Enhancing Phylogenetic Independent Contrasts: Addressing Abrupt Evolutionary Shifts with Outlier- and Distribution-Guided Correlation

Zheng-Lin Chen, Rui Huang, Hong-Ji Guo and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, 100875, China

*Correspondence to be sent to: *Deng-Ke Niu, College of Life Sciences, Beijing Normal University, Beijing, 100875, China*; Emails: dkniu@bnu.edu.cn, dengkeniu@hotmail.com

ABSTRACT

Phylogenetic comparative methods are indispensable for analyzing cross-species data while accounting for evolutionary relationships. Traditional methods like phylogenetically independent contrasts (PIC) and phylogenetic generalized least squares (PGLS) are well-suited for gradual evolution under Brownian motion (BM) or similar assumptions. However, their effectiveness diminishes under abrupt evolutionary shifts or heterogeneous evolutionary processes. We introduce a hybrid framework, outlier-guided correlation (OGC) for large datasets and outlier- and distribution-guided correlation (ODGC) for small datasets, collectively termed O(D)GC, which integrates Pearson and Spearman correlation analyses to handle outliers and non-normal data dynamically. Using simulations across diverse evolutionary scenarios, we compared PIC-O(D)GC with a comprehensive range of methods, including eight robust regression approaches (PIC-MM, PIC-L1, PIC-S, PIC-M, and their PGLS counterparts); PGLS optimized using five evolutionary models: BM, lambda, Ornstein–Uhlenbeck random (OU-random), OU-fixed, and Early-burst; Corphylo (an OU-based method); PIC-Pearson; and two advanced models, phylogenetic generalized linear mixed models (PGLMM) and multi-response phylogenetic mixed models (MR-PMM). Our results demonstrate that PIC-O(D)GC and PIC-MM consistently exhibit superior performance under conditions with evolutionary shifts, maintaining low false positive rates and high accuracy. In no-shift scenarios, PIC-O(D)GC and PIC-MM perform comparably to other reliable methods, underscoring their adaptability and robustness across varied datasets. Furthermore, PIC-O(D)GC offers computational efficiency and flexibility, making it particularly suitable for large-scale datasets. These findings underscore the value of adaptive, nonparametric frameworks like PIC-O(D)GC in complementing traditional phylogenetic approaches. By bridging computational simplicity with methodological robustness, O(D)GC provides a versatile tool for trait correlation analysis in both stable and dynamic evolutionary

contexts.

KEYWORDS

Phylogenetic comparative methods; Phylogenetically independent contrasts (PIC); Outlier-guided correlation (OGC); Abrupt evolutionary shifts; Robust regression; Trait correlation; phylogenetic autocorrelation; Adaptive frameworks.

INTRODUCTION

Phylogenetically closer species tend to exhibit more similar traits due to their shared ancestry, resulting in phylogenetic autocorrelation among samples (Felsenstein 1985, Garamszegi 2014, Cornwallis and Griffin 2024). Ignoring this autocorrelation and using conventional statistical methods can lead to erroneous conclusions. To address this challenge, evolutionary biologists have developed a range of phylogenetic comparative methods that explicitly account for these dependencies (Grafen 1989, Lynch 1991, Garamszegi 2014, O'Meara 2016, Cornwallis and Griffin 2024).

The earliest method was phylogenetically independent contrasts (PIC) (Felsenstein 1985). The core principle behind PIC is that, while species at the tips of a phylogenetic tree are interdependent due to shared ancestry, the evolutionary changes along each branch can be treated as statistically independent events. In this context, PIC refers to the contrast values calculated between species pairs along the phylogenetic tree, inferring evolutionary changes along pairs of branches and standardizing these changes by branch length (or time) to generate transformed data points. These PIC values, typically subjected to ordinary least squares (OLS) regression, provide insight into the relationships between traits. This approach makes the evolutionary data more suitable for conventional statistical methods by removing the phylogenetic non-independence. Thus, PIC can refer both to the contrast values calculated from the phylogeny and to the subsequent OLS regression analysis applied to these values.

PIC assumes that traits evolve according to a Brownian motion (BM) model, where trait variation accumulates gradually and randomly over time. Strictly speaking, OLS regression of PIC is a special case of the phylogenetic generalized least squares (PGLS) method under the conditions of perfect BM (Rohlf 2001). PGLS extends this framework by incorporating a variance-covariance matrix for the residuals in the generalized linear model to account for phylogenetic autocorrelation among samples (Martins and Hansen 1997, Pagel 1997, Pagel

1999). This adjustment helps reduce false positives. Although some PGLS models, such as Pagel's λ, Ornstein–Uhlenbeck (OU), and Early-burst (EB) models, can handle heterogeneous evolutionary rates (Hansen 1997, Freckleton et al. 2002, Harmon et al. 2010, Ho and Ane 2014), they fundamentally assume gradual trait changes (Landis et al. 2013). This assumption can be problematic when dealing with evolutionary jumps caused by singular events.

During biological evolution, certain branches may experience rare and dramatic trait changes (Maddison and FitzJohn 2015). Increasingly, studies and data have shown that abrupt evolutionary shifts, or evolutionary jumps, have a significant impact on biodiversity evolution (Landis and Schraiber 2017, Gao and Wu 2022, Smith et al. 2023, Sumner et al. 2023). A typical example of this is Felsenstein's worst-case scenario (Felsenstein 1985, Uyeda et al. 2018), where two traits undergo a significant jump near the root of the phylogenetic tree while evolving independently on other branches. Even if this dramatic change occurred only once, it could still lead to erroneous conclusions of significant correlation between the two traits using ordinary linear regression on PIC values or PGLS models (Uyeda et al. 2018).

Recently, more studies have focused on identifying jump events and providing more reasonable explanations for the processes underlying trait evolution. Uyeda et al. (2018) underscore the need for models like the singular events model, which can accommodate the complexities and unpredictabilities inherent in evolutionary biology. Some studies suggest that jump events cause trait changes to exhibit a “fat-tails” distribution, using Levy processes to handle this distribution (Landis et al. 2013, Duchen et al. 2017). Others propose that abrupt shift events lead to outliers, using robust regression to reduce model sensitivity to violations (Slater and Pennell 2014, Adams et al. 2024). Additionally, phylogenetic generalized linear mixed models (PGLMM) and Multi-response phylogenetic mixed models (MR-PMM) offer flexible frameworks to address the complexities of phylogenetic data, including diverse data types and random effects (Ives and Helmus 2011, Westoby et al. 2023, Halliwell et al. 2024).

PGLMM models both fixed and random effects to account for phylogenetic autocorrelation, while MR-PMM extends this capability to multiple traits, providing a comprehensive approach to co-evolutionary analysis. Theoretically, they are capable of handling deviations from BM.

Evolutionary shifts, especially those occurring in branches close to the root of the phylogenetic tree, typically do not manifest directly as outliers in the trait data but instead appear as outliers in the PIC values. As a result, robust statistical methods designed to handle outliers in the original trait data may not be as effective in addressing evolutionary shifts. Adams et al. (2024) demonstrated that when four different estimators (L1, S, M, and MM) were integrated into both PGLS and PIC analysis, methods like PIC-MM and PIC-S performed significantly better in detecting evolutionary relationships, especially when accounting for outliers. Thus, computing PIC values is essential for effective analysis of the relationships between traits with evolutionary shifts.

Dramatic changes in some evolutionary branches can result in several outliers in the PIC dataset, distorting the statistical results of parametric tests. Nonparametric tests, which measure the strength and direction of association between two variables without assuming a specific distribution, can effectively handle data with outliers, reducing false positives. Therefore, we propose a hybrid framework, Outlier-Guided Correlation (OGC), to optimize the analysis of trait relationships under diverse evolutionary conditions. This framework integrates the existing PIC method with a dynamic selection mechanism that evaluates the presence of outliers in trait contrasts. For larger datasets, OGC dynamically applies Pearson or Spearman correlation based on the presence of outliers. For smaller datasets, where non-normality can significantly impact results, the framework incorporates both normality testing and outlier detection, referred to as Outlier- and Distribution-Guided Correlation

(ODGC). Together, these approaches form a unified framework collectively referred to as O(D)GC, designed to address the challenges posed by varying data conditions.

Our analysis demonstrates that PIC-O(D)GC effectively reduces false positives in the presence of strong abrupt shifts, while maintaining high accuracy. When evolutionary shifts are weak or absent, PIC-O(D)GC performs comparably to established phylogenetic methods, including PGLS (optimized across five evolutionary models), PIC-Pearson, robust regression methods, Corphylo (Zheng et al. 2009), PGLMM, and MR-PMM. Its performance closely aligns with that of PIC-MM, with the added advantage of computational simplicity. This consistency across a wide range of evolutionary scenarios highlights O(D)GC as a reliable and scalable tool for analyzing trait correlations in phylogenetic datasets, complementing other robust methods such as PIC-MM.

MATERIALS AND METHODS

Simulations of Phylogenetic Trees

To account for the influence of phylogenetic tree topology, we simulated two types of phylogenetic trees. The first type is a fixed, fully balanced phylogenetic tree, where all simulations were conducted using the same balanced dichotomous tree. The second type consists of randomly generated phylogenetic trees based on a pure birth model, where the ancestral root splits into two descendant subtrees of equal size. Asymmetry in each subtree is introduced by varying the branching rates within each subtree, generating different tree topologies for each simulation to explore diverse phylogenetic scenarios across replicates.

We set the sample size to 128 species and conducted simulations using both fixed and random trees. Additionally, to investigate the performance of each model under small sample conditions, we conducted a supplementary analysis with a sample size of 16 species, using only the fixed tree topology.

