Title: Detecting mosaic patterns in phenotypic disparity

Caroline Parins-Fukuchi

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109.

Keywords: mosaic evolution, continuous traits, tempo and mode, macroevolution, phylogenetics.

Contents: Abstract, Introduction, Materials and Methods, Results/Discussion, Supplementary Information, Bibliography, Figures 1-8, Online Appendix.

Word count: 5639

Abstract

Understanding the complex patterns underlying phenotypic diversification across the tree of life has long been a fundamental aim in evolutionary biology. The modern evolutionary synthesis was characterized by the integration of disparate biological fields to better understand the diverse processes that drive phenotypic change. The centrality of the 'mosaic' evolutionary patterns that emerge from these diverse processes is a key feature of post-synthesis thought. Despite this fundamental importance, researchers have been limited in their ability to explore such complex patterns in nature analytically. However, recent advances in data collection provide new potential for investigating broad patterns shaping variation across the tree of life. These possibilities emphasize the need for new comparative approaches that will facilitate the examination of mosaic change. In this study, I introduce a novel comparative framework that accommodates the mosaic evolutionary processes that shape whole organisms. The approach recasts the comparative method to harness patterns in phenotypic disparity to identify macroevolutionary suites of continuous traits. Through demonstrations on simulated and empirical data, I demonstrate the utility of this new framework. This framework offers a first step toward evaluating mosaic variability comprehensively across whole organisms — an integrating goal of the modern synthesis.

Introduction: Characterizing the ways in which phenotypic disparity and evolutionary rates differ across lineages and throughout time has long been a central goal in evolutionary biology. Shifts in the rate of phenotypic change often coincide with the rapid diversification and ecological innovation of diverse lineages (Rabosky and Adams 2012; Rabosky *et al.* 2013). In addition, studying fluctuations in the tempo of evolutionary change can shed light on our knowledge of important evolution processes, such as adaptation and evolutionary constraint. A more recent body of work has focused on the development and application of phylogenetic comparative methods (PCMs) that identify patterns of phenotypic change using phylogenies (Harvey and Pagel 1991; Hansen 1997a; Butler and King 2004; O'Meara *et al.* 2006; Eastman *et al.* 2011; Beaulieu *et al.* 2012). Modern PCMs enable researchers to infer shifts in evolutionary rate, constraint, and disparity throughout time by applying stochastic models of trait evolution to comparative phenotypic data using phylogenetic trees. These approaches have helped to answer important questions concerning the tempo and mode of phenotypic evolution across deep timescales (Harmon *et al.* 2003; Scales *et al.* 2009; Harmon *et al.* 2010; Rohlfs *et al.* 2013; Slater 2013).

Paleobiological and comparative studies have typically examined univariate evolutionary patterns. This has perhaps grown from the tendency of early foundational work to document evolutionary patterns using single key traits, such as molar shape (Simpson 1944; Gingerich 1974). Although often limiting the ability to ask detailed evolutionary questions at broad scales, this framework has been effectively leveraged to alternately explore specific case studies (Scales *et al.* 2009), or general questions using single gross morphological traits, such as body mass, as an approximation for overall phenotypic variation (Harmon *et al.* 2003; Beaulieu *et al.* 2007; Harmon *et al.* 2010; Burbrink and Pyron 2010; Rabosky *et al.* 2013; Zanne *et al.* 2014; Bokma *et al.* 2015; Landis and Schraiber 2017).

Despite these past successes, a central goal in post-synthesis evolutionary biology has been to evaluate patterns across anatomical regions and at different phenotypic levels. Such broad investigations require approaches that can accommodate multiple characters. Recent contributions have explored multivariate approaches that enable the analysis of multiple traits simultaneously (Adams 2014). These enable inference of trait models commonly employed in univariate PCMs among high-dimension data. Another set of methods statistically evaluates correlated evolutionary change between multiple traits along a phylogenetic tree (Revell and Harmon 2008; Caetano and Harmon 2018). These methods will be critical extensions to PCMs moving forward, given the current influx of large, publicly available databases of morphology (Boyer *et al.* 2016), and emerging approaches for the rapid, algorithmic quantification of variation from digital specimen images (Boyer *et al.* 2015; Pomidor *et al.* 2016).

One aspect that has been under-explored in a comparative analytical context has been the tendency for distinct subsets of traits to display unique patterns in evolutionary rate and relative disparity. The historical focus on univariate analysis in PCMs and paleobiology may have contributed to a general lack of recognition of the diversity of evolutionary patterns that can combine to organismal body plans. Nevertheless, mosaic evolution has been a fundamental concept in evolutionary biology since the modern synthesis in its acknowledgement of the reality that anatomical regions are often exposed to natural selection at differing magnitudes and directions at different times (Stebbins 1983). The mosaic concept dates to Dollo (Gould 1970), and has frequently been invoked qualitatively as a key factor driving divergent morphological adaptation in different anatomical regions across lineages (De Beer 1954; Cracraft 1970; Mayr 1970; Gould 1977b; Stanley 1979). Although examination in a quantitative context has been limited overall, several key studies have shown that mosaic patterns explain the emergence of

important traits in diverse case studies, such as structural variation in the brain across mammals (Barton and Harvey 2000), phenotypic and genomic diversity across angiosperms (Stebbins 1984), and the defining suite of morphological characters displayed by the hominin lineage (McHenry 1975; Gould 1977a; Holloway and Post 1982).

In addition to being of fundamental biological interest, mosaic patterns have long been argued to present unique challenges when inferring phylogeny from morphological characters (Farris 1971). This concern has recently been reasserted by Goloboff and colleagues (2018), who suggest that heterogeneity in relative disparity, as measured by phylogenetic branch lengths, displayed across separate suites of morphological characters confounds phylogenetic inference from morphological traits using Bayesian and maximum likelihood approaches. These concerns parallel the conclusions of important recent studies that validate the prevalence of mosaic patterns in paleobiological and comparative data by manually testing the variability in evolutionary mode that can occur across large morphological datasets (Hopkins and Lidgard 2012; Felice and Goswami 2018). These studies demonstrate the urgent need for overdue extensions to existing uni- and multivariate PCMs that accommodate mosaic patterns in large phenotypic datasets. Methods that separate traits according to overall rate have long been available for phylogenetic inference (Yang 1996; Schraiber *et al.* 2013). However, these approaches do not adequately address mosaic patterns, which focus more on heterogeneity in relative disparity across lineages rather than absolute rate. Despite the importance, there has not yet been a computational approach that algorithmically partitions traits according to their patterns in disparity.

In this paper, I present a novel method that identifies suites of continuous traits displaying shared patterns in disparity to reconstruct the mosaic trends that have shaped organismal phenotypes. The mosaic character suites identified by the approach are defined by partitions of

traits that are best explained by shared phylogenetic branch lengths, measured in units of average disparity, along a fixed topology. The phylogenetic models underlying the construction of mosaic character suites, by representing the accummulation of disparity across lineages, thus provide information on relative rates of evolution. After introducing the method, I evaluate its performance using simulated data. I also present an analysis of an empirical dataset of developmental ossification times complied by Rose (2003). This dataset has previously been leveraged to explore mosaic heterogeneity in evolutionary pattern (Germain and Laurin 2009; Laurin 2014) and so is well suited as an empirical test of the method that I introduce here.

Methods and Materials

Code and data availability

The approach described below is implemented in a program called *greedo*. It is available freely on Github at (links are available from the journal office). All analyses on simulated and empirical data were performed using this program. Scripts and data used for the simulated and empirical analyses are also available on Github (links available from journal office).

Partitioning traits into mosaic suites

The method described here combines several unsupervised learning strategies to partition traits into separate mosaic character suites, with each possessing its own set of phylogenetic branch lengths expressed in units of disparity. These strategies are applied in sequence (Fig. 1), with the goal of identifying the configuration that yields the lowest AIC score.

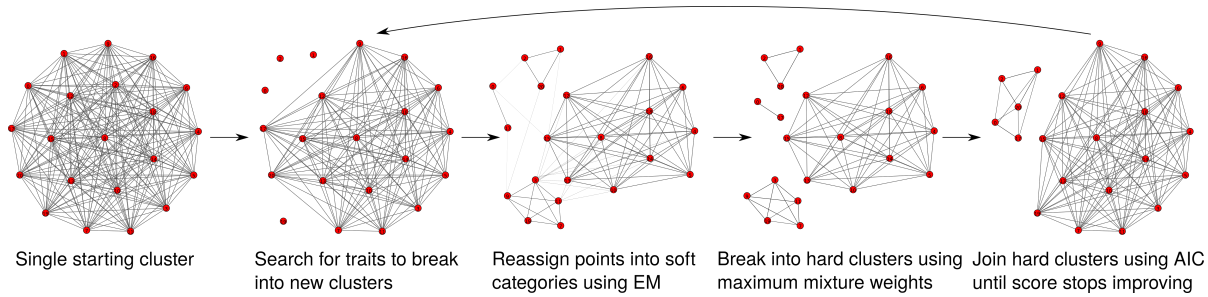

Figure 1. Search procedure to identify mosaic suites.

