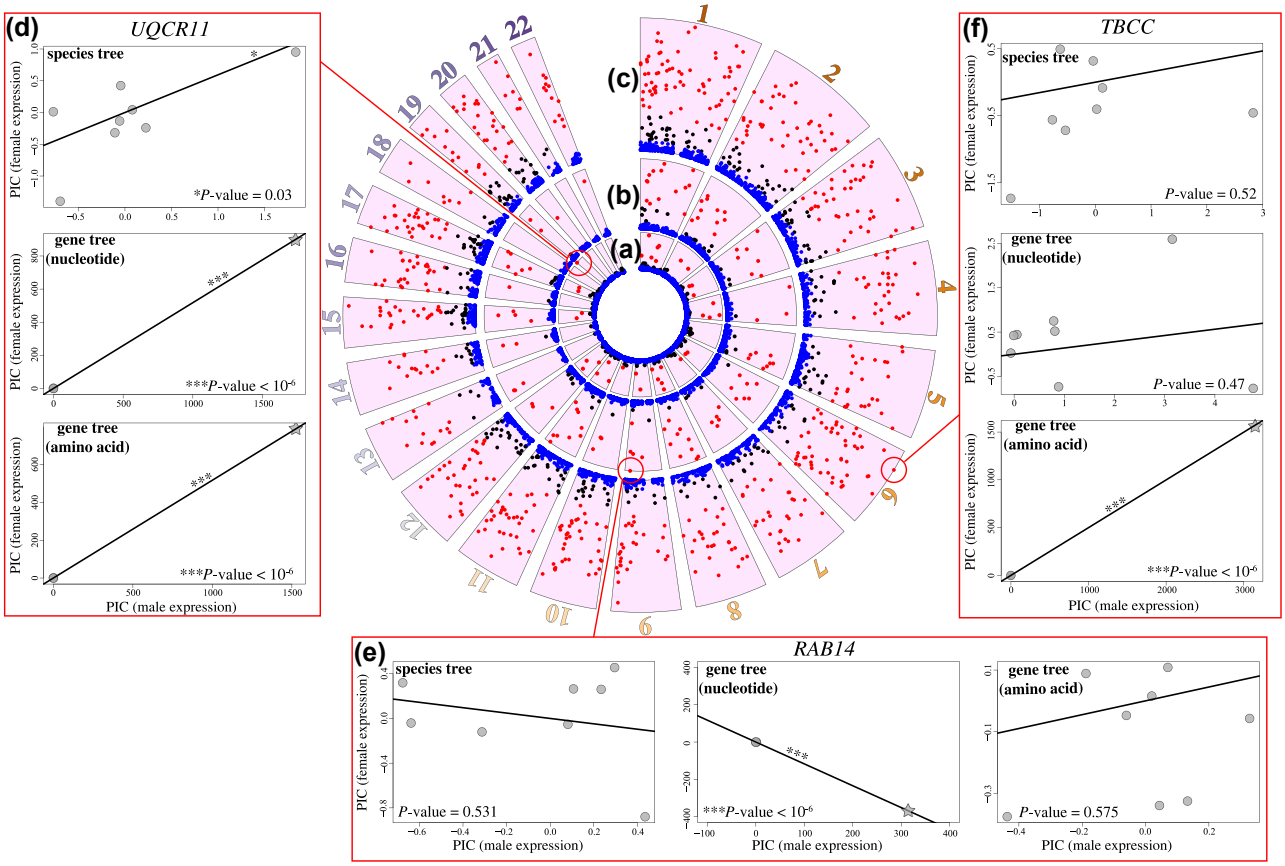

Fig. 6. Tree choice matters when testing female–male expression associations across species. Results shown across 22 autosomes for heart tissue expression measurements of 4,068 genes with measurable expression across species, with computed distance statistics d_S (inner track a), d_N(middle track b), and d_A (outer track c) based on L2-based phylogenetic regression. Empirical case studies comparing phylogenetic regression based on the species tree (d), nucleotide gene tree (e), and amino acid gene tree (f) are shown for analyses with anomalously high d_S, d_N, and d_A, respectively. Colors of points in circos plot (a to c) indicate relative level of divergent *P*-values, with blue indicating not significant, black indicating *P*-value <0.05, and red indicating strong outliers with *P*-value $<1.229 \times 10^{-5}$ after applying Bonferroni correction (Bonferroni 1936). Points depicted as gray stars indicate evidence of singular phylogenetic outliers found in specific analyses (d to f).

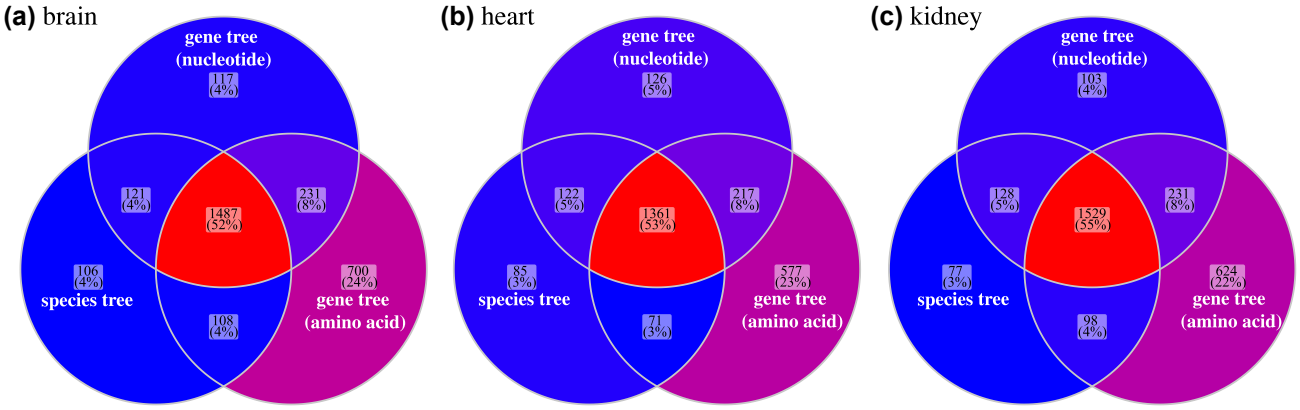

Fig. 7. Venn diagrams displaying the percentage of overlap in statistically significant genes for brain (a), heart (b), and kidney (c) expression levels in a mammalian dataset based on phylogenetic regression applied by assuming the species tree (left circles), nucleotide gene tree (top circles), or amino acid gene tree (right circles). Colors indicate the relative percentage of statistically significant genes across analyses.

