1 **Title:** Detecting mosaic patterns in phenotypic disparity

2 Caroline Parins-Fukuchi

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

4 **Keywords:** modularity, continuous traits, tempo and mode, macroevolution, phylogenetics.

5 Contents: Abstract, Introduction, Materials and Methods, Results/Discussion, Supplementary

6 Information, Bibliography, Figures 1-4, Table S1.

7 Word count: 5042

ABSTRACT

Understanding the patterns underlying phenotypic diversification across the tree of life has long been a fundamental aim in evolutionary and comparative biology. Classic and recent work has demonstrated both the wide variability in evolutionary rate throughout time and across lineages and the importance of characterizing these patterns in explaining the evolutionary proceses that generate biological diversity. A less extensive literature has shown that this variability extends to different aspects of phenotype, with separate suites, or modules, of traits within organisms showing different, "mosaic" patterns in rate and disparity across species. A merging of these two perspectives would identify modules of traits that display similar mosaic patterns in evolutionary tempo and mode. However, tools to do so have been limited. In this study, I introduce a new method for the identification of suites, or modules, of continuous traits that display shared patterns in evolutionary disparity across lineages. The approach defines a separate model of evolutionary disparification for each module defined by a phylogeny with branch lengths proportional to disparity. Module memberships and the number of modules are inferred using a greedy hill climbing approach that combines several different strategies to the unsupervised learning of classification and mixture models.

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 emergence of charismatic taxa, driving ecological differentiation between lineages. Early studies examined rates of change in a small number of traits, and identified a broad range of patterns of phenotypic change as lineages diverge (Simpson 1944; Stanley 1979).

A more recent body of work has focused on the development of statistical methods that identify patterns of phenotypic change using phylogenies (Harvey and Pagel 1991; Hansen 1997; Butler and King 2004; O'Meara *et al.* 2006; Beaulieu *et al.* 2012). These approaches have helped to reveal large-scale evolutionary trends across major lineages. These studies have frequently focused on increasing the phylogenetic and temporal scale compared to previous studies by focusing on only one or a small number of phenotypic characters either in isolation, or taken as a proxy for phenotype. For instance, although several studies have examined evolutionary rates using more comprehensive morphometric datasets (Rabosky and Adams 2012), adult body size is more commonly used in studies of animal taxa as a proxy for more detailed morphological measurements to characterize general patterns over deep timescales (Harmon *et al.* 2003, 2010; Burbrink and Pyron 2010; Rabosky *et al.* 2013; Bokma *et al.* 2015; Landis and Schraiber 2017). In plants, researchers often examine associations between a small number of key traits (Ree and Donoghue 1999; Beaulieu *et al.* 2007; Zanne *et al.* 2014).

The work described above has contributed greatly to both the empirical and conceptual understandings of patterns in the tempo and mode of phenotypic evolution across large and small timescales. Nevertheless, the typical focus on only a small number of characters has left open major questions surrounding the variation in pattern across body plans. Mosaic evolution is

expected to underlie most phenotypic change given the understanding that different traits are often exposed to selective pressures at different times. Researchers have argued for the ability of mosaic patterns to explain the emergence of structural variation in the brain across mammals (Barton and Harvey 2000), phenotypic and genomic diversity across angiosperms (Stebbins 1984), and the unique suite of morphological characters displayed by humans (McHenry 1975; Gould 1977; Holloway and Post 1982). However, despite their prevalence, mosaic evolutionary patterns have remained underexplored.

Biological modularity is a related concept that describes the tendency for suites of traits to contribute to a shared pattern or function, and has been explored at several phenotypic levels, including morphology (Cheverud 1982; Goswami 2006; Goswami *et al.* 2009), development (Wagner and Altenberg 1996), and gene expression (Brawand *et al.* 2011). Modularity can describe several different aspects of genotype and phenotype. Borrowing terminology from Wagner *et al.* (2007), the multivariate comparative approaches described above and other studies in morphology (Goswami 2006) are often focused on identifying 'variational' modules, or suites of traits that covary. Molecular studies often focus on 'functional' modules, or suites of features that contribute to some shared biological function. Developmental modules have also been explored, both on their own (Laurin 2014), and in association with variational morphological modules (Goswami *et al.* 2009).