Each mosaic character suite that defines the classification model contributes to the likelihood independently. The log-likelihoods of each of the traits belonging to character suite *j* are calculated under the associated branch lengths and added to yield an independent log-likelihood. The log-likelihood of the trait matrix, $\mathrm{LL}_{\mathrm{classification}}$, is calculated by summing the log-likelihoods of all *k* character suites:

$$LL_{classification} = \sum_{j=1}^{k} LL_j \tag{1}$$

Tree model

Each mosaic character suite of the classification model is defined by a phylogeny where the topology is fixed, but its branch lengths are free to vary and calculated from the constituent traits. Branch lengths are expressed in units of disparity and are calculated using a Brownian model of evolution. The distribution underlying the traits belonging to each partition are assumed to be multivariate Gaussian, with variances between taxa defined by the product of their evolutionary distance measured in absolute time and the instantaneous rate parameter (σ^2). The phylogenetic comparative methods literature often estimates σ^2 alone by assuming a fixed timescale given by branch lengths that have been scaled to absolute time using a molecular clock model. However, here the absolute times are assumed to be unknown, and the rate and time parameters are

confounded with one another. Thus, branch lengths are expressed in units of average morphological disparity, or variance per trait.

This parameterization differs from the typical use of stochastic continuous trait models in comparative studies. Traditional approaches generally start by applying single rate (clock-like) Brownian models to chronograms, comparing the fit to more complex models. For instance, σ^2 is sometimes allowed to vary locally in distinct subclades to yield 'multi-rate' BM models (O'Meara *et al.* 2006; Eastman *et al.* 2011; Thomas and Freckleton 2012). Expanding upon this, more parameter rich models have also been developed, such as Ornstein-Uhlenbeck (OU) (Hansen 1997b) and jump-diffusion (JD), or Lévy, processes (Landis *et al.* 2012). OU models expand upon standard BM by introducing a term, α, that generates a stabilizing force that constrains movement around an optimal trait value, while JD processes contain terms describing the frequency of jumps in character spaces. OU models are often used similarly to multi-rate BM, by allowing optimal trait values to vary across the tree (Butler and King 2004; Beaulieu *et al.* 2012). JD models relax single-rate BM by allowing sudden jumps in mean trait values (Eastman *et al.* 2013; Landis and Schraiber 2017).

The parameterization of the tree and branch lengths used here is continuous equivalent to that encountered in phylograms reconstructed from molecular and discrete morphological data. These express branch lengths in units of substitutions per site by similarly confounding the rate and time parameters. As used here, this approach has several benefits over the more common comparative approaches described above. Although it may be possible in principle to develop a similar approach that assigns traits to suites associated with separate OU models fit to chronograms with different adaptive regimes, or best explained by distinct BM and JD processes, the tree model used here captures much of the same information (Fig. 2), while drastically simplifying inference

from both a statistical and computational standpoint. While multi-rate BM, OU, and JD models are often aimed to describe explicit evolutionary processes, such as environmental adaptation or quantum evolution, they may simply improve model fit over simpler models by accommodating heterogeneity in disparity accumulated across the tree. Several studies have demonstrated substantial limitations in the ability to identify the extended OU model parameters using typical comparative datasets (Ho and Ané 2014; Cressler *et al.* 2015; Cooper *et al.* 2016), complicating the use and evolutionary interpretability of these models in empirical studies.

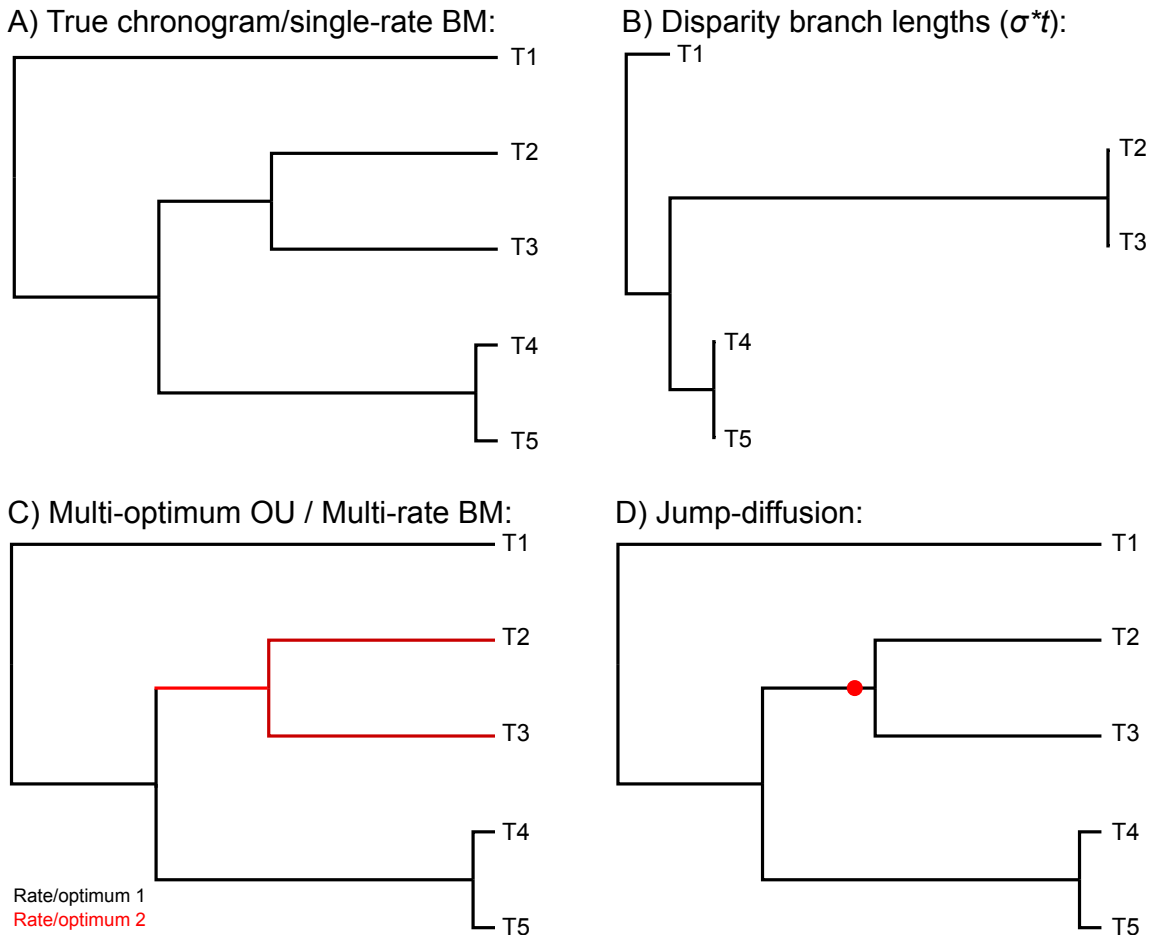

Figure 2. Fitting increased disparity observed in taxa T2 and T3. A) Single rate BM fit to a chronogram assumes clock-like evolution, ignoring both positive jumps in disparity and stasis. B) Expressing branch lengths in units of disparity under the phylogenetic Brownian parameterization

used here captures both increased disparity and stasis. C) Fitting multiple Brownian rates or shifts in adaptive optima or D) assuming the presence of a single jump in optimum trait value may explain the same pattern with more model parameters. Note that panels B, C, and D all display similar information, while panel A assumes clock-like evolution.

The likelihood is calculated in a recursion from the tips to the root after Felsenstein (1973). Full derivations of the likelihood and algorithm are also given by Felsenstein (1981) and Freckleton (2012), and summarized briefly here. The tree likelihood is computed from the phylogenetic independent contrasts (PICs) using a 'pruning' algorithm. Each internal node is visited in a postorder traversal, and the log-likelihood, L_{node} is calculated as univariate Gaussian, with a mean equal to the contrast between the character states, x_1 and x_2 at each subtending edge and variance calculated as the sum of each child edge, v_1 and v_2:

$$L_{node} = \frac{1}{2} * \frac{\log(2\pi) + \log(v_1 + v_2) + (x_1 - x_2)^2}{v_1 + v_2} \tag{2}$$

The PIC, x_{internal}, is calculated at each internal node and used as the character representing the internal node during the likelihood computation at the parent node. The edge length of the internal node, v_{internal} is also extended by averaging the lengths of the child nodes.

$$x_{internal} = \frac{(x_1 * v_2) + (x_2 * v_1)}{v_1 + v_2} \tag{3}$$

$$v_{internal} = v_{internal} + \frac{(v_1 * v_2)}{(v_1 + v_2)} \tag{4}$$

The total log-likelihood of the tree, L_{tree} is calculated by summing the log-likelihoods calculated at each of the n internal nodes.

$$L_{tree} = \sum_{node=1}^{n} L_{node} \tag{5}$$

Branch lengths are estimated by iteratively solving the analytical solution to the maximum likelihood (ML) branch lengths for a 3-taxon star topology. In this procedure, the tree is treated as unrooted. Picking a single internal node, PICs are calculated to each of the three connected branches. These are treated as 'traits' at the tips of a three-taxon tree. The edge lengths of the pruned tree (v_i) are then computed analytically using the MLE solutions for a three taxon tree (Felsenstein 1981). This procedure is performed on all of the internal nodes. This process is iterated until the branch lengths and the likelihoods converge, yielding a local optimum of the likelihood function. The algorithm and derivation of the 3-taxon ML solutions are given a detailed explanation by Felsenstein (1981).