rather than known trees, we still found relatively lower false positive rates when L1 phylogenetic regression (Fig. 9d), albeit to a lesser degree. Reflecting on our empirical case studies, we found several interesting differences between robust L1-based and conventional L2-based regression when assuming different trees (Fig. 10). In several examples, L1-based regression yielded comparatively smaller *P*-values (i.e. higher significance), sometimes leading to the inferences of statistically significant relationships not identified by L2-based regression (Fig. 10a to c). In others, L1-based regression returned

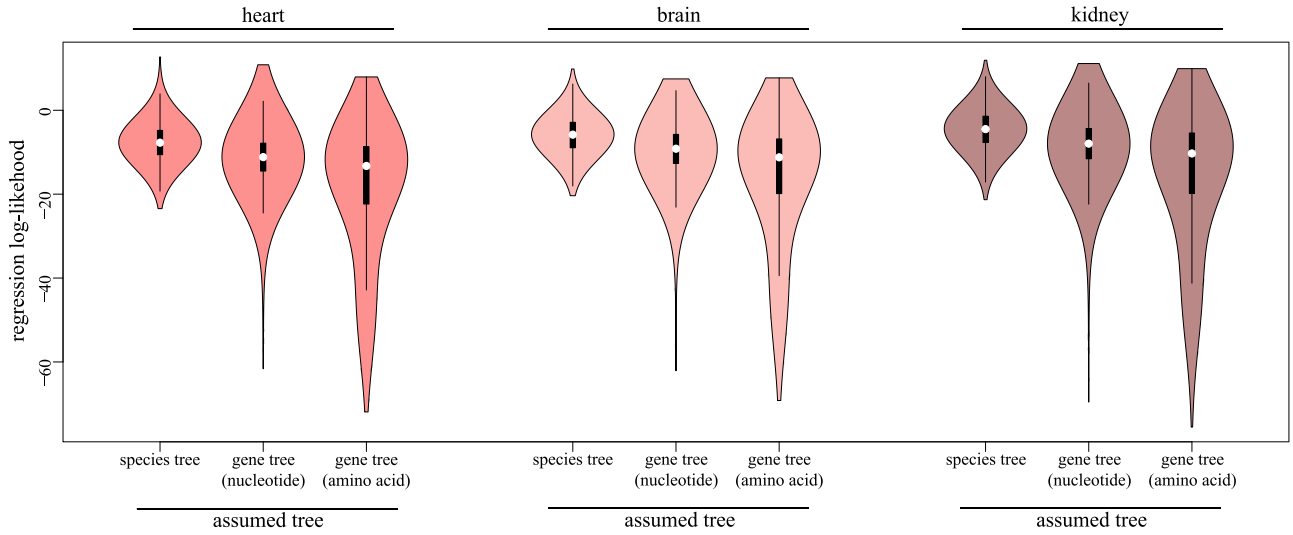

Fig. 8. Violin plots summarizing the distributions of model fit measured by log-likelihood for phylogenetic regression applied to gene expression from a mammalian dataset. Results shown across tissues (heart, brain, and kidney) and the three regression strategies that assume either the species tree, nucleotide gene tree, or amino acid gene tree.

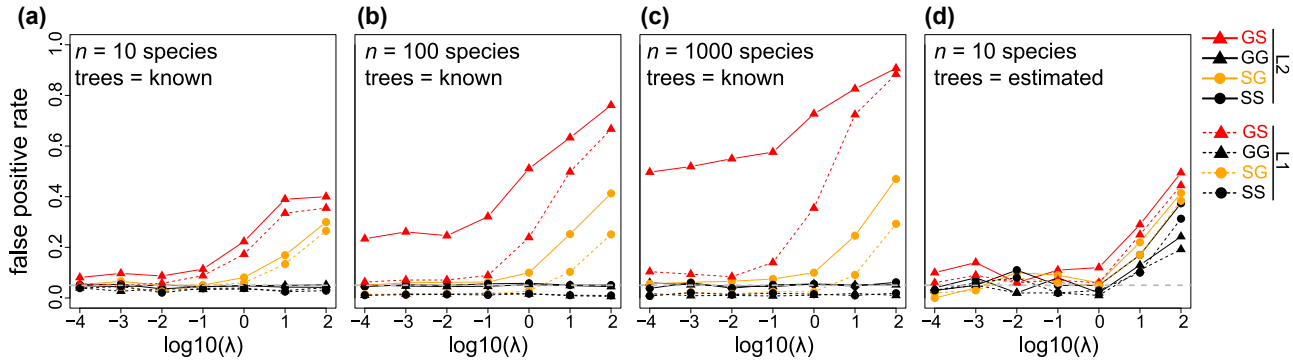

Fig. 9. Can robust estimators help? Results showing estimated false positive rates when using known trees with 10 species (a), 100 species (b), and 1,000 species (c) for robust L1-based regression (dashed lines) alongside standard L2-based regression (solid lines) under birth–death simulations with birth rate λ, death rate $\lambda/2$, and root age of 10 coalescent units. Estimated false positive rates are also shown for L1- and L2-based regression with estimated trees for our simulation case study with $n = 10$ species (d). Horizontal solid gray lines mark the typically accepted false positive rate of 0.05.

comparatively larger *P*-values, such that it did not find evidence of a significant relationship that was, however, inferred by L2-based regression (Fig. 10d to f).