Several researchers have contributed statistical approaches for geometric variables describing morphological shape, which are generally measured in multiple covarying dimensions (Adams 2014a). These approaches can be used to statistically evaluate known differences in evolutionary rate in predefined suites of continuous traits in a likelihood framework (Revell and Harmon 2008; Adams 2014b). This work has been a major benefit to researchers seeking to examine patterns in

variation of morphological shape. However, these methods can be impractical in several different situations. For instance, the boundaries dividing suites traits are often unknown, and so searching for suites of traits with shared signal in evolutionary rate or disparity may present unique insights. The focus of these methods on explicitly estimating rate also imposes the neeed to scale branch lengths to absolute time, which can create error and bias upon downstream analyses (Title and Rabosky 2016). A framework that characterizes the evolutionary structure and modularity underlying large phenotypic datasets using shared disparification patterns may be a useful complement to existing approaches by providing a point of reference that is not subject to the challenges involved in dating analyses or full multivariate estimation.

In this paper, I present a new method that identifies modules of continuous traits displaying shared patterns in disparity to reconstruct and characterize the mosaic trends that have shaped their evolution by forming suites of characters that are best explained by shared phylogenetic branch lengths along a fixed topology. After introducing the method, I evaluate its performance using simulated data. I also present an analysis of an empirical dataset of developmental traits complied by Rose (2003). This dataset has been analyzed previously for both modularity (Laurin 2014), and rate heterogeneity (Germain and Laurin 2009), and so is well suited to a re-examination using the method introduced here.

The approach is a novel contribution to the existing landscape of phenotypic modularity studies in both its utility and interpretation. Unlike previous approaches, which typically focus on variational modules, my method identifies ‘evolutionary’ modules defined by suites of characters displaying shared patterns in disparity across lineages. Importantly, the functionality of my approach differs from most previous work on modularity by offering a framework for machine-guided identification and delmitation of modules. Previous work has generally focused

on the statistical validation of modules specified by researchers *a priori*, with very little focus on ways of quantitatively delmiting modules among traits. Laurin's (2014) approach also delmits modules in phenotypic data, but my method is, to my knowledge, the only existing approach that identifies modules using a likelihood-based, phylogenetic framework.

Methods and Materials

Implementation

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.

Partitioning traits into modules

The method described here combines several unsupervised learning strategies to partition traits into separate modules, 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.

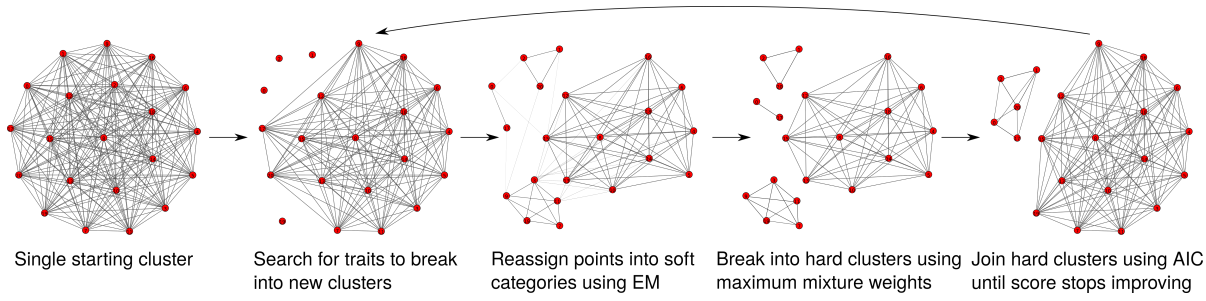


Figure 1. Search procedure to identify evolutionary modules.

Each component that defines the classification model contributes to the likelihood independently. The log-likelihoods of each of the traits belonging to component *j* are calculated under the component branch lengths, and added to yield the component log-likelihood. The

log-likelihood of the trait matrix, $LL_{classification}$, is calculated by summing the log-likelihoods of all *k* components

$$LL_{classification} = \sum_{j=1}^{k} LL_j \quad (1)$$

The details of the underlying phylogenetic Brownian model and the likelihood calculation follow Felsenstein (1981) and Parins-Fukuchi (2018) and are summarized in the supplement. Since the number of components is allowed to vary during the search, likelihoods are compared using the Akaike Information Criterion (AIC) to accommodate the difference in parameter count.