Search procedure

All traits start in a single shared partition. From here, traits that exhibit an improved likelihood in their own partition compared to their current placement are split into their own cluster. A penalty is imposed that is proportional to the difference in size between the existing partitions, as measured in the number of constituent traits. This functions similarly to a Dirichlet process prior by forcing traits to prefer belonging to partitions with a larger number of traits. This constraint is designed to prioritize traits with strongly divergent signal and to discourage overfitting of the clustering model. As a result, only traits with a strong preference for the new component over the existing component are selected. This step is repeated either until the number of occupied categories reaches a user-specified maximum threshold, or there are no more traits left to separate.

From here, the problem is temporarily recast as a finite mixture model, with the number of

components corresponding to the user-specified value. First, membership weights are calculated for each trait-component pair as the probability of the trait (x_i) belonging to each *j* of *K* components. This value is calculated for each component as the proportion of the likelihood of x_i (L_{ij}) under the corresponding set of branch lengths relative to the summed likelihoods of x_i under all *K* components.

$$P(x_i|K_j) = \frac{L_{ij}}{\sum_{k=1}^{K} L_{ik}} \tag{6}$$

Expectation-maximization (EM) (Dempster *et al.* 1977) is performed to update the mixture weights and the branch length parameters. The branch lengths associated with each component are updated as part of the mixture model, with each site in the matrix contributing to the branch lengths in each component according to the weights defined above. During this step, the model could be thought of as a variation of a typical multivariate Gaussian mixture model, where the covariance matrix is constrained to reflect the structure of a phylogenetic tree, since the phylogenetic Brownian model yields a multivariate Gaussian likelihood function.

Once the mixture model has been updated for several iterations, the components are broken into hard clusters, with the assignment for each site chosen to be the component with the maximum mixture weight. This arrangement is then reduced in an agglomerative manner. At each step of this procedure, the pair of components that results in the greatest improvement in AIC, calculated using the classification likelihood defined above, is merged. This hierarchical merging continues until either the AIC score cannot be further improved, or only a single component is left. The entire procedure is then repeated from this reduced configuration for a user-specified number of iterations.

Stochastically varying mosaic patterns

To examine the strengths and shortcomings of the method in detecting multiple suites of traits that vary stochastically, I performed tests using simulated datasets. A single topology of 20 taxa was simulated under a pure-birth model. For each partition of continuous traits, a new set of branch lengths was generated by drawing randomly from either a gamma or exponential distribution, then simulated under Brownian motion. The rate parameter of the Brownian process was set to 1 across the entire tree so that the matrices reflected the scale and heterogeneity of rates resulting from the altered branch lengths. Each trait matrix contained a single partition simulated under the original ultrametric branch lengths. The randomly drawn branch lengths were intended to mimic the differing rates of evolution that can be experienced by different lineages during evolutionary divergence, with the ultrametric branch lengths reflecting clock-like evolution (Fig. S1). This procedure resulted in highly complex simulated datasets that tested the ability of the method to detect mosaic structure of high dimensionality. All trees and traits were simulated using the phytools package in R (Revell 2012a).

Using this procedure, datasets comprised of 2, 3, and 4 partitions of 50 continuous traits each were generated. All traits were rescaled to a variance of 1. I ran *greedo* on these datasets to attempt to reconstruct these partitions. The maximum number of clusters for these runs was set to half the number of traits in each matrix.

Power analysis and detection of relative rate shifts

I performed another simulation experiment to evaluate the statistical power and limits of the method under more controlled conditions. To evaluate the performance of the method computationally, I expanded the size of the simulated trees to 100 taxa. In this trial, I simulated datasets using a procedure modeled after Eastman *et al.* (2011). First, I simulated 100 unique

pure-birth trees. I then altered the branch lengths to all be equal, with the trees scaled to a total height of 1. This yielded a set of 100 'equal rates' trees that displayed uniform disparity across all branches. To simulate heterogeneity in evolutionary pattern, I then randomly selected once clade in each of the 100 trees, and multipled the branch lengths by factors of 8, 16, 32, and 64. Like in the tests performed by Eastman *et al.* (2011), these rates were inherited across the entire clade. Clades randomly chosen for shifts were constrained to those containing at least 10, and no more than 90, terminal taxa. This resulted in a set of 400 'rate shift' trees of differing magnitude. I then compiled datasets by simulating 50 traits on the trees with equal branch lengths and the trees displaying rate shifts, respectively, with the goal to ascertain the ability of the method to separate traits displaying a shift in rate from those simulated under uniform rates. The 400 resulting datasets contained 100 traits, half of which were simulated along an equal rates tree, and the other half along a rate shift tree of one of the four shift magnitudes. To examine the prevalence of type 1 error, I also simulated datasets of 100 traits along each of the equal branch length trees to test whether the method correctly identified only one suite of traits. Like above, all traits were scaled to a variance of 1 to ensure that the method was correctly identifying heterogeneous patterns in relative disparity, rather than finding heterogeneity reflected in empirical variance, which would unfairly bias results in its favor. This step was also performed in the test described below.

Distinguishing rate shifts among covarying traits

In nature, many continuous traits are expected to covary. Such covariance has been suggested as a challenge to phylogenetic analyses (Felsenstein 1988) by increasing complexity and bias in morphological datasets. To test the behavior of my approach when detecting mosaic patterns among traits that form covarying modules, I performed an additional experiment. I simulated covarying datasets along the 16x and 64x rate shift trees generated above. Again, 50 traits each

were simulated along the equal lengths and rate shift trees, respectively, but with the traits separated into two covarying sets of 25 traits. These were compiled into datasets of 100 traits, displaying two distinct evolutionary patterns (equal and shifted rates), and four separate covarying modules. To examine the effect of covariance on type 1 error, I also simulated datasets of 100 traits along the equal rates trees, as above, but separated into four covarying modules. The strength of covariance is likely to affect results, and so data were simulated at correlational intervals of 0.1 (weak), 0.5 (moderate), and 0.9 (strong). Since the approx'ach introduced here does not explicitly model evolutionary or phenotypic integration, this test examined the extent to which these common patterns confound the identification of mosaic patterns in relative disparity. All simulated datasets were generated using the 'fastBM()' function in phytools (Revell 2012b).

Evaluation of reconstruction accuracy

I used the adjusted Rand index (ARI) to evaluate the accuracy of the inferred partitionings (Hubert and Arabie 1985) against the true partitionings. The RI measures congruence by counting the pairs of elements that either occupy the same or different clusters in both of the two clusterings, and calculating the proportion of this value relative to all of the possible permutations of elements. As a result, the RI can range from 0, indicating total disagreement, and 1, indicating total agreement. The ARI corrects for the propensity for elements to be randomly placed within the same cluster, with a value of 0 indicating a result indistinguishable from a random assignment of elements, and 1 indicating complete congruence. The metric takes negative values when a clustering is worse than random.

Empirical analysis

To examine the performance of the method on empirical data, I analyzed a dataset comprised of the ossification sequences of 21 cranial bones sampled across 21 taxa obtained from Laurin

(2014), initially assembled by Rose (2003). Laurin identified suites of traits displaying distinct phylogenetic patterns using an 'evolutionary' principal component analysis (PCA) and also performed a distance-based hierarchical clustering of the data, making these data well-suited to a test of the method introduced here. I was interested in evaluating the ability of my new approach to detect patterns in relative, rather than absolute, rates across lineages, and so I standardized the variance between the traits to 1. The tree presented in the original study was obtained from the authors and used to perform the comparative analyses here.

Results and Discussion

Simulated stochastically varying rates

The method is generally able to recover the structure of the simulated datasets. The number of inferred mosaic suites is usually correct, and ARI values are typically well above random. The two-partition analyses are very accurate, with high ARI values, and nearly always correctly identifying the correct number of clusters. The three- and four-partition analyses were less accurate, but still yield results much higher than random, and typically recovering the correct number of character suites. ARI indices achieved for the three- and four- partition analyses are similar to results from simulated data using general Gaussian clustering approaches, such as Gibbs sampling under a Dirichlet process (Dahl 2006).