Discussion

A choice of trees is always required when conducting phylogenetic regression. Yet deciding on a particular tree is often difficult and unlikely to become easier anytime soon. Collectively, our analyses underscore these challenges and expand our understanding of potential pitfalls of incorrect choices. To summarize, assuming the wrong tree may lead us to overestimate associations between traits that are truly independent—regardless of whether we are considering shallow or deep trees, few or many species, simple or complex architectures, or known or estimated trees. That is, tree choice matters.

Our study is the first to present these findings for phylogenetic regression, and thus we focus on mismatch resulting from gene tree-species tree discordance—a topic that has held our field captive for decades. Perhaps Hahn and Nakhleh (2016) stated it best: "The problems caused by ignoring variation in gene tree topologies are manifest because these genes underlie variation in the traits we are studying". Examining other potential sources of conflict (e.g. recombination, selection, and introgression) is arguably a worthwhile next step to incorporate other realistic processes encountered in empirical data. Moreover, future studies with expanded simulations will help us better understand the simultaneous effects of tree mismatch and estimation error (i.e. expanding results shown in Fig. 4), though it is worth emphasizing the computationally intensive and expensive demands of multilayered analyses spanning simulations and inferences of species trees, gene trees, sequence alignments, and trait evolution.

Mirroring similar evolutionary analyses (Hahn and Nakhleh 2016; Mendes and Hahn 2016; Guerrero and Hahn 2018; Mendes et al. 2018, 2019; Hibbins et al. 2020, 2023), tests of trait associations are sensitive to tree conflict. Speciation rate was an important factor in determining the degree of severity, with faster rates (yielding higher false positive rates). Because the expected length of internal branches (i.e. time between speciation events) is inversely related to speciation rate, faster rates yield shorter internal branches, which in turn can amplify phylogenetic conflict by providing less time for coalescent events in ancestral branches. This phylogenetic conflict is also reflected in the distance between the true and assumed tree, such that larger Robinson–Foulds and Hellinger distances were associated with higher false positive rates. Increasing the sample size (i.e. increasing the number of

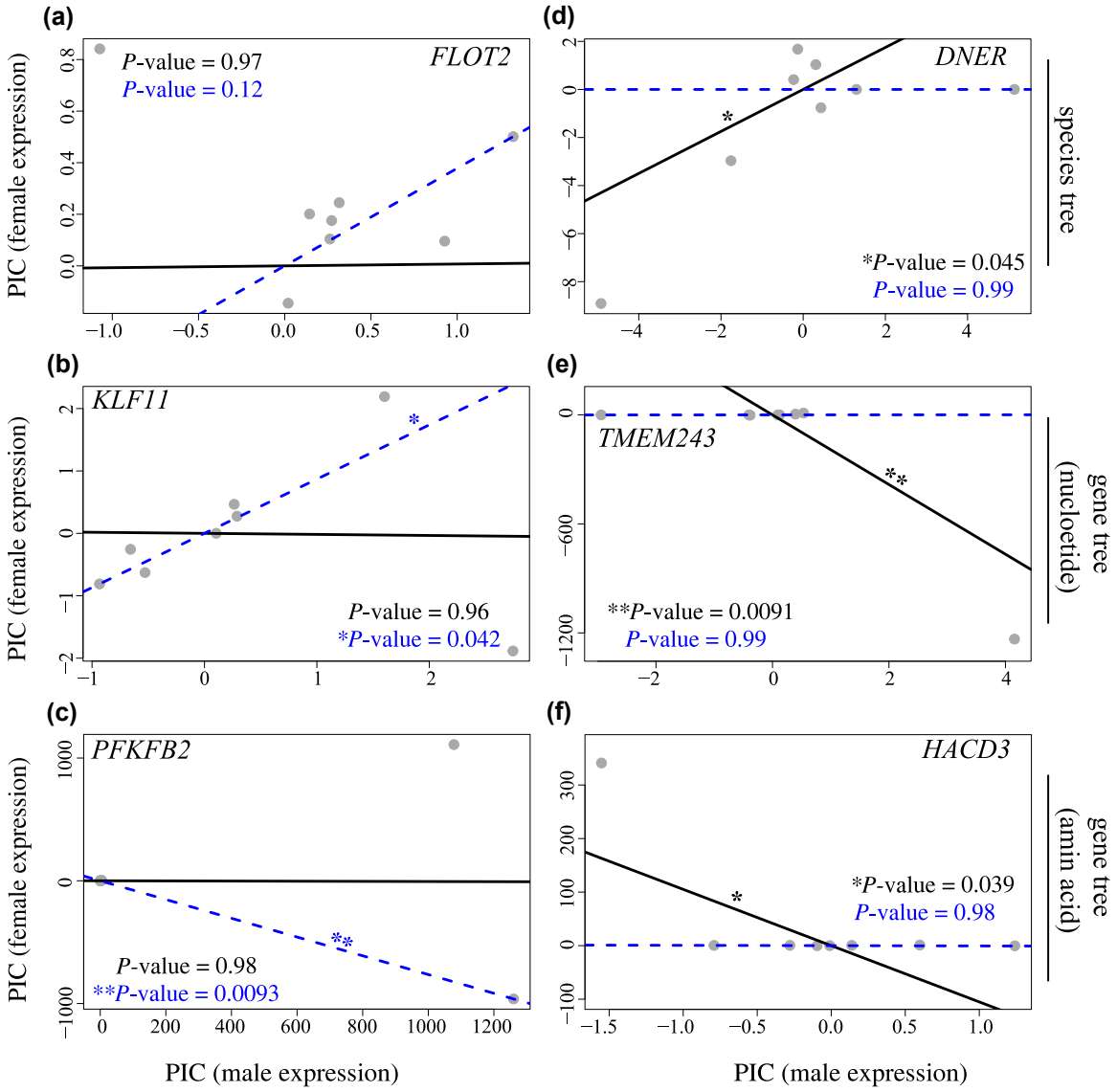

Fig. 10. Can robust estimators help with tree mismatch? Empirical examples from the mammalian gene expression data contrasting differences between standard L2-based (black lines) and robust L1-based (blue dashed lines) regression using the species tree (top row), nucleotide gene trees (middle row), and amino acid gene trees (bottom row).