144

145 **Simulated Data with Abrupt Shifts**

146 Two traits (X_1 and X_2) were independently simulated on a phylogenetic tree, with $X_2 = 0 \times X_1$
147 $+ e$, where both X_1 and e underwent an abrupt shift of equal magnitude on one of the branches
148 closest to the root, while on all other branches they evolved independently under the BM
149 model with a variance of 1 (Fig. 1). The shift magnitude ranged across values from 4^{-3} to 4^{7},
150 including both up-sizing and down-sizing shifts. Specifically, $4^{0} = 1$ represents no shift (i.e.,
151 no change in magnitude), while positive exponents correspond to increases in trait value, and
152 negative exponents correspond to decreases in trait value. Each shift gradient was simulated
153 1000 times, resulting in a total of 11,000 simulations. Despite the two traits being designed as
154 independent, significant correlations occasionally emerged, likely due to random noise
155 occurring in the same direction.

156

157 **Simulated Data without Trait Jumps**

158 We simulated two traits (X_1 and X_2) on the tree with the relationship $X_2 = X_1 + e$, where e
159 represents a noise term. By adjusting the variance of the noise term e, we varied the correlation
160 strength between X_1 and X_2. Referencing Revell (2010), our simulations included the
161 following scenarios:

162 1. **BM1 & BM1+BM2**: X_1 evolves under a BM process with a variance of 1, while e
163 follows another BM process with a variance of d, leading to X_2 representing the
164 combined BM1 and BM2 processes (BM1+BM2).
165 2. **BM & BM+Norm**: X_1 evolves under a BM process with a variance of 1, and e follows
166 a normal distribution (mean = 0, variance = d), resulting in X_2 combining the BM
167 process and the normal distribution (BM+Norm).

3. **Norm & Norm+BM**: X_1 follows a normal distribution (mean = 0, variance = 1), and e evolves under a BM process (variance = d), making X_2 represent the combination of the normal distribution and the BM process (Norm+BM).

The d takes values of 4^{-3}, 4^{-2}, 4^{-1}, 4^{0}, 4^{1}, 4^{2}, 4^{3}, 4^{4}, 4^{5}, 4^{6}, and 4^{7} in all the three scenarios. For each scenario and value of d, we conducted 1000 replicate simulations, totaling 33,000 simulations for each tree type.

This study did not include the fourth scenario from Revell (2010), Norm1 & Norm1+Norm2, in which Trait X_1 follows a normal distribution with a mean of 0 and a variance of 1, while X_2 evolves as X_1 plus a noise term e, where e follows a normal distribution with a mean of 0 and a variance of d. This scenario results in simulated data with little to no phylogenetic correlation, which, though useful as a control for the absence of phylogenetic influence, was not included in this study. While the first three simulation scenarios primarily model data with phylogenetic autocorrelation, large noise term variances in these scenarios can also lead to data lacking phylogenetic autocorrelation, making a separate control scenario unnecessary.

Statistical Methods Compared

To evaluate the performance of PIC-OGC examining the evolutionary relationship between X_1 and X_2 amid an abrupt evolutionary jump, we compared it with several previously established phylogenetic comparative methods:

1. **PIC-Pearson**: Pearson correlation analysis of PIC values, equivalent to OLS regression of PIC, which is commonly used in phylogenetic comparative analysis.
2. **PGLS**: Using five evolutionary models, BM (Felsenstein 1985), Pagel's λ (Pagel 1999), OU fixed, OU random (Hansen 1997), and EB (Harmon et al. 2010).

3. **Robust phylogenetic regression with both PIC and PGLS frameworks:** For each framework, we utilized four robust estimators: L1, S, M, and MM. These were implemented as PIC-L1, PIC-S, PIC-M, and PIC-MM for PIC-based methods, and PGLS-L1, PGLS-S, PGLS-M, and PGLS-MM for PGLS-based methods (Adams et al. 2024).
4. **Corphylo**: A method that calculates Pearson correlation coefficients for multiple continuous traits, allowing users to incorporate phylogenetic signal and measurement errors into the model (Zheng et al. 2009). This method assumes an OU process for trait evolution, making it particularly suitable for traits that deviate from the BM model. Corphylo serves a dual purpose: it enables both the detection of phylogenetic signals in individual traits and the examination of correlations between multiple traits. A key feature of Corphylo is its ability to include measurement errors of trait data, potentially improving accuracy under real-world conditions. While our simulations did not explicitly account for measurement errors, this functionality highlights its potential robustness in empirical applications.
5. **PGLMM** (Ives and Helmus 2011).
6. **MR-PMM** (Westoby et al. 2023, Halliwell et al. 2024).

Outlier- and Distribution-Guided Phylogenetic Correlation Analysis

To address the influence of evolutionary shifts and outliers in phylogenetic comparative analysis, we introduced a novel framework: PIC-OGC. In this method, the choice between Pearson and Spearman correlation analysis is guided by the identification of outliers within the PIC data. The process starts by calculating the PICs for each trait using the phylogenetic tree and determining their absolute deviations from the median PIC of the dataset. A PIC is defined as an outlier if its absolute deviation from the median exceeds $T_i = n \times MAD$, where T_i

represents the threshold for detecting outliers, n is a sample-size-adjusted factor, and MAD (Median Absolute Deviation) is the median of the absolute deviations from the median PIC of the dataset (Wilcox 2003, Rindskopf and Shiyko 2010, Leys et al. 2013).

For a normal distribution, MAD is approximately equal to 0.67449 times the standard deviation (σ) (Iglewicz and Hoaglin 1993, Wilcox 2003). The expected coverage of the data under different multiples of MAD (Rindskopf and Shiyko 2010) is shown in Table 1, which also provides the corresponding expected number of outliers. As the threshold increases, the data coverage percentage approaches 100%, leading to a sharp decline in the expected number of outliers. This reduction scales inversely with both the threshold and dataset size, reflecting how stricter thresholds provide greater specificity at the cost of identifying fewer outliers.

Table 1. Data coverage and expected number of outliers under different thresholds for detecting outliers.

Threshold ($n \times MAD$)	Coverage	Expected number of outliers	
		16 Species	128 Species
3	95.6%	0.7	6
4	99.26%	0.1	0.9
5	99.94%	0.009	0.08
6	99.994%	0.0009	0.008
7	99.9995%	0.00008	0.0006

This table presents the expected number of outliers and the corresponding data coverage percentages for various outlier detection thresholds, defined as multiples of the Median Absolute Deviation (MAD), under the assumption of a normal distribution. The expected number of outliers is calculated for datasets with 15 and 127 phylogenetic independent

234 contrasts (PIC values), corresponding to phylogenetic trees with 16 and 128 species,
235 respectively.
236
237 For our datasets with 16 species (15 PIC values) and 128 species (127 PIC values), we set
238 thresholds at 5×*MAD* and 6×*MAD*, respectively, to ensure that only data points with
239 substantial deviations are identified as outliers, thereby minimizing the risk of misclassifying
240 data points that do not significantly impact the performance of statistical analysis methods.
241 Using these thresholds, the expected number of falsely identified outliers for normal data is
242 less than one per hundred simulations, ensuring high specificity in identifying deviations
243 caused by evolutionary shifts or other phenomena beyond normal variation.
244 To further evaluate the robustness of our approach, we repeated the study using a more
245 stringent threshold to define outliers: 6×*MAD* for the datasets with 16 species and 7×*MAD* for
246 the datasets with 128 species. Although minor variations in specific metrics were observed
247 under these stricter thresholds, the overall ranking of the performance of the compared
248 methods was consistent (see the Results section for details).
249 If either of the two traits, X_1 or X_2 in a dataset, is identified as containing an outlier, the
250 entire dataset is flagged as outlier-containing. To further clarify, we considered an example
251 arising from random variation. For instance, when a data point exhibits extreme deviations,
252 such as 11×*MAD* for X_1 and 1×*MAD* for X_2, Pearson correlation may result in a false negative
253 if such points are not appropriately flagged as outliers. Specifically, if a true positive
254 correlation exists between X_1 or X_2 but the outlier is not removed, the extreme value can distort
255 the Pearson correlation coefficient, reducing its sensitivity and leading to an incorrect
256 conclusion of no correlation. Conversely, if no true correlation exists between X_1 and X_2,
257 assigning such points to Spearman correlation—even if they are incorrectly categorized as
258 outliers—would not inherently lead to false positives. This is because Spearman evaluates

rank-based monotonic relationships and is less influenced by the magnitude of extreme values unless the rankings of the data points coincidentally align in a way that suggests a spurious monotonic relationship. This example underscores the importance of defining outliers based on deviations from the data's central tendency, regardless of whether the deviation arises from evolutionary shifts or ordinary variation. Moreover, the design of OGC, which independently evaluates outliers for each trait, theoretically enables it to handle scenarios where only one trait experiences a shift. This adaptability ensures robust performance even in asymmetric shift conditions, highlighting its versatility across a broader range of evolutionary scenarios. Specifically, when no outliers are detected in either trait (X_1 or X_2), Pearson correlation is used to capture the linear relationship. Conversely, when outliers are identified in one or both traits, Spearman's rank correlation is employed to mitigate the influence of these extreme values, ensuring that the method remains robust against deviations from normality.

For larger datasets with 128 species (127 PIC values), Pearson correlation, despite its reliance on normality assumptions, is generally robust to minor deviations from normality due to the central limit theorem. In such cases, the impact of skewness and kurtosis is reduced by the large sample size, ensuring reliable correlation estimates. Therefore, for these datasets, the choice of Pearson or Spearman correlation was determined solely by the presence or absence of outliers.

For smaller datasets with 16 species (15 PIC values), both normality testing and outlier detection were conducted. This approach, referred to as PIC-ODGC, dynamically selects the appropriate correlation method based on data characteristics. If no outliers were detected and normality assumptions were satisfied, the Pearson correlation was used. In cases where outliers were present, or normality was violated, Spearman correlation was applied to minimize the impact of extreme values on the results.