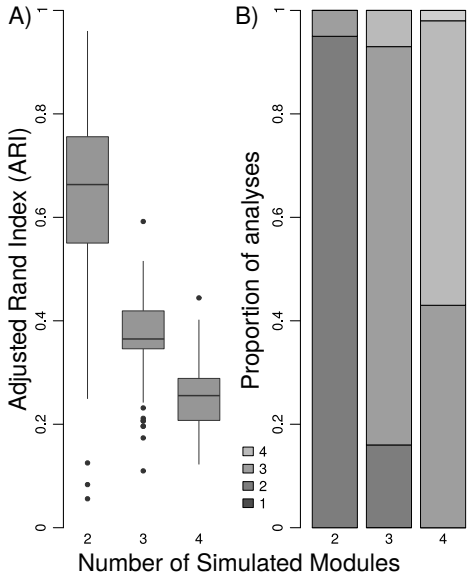

Figure 3. A) Adjusted Rand indices across reconstructions of simulated datasets. B) Number of clusters resulting from analyses of simulated data. Barplots are stacked to represent the proportion of replicates that resulted in *k* character suites. All violin and barplots are separated by true number of simulated character suites.

Despite the generally encouraging results from the simulated data, the trend toward decreasing accuracy when components are added suggests: 1) a limitation of the method in adequately exhausting the search space of component assignments or 2) a limitation in the power of the approach to detect subtle, stochastically varying differences in branch-wise evolutionary disparity. It may be helpful to initialize the analysis using results obtained from a less intensive approach, such as the evolutionary PCA developed by Laurin (2014).

Relative rate shifts in independently evolving traits

The method performs well in distinguishing traits evolved under a single rate shift from those under equal rates (Figs. 4a and 4b). The most subtle rate shift (8x faster) is detectable in most of the datasets, although mean accuracy is slightly lower than the 2-suite datasets tested above. This

difference is due to decreased power in detecting this more subtle heterogeneity. The method underfits the model in 8x rate shift datasets 30% of the time, by assigning all of the traits to a single shared character suite (Fig. 4b). Power to detect mosaic structure increases with the magnitude of the shift, with the 16x datasets correctly binning the traits into two character suites 95% of the time. The 32x and 64x datasets always correctly identify the number of character suites. Accuracy also increases with the strength of the rate shift. While ARI values in the 8x dataset are all acceptably high, reconstruction error approaches zero as the magnitude of the rate shifts are increased to 64x.

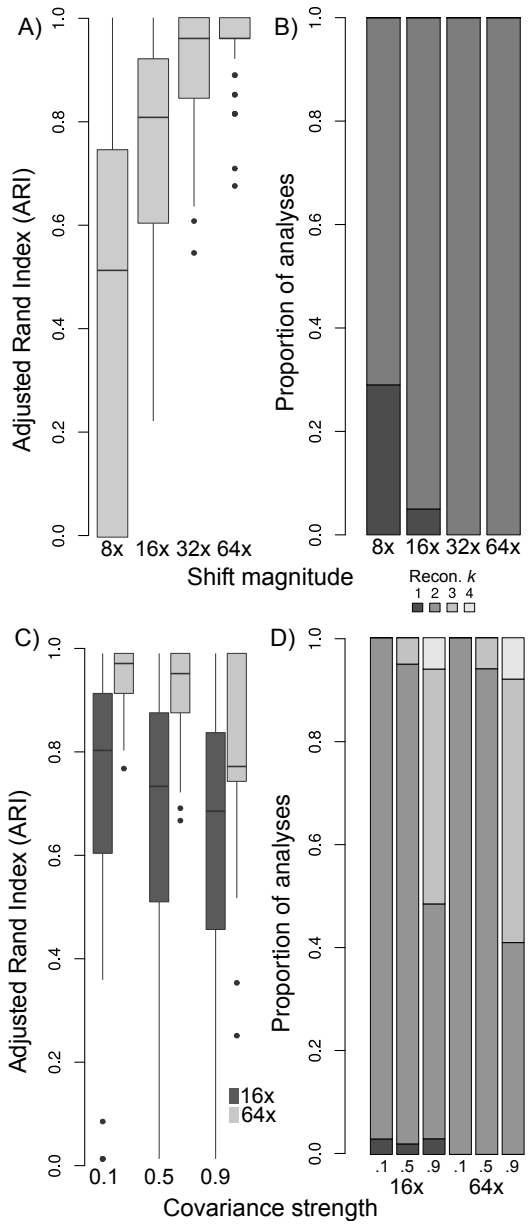

Figure 4. A) Adjusted Rand indices from reconstruction of datasets comprised of mosaic suites with one simulated under equal rates across branches, and the other displaying a randomly placed rate shift at magnitudes of 8, 16, 32, and 64. B) Stacked barplots to show the frequency that each of *k* clusters are inferred from the simulated datasets. C) and D) show the same metrics as above, summarized from datasets containing simulated covarying modules of characters.

Relative rate shifts in covarying traits

Evolutionary covariance across traits has a detectable, but not overwhelmingly deleterious effect on the capability of the approach in categorizing the number of character suites present in the data (Figs. 4c and 4d). When the simulated modules weakly covary (0.1), the method performs comparably to the independently evolving examples. However, accuracy decreases with the strength of the covariance, with reconstruction being notably less accurate in the strongly covarying datasets. This decrease in accuracy at high levels of covariance can be explained by the tendency for the approach to overfit the datasets by inferring the presence of more than two character suites in 59% of cases (Fig. 4d).

Type 1 error

As with the multi-rate covarying examples above, the method has a tendency to overfit datasets that display high levels of covariance. Nevertheless, the method fits the data correctly in the presence of no and weak covariance (Fig. 6). As with above, results generated from datasets suspected to display particularly high levels of covariance should be inspected to ensure that mosaic suites are not overly similar in their resultant branch lengths. As a more immediately tenable alternative to full multivariate estimation, future extensions would allow the ability for users to specify covarying modules of traits using a conventional approach (Goswami and Finarelli 2016), to ensure that strongly covarying sets of traits are placed together. And nevertheless, even in the absence of these tools, it should be noted that the extremely high and rampant levels of covariance that were needed to induce overfitting in both the type 1 and type 2 tests may be unlikely in certain types of empirical datasets.

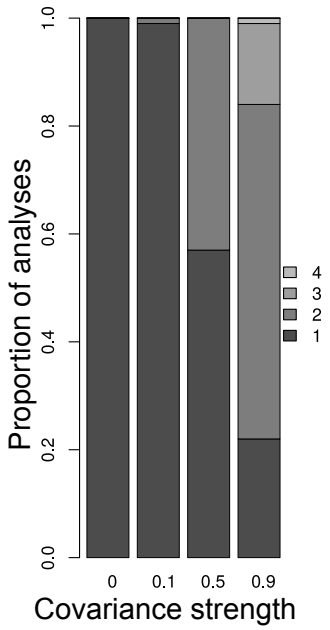

Figure 5. Type 1 error. Reconstructions show no type 1 error when traits are evolved independently, or display weak covariance. Type 1 error increases at moderate and high levels of covariance, as the method overfits biased patterns to place covarying modules in superfluous mosaic suites.

Empirical analysis

Four separate runs each yielded different partitionings into two mosaic suites. All arrangements overlap in their assignments, and the AIC scores are all close to one another. To visualize the overall support for the categorization of each trait across partitionings, I calculated the AIC weight of each model (Burnham and Anderson 2002). The AIC weight of model *i*, w_i can be interpreted as its probability of being the best model among a set of *K* candidates.

$$w_i = \frac{L_i^{rel}}{\sum_{k=1}^{K} L_k^{rel}} \tag{7}$$

where L_i^{rel} is the relative likelihood of model *i*:

$$L_i^{rel} = \exp(-0.5(AIC_i - AIC_{min})) \tag{8}$$

These weights were used to visualize the the strengths of the support connecting traits across all the four best partitionings in a graph (Fig. 2b). An edge was drawn between traits *i* and *j* if they occurred in the same component in any of the four results, with a weight given by the summed AIC weights of all of the configurations where *i* and *j* occur in the same character suite. The maximum weight possible is 1.0, when traits *i* and *j* share a suite in all of the configurations. The resulting graph suggests that traits 0, 1, 2, 17, and 20 all form a suite, with the rest of the traits sharing a separate suite. This result is very close to the pattern reconstructed by Laurin (2014) using an 'evolutionary PCA' approach, differing only in the assignment of the stapes (Table S1). This is likely because the approach developed by Laurin seeks to join traits that display similar values in phylogenetic independent contrasts (PICs), and so focuses on patterns similar to the phylogenetic Brownian likelihood model used here.