Search procedure

All traits start in a single shared partition. From here, traits that exhibit an improved likelihood in their own component compared to the single partition are broken into new components. To prioritize the separation of traits with especially strong divergent signal, a penalty is imposed that is proportional to the difference in size between the existing components. 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.

$$P(x_i|K_j) = \frac{L_{ij}}{\sum_{k=1}^{K} L_{ik}} \quad (2)$$

Expectation-maximization (EM) (Dempster *et al.* 1977) is performed to update the mixture weights and the branch length parameters. The branch lengths of 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 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. None of the steps impose a minimum constraint on the size of each cluster, and so clusters could range in size from including all of the traits to only one trait (although the latter case was not encountered in

Simulated data

To examine the strengths and shortcomings of the method, 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 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). All trees and traits were simulated using the phytools package in R (Revell 2012).

Using this procedure, matrices 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.

I used the adjusted Rand index (ARI) to evaluate the accuracy of the inferred partitionings (Hubert and Arabie 1985) against the true partitionings. ARI is a version of the Rand index (RI) (Rand 1971) that has been corrected for chance. 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 occupy the same cluster due to chance, with a value of 0 indicating a result indistinguishable from a random assignment of elements, and 1 indicating complete congruence, and also 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 obtained from Laurin (2014), initially assembled by Rose (2003). Laurin identified developmental modules using an 'evolutionary' Principal Components 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. Using my new approach, identified modules might be thought of as 'evolutionary developmental' modules, since the dataset is comprised of developmental sequences.

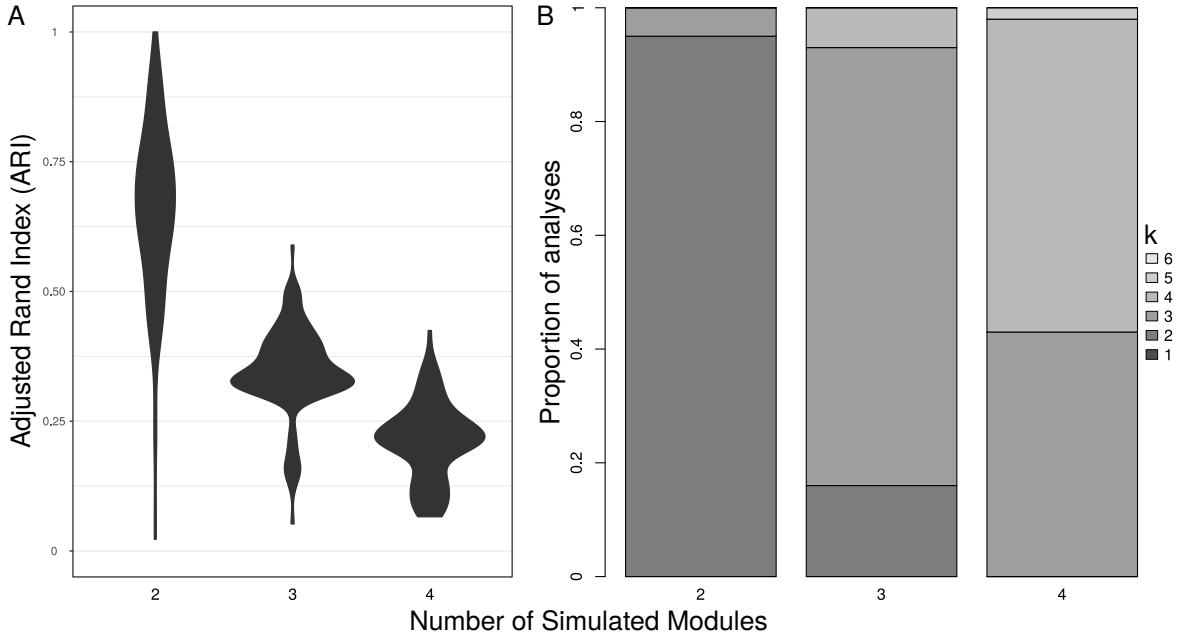
In his original analysis, Laurin (2014) fixed the developmental traits between the interval 0-1. However, this transformation yields data that display different empirical variance across taxa. This reflects the results of Germain and Laurin (2009), who demonstrated drastic variability (100x) in absolute rate across these characters. To prepare the data for the calculation of phylogenetic branch lengths, which assume traits of equal variance, I standardized the variance between the traits to 1. As a result, the analyses of disparity reflect relative, rather than absolute, ossification times. Importantly, differences in branch lengths across modules should thus be interpreted as reflecting variation in relative, rather than absolute disparity. The tree used for comparative analyses in the original study was used to calculate branch lengths (supplementary data).