species) only made the situation worse, and yet small trees were certainly not immune. Statistical power to detect true trait associations appeared less affected by tree mismatch than false positive rates, though future studies will be needed to better understand some of the patterns uncovered here. This phenomenon of increasing true positive rate in some scenarios (Fig. 5) likely reflects the compounding effects of tree-trait mismatch and, which each contributes to signals of statistical associations between traits.

Another surprising trend emerged when specifically contrasting the two mismatched models: false positive rates tended to be higher with GS than SG. That is, incorrectly assuming a species tree for traits simulated under a gene tree tended to be worse than the opposite. Neither mismatched model performed well, and yet our findings suggest that assuming an incorrect gene tree may represent a potential lesser of two evils in our explored scenarios. Dissecting this pattern further by comparing PIC magnitudes for tree cherries (i.e. nodes with exactly two extant descendants) provided evidence of increasingly larger contrasts for GS than GG regression (supplementary fig. S6, Supplementary Material online). Our simulations with complex architectures continued these results: incorrectly assuming the species tree nearly always amplified false positive rates, as did assuming an incorrect set of gene trees that were unrelated to studied traits. Artificially short branch lengths will inflate the influence of affected contrasts (Stone 2011), which is relevant to our findings here because branches in the species tree tend to be shorter than those of embedded gene trees.

Clearly, the reliability of an assumed tree is a major determinant of the reliability of an evolutionary hypothesis test. In an ideal world, one would always match the tree to the trait perfectly (i.e. GG and SS), but this is neither always possible nor probable. When designing this study, we first focused on using known trees to isolate and understand the behavior of mismatched regression. We then realized that we needed to consider an elephant in the room: in practice, phylogenetic regression is conducted using estimated rather than known trees. Of course, we seldom (if ever) estimate a phylogeny to perfection, and our findings argue for increased vigilance against both tree mismatch and estimation error. Gene tree

discordance was generally high in our simulation experiments (supplementary fig. S7, Supplementary Material online), which also likely influenced the accuracy of estimated species trees across the range of speciation rates explored here (supplementary figs. S8 and S9, Supplementary Material online). Future simulation studies seeking to fully explore the scope and scale of tree mismatch and estimation error are likely to be valuable and yet quite demanding. Though focused for computational feasibility, our simulations nonetheless argue that mismatch and estimation error are important, and we found evidence of alarming biases in the presence of both.

Building on our simulation-based investigations, we explored a functional genomic dataset to investigate impacts of tree choice when modeling female–male expression relationships across species. Because gene sequence and expression divergence are correlated (Duret and Mouchiroud 2000; Pál et al. 2001; Subramanian and Kumar 2004; Lemos et al. 2005; Assis and Kondrashov 2014), expression is typically assumed to evolve more or less according to an associated local gene tree. Thus, we used our empirical case study as a best-case scenario in which one might have a fighting chance of matching tree with trait. Perhaps most apparent in these analyses is the potential for stark differences in regression significance depending on the assumed tree. In this case, the empirical findings paint a somewhat different picture than what we observed from our simulations, showing that assuming the species tree was often most conservative. Though it unfortunately can be difficult to achieve synchrony between simulated and empirical results, we suspect that a number of factors could be at play here, including the complexities of regulatory mechanisms contributing to gene expression evolution, as well as the estimation of both gene and species trees. Clearly, the choice of a tree matters even in these scenarios. Our comparisons of tree distances may help explain some of these findings, as the lowest distances were observed between nucleotide gene trees and the species tree (supplementary fig. S10, Supplementary Material online). Likewise, quantile–quantile (QQ) plot comparison of *P*-value distributions also suggests differences in inferred significance based on the tree chosen for phylogenetic regression (supplementary fig. S11, Supplementary Material online). Given that mutation and recombination rates are on similar scales in mammals (McVean et al. 2004; Keightley and Eyre-Walker 2007; Kong et al. 2010), the propensity for intragenic recombination events is likely, violating another standard assumption of phylogenetic inference.

In light of our findings, it is interesting to consider the mechanisms underlying variation in traits and their phylogenetic architectures. Popular models of continuous trait evolution based on extensions of Brownian motion are designed to capture phenomena affecting the mean and variance of traits within a lineage (Felsenstein 1988; Revell and Harmon 2008; Blomberg et al. 2020). Thus, assuming the overall species tree might be justifiable for traits that adhere to canonical assumptions of quantitative genetic models. Recently, studies have also argued for more mechanistic frameworks in which traits are encoded by architectures composed of a single or perhaps multiple gene trees under a neutral model of evolution (Hibbins et al. 2023; Schraiber et al. 2024). Natural selection, however, acts directly on variation in traits, and therefore indirectly on the genealogical history and architecture encoding the traits (Lande 1976). How PCMs behave under such conditions remains an open question, and models that incorporate the ancestral selection graph may prove helpful here, though it is worth noting the computational difficulties involved (Krone and Neuhauser 1997; Brandt et al. 2024).