Benchmark for True Trait Correlations in Each Simulation

We aimed to compare the performance of different methods in detecting trait correlations while minimizing false-positive relationships between X_1 and X_2. An effective method should not only minimize false positives but also retain the ability to detect genuine, albeit rare, correlations that might arise due to random noise. Striking this balance is critical, as a method overly focused on avoiding false positives may miss true, meaningful correlations.

In the simulations without evolutionary shifts, traits X_1 and X_2 were simulated along a phylogenetic tree, with their relationships defined as $X_2 = X_1 + e$, where e represents a noise term. If the simulated formula is directly assumed to imply a significant positive correlation between X_1 and X_2, the expected correlation may diminish or even disappear when the variance of the noise term e is large, due to the overwhelming influence of noise. An accurate method, in this case, would identify the dataset as lacking a significant correlation, which contradicts the original assumption. Such discrepancies would label the method as producing a false negative result. Conversely, when the simulation explicitly sets no relationship between the two traits, random noise might generate spurious correlations. If a method accurately detects such correlations, it would erroneously be marked as producing a false positive result, penalizing its evaluation unfairly. These possibilities highlight the importance of carefully assessing correlations rather than assuming them based solely on the model structure. Similar concerns have been raised by Ives (2022), who emphasized that random errors in correlated data are not mere noise but can obscure or even invert the expected relationships between traits, particularly under conditions of high variance.

To define a robust benchmark for assessing true trait correlations, we adopted a phylogenetically informed approach. True evolutionary relationships between X_1 and X_2 were inferred by conducting traditional correlation tests on the normalized changes $\Delta X_1/L$ and $\Delta X_2/L$, where L is the branch length. The changes themselves are independent across branches,

ensuring that conventional correlation tests can provide reliable evaluations of trait relationships. The normalization by branch length serves to account for differences in evolutionary time scales, making the comparisons more meaningful and biologically interpretable.

Unlike our previous approach (Chen et al. 2023, Chen and Niu 2024), which relied on normality testing to select between Pearson and Spearman correlations, this study adopts an outlier-driven strategy. Outliers were identified as PIC values whose absolute deviation from the median exceeded a threshold $T_i = n \times MAD$, where n was adjusted based on the dataset size. This threshold was consistent with the one used in OGC, ensuring comparability in outlier detection across benchmarks. When no outliers were detected, Pearson correlation was applied; otherwise, Spearman rank correlation was used. This refined approach not only reduces the influence of extreme values but also ensures that true evolutionary relationships between traits with evolutionary shifts are accurately captured.

For smaller datasets with 16 species (15 PIC values), the benchmark also incorporated normality testing alongside outlier detection to better account for the data's characteristics. The same threshold applied in ODGC was used to define outliers, maintaining consistency in detection criteria across different analytical frameworks. If no outliers were identified and the normality assumption was satisfied, the Pearson correlation was used to assess the true evolutionary relationships. Conversely, in cases where outliers were present, or normality was violated, Spearman correlation was employed to mitigate the influence of extreme values. This dual consideration of outlier presence and normality ensured that the benchmark remained robust and adaptable to the unique challenges posed by small datasets.

Evaluating Concerns of Benchmark Bias in PIC-O(D)GC Assessment

Despite the robustness of our benchmark, concerns remain about potential biases when assessing PIC-O(D)GC because the benchmark shares methodological similarities with the PIC-O(D)GC framework itself. For example, both rely on outlier detection using a shared threshold (e.g., $T_i = n \times MAD$) and employ the same decision-making process between Pearson and Spearman correlation based on the presence of outliers or normality testing. Such alignment raises the possibility that the benchmark could inadvertently favor PIC-O(D)GC by design. If PIC-O(D)GC exhibits superior performance relative to other methods is evident, it is prudent to question whether this advantage arises from genuine methodological robustness or inherent alignment with the benchmark criteria.

To address this concern, we implemented a second benchmark that is entirely independent of PIC-O(D)GC's design. In this evaluation framework, we directly used the simulated formula to assume the presence or absence of correlations between X_1 and X_2. Specifically, when the relationship was defined as $X_2 = X_1 + e$, we treated X_1 and X_2 as correlated. Conversely, when the simulation explicitly set X_1 and X_2 as uncorrelated (e.g., $X_2 = 0 \times X_1 + e$), we used the absence of a correlation as the gold standard.

Although the second benchmark introduces limitations—namely, spurious correlations caused by random noise could mislabel a method's correct detection as a false positive—its influence on all methods being compared remains consistent. This uniform impact ensures a logically unbiased assessment of relative performance, allowing us to validate PIC-O(D)GC's performance under a complementary and independent criterion.

By applying both benchmarks, we aimed to mitigate potential bias and provide a more comprehensive evaluation of the methods under consideration. In the Results section, unless otherwise specified, the comparisons of different methods are based on the first benchmark's determination of whether trait correlations exist within the data.

Programs and Packages

All simulations and analyses were performed in the R environment (version 4.0.2) (R Core Team 2020) on a Linux system. The phytools package (version 2.3.0) (Revell 2024) was used for tree generation. The geiger *package* (version 2.0.11) (Pennell et al. 2014) was used for simulations. Normality tests were conducted using the Shapiro-Wilk test (Royston 1992) via the *shapiro.test()* function from the *stats* package (version 4.3.1). Pearson and Spearman correlation analyses were performed using the *cor.test()* function from the *stats* package. PICs were calculated using the *pic()* function from the ape package (version 5.8) (Paradis and Schliep 2019). All PGLS regressions were conducted using the *phylolm()* function from the *phylolm* package (version 2.6.2) (Ho and Ane 2014). Corphylo was conducted using the *corphylo()* function from the *ape* package (version 5.8) (Zheng et al. 2009). PGLMM and MR-PMM were both conducted using the *MCMCglmm* function from the *MCMCglmm* package (version 2.36) (Hadfield and Nakagawa 2010).

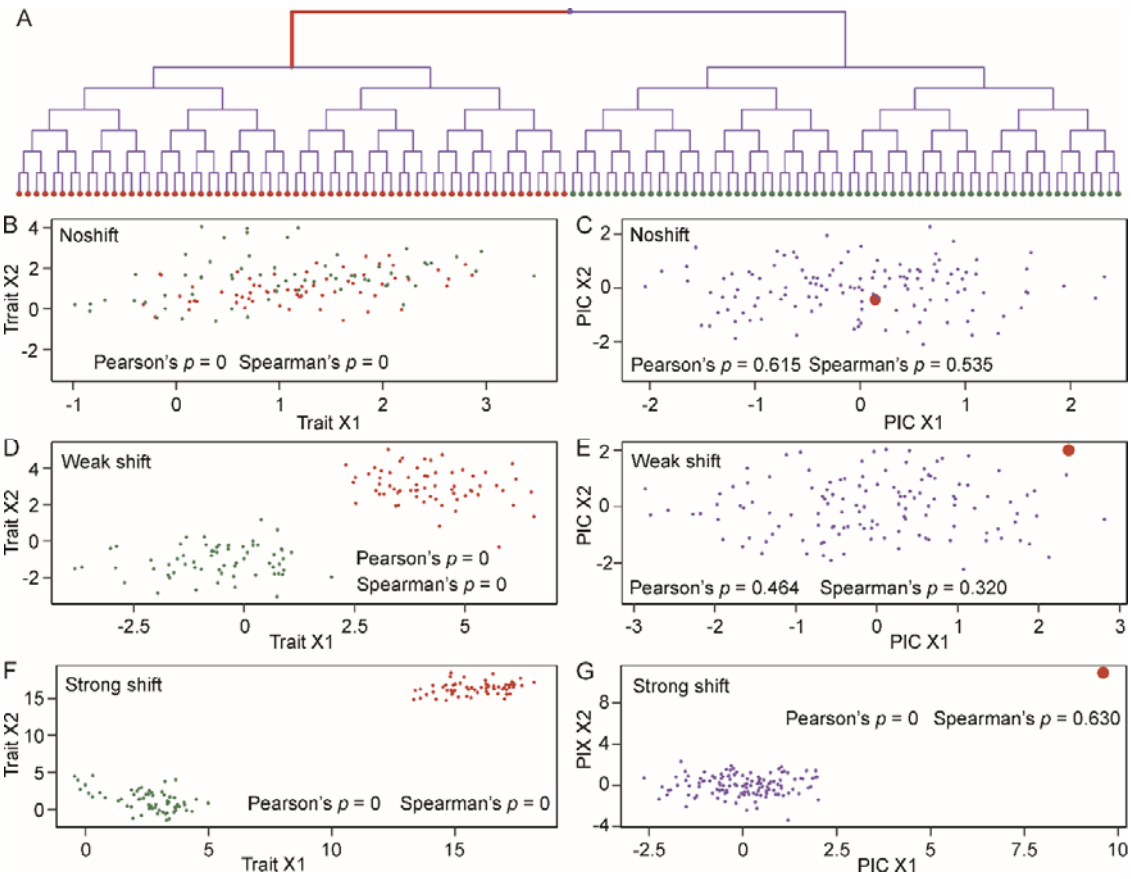
RESULTS

Impact of Abrupt Evolutionary Shifts on Data Patterns and Correlation Analysis

To assess PIC-OGC's ability to handle abrupt evolutionary shifts, we simulated two traits on a balanced phylogenetic tree with 128 species. Each trait underwent a shift near the root, while other branches evolved independently under a BM model (variance = 1, Fig. 1A). Shift magnitudes ranged from 4^{-3} to 4^{7}, including both increases and decreases, with 4^{0} representing no shift. From these simulations, we selected three datasets as case studies to examine how shifts affect both the overall trait patterns and the resulting PIC contrasts.