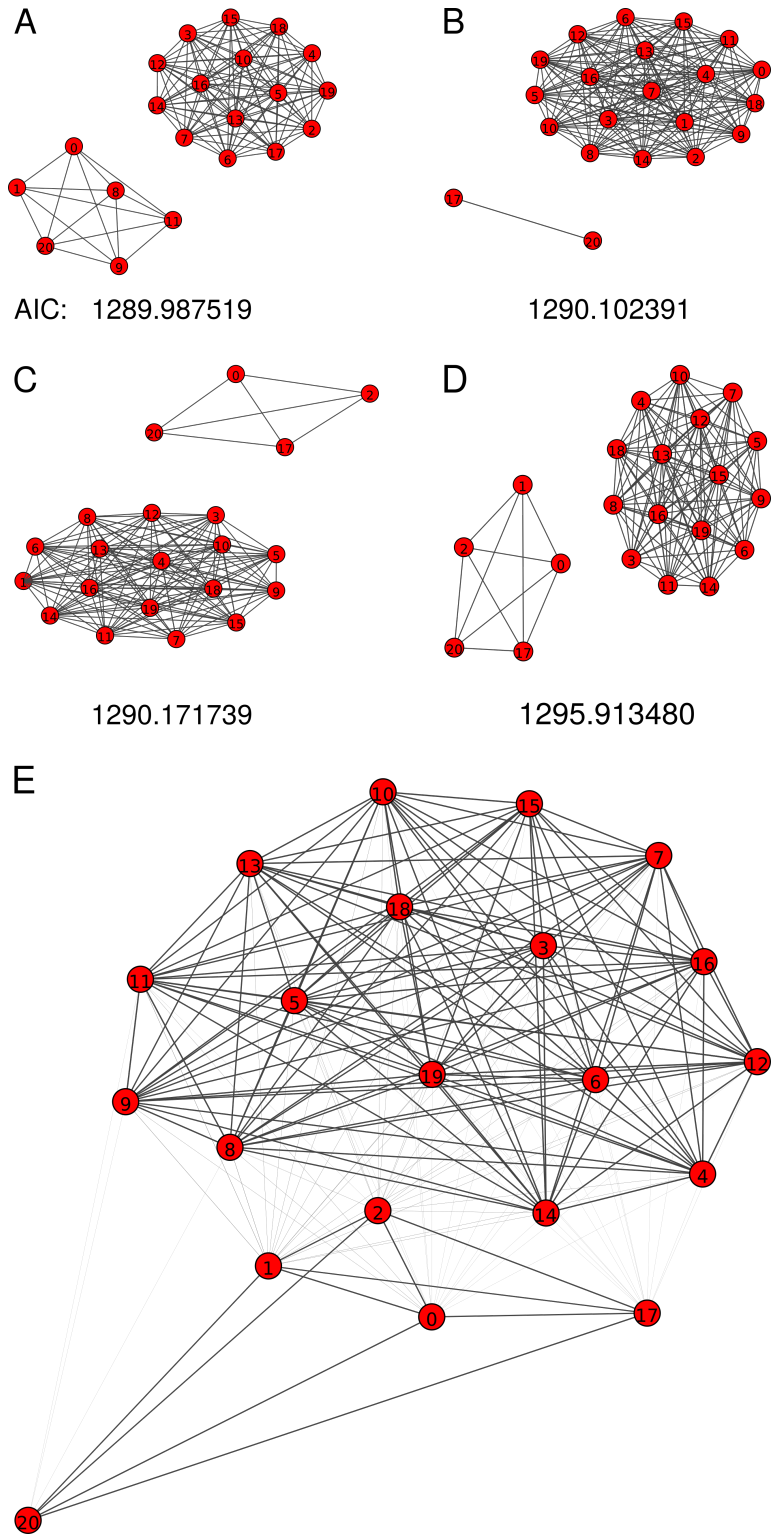

Figure 6. A-D) Four best configurations with AIC scores. E) Weighted graph calculated by summing the AIC weights associated with the each model to form edges and edge weights. All

graphs were drawn using the "lgl" format implemented in igraph.

In his original study, Laurin (2014) also performed an exploratory hierarchical clustering of the ossification times and found substantial differences in structure as compared to that revealed by his evolutionary PCA approach. This discordance occurs because the the PCA considers covariance between PIC values, while the hierarchical clustering only reflects shared similarity in absolute value. Like the original study, the results here differ substantially from the pattern resulting from the exploratory hierarchical clustering performed by Laurin, instead aligning very closely to the PCA approach. This is reassuring for the performance of my method, as Laurin considered the evolutionary PCA to yield the correct answer, and the hierarchical clustering to demonstrate the inadequacy of raw similarity in delimiting meaningful patterns (Laurin, pers. comm.). Although they differ in the specific criteria used to identify character suites, the similarity in results between my and Laurin's method are not surprising. Laurin's PCA method identifies structure from patterns in covariance that have been corrected for phylogenetic non-independence, whereas my method identifies a minimally complex set of models, defined by phylogenies with non-negative branch lengths. The trees describing each suite in my method may be thought of as representing disparity between taxa as patristic distances. And so, although they differ in formulation and statistical paradigm, both approaches share overlap in their treatment of phenotypic variation.

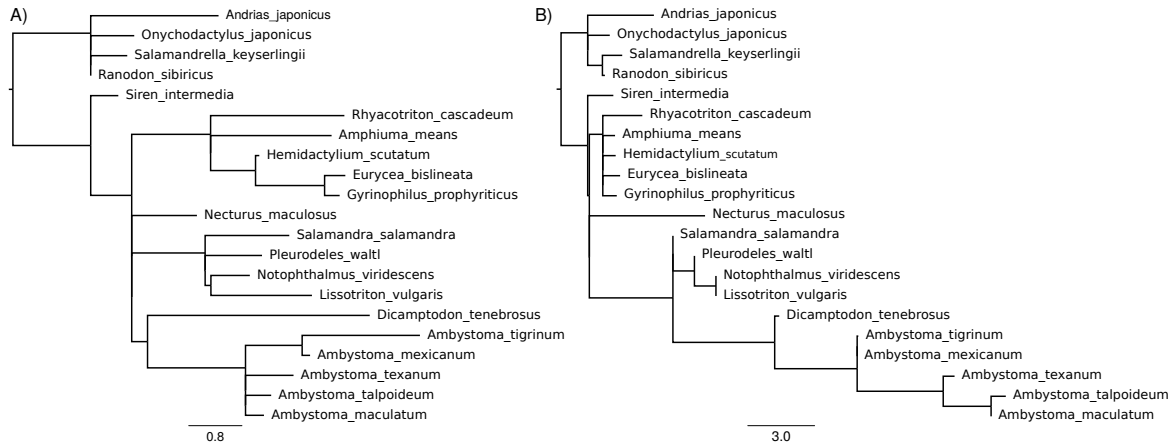

Figure 7. Branch lengths reconstructed from traits contained within: A) suite 0 (Table S1); B) suite 1 (Table S1) display distinct patterns in disparity.

Utility in transcriptomic studies

In addition to morphological phenotypes, the method described here may be useful in identifying mosaic suites among molecular phenotypic traits, such as normalized gene expression levels. Expression data have been increasingly examined in a comparative, phylogenetic context, but previous studies have not had a meaningful way in which to partition sets of genes. As a result, researchers typically fall back on methods such as binning all genes expressed in the transcriptome together into a single analysis (Chaix *et al.* 2008), binning genes based upon functional pathways (Schraiber *et al.* 2013), and using non-phylogenetic clustering approaches (Brawand *et al.* 2011). The method described here may benefit such studies by identifying the major axes of variation in evolutionary pattern across transcriptomic datasets.

Utility to phylogenetic comparative methods

By identifying suites of characters that display similar patterns in disparity across lineages, my approach seeks to integrate existing work that takes a broad view of the tempo and mode of phenotypic evolution with under-examined patterns in mosaic evolution. Although the tendency

for different traits to evolve according to different patterns is expected and well documented (Stanley 1979; Stebbins 1984), there has not yet been an approach that explicitly incorporates phylogeny to reveal the complex mosaic of patterns in divergence underlying the evolution of phenotypes. The analyses of simulated and empirical data showed the capability of my new method to identify biologically meaningful suites of continuous traits that reflect differences in their patterns in disparity across taxa.

Although Brownian motion is often interpreted in comparative analyses as a neutral process of phenotypic change reflecting genetic drift (e.g., Butler and King 2004) occurring under a single rate, the parameterization used in my approach is more general. As in previous approaches (Felsenstein 1981), rate and time are confounded with one another. As a result, a long branch representing high disparity to adjacent lineages could reflect either a fast rate, or a long time of divergence. As a result, a tree with heterogeneity in branch lengths could express variation in evolutionary rates across lineages, or tips that were sampled at different points in time. Since phenotypic disparity can be generated by a broad range of processes at the population level, the phylogenetic Brownian model used here does not assume that the traits are selectively neutral. By representing branch lengths in terms of phenotypic disparity, my method is able to capture much of the same information from the data that is sought by more parameter-rich PCMs (Fig. 2). As a result, the approach could be form an analytical basis for delimiting suites of traits evolving under distinct evolutionary regimes, such as those found manually by Hopkins and Lidgard (2012). The suites recovered using this method could then be placed into a context more similar to existing PCMs by mapping the disparity branch lengths to a time-scaled phylogeny. This would enable the exploration of absolute rates throughout time, facilitating the exploration of patterns in a framework similar to existing PCMs.