Flaws are often much easier to find than solutions; seldom is it satisfying to simply point out issues without offering at least a hope of a remedy. We found that to be the case here. While the primary purpose of this study was to provide a first perspective on the dangers of tree mismatch, we also explored the promise of robust phylogenetic regression, which improved inferences for both known and estimated trees. Additionally, we illustrated several examples of large differences between *P*-values obtained with L1- and L2-based regression, most of which altered conclusions about tested relationships between male and female expression. Our findings suggest that robust estimators might provide a potential, albeit imperfect, solution to some issues raised by tree mismatch. We can say with confidence that robust phylogenetic regression was never meant to be a panacea for all ailments that might afflict PCMs. Progress—not perfection—is the goal, and more studies are needed to explore the possibilities and space of phylogenetic mismatch and the potential for different types of robust estimators with different types of model violations. Comparisons of log-likelihoods of matched and mismatched models also may hold clues for comparing phylogenetic regression model fit (Fig. 8 and supplementary fig. S12, Supplementary Material online). Future studies that employ both robust estimators and other recent advances in phylogenetic modeling (e.g. phylogenomic comparative methods; Hibbins et al. 2023) may prove helpful in this context. Additionally, strategies for addressing evolutionary uncertainty (de Villemereuil et al. 2012; Fuentes-G et al. 2020; Bastide et al. 2021; Zhang et al. 2024) may be promising, though such approaches are not widely applied for such purposes and may still be sensitive to hemiplasy (Hahn and Nakhleh 2016; see Supplementary Case Study section and supplementary fig. S13, Supplementary Material online). Altogether, our findings underscore the difficulties of phylogenetic regression with uncertain trees and call for increased vigilance against phylogenetic mismatch—whether due to ILS, estimation error, or otherwise.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgements

We thank Joshua Schraiber and two anonymous reviewers for their helpful comments. This research was also supported by the Arkansas High Performance Computing Center, which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission, as well as Research Computing at Florida Atlantic University.

Funding

This work was supported by the National Science Foundation grants DBI-2130666, DEB-2302258, DEB-2392257, and BCS-2001063, the National Institutes of Health grants R35 GM142438 and R35 GM128590, start-up funds provided by the University of Arkansas, and a grant from the Arkansas Bioscience Institute.

Conflict of Interest

The authors of this study do not report any conflicts of interest.

Data Availability

The data underlying this article are available in the article and in its online Supplementary material.

References

Adams DC. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. *Syst Biol*. 2013:62(2):181–192. https://doi.org/10.1093/sysbio/sys083.

Adams RH, Blackmon H, DeGiorgio M. Of traits and trees: probabilistic distances under continuous trait models for dissecting the interplay among phylogeny, model, and data. *Syst Biol*. 2021:70(4):660–680. https://doi.org/10.1093/sysbio/syab009.

Adams RH, Blackmon H, Reyes-Velasco J, Schield DR, Card DC, Andrew AL, Waynewood N, Castoe TA. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome*. 2016:59(5):295–310. https://doi.org/10.1139/gen-2015-0124.

Adams RH, Cain Z, Assis R, DeGiorgio M. Robust phylogenetic regression. *Syst Biol*. 2024:73(1):140–157. https://doi.org/10.1093/sysbio/syad070.

Adams RH, Schield DR, Card DC, Castoe TA. Assessing the impacts of positive selection on coalescent-based species tree estimation and species delimitation. *Syst Biol*. 2018:67(6):1076–1090. https://doi.org/10.1093/sysbio/syy034.

Al-Kahtani MA, Zuleta C, Caviedes-Vidal E Jr, Garland T. Kidney mass and relative medullary thickness of rodents in relation to habitat, body size, and phylogeny. *Physiol Biochem Zool*. 2004:77(3):346–365. https://doi.org/10.1086/420941.

Assis R. Lineage-specific expression divergence in grasses is associated with male reproduction, host-pathogen defense, and domestication. *Genome Biol Evol*. 2019:11(1):207–219. https://doi.org/10.1093/gbe/evy245.

Assis R, Kondrashov AS. Conserved proteins are fragile. *Mol Biol Evol*. 2014:31(2):419–424. https://doi.org/10.1093/molbev/mst217.

Avise JC, Robinson TJ. Hemiplasy: a new term in the lexicon of phylogenetics. *Syst Biol*. 2008:57(3):503–507. https://doi.org/10.1080/10635150802164587.

Bastide P, Ho LST, Baele G, Lemey P, Suchard MA. Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. *Ann Appl Stat*. 2021:15(2):971–997. https://doi.org/10.1214/20-AOAS1419.

Bastide P, Soneson C, Stern DB, Lespinet O, Gallopin M. A phylogenetic framework to simulate synthetic interspecies RNA-seq data. *Mol Biol Evol*. 2023:40(1):269. https://doi.org/10.1093/molbev/msac269.

Beaulieu JM, Jhwueng D, Boettiger C, O'Meara BC. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. *Evolution*. 2012:66(8):2369–2383. https://doi.org/10.1111/j.1558-5646.2012.01619.x.

Bertram J, Fulton B, Tourigny J, Pena-Garcia Y, Moyle LC, Hahn MW. CAGEE: computational analysis of gene expression evolution. *Mol Biol Evol*. 2023:40(5):5. https://doi.org/10.1093/molbev/msad106.

Blanchette M, Green ED, Miller W, Haussler D. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*. 2004:14(12):2412–2423.

Blomberg SP, Garland T Jr, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*. 2003:57(4):717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M. Independent contrasts and PGLS regression estimators are equivalent. *Syst Biol*. 2012:61(3):382–391. https://doi.org/10.1093/sysbio/syr118.

Blomberg SP, Rathnayake SI, Moreau CM. Beyond Brownian motion and the Ornstein-Uhlenbeck process: stochastic diffusion models for the evolution of quantitative characters. *Am Nat*. 2020:195(2):145–165. https://doi.org/10.1086/706339.