As shown in Figures 1B, 1D, and 1F, an abrupt shift in trait evolution causes the raw trait values to form two distinct clusters in the coordinate space, with the separation between these clusters increasing as the shift magnitude becomes larger. In the PIC coordinate system,

383 the shift manifests as an outlier in the PIC data, with the degree of outlier deviation
384 becoming more pronounced as the shift magnitude increases (Figs. 1C, 1E, and 1G).
385 When phylogenetic autocorrelation is ignored, directly applying Pearson or Spearman
386 correlation to the raw trait data frequently results in false positives, regardless of the shift
387 magnitude (Figs. 1B, 1D, and 1F).In contrast, controlling for phylogenetic autocorrelation by
388 using PIC values instead of raw trait data eliminates false positives when there is no shift or
389 when the shift is small (Figs. 1C and 1E). However, under large shifts, Pearson correlation
390 becomes highly sensitive to the outlier, leading to a false positive result and erroneously
391 inferring a significant positive correlation between X_1 and X_2, akin to non-phylogenetic
392 methods. Conversely, the Spearman correlation applied to the PIC data under large shifts
393 avoids false positives by correctly detecting no significant correlation between the two traits
394 (Fig. 1G).



395

Figure 1. Effects of abrupt shifts on data pattern and correlation analysis. (A) A balanced phylogenetic tree with 128 species, with the thick red line indicating an abrupt shift. (B, C) Simulation outcomes without any shifts (1-fold shift). (D, E) Simulation outcomes with a 4-fold shift. (F, G) Simulation outcomes with a 16-fold shift. The phylogenetically independent contrasts (PIC) values affected by the evolutionary shift are highlighted as enlarged red dots. The simulations illustrate that significant evolutionary shifts lead to the emergence of outliers in the PIC values. When $3\times MAD$ (Median Absolute Deviation) is used as the threshold for detecting outliers, all three PIC datasets contain identifiable outliers. In the weak-shift PIC dataset (E), the red dot is recognized as an outlier; however, it is not the point with the greatest deviation from the median. In the strong-shift PIC dataset (G), the red dot represents the most deviated point, with X_1 at $14\times MAD$ and X_2 at $15\times MAD$. This figure highlights the performance of traditional correlation methods and the varying effectiveness of parametric and nonparametric correlation analyses of PIC under different shift magnitudes. The data used to plot this figure are available in Supplementary Table S1

Refining Method Selection for Comparison with PIC-O(D)GC

Directly comparing all 18 parameter-model combinations from the methods listed in the Materials and Methods section would lead to an overly complex and unclear presentation, masking critical differences between approaches. To ensure a clear and focused evaluation of PIC-O(D)GC against other phylogenetic comparative methods, we refined the comparison by selecting only the best-performing models within each category.

For PGLS, the Akaike information criterion (AIC) was used to identify the optimal evolutionary model—BM, λ, OU fixed, OU random, or EB—for each dataset. Only the best-fitting PGLS model for each dataset was included in the final analysis.

For robust phylogenetic regression, which encompasses two methodological types (PGLS- and PIC-based regression) and four robust estimators (L1, S, M, and MM), we first evaluated their performance using simulations on the 128-species fixed tree.

Facing evolutionary shifts, PIC-MM stands out as the best method, consistently maintaining low false positive rates across all shift gradients, with rates remaining stable within narrow bounds of 0.046–0.062, demonstrating remarkable robustness and reliability (Fig. 2A). PIC-S also showed stability across different shift magnitudes, with false positive rates ranging from 0.098 to 0.103. However, PIC-S consistently exhibited higher false positive rates compared to PIC-MM, indicating inferior control over false positives despite its stability. Other methods, whether based on PIC or PGLS, showed acceptable performance under weak shifts (shift ≤ 4), with lower false positive rates. However, their false positive rates sharply increased to 1 under stronger shifts (shift > 4), indicating their instability in managing abrupt evolutionary changes effectively. Interestingly, PIC-L1 and PGLS-L1 exhibited lower false positive rates than PIC-MM under weak shifts (shift ≤ 4), suggesting that the L1 estimator may have potential advantages in specific contexts.

Accuracy reflects the combined influence of false positive and false negative rates by measuring the proportion of correct predictions (true positives and true negatives), offering a holistic view of the methods' overall performance. Figure 2B confirms that PIC-MM is the most reliable method, achieving both high accuracy and stability across all evolutionary shift gradients. PIC-S follows as the second-best performer, while PIC-L1 and PGLS-L1 slightly excel under low shift conditions. The trends in accuracy largely align with the false positive rates in Figure 2A, underscoring the strong correlation between these two metrics for robust methods.

Figures 2C-2E show the accuracy of various methods across three no-shift evolutionary scenarios: BM1 & BM1+BM2, BM & BM+Norm, and Norm & Norm+BM. PIC-MM is

generally among the best-performing methods, though it is occasionally slightly outperformed in specific cases, such as under high variance in the BM & BM+Norm scenario, where PIC-L1 and PGLS-L1 show a clear, though modest, advantage. PIC-S, on the other hand, is no longer the second-best performer; in fact, it ranks among the worst in BM1 & BM1+BM2 and Norm & Norm+BM, showing considerable performance degradation under these conditions.

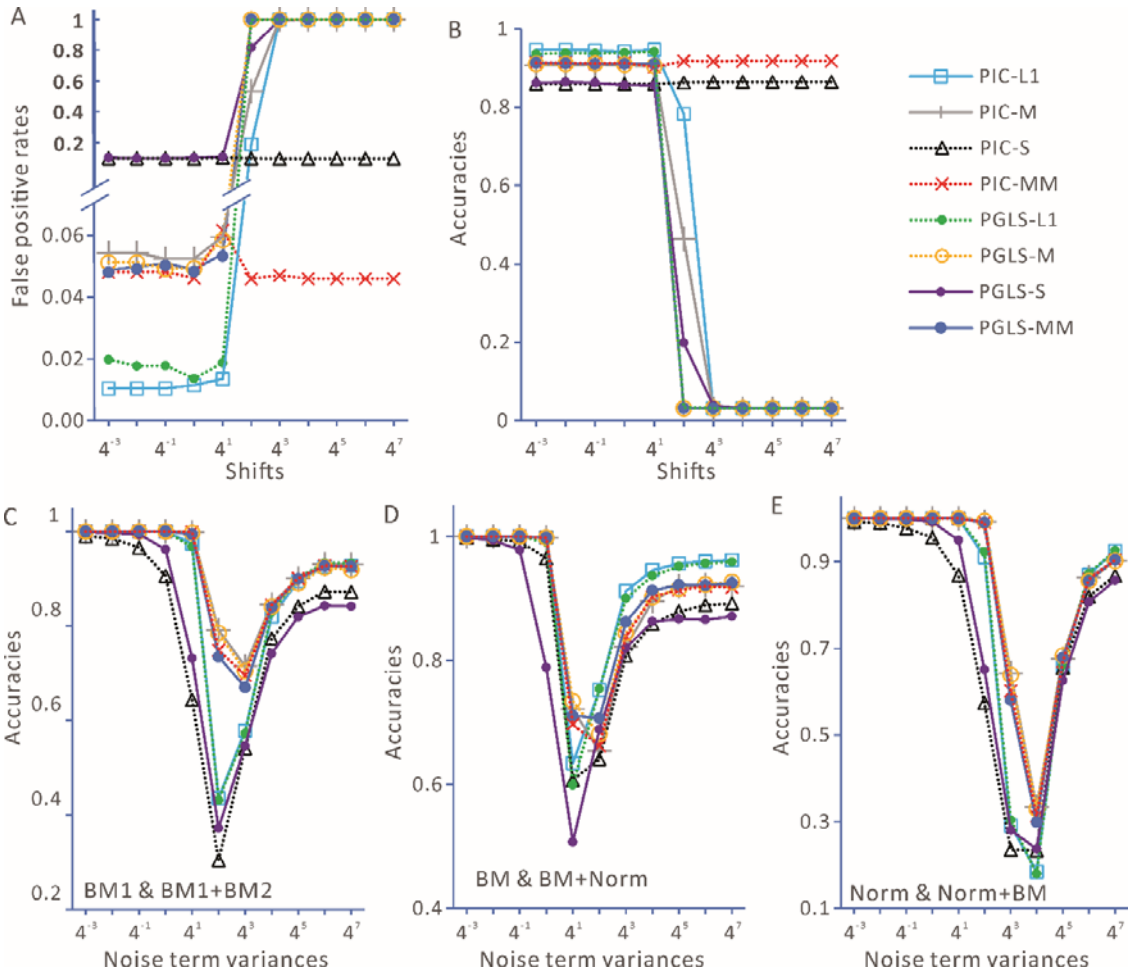


Figure 2. Performance of eight robust phylogenetic regression methods. This figure shows their false positive rates (A) and accuracies (B) across evolutionary shift gradients and their accuracies across three simulation scenarios without abrupt shifts (C-E). The datasets were simulated using a fixed-balanced tree of 128 species. The data used to plot this figure are available in Supplementary Table S2. Please refer to the Materials and Methods section for