Results and Discussion

Simulated data

The method is generally able to recover the structure of the simulated datasets. The number of inferred modules is typically close to the true number, and ARI values are typically well above random. The two-partition analyses are very accurate, with high ARI values, and nearly always

198 correctly identifying the correct number of clusters. The three- and four-partition analyses were
199 less accurate, but still yield results much higher than random, and typically recovering the correct
200 number of modules. ARI indices achieved for the three- and four- partition analyses appear
201 comparable to results from simulated data using more general clustering approaches, such as
202 Gibbs sampling under a Dirichlet process (Dahl 2006).



203

204 **Figure 2.** A) Adjusted Rand indices across reconstructions of simulated datasets. B) Number of
205 clusters resulting from analyses of simulated data. Barplots are stacked to represent the frequency
206 of each reconstructed *k*. All violin and barplots are separated by the number of modules in the
207 simulated datasets.

208 Despite the generally encouraging results from the simulated data, the trend toward
209 decreasing accuracy when components are added suggests either a limitation of the method in
210 adequately exhausting the search space of component assignments or a limitation in the power of
211 the simulated datasets in displaying sufficient signal across taxa. Since the primary steps of the
212 search alternate between greedy EM and hierarchical approaches, each iteration identifies a local

peak in the likelihood surface. As categories are added, it is possible that the added heterogeneity causes the surface becomes more peaky by adding permutations of locally optimal configurations. Because of this tendency, it might be useful to average across a set of best-supported configurations rather than rely on a single point estimate. It may also be helpful to initialize the analysis using results obtained from a less intensive approach, such as the evolutionary PCA developed by Laurin (2014). This may improve performance by requiring the search to traverse less of the likelihood surface, decreasing the chances of becoming stuck at a peak distant from the globally optimal configuration.

Empirical analysis

Four separate runs each yielded different partitionings into two modules. 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.

$$w_i = \frac{L_i^{rel}}{\sum_{k=1}^{K} L_k^{rel}} \tag{3}$$

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

$$L_i^{rel} = \exp(-0.5(AIC_i - AIC_{min})) \tag{4}$$

These weights were used to visualize the the strengths of the connections between 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 module. The maximum weight possible is 1.0, when traits *i* and *j* share a module in all of the configurations. The resulting graph suggests that traits 0, 1, 2, 17, and 20 all form a module, with the rest of the traits sharing a separate module. This result is very close to the pattern in modularity reconstructed by Laurin (2014) using an 'evolutionary PCA' approach, differing only in the assignment of the stapes (Table S1). The similarity of the empirical results to those of the original study demonstrate the capability of the new approach to identify meaningful modules in biological data.

