

1           Phylofactorization: a graph-partitioning  
2           algorithm to identify phylogenetic scales of  
3                           ecological data

4  
5                           December 15, 2017

6     Alex D. Washburne<sup>1</sup>, Justin D. Silverman<sup>2,3</sup>, James T. Morton<sup>4,5</sup>, Daniel J.  
7     Becker<sup>1</sup>, Daniel Crowley<sup>1</sup>, Sayan Mukherjee<sup>3,6</sup>, Lawrence A. David<sup>3</sup>, Raina K.  
8                           Plowright<sup>1</sup>

9     **Affiliations:** <sup>1</sup>Department of Microbiology and Immunology, Montana State  
10    University, Bozeman MT, 59717, USA

11     <sup>2</sup>Program for Computational Biology and Bioinformatics, Duke University,  
12    Durham NC, 27708, USA

13     <sup>3</sup>Center for Genomic and Computational Biology, Duke University, Durham  
14    NC, 27708, USA

15     <sup>4</sup>Department of Computer Science, University of California San Diego, La  
16    Jolla CA, 92037, USA

17     <sup>5</sup>Department of Pediatrics, University of California San Diego, La Jolla CA,  
18    92037, USA

19     <sup>6</sup>Department of Statistical Science, Mathematics, and Computer Science,  
20    Duke University, Durham NC, 27708 USA

## 21 Abstract

22 The problem of pattern and scale is a central challenge in ecology [27]. The  
23 problem of scale is central to community ecology, where functional ecological  
24 groups are aggregated and treated as a unit underlying an ecological pattern,  
25 such as aggregation of “nitrogen fixing trees” into a total abundance of a trait  
26 underlying ecosystem physiology. With the emergence of massive community  
27 ecological datasets, from microbiomes to breeding bird surveys, there is a need to  
28 objectively identify the scales of organization pertaining to well-defined patterns  
29 in community ecological data.

30 The phylogeny is a scaffold for identifying key phylogenetic scales associ-  
31 ated with macroscopic patterns. Phylofactorization was developed to objec-  
32 tively identify phylogenetic scales underlying patterns in relative abundance  
33 data. However, many ecological data, such as presence-absences and counts,  
34 are not relative abundances, yet the logic of defining phylogenetic scales under-  
35 lying a pattern of interest is still applicable. Here, we generalize phylofactor-  
36 ization beyond relative abundances to a graph-partitioning algorithm for traits  
37 and community-ecological data from any exponential-family distribution.

38 Generalizing phylofactorization yields many tools for analyzing community  
39 ecological data. In the context of generalized phylofactorization, we identify  
40 three phylogenetic factors of mammalian body mass which arose during the K-  
41 Pg extinction event, consistent with other analyses of mammalian body mass  
42 evolution. We introduce a phylogenetic analysis of variance which refines our  
43 understanding of the major sources of variation in the human gut. We employ  
44 generalized additive modeling of microbes in central park soils to confirm that  
45 a large clade of Acidobacteria thrive in neutral soils. We demonstrate how to  
46 extend phylofactorization to generalized linear and additive modeling of any  
47 dataset of exponential family random variables. We finish with a discussion

48 of how phylofactorization produces a novel species concept, a hybrid of a phy-  
49 logenetic and ecological species concepts in which the phylogenetic scales and  
50 units of interest are defined objectively by defining the ecological pattern and  
51 partitioning the phylogeny into clades based on different contributions to the  
52 pattern. All of these tools can be implemented with a new R package available  
53 online.

## 54 **Keywords**

55 Phylofactorization, phylogeny, microbiome, ecological data, big data, graph par-  
56 titioning, dimensionality reduction

## 57 **Introduction**

58 The problem of pattern and scale is a central problem in ecology [27]. Ecosys-  
59 tem physiology, species abundance distributions, epidemics, ecosystem services  
60 of animal-associated microbial communities and more ecological patterns of in-  
61 terest are often the result of processes operating at multiple scales. Tradition-  
62 ally, the “scales” of interest are space, time, and levels of ecological organization  
63 ranging from individuals to populations to ecosystems. Prediction of spatial  
64 variation over millimeters, meters, or kilometers changes the processes driving  
65 patterns observed. Predicting climatic and weather patterns over days, years,  
66 or millennia requires different data, processes and models. Predicting the col-  
67 lective behavior of a school of fish requires interfacing individual behavior with  
68 interaction networks of those individuals [25] whereas predicting the ability of  
69 a forest to act as a carbon sink requires interfacing weather, nutrient cycles,

70 and competition between trees with different traits, such as nitrogen fixation  
71 [11]. Understanding emergent infectious diseases requires interfacing processes  
72 over scales ranging from animal population dynamics, reservoir epizootiology,  
73 and human epidemiology [37]. Ecological theory requires interfacing phenom-  
74 ena across scales believed to be important, and continually updating our beliefs  
75 about which scales are important to interface.