456 additional details on the simulation setup, variance gradients, and an overview description of
457 the eight methods.
458
459 When analyzing randomly generated 128-species trees and 16-species fixed trees,
460 PIC-MM consistently demonstrates the strongest ability to handle evolutionary shifts. Across
461 the three simulation scenarios without evolutionary shifts, PIC-MM performs at least as well
462 as, or not significantly worse than, the other methods (Supplementary Table S2). Even when
463 the outlier detection threshold was adjusted from $6\times MAD$ to $7\times MAD$ for 128-species datasets,
464 PIC-MM maintained its top performance across both random and fixed trees (Supplementary
465 Table S3). These results highlight PIC-MM's robustness and versatility, making it a reliable
466 tool for analyzing phylogenetic data under both stable and dynamic evolutionary conditions.
467 These findings align with the conclusions of Adams et al. (2024). Consequently, PIC-MM was
468 selected as the representative robust regression method for comparison with PIC-O(D)GC.
469 Through this optimization and selection process, we reduced the total number of methods
470 compared with PIC-O(D)GC to six, facilitating a more focused and interpretable evaluation.
471
472 **PIC-OGC: Robust Performance Across Evolutionary Shifts**
473 We compared the performance of PIC-OGC with six other methods using simulations on the
474 128-species fixed tree. As shown in Figure 3A, there are significant differences in false
475 positive rates among the methods. Under weak shifts (shift ≤ 4), PIC-OGC consistently
476 exhibits the same false positive rates as PIC-Pearson, which is equivalent to traditional
477 PIC-OLS regression. These rates are the lowest among the seven methods, and this similarity
478 can be attributed to the fact that weak or downsizing evolutionary shifts do not generate
479 outliers in the PIC datasets. However, when the shift becomes stronger (> 4), the false positive
480 rate of PIC-Pearson sharply increases to 1, while PIC-OGC maintains a stable false positive

rate of 0.039. This result demonstrates that PIC-OGC is significantly more robust against evolutionary shifts than traditional PIC regression methods. PIC-MM also shows robustness against evolutionary shifts but has slightly higher false positive rates than PIC-OGC. Corphylo shows similar performance to PIC-OGC and PIC-Pearson under weak shifts (shift ≤ 4), but its false positive rates experience a sharp increase to 1 under stronger shifts, similar to PIC-Pearson. PGLMM and PGLS also show low false positive rates under weak shifts (shift ≤ 4) but experience sharp increases to 1 under stronger shifts (shift > 4). MR-PMM performs the worst, with the highest or among the highest false positive rates across 10 out of 11 shift gradients.

Figure 3B displays the accuracy of the same seven methods under the same 11 evolutionary shift gradients. Accuracy here serves as a comprehensive metric, reflecting both false positive and false negative rates. Notably, PIC-OGC demonstrates the highest accuracy or shares the top position across all shift gradients. PIC-MM also performs exceptionally well, with accuracy very similar to PIC-OGC. The other methods show acceptable performance under weak shifts but exhibit unacceptably low accuracy under stronger shifts. Despite PGLS results being derived from selecting the best-fitting model from five candidate models for each simulation, its performance still falls significantly short of PIC-OGC and PIC-MM.

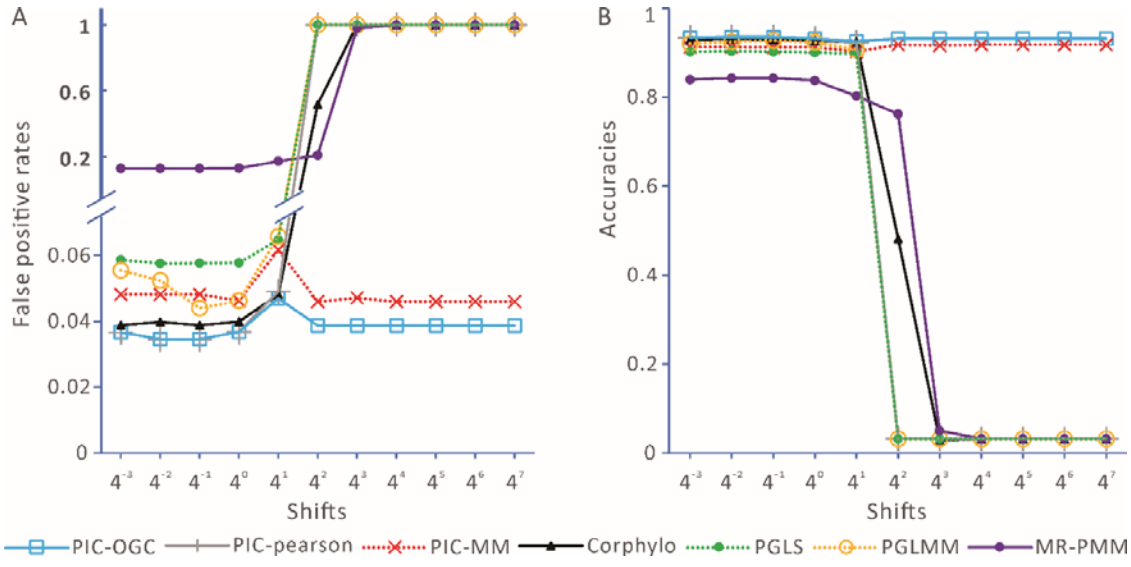


Figure 3. False positive rates (A) and accuracies (B) of seven methods on simulations across evolutionary shift gradients simulated on a fixed balanced tree of 128 species. The data used to plot this figure, including false positive rates and accuracy metrics, are available in Supplementary Table 4. Please refer to the Materials and Methods section for additional details on the simulation setup, variance gradients, and an overview description of the seven methods.

Fixed trees represent a fully balanced phylogenetic structure, where all simulations are conducted using the same balanced dichotomous tree. This consistent topology ensures uniformity in the branching structure, resulting in homogenized data. In contrast, random trees are generated based on a pure birth model, introducing asymmetry by varying the branching rates within each subtree. This variability results in diverse tree topologies across simulations, reflecting a wider range of phylogenetic scenarios and introducing additional heterogeneity in the data. In our analysis with 128-species random trees (Supplementary Table S4), most methods showed performance very similar to their results under fixed tree simulations, with PIC-OGC and PIC-MM consistently demonstrating the best performance overall. However, the slight yet consistent edge that PIC-OGC held over PIC-MM across all shift gradients under fixed tree simulations was no longer observed with random trees. Instead, each method exhibited slight advantages under specific gradients, emphasizing their complementary strengths in handling the complexities introduced by varying phylogenetic structures. Additionally, MR-PMM exhibited a significant improvement under low shift gradients (shift ≤ 4). While it had the worst performance with fixed trees, it became the method with the lowest false positive rate and highest accuracy in weak conditions, outperforming all other methods.

All the above results were obtained using a $6\times MAD$ threshold to define outliers. Using a $7\times MAD$ threshold yielded highly similar results, with PIC-OGC and PIC-MM remaining the top-performing methods (Supplementary Table S5).

PIC-OGC: Strong Performance in Simulations Without Abrupt Shifts

The analysis here focuses on simulations conducted on a 128-species fixed tree, where two traits exhibit a relationship of $X_2 = X_1 + e$, with e representing random noise varying across 11 gradients of variance displayed on the x-axis of Figure 4. These simulations aim to evaluate the methods' performance in detecting relationships between evolutionary traits under conditions without abrupt shifts.

In the BM1 & BM1+BM2 scenario (Fig. 4A), the three PIC-based methods (PIC-OGC, PIC-MM, and PIC-Pearson), along with Corphylo, PGLS, and PGLMM, maintain comparably high accuracy across all variance gradients, demonstrating their robustness in handling traits with strong phylogenetic signals. MR-PMM, however, exhibits a distinct pattern, showing a significant decline in accuracy at intermediate variance gradients while maintaining comparable performance at both low and high variance levels.

In the BM & BM+Norm scenario (Fig. 4B), the three PIC-based methods (PIC-OGC, PIC-MM, and PIC-Pearson) and Corphylo clearly outperform MR-PMM, PGLS, and PGLMM.

In the Norm & Norm+BM scenario (Fig. 4C), all methods achieve very high accuracy (close to 1) at low variance gradients ($\leq 4^2$), except for Corphylo, which underperforms at the lowest variance gradient (4^{-3}). As variance increases to moderately high gradients, the accuracies of the six methods—PIC-OGC, PIC-MM, PIC-Pearson, Corphylo, PGLS, and PGLMM—decline significantly and consistently. In contrast, MR-PMM shows a unique

pattern distinct from the other methods, with neither a clear advantage nor disadvantage in overall performance but demonstrating context-specific strengths.

Across all three simulation scenarios, the three PIC-based methods (PIC-OGC, PIC-MM, and PIC-Pearson) consistently exhibit near-optimal performance, highlighting their robustness in handling phylogenetic traits without abrupt shifts.

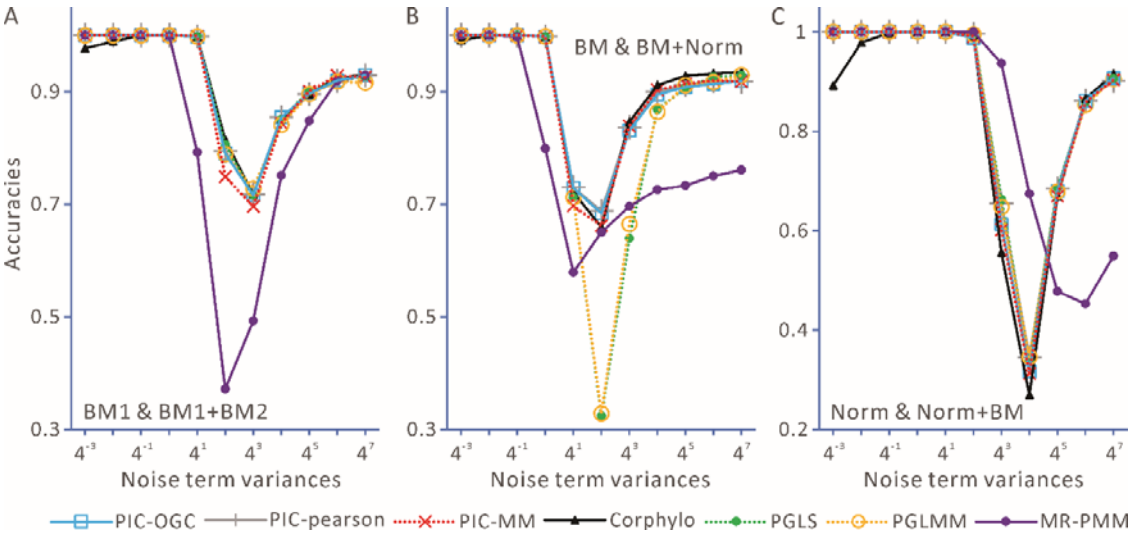


Figure 4. Accuracy of seven methods on simulations without abrupt shifts. Simulations were conducted on a fixed balanced tree of 128 species, with relationships between traits defined as $X_2 = X_1 + e$, where e represents random noise, across three scenarios: (A) BM1 & BM1+BM2, (B) BM & BM+Norm, and (C) Norm & Norm+BM. The data used to plot this figure are available in Supplementary Table S4. Please refer to the Materials and Methods section for additional details on the simulation setup, variance gradients, and an overview description of the seven methods.