Previous studies have shown that morphological (Lynch 1990) and gene expression phenotypes (Yang *et al.* 2017) often display patterns in rate that are not easily distinguishable from conservative evolutionary forces such as genetic drift and stabilizing selection. Nevertheless, comparative analysis of key traits used in classic studies (Simpson 1944; Gingerich 1983, 1993) have shown that certain features can show substantial variation in rate across lineages that can provide crucial evolutionary insights. As phenotypic datasets increase in scale, it may become increasingly likely for the signature underlying interesting patterns in tempo and mode to become swamped by large numbers of conservatively evolving traits. By separating traits according to their implied patterns in disparity, the approach introduced will help to identify less conservatively evolving traits that may be of particular interest for more detaileds comparative analyses.

Utility for phylogenetic inference and divergence time estimation

In addition to being comparatively interesting in their own right, the mosaic suites identified by my approach will be useful in methods for inferring phylogeny and divergence times. Several recent articles have demonstrated the strong potential for continuous traits as an alternative to discrete traits when reconstructing phylogeny (Parins-Fukuchi 2018b,a) and divergence times (Alvarez-Carretero *et al.* 2018). Despite these successes, one persistent issue in morphological phylogenetics more generally has been the difficulty in accommodating the expected heterogeneity in branch lengths across traits. This has been argued to be a foundational limitation to model-based approaches by cladists when demonstrating the prevalence of these patterns across empirical datasets (Goloboff *et al.* 2018). Error stemming from the effect of mosaic branch lengths may be expected to exhibit an even greater effect on divergence time estimation, which depends on calculating rates from branch-wise patterns in disparity.

Although long postulated to be a problem in phylogenetic inference, there has not yet been a

computational approach that delimits suites of traits displaying similar patterns in branch lengths. The approach introduced here, if implemented in existing approaches for phylogenetic reconstruction and divergence time estimation, would help to alleviate this source of error. Identifying suites of traits based on patterns in relative evolutionary rate may also provide a framework through which to separate clock-like characters from those displaying more erratic patters. Filtering character data based on conformity to clock-like patterns improves divergence time estimation in molecular data (Smith *et al.* 2018), and so would also be beneficial in morphological data.

Overfitting, covariance, and the true number of clusters

The approach that I introduce here does not explicitly model covariance among characters. However, when data are simulated as sets of strongly covarying modules, the method infers multiple character suites, despite having been evolved along the same set of branch lengths. This result should not be taken to suggest that the method is well-suited to identifying covarying modules of traits in addition to the mosaic suites that it seeks to discover. Instead, users should be aware that particularly high levels of covariance may lead to slight overfitting, by erroneously splitting up mosaic suites into multiple categories. Ideally, future extensions to this method will explicitly take covariance into account by incorporating older or recent approaches to identifying evolutionary covariance across multiple continuous traits (Felsenstein 1973). And despite the error encountered at high levels of covariance, the method performs well in the presence of this common pattern by correctly inferring the co-occurrences of traits in the same cluster, despite the fact that suites may be broken up in more extreme cases. As a result, this behavior is biased in a direction that is preferable when attempting to infer the number of clusters in a dataset, which is generally either unknown, or ambiguous. While it is simple to manually combine clusters that

have been overfit due to bias, underfitting causes users to miss potentially valuable structure in the data.

The ‘overfitting’ effect observed in the simulation experiment is predictable. The method that I introduce here is fundamentally a clustering method in design and implementation. Identification of the ’correct’ number of clusters can often become a philosophical problem when analyzing complex datasets, such as the covarying modules that comprise mosaic suites above (Hennig 2015). Although multiple modules may be evolved according the same pattern in branch lengths, stochastic variation between separate instantiations of random walks may create real differences in the resulting evolutionary signal between them. This stochastic effect, compounded by the large number of highly correlated traits within each covarying module, (Fig. 8c) generates strongly informative signal favoring precise differences from the generating set of branch lengths. Put simply for the user, larger and more strongly covarying modules of traits will increase the probability of encountering this effect. However, this behavior might be viewed as a benefit, by demonstrating the method’s ability to detect small differences from covarying data that result from stochastic variation in macroevolutionary processes. Researchers who are interested in both macroevolutionary patterns and morphological integration should employ my method in conjunction with an approach explicitly designed to recover integration patterns. This is because the simulations demonstrate that, although the method can identify strongly covarying modules, it will miss weak and moderate covariance among traits. This property is implied within the aims of the method, and so researchers should be thoughtful in parsing through these issues.

Morphological integration, mosaic evolution, and geometric morphometrics

The method that I have introduced here is designed to identify shared mosaic patterns in evolutionary disparity, or relative rates across lineages. This aim is related to, but distinct from, an

existing body of literature that aims to identify correlational patterns in morphological integration and modularity (Olson and Miller 1958; Cheverud 1982; Goswami 2006; Goswami *et al.* 2009). The mosaic suites identified by this study are a superset of the modules identified in typical morphological integration studies (Fig. 8). Strong covariance between traits implies the presence of a shared underlying pattern of macroevolutionary disparity (Fig. 8c) (Felice *et al.* 2018). However, as is demonstrated by the simulation experiment performed here, independent modules of covarying traits may share macroevolutionary patterns. This has also been demonstrated empirically by Felice and Goswami (2018) in avian cranial modules, by revealing both mosaic and shared patterns between integrated modules. Although the method that I present here is capable of 'accidentally' discovering covarying modules of traits by overfitting biased patterns that result from strong and rampant covariance, future extensions of the method should more rigorously explore this interaction. One avenue would be to incorporate extensions to the likelihood model that explicitly account for covariance among traits (Adams 2014; Alvarez-Carretero *et al.* 2018). Alternatively, it would be more immediately managable to apply the approach to data that have already been explored for integration. This would enable the algorithmic exploration of mosaic patterns among a large number of pre-identified covarying modules of traits, as has been done previously to infer phylogeny and mosaic evolution from geometric data (González-José *et al.* 2008; Felice and Goswami 2018). This step will be important when applying the method to geometric morphometric data, which are often collected to exhibit high levels of integration and covariance, but may be less necessary when exploring patterins in data that represent higher-level morphological structures, such as the Laurin ossification dataset explored here, or large sets of linear measurements and shape ratios. Emerging methods that algorithmically segment CT scans of whole organisms (Dunmore *et al.* 2018) and

quantify shape differences in specimen images often provide a small number of independent metrics for each recovered segment (Pomidor *et al.* 2016). The method introduced here would be useful in harnessing such data to identify mosaic patterns across the entire organisms.

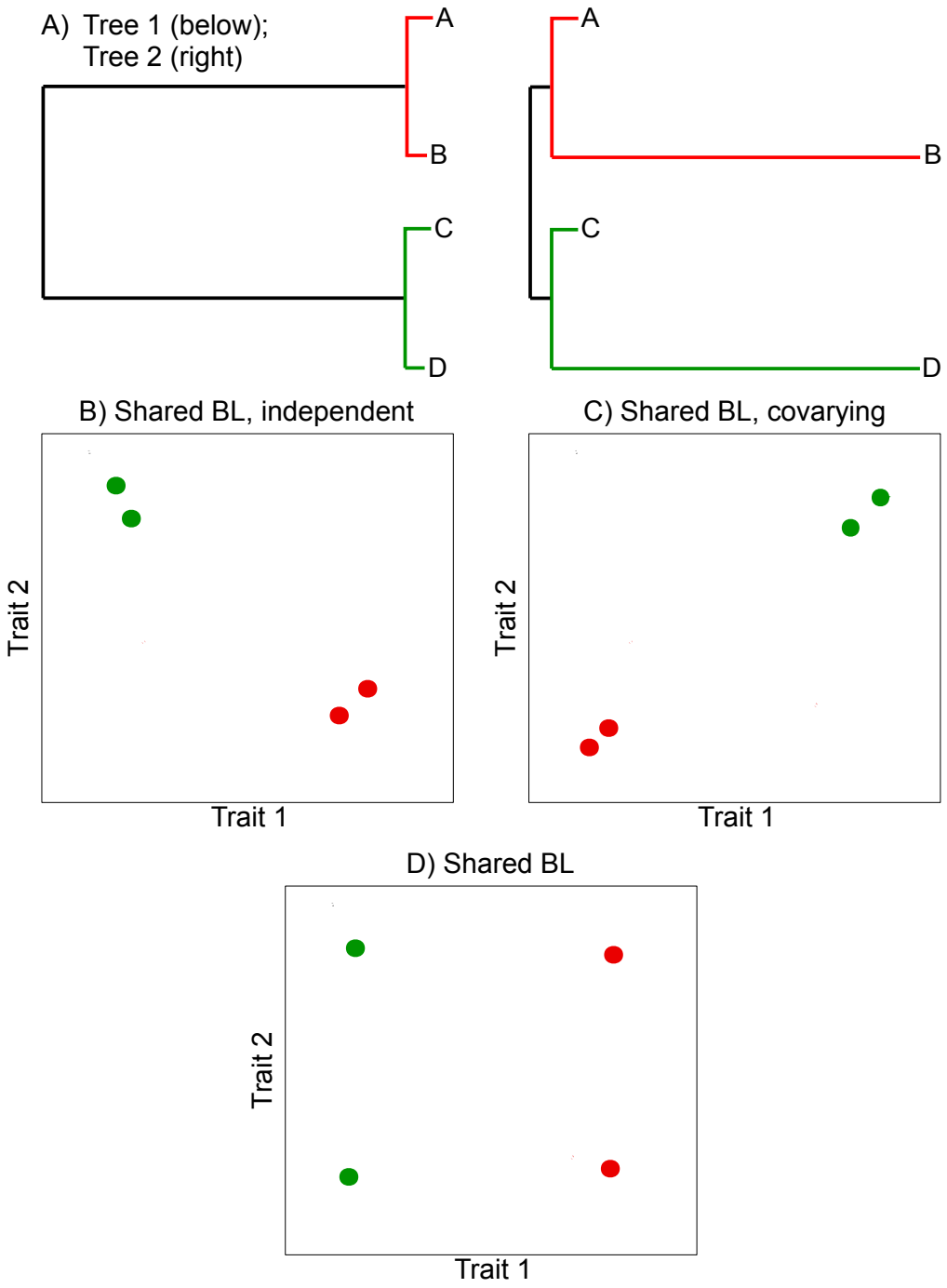

Figure 8. Expected patterns between traits with and without integration. A) Two possible patterns