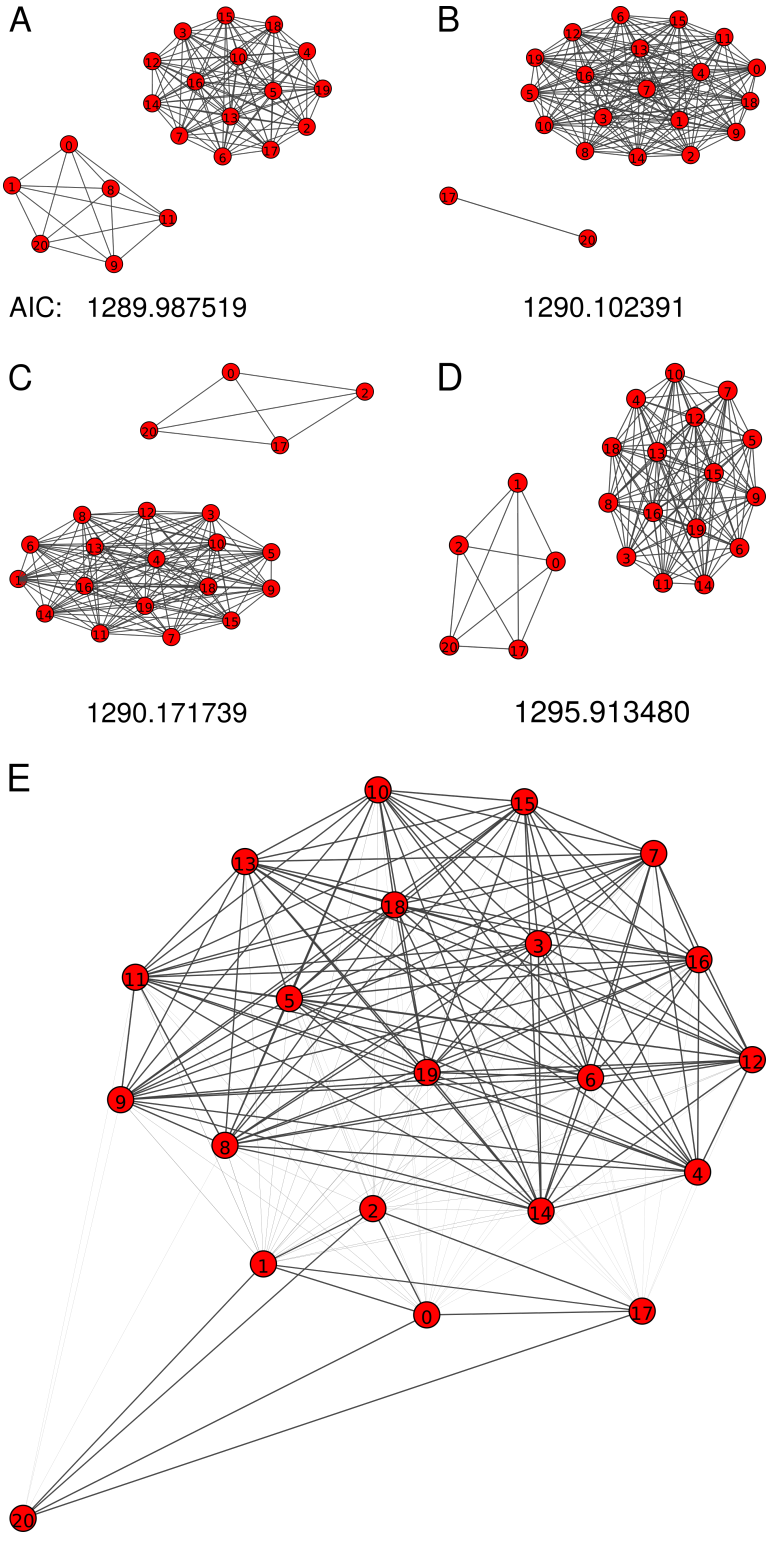


Figure 3. A-D) Four best configurations with AIC scores. E) Weighted graph calculated by summing the AIC weights associated with the each model to form edges and edge weights. All

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

242 In his original study, Laurin (2014) also performed an exploratory hierarchical clustering of
243 the developmental sequences and found substantial differences in structure as compared to that
244 revealed by his evolutionary PCA approach. The discordance between results achieved from each
245 method occurs because the the PCA considers covariance, while the hierarchical clustering only
246 reflects shared similarity in absolute value. Like the original study, the results here differ
247 substantially from the pattern resulting from the exploratory hierarchical clustering performed by
248 Laurin, instead aligning very closely to the PCA approach. This is reassuring for the performance
249 of my method, as Laurin considered the evolutionary PCA to yield the correct answer, and the
250 hierarchical clustering to demonstrate the inadequacy of similarity in delimiting meaningful
251 modules (Laurin, pers. comm.). Although they differ in the specific criteria used to identify
252 modules, the similarity in results between my and Laurin's method are not surprising. Laurin's
253 PCA method identifies structure from patterns in covariance that have been corrected for
254 phylogenetic non-independence, whereas my method identifies a minimally complex set of
255 models, defined by phylogenies with non-negative branch lengths. The trees describing each
256 module in my method may be thought of as representing disparity between taxa as patristic
257 distances. And so, although they differ in formulation and statistical paradigm, both approaches
258 are similar in their treatment of phenotypic variation. The method described here may be useful
259 as a complement to existing approaches of modularity by achieving similar results to other
260 evolutionary focused approaches, but benefiting from its placement in a likelihood and
261 information-theoretic framework, such as the ability to compare and average models.