Results from simulations on the 128-species random trees show overall trends consistent with those observed on the fixed tree, reaffirming the strong performance of the three PIC-based methods across diverse phylogenetic topologies (Supplementary Table S4). While their overall performance remains excellent, PIC-OGC shows a slight drop at two specific

variance gradients (4^4 and 4^5) of the last simulation scenario (Norm & Norm+BM), where it no longer ranks among the top three methods. These results indicate a minor, context-dependent limitation of PIC-OGC in random tree simulations.

The above results on methods' accuracies detecting relationships between evolutionary traits under conditions without abrupt shifts were obtained using a $6\times MAD$ threshold to define outliers. Using a $7\times MAD$ threshold yielded highly similar results, with the three PIC-based methods remaining the top-performing methods (Supplementary Table S5).

Performance of the Methods in Detecting Relationships Beyond BM Assumptions

Given that the computation of PIC values relies strictly on the assumption of a BM model, it is logically problematic for non-BM data. As a result, PIC-Pearson, which uses Pearson correlation analysis on PIC values, is excluded from further comparisons in such cases because its theoretical foundation no longer holds for non-BM data. However, PIC-MM and PIC-OGC, despite being rooted in PIC calculations, are retained for further analysis due to their strong performance in earlier results (Figs. 3–4), raising the question of whether their robustness extends to non-BM conditions. PGLS is included because it is a more flexible framework that can adapt to different evolutionary models by incorporating the appropriate covariance structure, making it suitable for non-BM data. Corphylo is retained as it assumes an OU process for trait evolution, making it suitable for data that deviate from BM assumptions. PGLMM and MR-PMM are also retained because they do not strictly rely on BM assumptions, and their flexibility in modeling phylogenetic relationships and handling complex data may allow them to outperform PIC-based methods under non-BM conditions.

Garland et al. (1992) clarified that the application of PIC is not strictly limited to gradual, clock-like evolution as modeled by BM. Instead, PIC remains applicable to non-BM scenarios, provided that branch lengths and evolutionary models are appropriately transformed to

account for deviations from BM assumptions. Building on this foundational flexibility, our study evaluates the performance of PIC-OGC and PIC-MM under untransformed conditions, testing whether these methods can deliver reliable results even without adjustments to branch lengths or models.

To classify simulated data as either BM or non-BM condition, we calculated the AIC values from PGLS regression models fitted under five evolutionary models: BM, lambda, OU random, OU fixed, and EB. The data were classified as BM condition if the BM model yielded the lowest AIC; otherwise, they were considered non-BM condition.

Generally, the two PIC-based methods, PIC-OGC and PIC-MM, maintained their superiority under non-BM conditions involving evolutionary shifts (Figs. 5A–5B). PIC-OGC performed exceptionally well, achieving almost the lowest false positive rates and highest accuracy across all shift gradients. PIC-MM also demonstrated strong overall performance, though it exhibited a high false positive rate and low accuracy at a specific shift gradient (4). Despite being first optimized for performance by selecting the best-fitting evolutionary model, PGLS exhibited the weakest performance among retained methods under non-BM conditions, showing high false positive rates and low accuracy, particularly at higher shift gradients.

In scenarios without evolutionary shifts (Fig. 5C), none of the six retained methods showed a consistent advantage or disadvantage. Performance varied across different variance levels, with no single method consistently emerging as the clear best or worst. Similar trends were observed when using random trees (Supplementary Table S6) or adjusting the threshold for defining outliers (Supplementary Table S7), reinforcing the robustness of these conclusions across varying tree structures and outlier detection criteria.

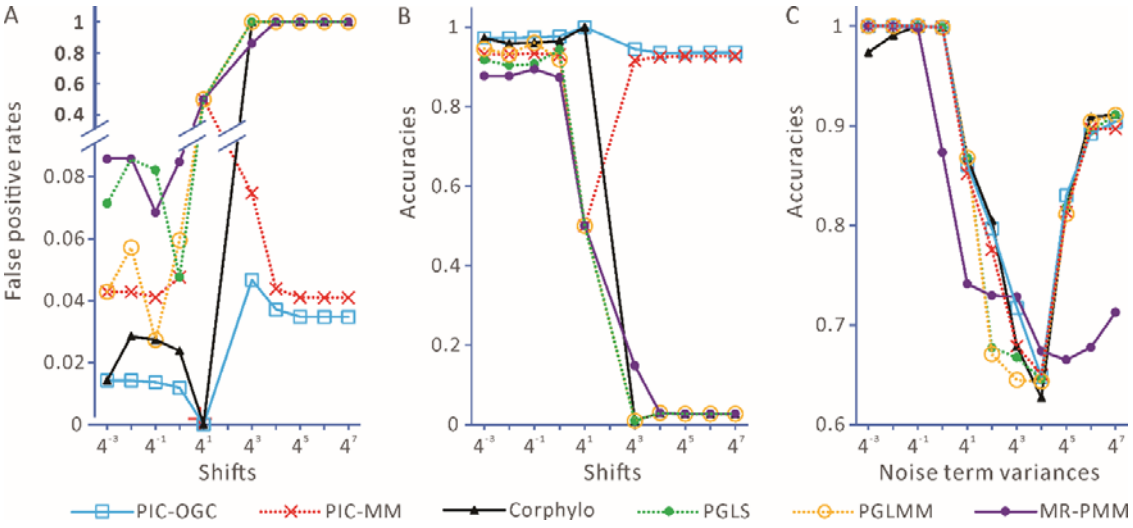


Figure 5. Performance of six methods across datasets that deviate from BM assumptions. (A). The false positive rates of six methods under datasets with evolutionary shifts. (B). The accuracy of the methods under the same conditions. (C) combines results from datasets without evolutionary shifts (three no-shift simulation scenarios) under non-BM conditions. All the datasets were simulated using a fixed-balanced tree of 128 species. They were classified as BM or non-BM conditions based on AIC values from PGLS regression models fitted under five evolutionary models (BM, lambda, OU random, OU fixed, and EB). Data were classified as BM condition if the BM model had the lowest AIC; otherwise, they were considered non-BM condition. The data used to plot this figure are available in Supplementary Table S6. Similar trends were observed when analyzing randomly generated 128-species trees (Supplementary Table S6). Please refer to the Materials and Methods section for additional details on the simulation setup, variance gradients, and an overview description of the six methods.

Comparison of the accuracies under conditions with and without shifts (Fig. 3B vs Fig. 4A–C; Fig. 5B vs Fig. 5C; see also Supplementary Tables S2–S7) reveals that methods other

than PIC-OGC and PIC-MM effectively handle low shift and low variance conditions but generally perform poorly under medium-to-high shift or variance conditions. Notably, their ability to cope with medium-to-high variance is significantly better than their performance under medium-to-high shifts.

PIC-ODGC: Robust Performance with Small Samples

For small samples consisting of 16 species (15 PIC values), the choice between Pearson and Spearman correlation is influenced not only by the presence or absence of outliers but also by whether the data follow a normal distribution. To address this, we enhanced PIC-OGC to PIC-ODGC, incorporating both outlier detection and normality testing.

Using a fixed tree with 16 species, we replicated the analyses conducted with the 128-species fixed tree. Two thresholds for defining outliers, $5\times MAD$ and $6\times MAD$, were applied. The conclusions drawn from the 128-species dataset regarding PIC-OGC and PIC-MM were consistently confirmed in the 16-species dataset, with PIC-ODGC and PIC-MM emerging as the top-performing methods (see Supplementary Tables S2–S7). This consistency highlights the robustness of these methods, even under small sample conditions.

Addressing Concerns about PIC-Based Performance

PIC-O(D)GC and PIC-MM, being rooted in PIC calculations, might initially raise concerns about whether their observed strong performance stems from a bias due to the gold standard's reliance on calculating $\Delta X_1/L$ and $\Delta X_2/L$, a process that closely mirrors the PIC estimation procedure. This concern is particularly pronounced for PIC-O(D)GC, as it also employs the same outlier detection thresholds and criteria for selecting Pearson or Spearman correlation.

To address potential biases in the original benchmark, we validated PIC-O(D)GC and PIC-MM using a second, entirely independent benchmark. Unlike the first benchmark, this alternative framework does not rely on any assumptions or calculations related to PIC or any specific method. Instead, it directly determines the presence or absence of correlations based on the simulated formula: X_1 and X_2 are treated as correlated when defined as $X_2 = X_1 + e$, and uncorrelated when defined as $X_2 = 0 \times X_1 + e$. Although this benchmark may classify spurious correlations caused by random noise as false positives, its design is method-agnostic, ensuring no bias toward or against any specific approach.

Using the second benchmark, we compared the methods in their false positive rates under shift conditions and false negative rates under no-shift conditions.

First, we observed no substantial differences in the performance of various methods between 128-species random and fixed trees.