in disparity represented by differences in disparity-scaled branch lengths. B) Values of two traits generated according to the pattern implied by tree 1 in panel A, but evolved independently. C) Two traits generated according to tree 1, but with strongly covarying evolutionary paths. D) Two traits generated independently and distinctly according to tree 1 and 2. Note that the covariance in panel C implies a shared pattern in disparity, while the examples in panels B and D do not.

Conclusions

While there have been fundamental and seminal works sketching out major historical trends in morphological disparification, most offer only a glimpse into the diversity of patterns that shape separate anatomical regions and operate at different phenotypic levels that extend from the gene to the environment. Although a foundational concept in post-synthesis evolutionary biology, researchers have been limited in their ability to explore mosaic evolution by both the difficulty in assembling comprehensive phenotypic datasets, and the lack of computational methods to handle comparative problems of significant breadth. Emerging approaches that can quantify morphology across entire organisms along with the increasing availability of large-scale transcriptomic and environmental datasets offer the unprecedented opportunity to develop a synthetic evolutionary picture that includes complex mosaic patterns operating at separate phenotypic levels and timescales. Methods like the one introduced here will be critical to these developments by facilitating the reconstruction of diverse mosaic patterns that have shaped evolutionary variation at different phenotypic levels across the tree of life, fulfilling the promises of the modern synthesis.

Acknowledgements

I would like to thank M Laurin, SD Smith, SA Smith, N Walker-Hale, Associate Editor DC Collar, and three anonymous reviewers for comments that greatly improved the manuscript.

Online-only Appendix

Index	Bone	Laurin 2014 suite	*greedo* suite
0	coronoid	1	1
1	vomer	1	1
2	palatine	1	1
3	dentary	0	0
4	premaxilary	0	0
5	prearticular	0	0
6	squamosal	0	0
7	parasphenoid	0	0
8	frontal	0	0
9	parietal	0	0
10	pterygoid	0	0
11	exoccipital	0	0
12	maxilla	0	0
13	quadrate	0	0
14	opisthotic	0	0
15	prefrontal	0	0
16	prootic	0	0
17	stapes	0	1
18	orbitosphenoid	0	0
19	nasal	0	0

Index	Bone	Laurin 2014 suite	*greedo* suite
20	septomaxilla	1	1

Table S1. Character suite assignments from original study (Laurin 2014) and the weighted graph in Fig. 3e. Suites are given arbitrary labels that are consistent for both studies. The two arrangements differ only in the assignment of the stapes developmental sequence.

Scale and rate

The approach described here seeks to identify suites of traits sharing similar patterns in relative evolutionary disparity across lineages. Continuous traits displaying greater empirical variances will display higher absolute rates of evolution when modeled under Brownian motion. As a result, I normalized the variances of the datasets used in the simulation experiments and empirical case study. Nevertheless, alteration of the scale of continuous traits may often change the interpretation of results, and so should be performed thoughtfully. In cases where phenotypes are quantified using a single, shared set of units, such as with geometric morphometric data, standardization of the variances across traits erases information characterizing absolute evolutionary rate. In such carefully constructed datasets, including the matrix of developmental sequences used in the empirical example above, researchers may wish to quantify differences in absolute evolutionary rate across characters. For instance, using the same dataset, Germain and Laurin (2009) demonstrated substantial variability in absolute rate across traits. Study of absolute and relative rates can each yield unique insights into evolutionary processes, and so the scaling of traits should be considered carefully. I did not explore inference of heterogeneity in absolute rates here because it is an easier clustering problem, and thus would have been an overly favorable test

of the method that I introduced. In addition, approaches that identify heterogeneity in absolute rate have been well explored in the molecular phylogenetics literature (Yang 1996), and successfully ported to continuous traits (Schraiber *et al.* 2013).

Information criteria and overfitting

In the analyses performed here, I exclusively used the AIC, in lieu of the corrected version, AICc, and the Bayesian Information Criterion (BIC). Previous authors have suggested that the AICc should be generally preferred to the uncorrected version (Burnham and Anderson 2002). My preference for the AIC was driven by several factors. The number of clusters is generally completely unknown prior to the analysis, and perhaps more importantly, there is generally no single 'true' clustering underlying the mosaic evolutionary patterns sought by the method. As a result, it might generally be preferable in the context of addressing comparative questions to identify a small number of spurious components in the final configuration than to ignore important biological variation that could be missed due to the steeper penalty imposed by the AICc. The analyses here support this justification. The simulated analyses show that, when AIC is used, overestimating the number of components is not a major problem (Fig. 2). In addition, the results of the empirical analysis suggest that more coherent patterns emerge when several well-supported configurations are averaged. If spurious partitions are encountered in some arrangements, averaging over the results should generally reveal reasonably strong connections between points occupying overfit components.

Although BIC has been used successfully to select the number of components in mixture models (Fraley and Raftery 1998, 1999), I preferred the behavior and basis of AIC for these analyses. BIC assumes that the true model is within the set of candidate models, and so can be sensitive to model-misspecification (Wagenmakers and Farrell 2004). This assumption is

incompatible with the goals of my method, which does not seek to identify a single 'true' configuration, but instead characterize the major axes of heterogeneity in disparity across lineages. This goal is more consistent with AIC, which simply seeks to identify the model that yields the lowest amount of information loss relative to the dataset. Despite my preference for AIC in the analyses presented here, AICc or BIC may be more appropriate in other situations. As such, researchers should be thoughtful in their choice of information criterion when performing the approach introduced here.

References

Adams, D. C. 2014. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution*, 68(9): 2675–2688.

Alvarez-Carretero, S., Goswami, A., Yang, Z., and dos Reis, M. 2018. Bayesian estimation of species divergence times using correlated quantitative characters. *bioRxiv*, page 441105.

Barton, R. A. and Harvey, P. H. 2000. Mosaic evolution of brain structure in mammals. *Nature*, 405(6790): 1055.

Beaulieu, J. M., Moles, A. T., Leitch, I. J., Bennett, M. D., Dickie, J. B., and Knight, C. A. 2007. Correlated evolution of genome size and seed mass. *New Phytologist*, 173(2): 422–437.

Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. C. 2012. Modeling stabilizing selection: expanding the ornstein–uhlenbeck model of adaptive evolution. *Evolution*, 66(8): 2369–2383.

Bokma, F., Godinot, M., Maridet, O., Ladevèze, S., Costeur, L., Solé, F., Gheerbrant, E., Peigné, S., Jacques, F., and Laurin, M. 2015. Testing for depéret's rule (body size increase) in mammals using combined extinct and extant data. *Systematic biology*, 65(1): 98–108.

Boyer, D. M., Puente, J., Gladman, J. T., Glynn, C., Mukherjee, S., Yapuncich, G. S., and Daubechies, I. 2015. A new fully automated approach for aligning and comparing shapes. *The Anatomical Record*, 298(1): 249–276.

Boyer, D. M., Gunnell, G. F., Kaufman, S., and McGeary, T. M. 2016. Morphosource: Archiving and sharing 3-d digital specimen data. *The Paleontological Society Papers*, 22: 157–181.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., *et al.* 2011. The evolution of gene expression levels in

mammalian organs. *Nature*, 478(7369): 343.

Burbrink, F. T. and Pyron, R. A. 2010. How does ecological opportunity influence rates of speciation, extinction, and morphological diversification in new world ratsnakes (tribe lampropeltini)? *Evolution: International Journal of Organic Evolution*, 64(4): 934–943.

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

Butler, M. A. and King, A. A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.*, 164(6): 683–695.

Caetano, D. S. and Harmon, L. J. 2018. Estimating correlated rates of trait evolution with uncertainty. *Systematic Biology*, syy067.

Chaix, R., Somel, M., Kreil, D. P., Khaitovich, P., and Lunter, G. 2008. Evolution of primate gene expression: drift and corrective sweeps? *Genetics*.