76 For a novel or unfamiliar pattern, such as a change in microbial community  
77 composition along environmental gradients, how can one objectively identify  
78 the appropriate scales of ecological organization? In macroscopic systems, a  
79 researcher will use intuition derived from natural history knowledge to determine  
80 scales of interest. Models of how the presumably important natural history traits  
81 affect the pattern will be constructed, and the goodness of fit to the pattern of  
82 interest will be used as a metric for the successful identification of ecological  
83 scales/traits. However, for some patterns, such as the ecosystem physiology of  
84 the human microbiome, there is limited natural history knowledge to draw on to  
85 assist the decision of the appropriate scales of interest. There is a need for rules,  
86 algorithms and laws for the simplification, aggregation, and scaling of ecological  
87 phenomena.

88 A central feature of biological systems is the existence of a hierarchical as-  
89 semblage of entities, from genes to species, whose relationships and evolutionary  
90 history can be estimated and organized into a hierarchical tree. The estimated  
91 phylogeny contains edges along which mutations occur and new traits arise.  
92 When the phylogeny correctly captures the evolution of discrete, functional eco-  
93 logical traits underlying a pattern of interest, the phylogeny is a natural scaffold  
94 for simplification, aggregation, and scaling in ecological systems. Patterns such  
95 as the change of bacterial abundances following antibiotic exposure, whose func-  
96 tional ecological traits of antibiotic resistance are laterally transferred, can still

97 be simplified by constructing a phylogeny of the laterally transferred genes, such  
98 as the beta-lactamases[18], as a natural scaffold for defining the entities with  
99 different responses to antibiotics.

100 The phylogeny contains a hierarchy of possible scales for aggregation. Gra-  
101 ham et al. [17] develop the term “phylogenetic scale” to refer to the depth of the  
102 tree over which we aggregate information from a clade. Functional ecological  
103 traits often arise at different depths of the tree. Many ecological phenomena may  
104 be driven by traits not properly summarized or aggregated by “hedge-row” trim-  
105 ming of the phylogeny along a constant depth, but by identification of multiple  
106 phylogenetic scales, or grains, underlying an ecological pattern of interest. For  
107 example, the patterns of vertebrate abundances on land and water are simpli-  
108 fied by nested clades: tetrapods, cetaceans, Pinnipeds, etc. Identifying multiple  
109 phylogenetic scales associated with or driving an ecological pattern of interest  
110 requires general methods for partitioning the phylogeny into the grains with  
111 significantly different associations or contributions to the ecological pattern.

112 Phylofactorization [51] was developed to identify the phylogenetic scales in  
113 relative abundance (i.e. compositional) data by iteratively partitioning the phy-  
114 logeny and constructing variables corresponding to edges in the phylogeny and  
115 selecting variables which maximize an objective function. Phylofactorization  
116 of compositional data exploits a common transform from compositional data  
117 analysis [1], referred to as the isometric log-ratio transform [10, 9], which pro-  
118 vides a natural way to turn the phylogeny into a set of variables capable of  
119 identifying differences between clades. A coordinate in an isometric log-ratio  
120 transform aggregates relative abundances within clades by a geometric mean  
121 and contrasts clades through log-ratios of the clades’ geometric mean relative  
122 abundances. The isometric log-ratio transform is used to identify phylogenetic  
123 scales capturing large blocks of variation in relative-abundance data, with vari-

124 ables that correspond to edges along which hypothesized functional ecological  
125 traits arose.

126       However, many ecological data are more appropriately viewed as counts, not  
127 compositions. In this paper, we generalize phylofactorization to broader classes  
128 of data types by generalizing the logic of phylofactorization and the general  
129 problem of scale in ecology to a set of three operations: aggregation, contrast,  
130 and an objective function defined by the pattern of interest. With these opera-  
131 tions, phylofactorization can be defined as a graph-partitioning algorithm which  
132 avoids the nested dependence of hierarchies of clades and controls for previously  
133 identified phylogenetic scales. Generalizing the operations of aggregation and  
134 contrast in a phylogenetic graph-partitioning algorithm provides an explicit,  
135 theoretical framework defining place-holders for specific operations used a re-  
136 search endeavor. Furthermore, as points on the surface of a sphere are easily  
137 represented in spherical coordinates, ecological data at the tips of a phylogeny  
138 are easily represented with a change of variables made possible by aggregation  
139 and contrast, a set of variables we call the “contrast basis”. Phylofactorization  
140 is a versatile tool for identifying the phylogenetic scales underlying ecological  
141 patterns of interest across a range of patterns and data types.