Among the eight phylogenetically robust methods, PIC-MM consistently demonstrated the lowest false positive rates, reaffirming its superior position within this group (Supplementary Table S8). Under no-shift conditions with 128 species, PIC-MM performed similarly to three comparable methods (PIC-M, PGLS-M, and PGLS-MM), all significantly outperforming the other four methods. However, with 16 species, none of the eight methods exhibited clear superiority in terms of false negative rates.

For the false positive rates under shift conditions, PIC-OGC and PIC-MM performed almost identically, both achieving top performance with consistently low false positive rates (Supplementary Table S9). In the three no-shift evolutionary scenarios, neither PIC-OGC nor PIC-MM consistently achieved the lowest false negative rates, but both demonstrated generally acceptable performance. Specifically, MR-PMM exhibited significantly higher false negative rates in the BM1 & BM1+BM2 scenario and markedly lower false negative rates in the Norm & Norm+BM scenario, while the other methods, including PIC-OGC and

679 PIC-MM, showed no substantial differences. In the BM & BM+Norm scenario, PGLS and
680 PGLMM performed slightly better than the other methods.

681 For the 16-species fixed tree, PIC-ODGC's false positive rates were similar to those of
682 PIC-MM, outperforming all other methods (Supplementary Table S9). Under the three
683 no-shift evolutionary scenarios, none of the seven methods showed distinct advantages or
684 disadvantages.

685 Under non-BM conditions, the performance of various methods was largely consistent
686 across the 128-species fixed tree, 128-species random trees, and the 16-species fixed tree. In
687 scenarios with evolutionary shifts, PIC-O(D)GC and PIC-MM maintained stable and low
688 false positive rates, reaffirming their robustness in handling such conditions. However, under
689 no-shift conditions, the false negative rates of PIC-ODGC and PIC-MM were comparable to
690 those of other methods, showing no clear advantage or disadvantage.

691 The use of two entirely different benchmarks for defining the presence or absence of
692 correlations between X_1 and X_2 did not lead to differences in the ranking of method
693 performance. This consistency highlights that the advantages demonstrated by PIC-O(D)GC
694 and PIC-MM across various scenarios, including their stable and low false positive rates
695 under evolutionary shifts, are not merely a result of their apparent similarity to the first
696 benchmark. Instead, these strengths are driven by their inherent robustness and adaptability
697 in handling phylogenetic data under varying evolutionary conditions. Moreover, the
698 alignment of method performance trends under both benchmarks reinforces the reliability of
699 the first benchmark. This consistency suggests that the first benchmark is not biased in favor
700 of any specific method but rather serves as a fair and effective framework for evaluating
701 phylogenetic comparative methods.

702

703 ## DISCUSSION

Equivalence of Pearson Correlation and OLS in Context

This study employed Pearson correlation to evaluate PIC relationships, which is mathematically equivalent to the OLS regression commonly used in previous studies. Pearson correlation provides the same statistical significance and directionality as OLS regression, ensuring that our results are directly comparable with prior work. We chose to use Pearson correlation for its computational simplicity and consistency with other nonparametric methods, such as Spearman correlation, which is a key component of the PIC-O(D)GC framework. The equivalence between Pearson correlation and OLS regression ensures logical and statistical consistency with previous findings while reinforcing the relevance and robustness of our results within the context of phylogenetic comparative analyses.

Limitations of Traditional Data Transformations in Phylogenetic Correlation Analysis

A potential concern readers may raise is the complexity of the PIC-O(D)GC framework. Given its focus on handling non-normal data distributions and outliers, why not simply apply data transformations (e.g., logarithmic or square root transformations) to standardize the data before using the traditional PIC-OLS regression? Such transformations are well-known for their ability to address skewness and improve normality, potentially achieving comparable results with much less complexity.

While data transformations are a commonly used strategy, their application to PIC data in phylogenetic analyses faces several critical limitations:

1. Negative Values in PIC Data. PIC calculations often result in negative values, which cannot be directly transformed using logarithmic or square root functions. Adding a constant to shift the data into a positive range is a potential workaround, but this

approach introduces arbitrary biases, especially when the data span a wide range of values.

2. Ineffectiveness of Yeo-Johnson Transformation. The Yeo-Johnson transformation is designed to handle both positive and negative values, making it seemingly suitable for phylogenetic data. However, its effectiveness in achieving normality is severely limited for PIC data. In our prior studies using random tree simulations with 128 species, none of the datasets $\Delta X_1/L$ or $\Delta X_2/L$, showed normal distributions. Even after applying the Yeo-Johnson transformation, fewer than 1% of the datasets achieved normality (Chen and Niu 2024). Moreover, most phylogenetic trees in real-world research are non-balanced, closely resembling the random trees used in our simulations. Such trees inherently generate $\Delta X_1/L$ and $\Delta X_2/L$ values, as well as PIC data, that deviate significantly from normality. This highlights not only the inherent challenges of achieving normality in phylogenetic datasets but also the practical limitations of relying on transformations like Yeo-Johnson to address these challenges effectively.
3. Outlier Sensitivity. Traditional data transformations focus on normalizing distributions but do not explicitly address outliers, which can significantly impact the performance of correlation analyses. PIC-O(D)GC, by dynamically selecting Pearson or Spearman correlation based on the presence of outliers, provides a tailored approach that is inherently robust to extreme values.

Comparing Statistical Methods for Detecting Trait Correlations

Our study evaluated the performance of various statistical methods in detecting trait correlations under diverse evolutionary scenarios, including with and without abrupt shifts. PIC-O(D)GC and PIC-MM consistently demonstrated robust performance, particularly

753 excelling under shift scenarios by minimizing false positives and maintaining high accuracy.
754 In contrast, traditional methods like PIC-Pearson and PGLS, as well as robust variants such
755 as PGLS-MM, struggled under high shift gradients, highlighting their limitations in
756 challenging conditions.

757 In no-shift scenarios, PIC-MM performed comparably to other reliable methods, such as
758 PIC-L1, PGLS-L1, and Corphylo, highlighting the role of estimator choice within
759 regression-based methods. Meanwhile, PIC-O(D)GC maintained strong performance
760 through its hybrid correlation framework, independent of estimator-specific considerations.
761 The resilience of PIC-O(D)GC in adapting to outliers and non-normal data further supports
762 its utility across stable and dynamic datasets.

763

764 **Advantages of PIC-OGC: Computational Efficiency and Simplicity**

765 Among robust phylogenetic regression estimators, MM estimators were recommended by
766 Adams et al. (2024) for their strong performance across various scenarios. Consistent with
767 their findings, we observed that PIC-MM effectively controls false positive rates in datasets
768 with shifts. Similarly, PIC-O(D)GC exhibits comparable robustness, aligning closely with
769 PIC-MM in both false positive rate and accuracy.

770 However, the computational efficiency of PIC-O(D)GC offers a distinct advantage over
771 existing approaches, particularly the PIC-MM method. Both methods share an initial
772 computational complexity of $O(n)$ for calculating PIC values, where n represents the number
773 of taxa. The key difference lies in the subsequent steps:

774 1. PIC-MM employs iterative adjustments of model parameters, requiring k iterations to
775 converge. Each iteration involves recalculating residuals and adjusting weights,
776 resulting in a total complexity of $O(k \cdot n)$.

2. In contrast, PIC-OGC identifies outliers by evaluating the deviation of each PIC value from the median. If no outliers are detected, a parametric correlation method (Pearson) is applied; otherwise, a nonparametric method (Spearman) is used. The outlier detection step is computationally lightweight $O(n)$ and does not impact the overall complexity, which remains $O(n\log n)$, driven by the sorting step in Spearman's rank correlation.

PIC-OGC's lower computational complexity provides a clear advantage in scalability, making it particularly well-suited for large phylogenetic datasets. This efficiency, combined with its robustness to abrupt evolutionary changes, positions PIC-O(D)GC as a scalable and practical alternative to MM-based methods for phylogenetic comparative analyses.

For small datasets, the OGC framework is extended to PIC-ODGC by incorporating normality testing to address the increased sensitivity of parametric methods to non-normal distributions. The Shapiro-Wilk test, commonly used for normality assessment, adds an additional computational complexity of $O(n^2)$ due to pairwise comparisons. As a result, PIC-ODGC has a higher computational complexity compared to PIC-MM. However, this complexity is negligible in practice for small sample sizes, such as datasets with 16 species, where the additional computation does not impose a significant burden. While PIC-ODGC maintains the computational simplicity and robustness of PIC-OGC, the added step of normality testing introduces an additional layer of decision-making that may not always be necessary. Nonetheless, its adaptive framework ensures accuracy across diverse scenarios, making it a valuable extension of PIC-OGC for scenarios involving small datasets or heightened sensitivity to data distribution characteristics.

Another notable advantage of PIC-O(D)GC is its simplicity and accessibility. By dynamically adapting to the presence or absence of outliers and data normality, PIC-O(D)GC effectively balances robustness and interpretability. Its reliance on

straightforward statistical concepts, such as Spearman's rank correlation, makes it easier to understand and interpret, even for researchers with limited statistical training. By avoiding assumptions about data distribution, PIC-O(D)GC offers intuitive and transparent results, reducing the risk of misinterpretation and enhancing accessibility for a broader range of biologists.

Advantages of Correlation-Based Methods over Regression-Based Approaches

Correlation-based methods such as PIC-O(D)GC, Corphylo, and MR-PMM have notable advantages over regression-based approaches like PIC-OLS, PGLS, and PGLMM. The latter methods, similar to OLS regression, inherently assume a directional causal relationship, where the independent variable is considered error-free, or its errors are negligible (McArdle 1988, Osborne and Waters 2002, Casson and Farmer 2014, Jarantow et al. 2023). This assumption implies that the values of the independent variable are precise, while the dependent variable may include a degree of error. However, when the independent variable is subject to measurement error, this can lead to biased regression coefficients—a phenomenon known as "measurement error bias."