Boettiger C, Coop G, Ralph P. Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*. 2012:66(7):2240–2251. https://doi.org/10.1111/j.1558-5646.2011.01574.x.

Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. *Pubbl del R Ist Super di Sci Econ e Commericiali di Firenze*. 1936:8:3–62.

Borges R, Boussau B, Szöllosi GJ, Kosiol C. Nucleotide usage biases distort inferences of the species tree. *Genome Biol Evol*. 2022:14(1):evab290.

Brahmantio B, Bartoszek K, Yapar E. Bayesian inference of mixed Gaussian phylogenetic models. *arXiv:2410.11548*. 2024.

Brandt DY, Huber CD, Chiang CW, Ortega-Del Vecchyo D. The promise of inferring the past using the ancestral recombination graph. *Genome Biol Evol*. 2024:16(2):evae005. https://doi.org/10.1093/gbe/evae005.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, *et al*. The evolution of gene expression levels in mammalian organs. *Nature*. 2011:478:343–348.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. 2014:513(7518):375–381. https://doi.org/10.1038/nature13726.

Chen F, Li Z, Zhang X, Wu P, Yang W, Yang J, Chen X, Yang JR. Phylogenetic comparative analysis of single-cell transcriptomes reveals constrained accumulation of gene expression heterogeneity during clonal expansion. *Mol Biol and Evol*. 2023:40(5):5. https://doi.org/10.1093/molbev/msad113.

Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, Haerty W, Di Palma F, Regev A. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res*. 2019:29(1):53–63. https://doi.org/10.1101/gr.237636.118.

Cressler CE, Butler MA, King AA. Detecting adaptive evolution in phylogenetic comparative analysis using the Ornstein–Uhlenbeck model. *Syst Biol*. 2015:64(6):953–968. https://doi.org/10.1093/sysbio/syv043.

DeGiorgio M, Rosenberg NA. Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. *Theor Popul Biol*. 2016:110:12–24. https://doi.org/10.1016/j.tpb.2016.02.002.

Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 2009:24(6):332–340. https://doi.org/10.1016/j.tree.2009.01.009.

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evol Biol*. 2012:12(1):1–16. https://doi.org/10.1186/1471-2148-12-102.

Diaz-Uriarte R, Garland T Jr. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Syst Biol*. 1996:45(1):27–47. https://doi.org/10.1093/sysbio/45.1.27.

Diaz-Uriarte R, Garland T Jr. Effects of branch length errors on the performance of phylogenetically independent contrasts. *Syst Biol*. 1998:47(4):654–672. https://doi.org/10.1080/106351598260653.

Dimayacyac JR, Wu S, Pennell M. Evaluating the performance of widely used phylogenetic models for gene expression evolution. *Genome Biol Evol*. 2023:15(12):12. https://doi.org/10.1093/gbe/evad211.

Doña J, Johnson KP. Host body size, not host population size, predicts genome-wide effective population size of parasites. *Evol Lett*. 2023:7(4):285–292. https://doi.org/10.1093/evlett/qrad026.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, *et al*. The UCSC Genome Browser database: extensions and updates 2011. *Nucl Acid Res*. 2012:40(D1):D918–D923.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Natl Acad Sci U S A*. 2018:115(3):E409–E417. https://doi.org/10.1073/pnas.1707515115.

Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 2000:17(1):68–70. https://doi.org/10.1093/oxfordjournals.molbev.a026239.

Edwards SV. Is a new and general theory of molecular systematics emerging? *Evolution (N Y)*. 2009:63:1–19.