262 The graph-based model averaging approach was shown in the empirical analysis to be
263 particularly useful in distilling the information across multiple well-supported module

configurations to smooth over imperfections in optimization. The importance of this step on such a small dataset, with only two clear modules suggests its potential to improve upon single point estimates using larger datasets with more clusters. Further tests will be needed to determine whether the approach can improve estimation in and alleviate the challenges encountered in the more heterogeneous simulated datasets. This step may also be important in characterizing the complex signal often encountered in large empirical datasets. Although both my method and previous approaches using PCA both yield a single 'hard' classification of traits into modules, biological data can often display a complicated network of interactions that can undermine such point estimates. In addition to smoothing over challenges in traversing peaky likelihood surfaces, the model averaging approach used above may also help to accommodate the complex variation in empirical data by weighting and combining evidence from a set of well-supported candidate models.

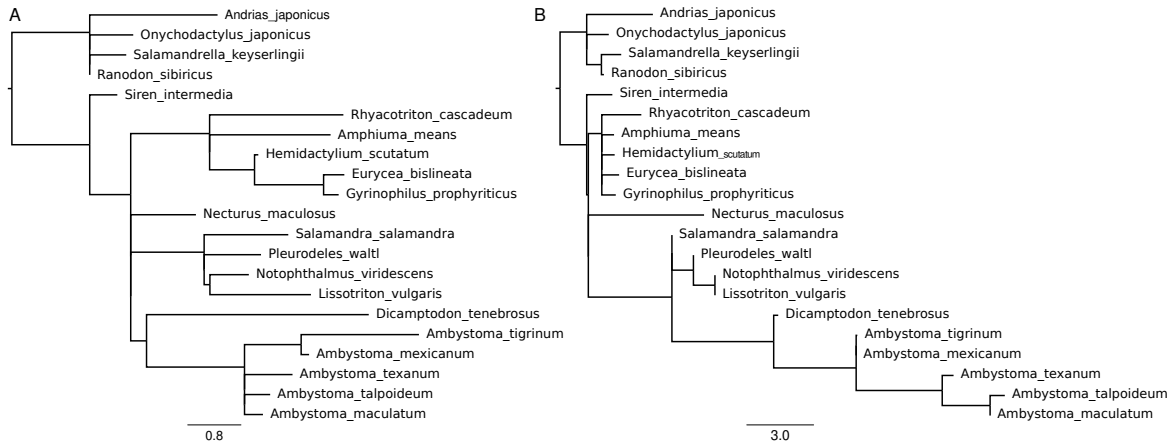


Figure 4. Branch lengths reconstructed from traits contained within: A) module 0 (Table S1); B) module 1 (Table S1).

The method that I introduce here identifies modules of continuous traits displaying similar patterns in evolutionary divergence. This may be useful in several different scenarios. As is stated

in the introduction, existing comparative studies tend to focus on only one or a small number of traits. Although this may in part be a result of the challenges in assembling large phenotypic datasets, another possible contributing factor may be the difficulty in performing tests and interpreting results across large numbers of traits. In these cases, the approach here might be useful as a preliminary, exploratory step by reducing large phenotypic datasets into a more tractable set of evolutionary modules. Traditional statistical comparative analyses could then be performed on the resulting modules rather than on single or arbitrarily joined groupings of traits. This approach may have the added benefit of increasing the amount of information from which to infer comparative models. As an alternative to the use of existing comparative methods, disparity branch lengths associated with each module may themselves yield sufficient information for evolutionary interpretation on their own. Reconstructed modules show very distinct patterns in lineage-wide disparity from one another (Fig. 4), and so may be useful in presenting a fine-scaled picture of the mosaic heterogeneity in pattern displayed across suites of characters.

The utility of my approach is distinct from most existing approaches to modularity. Most previous work exploring modularity has focused upon the statistical testing and validation of hypotheses of modularity specified *a priori* by the researcher by defining explicitly the constituent members of each module (Goswami 2006; Goswami *et al.* 2009). In contrast, the method that I introduce here detects and delimits modules automatically through a machine-driven search. This is more similar in purpose to the method developed by Laurin (2014), which also identifies modules, but differs in its explicit formulation in a model-based phylogenetic framework rather than the frequentist framework used in his approach, and the use of shared patterns in disparification as the basis of module delimitation rather than covariance.