In contrast, correlation-based methods do not require assumptions about the directionality of causal relationships and are unaffected by measurement errors in either variable. This characteristic makes correlation-based methods particularly robust in scenarios where the precision of independent variables cannot be guaranteed. While correlation-based approaches cannot fully address all statistical challenges, their focus on detecting associations rather than causation makes them more reliable under conditions where measurement errors are present in both traits.

826 **Limitations of Correlation-Based and Nonparametric Methods in Phylogenetic**
827 **Analysis**

828 While correlation-based methods like PIC-O(D)GC offer notable advantages, including
829 robustness to outliers and flexibility in handling non-normal data, they share certain
830 limitations inherent to this class of methods. These methods, by design, cannot capture
831 non-monotonic relationships, which limits their ability to detect more complex evolutionary
832 patterns. This constraint applies to all correlation-based approaches, as they inherently focus
833 on assessing monotonic associations between variables.

834 Moreover, correlation-based methods, including PIC-O(D)GC, are not equipped to
835 handle interactions or analyze multiple predictors simultaneously, a key capability of
836 regression-based approaches. For instance, regression methods such as PGLS can model
837 complex relationships between traits and account for confounding factors, providing a more
838 comprehensive framework for studying evolutionary mechanisms. By relying solely on
839 pairwise correlations, methods like PIC-O(D)GC are inherently limited in their ability to
840 unravel multi-trait co-evolutionary dynamics.

841 Another critical limitation of nonparametric methods, including PIC-O(D)GC, lies in
842 their reliance on testing null hypotheses. While rejecting null models is informative for
843 identifying significant associations, it offers limited insight into the underlying evolutionary
844 processes. Parameter-based methods, such as regression models, provide richer information
845 through the estimation of meaningful parameters. For example, regression coefficients (β) in
846 PGLS models directly quantify the influence of one trait on another, enabling researchers to
847 test mechanistic hypotheses about co-evolution and trait evolution rates.

848 Despite these limitations, correlation-based methods remain valuable tools for
849 addressing specific challenges in phylogenetic data analysis. They should not be viewed as
850 replacements for parametric models but rather as complementary approaches that provide

robust and computationally efficient solutions, particularly in scenarios with non-normal data distributions, outliers, or small sample sizes.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (grant number 31671321), with leftover resources used for occasional research expenses.

ACKNOWLEDGMENTS

Many sentences in this article were written with the assistance of ChatGPT-4 (OpenAI 2024). The authors thoroughly reviewed and revised each sentence and took full responsibility for the language and content of the entire article.

DATA AVAILABILITY

Supplementary Tables S1–S8 are provided in the file Supplemental_Tables.xlsx. The raw data for the simulated phylogenetic analyses, including trait correlation data for 16-species fixed trees, 128-species fixed trees, and 128-species random trees across shift and non-shift scenarios, are available in RawData.zip. The scripts used to generate the results presented in the main text— covering phylogenetic tree construction and trait data simulation (Simulation_01_FixedBalancedTree.R and Simulation_02_RandomTree.R), outlier detection (Analysis_05_MAD.R), regression model fitting (Analysis_01_PIC.R, Analysis_02_PGLS_RPR.R, Analysis_03_Corphylo.R, Analysis_04_PGLMM_MRPMM.R), are provided in Code.zip.

REFERENCES

Adams R., Cain Z., Assis R., DeGiorgio M. 2024. Robust phylogenetic regression. Syst. Biol., 73:140-157.

874 Casson R.J., Farmer L.D. 2014. Understanding and checking the assumptions of linear
875 regression: a primer for medical researchers. Clin. Exp. Ophthalmol., 42:590-596.
876 Chen Z.-L., Guo H.-J., Niu D.-K. 2023. Dependent variable selection in phylogenetic
877 generalized least squares regression analysis under Pagel's lambda model.
878 bioRxiv:2023.2005.2021.541623.
879 Chen Z.-L., Niu D.-K. 2024. Optimizing variable selection in phylogenetic eigenvector
880 regression for trait correlation analysis. bioRxiv:2024.2004.2014.589420.
881 Cornwallis C.K., Griffin A.S. 2024. A guided tour of phylogenetic comparative methods for
882 studying trait evolution. Annu. Rev. Ecol. Evol. Syst., 55:181-204.
883 Duchen P., Leuenberger C., Szilagyi S.M., Harmon L., Eastman J., Schweizer M., Wegmann
884 D. 2017. Inference of evolutionary jumps in large phylogenies using Levy processes. Syst.
885 Biol., 66:950-963.
886 Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat., 125:1-15.
887 Freckleton R.P., Harvey P.H., Pagel M. 2002. Phylogenetic analysis and comparative data:
888 A test and review of evidence. Amer. Natur., 160:712-726.
889 Gao Y., Wu M. 2022. Microbial genomic trait evolution is dominated by frequent and rare
890 pulsed evolution. Sci. Adv., 8:eabn1916.
891 Garamszegi L.Z. 2014. Modern Phylogenetic Comparative Methods and Their Application
892 in Evolutionary Biology: Concepts and Practice. Berlin, Springer.
893 Garland T., Jr, Harvey P.H., Ives A.R. 1992. Procedures for the analysis of comparative data
894 using phylogenetically independent contrasts. Syst. Biol., 41:18-32.

895 Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B Biol. Sci.,
896 326:119-157.
897 Hadfield J.D., Nakagawa S. 2010. General quantitative genetic methods for comparative
898 biology: phylogenies, taxonomies and multi-trait models for continuous and categorical
899 characters. J. Evol. Biol., 23:494-508.
900 Halliwell B., Holland B.R., Yates L.A. 2024. Multi-response phylogenetic mixed models:
901 concepts and application. bioRxiv:2022.2012.2013.520338.
902 Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation.
903 Evolution, 51:1341-1351.
904 Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings
905 W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E.,
906 Schluter D., Schulte Ii J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T.,
907 Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative
908 data. Evolution, 64:2385-2396.
909 Ho L.S.T., Ane C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait
910 evolution models. Syst. Biol., 63:397-408.
911 Iglewicz B., Hoaglin D. 1993. Volume 16: How to Detect and Handle Outliers. ASQ Quality
912 Press.
913 Ives A.R. 2022. Random errors are neither: On the interpretation of correlated data. Methods
914 Ecol. Evol., 13:2092-2105.

915 Ives A.R., Helmus M.R. 2011. Generalized linear mixed models for phylogenetic analyses of
916 community structure. Ecol. Monographs, 81:511-525.
917 Jarantow S.W., Pisors E.D., Chiu M.L. 2023. Introduction to the use of linear and nonlinear
918 regression analysis in quantitative biological assays. Curr. Protoc., 3:e801.
919 Landis M.J., Schraiber J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes.
920 Proc Natl Acad Sci U S A, 114:13224-13229.
921 Landis M.J., Schraiber J.G., Liang M. 2013. Phylogenetic analysis using Levy processes:
922 finding jumps in the evolution of continuous traits. Syst Biol, 62:193-204.
923 Leys C., Ley C., Klein O., Bernard P., Licata L. 2013. Detecting outliers: Do not use
924 standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc.
925 Psychol., 49:764-766.
926 Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology.
927 Evolution, 45:1065-1080.
928 Maddison W.P., FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation
929 tests for categorical characters. Syst. Biol., 64:127-136.
930 Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: A general
931 approach to incorporating phylogenetic information into the analysis of interspecific data.
932 Amer. Natur., 149:646-667.
933 McArdle B.H. 1988. The structural relationship: regression in biology. Can. J. Zool.,
934 66:2329-2339.

O'Meara B. 2016. Phylogenetic Comparative Method. In: Kliman RM editor. Encyclopedia of Evolutionary Biology. Oxford, Academic Press, p. 254-256.

OpenAI. 2024. ChatGPT-4. https://www.openai.com/.

Osborne J.W., Waters E. 2002. Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval., 8:2.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr., 26:331-348.

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

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526-528.

Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., Harmon L.J. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics, 30:2216-2218.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

Revell L.J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ, 12:e16505.

Rindskopf D., Shiyko M. 2010. Measures of Dispersion, Skewness and Kurtosis. In: Peterson P, Baker E, McGaw B editors. International Encyclopedia of Education (Third Edition). Oxford, Elsevier, p. 267-273.

956 Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: geometric
957 interpretations. Evolution, 55:2143-2160.
958 Royston P. 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and
959 Computing, 2:117-119.
960 Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simulation
961 increase power to detect early bursts of trait evolution. Syst Biol, 63:293-308.
962 Smith E.G., Surm J.M., Macrander J., Simhi A., Amir G., Sachkova M.Y., Lewandowska M.,
963 Reitzel A.M., Moran Y. 2023. Micro and macroevolution of sea anemone venom phenotype.
964 Nat. Commun., 14:249.
965 Sumner S., Favreau E., Geist K., Toth A.L., Rehan S.M. 2023. Molecular patterns and
966 processes in evolving sociality: lessons from insects. Philos Trans R Soc Lond B Biol Sci,
967 378:20220076.
968 Uyeda J.C., Zenil-Ferguson R., Pennell M.W. 2018. Rethinking phylogenetic comparative
969 methods. Syst. Biol., 67:1091-1109.
970 Westoby M., Yates L., Holland B., Halliwell B. 2023. Phylogenetically conservative trait
971 correlation: Quantification and interpretation. J Ecol, 111:2105-2117.
972 Wilcox R.R. 2003. 3 - Summarizing data. In: Wilcox RR editor. Applying Contemporary
973 Statistical Techniques. Burlington, Academic Press, p. 55-91.
974 Zheng L., Ives A.R., Garland T., Larget B.R., Yu Y., Cao K. 2009. New multivariate tests
975 for phylogenetic signal and trait correlations applied to ecophysiological phenotypes of nine
976 *Manglietia* species. Funct. Ecol., 23:1059-1069.

977