Felenstein J. *Inferring phylogenies*. Sunderland, MA: Sinauer associates; 2004.
Felsenstein J. Phylogenies and the comparative method. *Am Nat*. 1985:125(1):1–15. https://doi.org/10.1086/284325.
Felsenstein J. Phylogenies and quantitative characters. *Annu Rev Ecol Syst*. 1988:10(1):445–471. https://doi.org/10.1146/annurev.es.19.110188.002305.
Fuentes-G JA, Polly PD, Martins EP. A Bayesian extension of phylogenetic generalized least squares: incorporating uncertainty in the comparative study of trait relationships and evolutionary rates. *Evolution*. 2020:74(2):311–325. https://doi.org/10.1111/evo.13899.
Gardner JD, Organ CL. Evolutionary sample size and consilience in phylogenetic comparative analysis. *Syst Biol*. 2021:70(5):1061–1075. https://doi.org/10.1093/sysbio/syab017.
Gittleman JL, Luh HK. On comparing comparative methods. *Ann Rev Ecol Syst*. 1992:23(1):383–404. https://doi.org/10.1146/annurev.es.23.110192.002123.
Grafen A. The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci. 1989:326(1233):119–157. https://doi.org/10.1098/rstb.1989.0106.
Gu X. Understanding tissue expression evolution: from expression phylogeny to phylogenetic network. *Brief Bioinform*. 2016:17(2):249–254. https://doi.org/10.1093/bib/bbv041.
Guerrero RF, Hahn MW. Quantifying the risk of hemiplasy in phylogenetic inference. *Proc Natl Acad Sci U S A*. 2018:115(50):12787–12792. https://doi.org/10.1073/pnas.1811268115.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systemat Biol*. 2010:59(3):307–321.
Hahn MW, Nakhleh L. Irrational exuberance for resolved species trees. *Evolution*. 2016:70(1):7–17. https://doi.org/10.1111/evo.12832.
Hansen TF. Stabilizing selection and the comparative analysis of adaptation. *Evolution (N Y)*. 1997:51(5):1341–1351. https://doi.org/10.2307/2411186.
Harmon L. *Phylogenetic comparative methods: learning from trees*. North Charleston, SC: CreateSpace Independent Publishing Platform; 2019. ISBN-13:978-1719584463.
Harvey PH, Pagel MD. *The comparative method in evolutionary biology*. Oxford: Oxford university press; 1991.
He C, Liang D, Zhang P. Asymmetric distribution of gene trees can arise under purifying selection if differences in population size exist. *Mol Biol Evol*. 2020:37(3):881–892. https://doi.org/10.1093/molbev/msz232.
Hensen N, Bonometti L, Westerberg I, Brännström IO, Guillou S, Cros-Aarteil S, Calhoun S, Haridas S, Kuo A, Mondo S, *et al*. Genome-scale phylogeny and comparative genomics of the fungal order Sordariales. *Mol Phys Evol*. 2023:189:107938. https://doi.org/10.1016/j.ympev.2023.107938.
Hibbins MS, Breithaupt LC, Hahn MW. Phylogenomic comparative methods: accurate evolutionary inferences in the presence of gene tree discordance. *Proc Natl Acad Sci U S A*. 2023:120(22):e2220389120. https://doi.org/10.1073/pnas.2220389120.
Hibbins MS, Gibson MJS, Hahn MW. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *Elife*. 2020:9:e63753. https://doi.org/10.7554/eLife.63753.
Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res*. 2011:21(3):349–356. https://doi.org/10.1101/gr.114751.110.
Jiang X, Assis R. Population-specific genetic and expression differentiation in Europeans. *Genome Biol Evol*. 2020:12(4):358–369. https://doi.org/10.1093/gbe/evaa021.
Kamilar JM, Cooper N. Phylogenetic signal in primate behaviour, ecology and life history. *Philos Trans R Soc B: Biol Sci*. 2013:368(1618):20120341. https://doi.org/10.1098/rstb.2012.0341.
Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*. 2007:177(4):2251–2261. https://doi.org/10.1534/genetics.107.080663.
Koch H, DeGiorgio M. Maximum Likelihood estimation of species trees from gene trees in the presence of ancestral population structure. *Genome Biol Evol*. 2020:12(2):3977–3995. https://doi.org/10.1093/gbe/evaa022.
Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, *et al*. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*. 2010:467(7319):1099–1103. https://doi.org/10.1038/nature09525.
Krone SM, Neuhauser C. Ancestral processes with selection. *Theor Popul Biol*. 1997:51(3):210–237. https://doi.org/10.1006/tpbi.1997.1299.
Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 2007:56(1):17–24. https://doi.org/10.1080/10635150601146041.
Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Mol Biol Evol*. 2014:31(8):2004–2017. https://doi.org/10.1093/molbev/msu186.
Lande R. Natural selection and random genetic drift in phenotypic evolution. *Evolution*. 1976:1(2):314–334. https://doi.org/10.2307/2407703.
Leaché AD, Harris RB, Rannala B, Yang Z. The influence of gene flow on species tree estimation: a simulation study. *Syst Biol*. 2014:63(1):17–30. https://doi.org/10.1093/sysbio/syt049.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 2005:22(5):1345–1354. https://doi.org/10.1093/molbev/msi122.
Liu L, Xi Z, Wu S, Davis CC, Edwards SV. Estimating phylogenetic trees from genome-scale data. *Ann N Y Acad Sci*. 2015:1360(1):36–53. https://doi.org/10.1111/nyas.12747.
Liu L, Yu L. Phybase: an R package for species tree analysis. *Bioinformatics*. 2010:26(7):962–963. https://doi.org/10.1093/bioinformatics/btq062.
Long C, Kubatko L. The effect of gene flow on coalescent-based species-tree inference. *Syst Biol*. 2018:67(5):770–785. https://doi.org/10.1093/sysbio/syy020.
Maddison WP. Gene trees in species trees. *Syst Biol*. 1997:46(3):523–536. https://doi.org/10.1093/sysbio/46.3.523.
Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol*. 2015:64(1):127–136. https://doi.org/10.1093/sysbio/syu070.
Martins EP. Phylogenies, spatial autoregression, and the comparative method: a computer simulation test. *Evolution*. 1996:50(5):1750–1765. https://doi.org/10.2307/2410733.
Martins EP, Garland T Jr. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution*. 1991:45(3):534–557. https://doi.org/10.2307/2409910.
Martins EP, Hansen TF. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat*. 1997:149(4):646–667. https://doi.org/10.1086/286013.
Mazel F, Davies TJ, Georges D, Lavergne S, Thuiller W, Peres-Neto PR. Improving phylogenetic regression under complex evolutionary models. *Ecology*. 2016:97(2):286–293. https://doi.org/10.1890/15-0086.1.
McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science*. 2004:304(5670):581–584. https://doi.org/10.1126/science.1092500.
Mendes FK, Fuentes-González JA, Schraiber JG, Hahn MW. A multispecies coalescent model for quantitative traits. *Elife*. 2018:7:e36482. https://doi.org/10.7554/eLife.36482.
Mendes FK, Hahn MW. Gene tree discordance causes apparent substitution rate variation. *Syst Biol*. 2016:65(4):711–721. https://doi.org/10.1093/sysbio/syw018.
Mendes FK, Hahn Y, Hahn MW. Gene tree discordance can generate patterns of diminishing convergence over time. *Mol Biol Evol*. 2016:33(12):3299–3307. https://doi.org/10.1093/molbev/msw197.