142       After defining phylofactorization as a graph-partitioning algorithm, we il-  
143 lustrate the generality of the algorithm through several examples. First, we  
144 show that two-sample tests, such as t-tests and Fisher’s exact test, are natural  
145 operations for phylofactorization - they first aggregate data from two groups  
146 through means, contrast the aggregates via a difference of means, and have nat-  
147 ural objective functions defined by their test-statistics. We illustrate the use of  
148 two-sample tests by performing phylofactorization of a dataset of mammalian  
149 body mass.

150       Then, we show how the phylogeny serves as a scaffold for changing variables

151 in biological data through a contrast basis which can be used to identify the  
152 phylogenetic scales providing low-rank, phylogenetically-interpretable represen-  
153 tations of a dataset. Defining the contrast basis allows us to introduce a phyloge-  
154 netic analog of principal components analysis - phylogenetic components anal-  
155 ysis - which identifies edges and dominant, phylogenetic scales differentiating  
156 species and explaining variance in a dataset. We perform phylogenetic compo-  
157 nents analysis on the American Gut microbiome dataset ([www.americangut.org](http://www.americangut.org))  
158 and reveal that some of the dominant clades explaining variation in the Ameri-  
159 can Gut correspond to clades within Bacteroides and Firmicutes, thereby pro-  
160 viding finer, phylogenetic resolution of a known, major axis of variation in  
161 human gut microbiomes found to be associated with obesity [47], age [31]  
162 and more. Another phylogenetic factor of variance in the American Gut is  
163 a clade of Gammaproteobacteria strongly associated with IBD, corroborating  
164 a recent study's use of phylofactorization to diagnose patients with IBD [49].  
165 The contrast basis can also be used with regression if the data assumed to  
166 be approximately normal, log-normal, logistic-normal or otherwise related to  
167 the normal distribution through a monotonic transformation. We illustrate  
168 regression-phylofactorization through a generalized additive model analysis of  
169 how microbial abundances change across a range of pH, Nitrogen, and Carbon  
170 concentrations in soils. The resulting contrast basis and its fitted values from  
171 generalized additive modeling yield a low-rank representation of biological big-  
172 data and translates to clear biological hypotheses aiming to identify the traits  
173 driving observed non-linear patterns of abundance across pH [39].

174 Datasets comprised of non-Gaussian, exponential family random variables  
175 can still be analyzed through regression-phylofactorization by using the gener-  
176 alized algorithm and implementing factor-contrasts in a multivariate generalized  
177 linear model as the contrast operation. We present, and demonstrate the power

178 of, three methods for generalized regression-phylofactorization in exponential  
179 family data. The first method is to use the contrast basis to obtain a low-rank  
180 approximation of coefficient matrices in multivariate generalized linear models.  
181 The second is a reduced rank regression model in which a phylogenetic factor,  
182 an explanatory variable indicating which side of an edge a species is found, is  
183 incorporated into regression and used to define objective functions based on  
184 the deviance or the magnitude of the coefficients for the factor-contrast. The  
185 third method aggregates exponential family data within clades to marginally  
186 stable distributions within the exponential family, and then performs a two-  
187 variable multivariate regression with a factor contrast as used in the second  
188 method. We simulate the asymptotic power of the last two methods, demon-  
189 strating that marginally-stable aggregation and factor-contrasts are a viable  
190 method for phylofactorization through generalized linear and additive models.  
191 We finish with a discussion of the challenges, and opportunities, for future devel-  
192 opment of phylofactorization, and provide an R package - phylofactor - available  
193 at <https://github.com/reptalex/phylofactor>.

## 194 Phylofactorization

195 Which vertebrates live on land, and which vertebrates live in the sea (Figure  
196 1a)? Most children have enough natural history knowledge to say “fish live in the  
197 sea”, thus correctly identifying one of the most important phylogenetic factors of  
198 land/sea associations in vertebrates. The statement “fish live in the sea” can be  
199 mathematically captured by noting that one edge in the vertebrate phylogeny  
200 separates “fish” from “non-fish” (Figure 1b). Partitioning the phylogeny along  
201 the edge basal to tetrapods can separate our species fairly well by land/sea asso-  
202 ciations. An algorithm for identifying that edge by land/sea associations alone,



203 without requiring detailed knowledge of macroscopic life and morphological and  
204 physiological traits, can correctly identify an edge along which functional ecolog-  
205 ical traits and life-history traits arose. Controlling for the previously identified  
206 edge, one might be able to identify the edges basal to Cetaceans and Pinnipeds,  
207 tetrapods which live in the sea (Figure 1b). Three edges can capture most of  
208 the variation in land/sea associations across potentially thousands of vertebrate  
209 species.