In addition to morphological and developmental phenotypes, the method described here may

be useful in identifying evolutionary modules 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), defining modules 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.

Evolutionary interpretation of modules

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 modules of continuous traits that reflect differences in their patterns in disparity across taxa (Fig. 4). The method will be a valuable tool moving forward to aid in the identification of such modules by providing a reasonable basis upon which to perform more detailed comparative tests.

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. By segmenting the 'phenome' into subsets of traits displaying similar patterns in disparity, approaches to the identification of evolutionary modularity such as that introduced here have the potential to improve resolution into patterns of phenotypic diversification by separating conservatively evolving traits from those experiencing fluctuations in rate in certain lineages. This can benefit downstream comparative analyses, for example, by preventing conservatively evolving traits from swamping the signal expressed by those more erratic in their evolutionary pattern.

Characterization of patterns in disparity and evolutionary rate across lineages and their diversity across different aspects of phenotype have long been two fundamental, overarching goals in comparative biology. Although a substantive literature has developed a strong framework through which to understand general patterns in the tempo and mode of phenotypic evolution, researchers have been somewhat limited in the ability to reconstruct the diversity of pattern encountered across large datasets. This can probably be attributed to challenges in both the acquisition and analysis of such datasets. However, recent advances are improving the accessibility of large phenotypic datasets ranging from the morphological to the molecular scales. For instance, new developments in increasingly high-throughput methods that quantify morphology (Chang and Alfaro 2015; Boyer *et al.* 2015) and the increasing efforts of natural history museums in digitizing specimens as 3D images will yield increasingly large datasets. The approach introduced here represents an early step toward tackling the analysis of such datasets, by compressing the information contained within into a more analytically tractable set of modules

that display similar patterns in disparity.

Scale and rate

The approach described here seeks to identify suites of traits sharing similar patterns in evolutionary divergence across lineages. Continuous traits displaying greater empirical variances will display higher absolute rates of evolution when modeled under Brownian motion, resulting in differences in the reconstructed tree height across traits. This numerical reality may lead the method to cluster together traits with similar scales of variability, when it is often more desirable to identify traits with similar relative divergence patterns. As a result, it may often be beneficial to transform matrices to standardize the variances across all the traits. Since the units of measurement of continuous traits are typically arbitrary, this transformation is not likely to introduce biases.

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, 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. Although not explored here, my approach has the flexibility to examine both absolute and relative disparity depending on whether or not variances have been standardized between traits. Identification of shared signal in relative disparity is a more challenging clustering problem, since the erasure of variation in

empirical variance creates a flatter likelihood surface, and so the analysis of appropriately measured and scaled traits for differences in absolute disparity is possible using my approach, and likely an easier problem than the examples presented here.

Phylogenetic signal and evolutionary patterns and processes

Previous approaches to characterizing modularity often emphasize the need to identify phylogenetic signal in the data to justify the use of phylogenetic comparative approaches (Laurin 2014). Although the approach introduced here uses phylogenies, data need not explicitly display phylogenetic signal for the approach to be useful. This is because the method uses a species tree assumed to reflect true divergences as a scaffolding to fit observed patterns of evolutionary divergence. Since the branch lengths used in this approach reflect disparity, and so can accommodate patterns ranging from very weak phylogenetic covariance (star-like topology), to the strong covariance expected under neutral, clock-like phenotypic change by altering the branch lengths.

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 ambiguous. 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.

Moving forward

396 Although the results of the empirical and simulated analyses are generally encouraging, they
397 also reveal substantial hurdles in the use of the method moving forward. Accuracy decreases with
398 the number of categories, indicating the need to develop and evaluate more refined approaches to
399 both the search procedure and in model averaging. Nevertheless, the ability of the graph-based
400 averaging procedure shown in the empirical analysis to improve the quality of the final result and
401 sort out overlapping, but conflicting information across a set of well supported configurations
402 increases the prospects for the method to handle increasingly large phenotypic datasets. Finally,
403 although possessing caveats, the approach that I introduce here represents a step forward in the
404 analysis of phenotypic data toward a more thorough integration of studies characterizing tempo
405 and mode and those identifying modules and mosaic patterns in evolution, and toward the
406 analytical tractability of large phenotypic datasets.