Mendes FK, Livera AP, Hahn MW. The perils of intralocus recombination for inferences of molecular convergence. *Philos Trans R Soc B*. 2019:374(1777):20180244. https://doi.org/10.1098/rstb.2018.0244.
Miles DB, Dunham AE. Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. *Annu Rev Ecol Syst*. 1993:24(1):587–619. https://doi.org/10.1146/annurev.es.24.110193.003103.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020:37(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.
Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008:5(7):621–628.
Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, *et al*. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res*. 2021:49(D1):D1046–D1057. https://doi.org/10.1093/nar/gkaa1070.
Nichols R. Gene trees and species trees are not the same. *Trends Ecol Evol*. 2001:16(7):358–364. https://doi.org/10.1016/S0169-5347(01)02203-0.
O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. Testing for different rates of continuous trait evolution using likelihood. *Evolution*. 2006:60:922–933.
Pagel M. Inferring evolutionary processes from phylogenies. *Zool Scr*. 1997:26(4):331–348. https://doi.org/10.1111/j.1463-6409.1997.tb00423.x.
Pagel M. Inferring the historical patterns of biological evolution. *Nature*. 1999:401(6756):877–884. https://doi.org/10.1038/44766.
Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics*. 2001:158(2):927–931. https://doi.org/10.1093/genetics/158.2.927.
Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019:35(3):526–528. https://doi.org/10.1093/bioinformatics/bty633.
Pardo L. *Statistical inference based on divergence measures*. Boca Raton, FL: Chapman and Hall/CRC; 2005.
Pei J, Wu Y. STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. *Bioinformatics*. 2017:33(12):1789–1797. https://doi.org/10.1093/bioinformatics/btx079.
Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. Geiger v2. 0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*. 2014:30(15):2216–2218. https://doi.org/10.1093/bioinformatics/btu181.
Pennell MW, Harmon LJ. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann N Y Acad Sci*. 2013:1289(1):90–105. https://doi.org/10.1111/nyas.12157.
Pinheiro J, Bates D, DebRoy S, Sarkar D, Heisterkamp S, Van Willigen B, Maintainer R. 2017. Package 'nlme'. Linear Nonlinear Mixed Effects Models. version, 3(1):274.
Revell LJ. Phylogenetic signal and linear regression on species data. *Methods Ecol Evol*. 2010:1(4):319–329. https://doi.org/10.1111/j.2041-210X.2010.00044.x.
Revell LJ, Harmon LJ. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evol Ecol Res*. 2008:10:311–331.
Revell LJ, Harmon LJ, Collar DC. Phylogenetic signal, evolutionary process, and rate. *Syst Biol*. 2008:57(4):591–601. https://doi.org/10.1080/10635150802302427.
Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981:53(1-2):131–147. https://doi.org/10.1016/0025-5564(81)90043-2.
Rohlf FJ. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*. 2001:55(11):2143–2160.
Rohlfs RV, Nielsen R. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Syst Biol*. 2015:64(5):695–708. https://doi.org/10.1093/sysbio/syv042.
Ross CF, Henneberg M, Ravosa MJ, Richard S. Curvilinear, geometric and phylogenetic modeling of basicranial flexion: is it adaptive, is it constrained? *J Hum Evol*. 2004:46(2):185–213. https://doi.org/10.1016/j.jhevol.2003.11.001.
Rousseeuw P, Yohai V. *Robust regression by means of S-estimators. Robust and nonlinear time series analysis*. New York: Springer; 1984. p. 256–272.
Sanford GM, Lutterschmidt WI, Hutchison VH. The comparative method revisited. *Bioscience*. 2002:52(9):830–836. https://doi.org/10.1641/0006-3568(2002)052[0830:TCMR]2.0.CO;2.
Schluter D. Uncertainty in ancient phylogenies. *Nature*. 1995:377(6545):108–109. https://doi.org/10.1038/377108a0.
Schraiber JG, Edge MD, Pennell M. Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. *PLoS Biol*. 2024:22(10):10. https://doi.org/10.1371/journal.pbio.3002847.
Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics*. 2004:1(4):274–286. https://doi.org/10.1186/1479-7364-1-4-274.
Slatkin M, Pollack JL. Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol*. 2008:25(10):2241–2246. https://doi.org/10.1093/molbev/msn172.
Solís-Lemus C, Yang M, Ané C. Inconsistency of species tree methods under gene flow. *Syst Biol*. 2016:65(5):843–851. https://doi.org/10.1093/sysbio/syw030.
Stadler T. Simulating trees with a fixed number of extant species. *Syst Biol*. 2011:60(5):676–684. https://doi.org/10.1093/sysbio/syr029.
Stone EA. Why the phylogenetic regression appears robust to tree misspecification. *Syst Biol*. 2011:60(3):245–260. https://doi.org/10.1093/sysbio/syq098.
Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*. 2004:168(1):373–381. https://doi.org/10.1534/genetics.104.028944.
Symonds MRE. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. *Syst Biol*. 2002:51(4):541–553. https://doi.org/10.1080/10635150290069977.
Tian Y, Kubatko LS. Distribution of coalescent histories under the coalescent model with gene flow. *Mol Phylo Evol*. 2016:105:177–192. https://doi.org/10.1016/j.ympev.2016.08.024.
Uyeda JC, Zenil-Ferguson R, Pennell MW. Rethinking phylogenetic comparative methods. *Syst Biol*. 2018:67(6):1091–1109. https://doi.org/10.1093/sysbio/syy031.
Villemereuil PD, Wells JA, Edwards RD, Blomberg SP. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evolut Biol*. 2012:12:1–16.
Wascher M, Kubatko LS. On the effects of selection and mutation on species tree inference. *Mol Phylogenet Evol*. 2023:179:107650. https://doi.org/10.1016/j.ympev.2022.107650.
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, *et al*. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010:329(5987):75–78. https://doi.org/10.1126/science.1190371.
Yu Y, Than C, Degnan JH, Nakhleh L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst Biol*. 2011:60(2):138–149. https://doi.org/10.1093/sysbio/syq084.
Yule GU. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. *Philos Trans R Soc Lond*. 1925:213(402–410):21–87. https://doi.org/10.1098/rstb.1925.0002.
Zhang R, Drummond AJ, Mendes FK. Fast Bayesian inference of phylogenies from multiple continuous characters. *Syst Biol*. 2024:73(1):102–124.