Cheverud, J. M. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. *Evolution*, 36(3): 499–516.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., and Freckleton, R. P. 2016. A cautionary note on the use of ornstein uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1): 64–77.

Cracraft, J. 1970. Mandible of archaeopteryx provides an example of mosaic evolution. *Nature*, 226(5252): 1268.

Cressler, C. E., Butler, M. A., and King, A. A. 2015. Detecting adaptive evolution in phylogenetic comparative analysis using the ornstein–uhlenbeck model. *Systematic biology*, 64(6): 953–968.

Dahl, D. B. 2006. Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, 201: 218.

De Beer, G. R. 1954. *Archaeopteryx lithographica: A study based upon the British Museum specimen*. Trustees of the British Museum.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38.

Dunmore, C. J., Wollny, G., and Skinner, M. M. 2018. Mia-clustering: a novel method for segmentation of paleontological material. *PeerJ*, 6: e4374.

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., and Harmon, L. J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, 65(12): 3578–3589.

Eastman, J. M., Wegmann, D., Leuenberger, C., and Harmon, L. J. 2013. Simpsonian 'evolution by jumps' in an adaptive radiation of anolis lizards. *arXiv preprint arXiv:1305.4216*.

Farris, J. S. 1971. The hypothesis of nonspecificity and taxonomic congruence. *Annual Review of Ecology and Systematics*, 2(1): 277–302.

Felice, R. N. and Goswami, A. 2018. Developmental origins of mosaic evolution in the avian cranium. *Proceedings of the National Academy of Sciences*, 115(3): 555–560.

Felice, R. N., Randau, M., and Goswami, A. 2018. A fly in a tube: Macroevolutionary expectations for integrated phenotypes. *Evolution*.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5): 471.

Felsenstein, J. 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, 35(6): 1229–1242.

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and*

Systematics, 19(1): 445–471.

Fraley, C. and Raftery, A. E. 1998. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8): 578–588.

Fraley, C. and Raftery, A. E. 1999. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2): 297–306.

Freckleton, R. P. 2012. Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, 3(5): 940–947.

Germain, D. and Laurin, M. 2009. Evolution of ossification sequences in salamanders and urodele origins assessed through event-pairing and new methods. *Evolution & development*, 11(2): 170–190.

Gingerich, P. D. 1974. Stratigraphic record of early eocene hyopsodus and the geometry of mammalian phylogeny. *Nature*, 248(5444): 107.

Gingerich, P. D. 1983. Rates of evolution: effects of time and temporal scaling. *Science*, 222(4620): 159–161.

Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. *Am. J. Sci.*, 293(A): 453–478.

Goloboff, P. A., Pittman, M., Pol, D., and Xu, X. 2018. Morphological datasets fit a common mechanism much more poorly than dna sequences and call into question the mkv model. *Systematic Biology*.

González-José, R., Escapa, I., Neves, W. A., Cúneo, R., and Pucciarelli, H. M. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in homo evolution. *Nature*, 453(7196): 775.

Goswami, A. 2006. Cranial modularity shifts during mammalian evolution. *The American*

Naturalist, 168(2): 270–280.

Goswami, A. and Finarelli, J. A. 2016. Emmli: A maximum likelihood approach to the analysis of modularity. *Evolution*, 70(7): 1622–1637.

Goswami, A., Weisbecker, V., and Sánchez-Villagra, M. 2009. Developmental modularity and the marsupial–placental dichotomy. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 312(3): 186–195.

Gould, S. J. 1970. Dollo on dollo's law: irreversibility and the status of evolutionary laws. *Journal of the History of Biology*, 3(2): 189–212.

Gould, S. J. 1977a. Bushes and ladders in human evolution. *Ever Since Darwin. Reflections in Natural History*, pages 56–62.

Gould, S. J. 1977b. *Ontogeny and Phylogeny*. Harvard University Press.

Hansen, T. F. 1997a. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, pages 1341–1351.

Hansen, T. F. 1997b. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5): 1341–1351.

Harmon, L. J., Schulte, J. A., Larson, A., and Losos, J. B. 2003. Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, 301(5635): 961–964.

Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., Near, T. J., *et al.* 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution: International Journal of Organic Evolution*, 64(8): 2385–2396.

Harvey, P. H. and Pagel, M. D. 1991. *The comparative method in evolutionary biology*, volume 239. Oxford University Press.

Hennig, C. 2015. What are the true clusters? *Pattern Recognition Letters*, 64: 53–62.

Ho, L. S. T. and Ané, C. 2014. Intrinsic inference difficulties for trait evolution with ornstein-uhlenbeck models. *Methods in Ecology and Evolution*, 5(11): 1133–1146.

Holloway, R. L. and Post, D. G. 1982. The relativity of relative brain measures and hominid mosaic evolution. In *Primate brain evolution*, pages 57–76. Springer.

Hopkins, M. J. and Lidgard, S. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proceedings of the National Academy of Sciences*, 109(50): 20520–20525.

Hubert, L. and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Landis, M. J. and Schraiber, J. G. 2017. Pulsed evolution shaped modern vertebrate body sizes. *Proceedings of the National Academy of Sciences*, 114(50): 13224–13229.

Landis, M. J., Schraiber, J. G., and Liang, M. 2012. Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. *Systematic biology*, 62(2): 193–204.

Laurin, M. 2014. Assessment of modularity in the urodele skull: an exploratory analysis using ossification sequence data. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322(8): 567–585.

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *The American Naturalist*, 136(6): 727–741.

Mayr, E. 1970. *Populations, species, and evolution: an abridgment of animal species and evolution*, volume 19. Harvard University Press.

McHenry, H. M. 1975. Fossils and the mosaic nature of human evolution. *Science*, 190(4213): 425–431.

Olson, E. C. and Miller, R. L. 1958. *Morphological Integration*. University of Chicago Press.

O'Meara, B. C., Ané, C., Sanderson, M. J., and Wainwright, P. C. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60(5): 922–933.

Parins-Fukuchi, C. 2018a. Bayesian placement of fossils on phylogenies using quantitative morphometric data. *Evolution*, 72(9): 1801–1814.

Parins-Fukuchi, C. 2018b. Use of continuous traits can improve morphological phylogenetics. *Systematic Biology*, 67(2): 328–339.

Pomidor, B. J., Makedonska, J., and Slice, D. E. 2016. A landmark-free method for three-dimensional shape analysis. *PloS one*, 11(3): e0150368.

Rabosky, D. L. and Adams, D. C. 2012. Rates of morphological evolution are correlated with species richness in salamanders. *Evolution: International Journal of Organic Evolution*, 66(6): 1807–1818.

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., and Alfaro, M. E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature communications*, 4: 1958.

Revell, L. J. 2012a. phytools: an r package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, 3(2): 217–223.

Revell, L. J. 2012b. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2): 217–223.

Revell, L. J. and Harmon, L. J. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, 10(3): 311–331.

Rohlfs, R. V., Harrigan, P., and Nielsen, R. 2013. Modeling gene expression evolution with an extended ornstein–uhlenbeck process accounting for within-species variation. *Molecular*

biology and evolution, 31(1): 201–211.

Rose, C. S. 2003. The developmental morphology of salamander skulls. *Amphibian biology*, 5: 1684–1781.

Scales, J. A., King, A. A., and Butler, M. A. 2009. Running for your life or running for your dinner: what drives fiber-type evolution in lizard locomotor muscles? *The American Naturalist*, 173(5): 543–553.

Schraiber, J. G., Mostovoy, Y., Hsu, T. Y., and Brem, R. B. 2013. Inferring evolutionary histories of pathway regulation from transcriptional profiling data. *PLoS Computational Biology*, 9(10): e1003255.

Simpson, G. G. 1944. *Tempo and mode in evolution*. Columbia University Press.

Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. *Methods in Ecology and Evolution*, 4(8): 734–744.

Smith, S. A., Brown, J. W., and Walker, J. F. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PloS one*, 13(5): e0197433.

Stanley, S. M. 1979. *Macroevolution, pattern and process*. Johns Hopkins University Press.

Stebbins, G. 1983. Mosaic evolution: An integrating principle for the modern synthesis. *Experientia*, 39(8): 823–834.

Stebbins, G. L. 1984. Mosaic evolution, mosaic selection and angiosperm phylogeny. *Botanical journal of the Linnean Society*, 88(1-2): 149–164.

Thomas, G. H. and Freckleton, R. P. 2012. Motmot: models of trait macroevolution on trees. *Methods in Ecology and Evolution*, 3(1): 145–151.

Wagenmakers, E.-J. and Farrell, S. 2004. Aic model selection using akaike weights. *Psychonomic*

Bulletin & Review, 11(1): 192–196.

Yang, J.-R., Maclean, C. J., Park, C., Zhao, H., and Zhang, J. 2017. Intra and interspecific variations of gene expression levels in yeast are largely neutral:(nei lecture, smbe 2016, gold coast). *Molecular biology and evolution*, 34(9): 2125–2139.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9): 367–372.

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G., McGlinn, D. J., O'Meara, B. C., Moles, A. T., Reich, P. B., *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature*, 506(7486): 89.