407 **Acknowledgements**

408 I would like to thank Michel Laurin and Stacey D. Smith for comments that greatly improved
409 the manuscript.

410 **Supplemental Information**

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

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

Tree model

Each component of the classification model describing the trait matrix 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).

The phylogenetic comparative methods literature often estimates σ alone by assuming a fixed timescale given by branch lengths that have been scaled to absolute time using a clock model. However, here the absolute times are assumed to be unknown, and the rate and time parameters are allowed to covary. As a result, branch lengths are expressed in units of Brownian variance (or $\sigma^2 t$). This describes the amount of divergence between taxa, and so can be interpreted as estimates of phenotypic disparity, averaged across all traits.

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:

$$L_{node} = \frac{1}{2} * \frac{\log(2\pi) + \log(v_1 + v_2) + (x_1 - x_2)^2}{v_1 + v_2} \tag{5}$$

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.

$$x_{internal} = \frac{(x_1 * v_2) + (x_2 * v_1)}{v_1 + v_2} \tag{6}$$

$$v_{internal} = v_{internal} + \frac{(v_1 * v_2)}{(v_1 + v_2)} \tag{7}$$

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

$$L_{tree} = \sum_{node=1}^{n} L_{node} \tag{8}$$

The branch lengths associated with each component are estimated using an Expectation-Maximization procedure that leverages 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) is 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) and summarized by me in a previous article (Parins-Fukuchi 2018).

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

Adams, D. C. 2014b. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. *Syst. Biol.*, 63(2): 166–177.

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.

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.

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

Chang, J. and Alfaro, M. E. 2015. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods Ecol. Evol.*

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

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.

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.

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

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.

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

524 Germain, D. and Laurin, M. 2009. Evolution of ossification sequences in salamanders and
525 urodele origins assessed through event-pairing and new methods. *Evolution & development*,
526 11(2): 170–190.

527 Gingerich, P. D. 1983. Rates of evolution: effects of time and temporal scaling. *Science*,
528 222(4620): 159–161.

529 Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. *Am. J. Sci.*, 293(A):
530 453–478.

531 Goswami, A. 2006. Cranial modularity shifts during mammalian evolution. *The American*
532 *Naturalist*, 168(2): 270–280.

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

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

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

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

542 Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L.,
543 Bryan Jennings, W., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., Near, T. J., *et al.* 2010.
544 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.

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.

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.

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.

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

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.

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.

568 Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the*
569 *American Statistical association*, 66(336): 846–850.

570 Ree, R. H. and Donoghue, M. J. 1999. Inferring rates of change in flower symmetry in asterid
571 angiosperms. *Syst. Biol.*, 48(3): 633–641.

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

574 Revell, L. J. and Harmon, L. J. 2008. Testing quantitative genetic hypotheses about the
575 evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, 10(3):
576 311–331.

577 Rose, C. S. 2003. The developmental morphology of salamander skulls. *Amphibian biology*, 5:
578 1684–1781.

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

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

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

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

586 Title, P. O. and Rabosky, D. L. 2016. Do macrophylogenies yield stable macroevolutionary
587 inferences? an example from squamate reptiles. *Syst. Biol.*, 66(5): 843–856.

588 Wagenmakers, E.-J. and Farrell, S. 2004. Aic model selection using akaike weights. *Psychonomic*
589 *Bulletin & Review*, 11(1): 192–196.

590 Wagner, G. P. and Altenberg, L. 1996. Perspective: complex adaptations and the evolution of

591 evolvability. *Evolution*, 50(3): 967–976.

592 Wagner, G. P., Pavlicev, M., and Cheverud, J. M. 2007. The road to modularity. *Nature Reviews*

593 *Genetics*, 8(12): 921.

594 Yang, J.-R., Maclean, C. J., Park, C., Zhao, H., and Zhang, J. 2017. Intra and interspecific

595 variations of gene expression levels in yeast are largely neutral:(nei lecture, smbe 2016, gold

596 coast). *Molecular biology and evolution*, 34(9): 2125–2139.

597 Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G.,

598 McGlinn, D. J., O'Meara, B. C., Moles, A. T., Reich, P. B., *et al.* 2014. Three keys to the

599 radiation of angiosperms into freezing environments. *Nature*, 506(7486): 89.