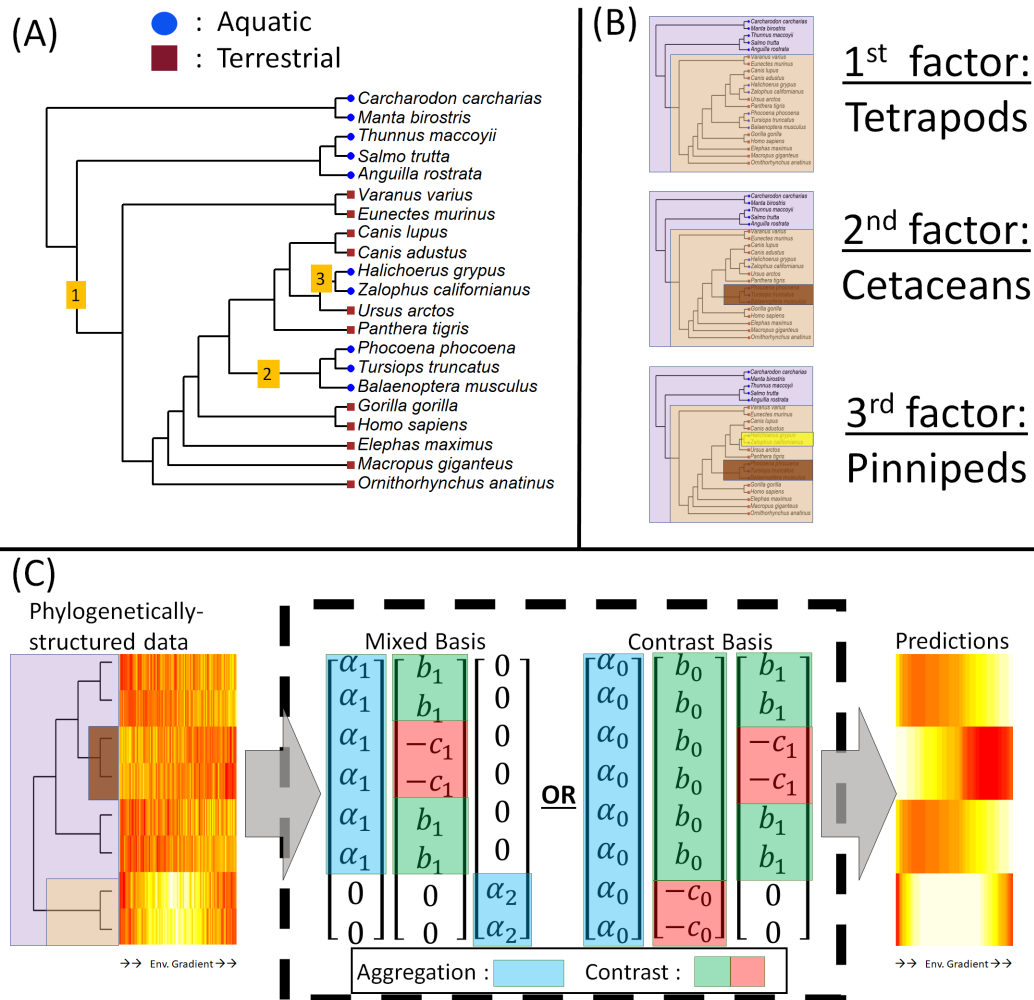


Figure 1: Phylofactorization aims to generalize the logic of how to simplify phylogenetically-structured datasets. (A) A dataset of vertebrate land/water associations can be simplified by partitioning the tree into the edges along which major traits arose. (B) The first phylogenetic factor of vertebrate land/water associations is the edge along which tetrapods arose - an edge along which lungs and limbs evolved that allowed colonization of land. Downstream factors can refine the original partitioning, and include the Cetaceans and Pinnipeds, among other edges along which adaptation to aquatic life arose among tetrapods. (C) Phylogenetic factorization generalizes this same logic for phylogenetically-structured data in which traits might not be known or their evolution easily modeled, including traits like a non-linear relationship between abundance and an environmental gradient. Phylogenetically-structured data can be partitioned through operations of aggregation and contrast. Pure aggregations (blue) are total abundances of a clade, whereas contrast (green/red) are statements of differences between two clades. Low-rank, phylogenetically-interpretable predictions of our data can be obtained through a mixed basis of a series of aggregations and contrasts, or a “contrast basis” in which there is a global aggregate partitioned in subsequent contrasts.

210 Ancestral state reconstruction of habitat association provides a well-known  
211 means of making such inferences. However, sometimes the desired “traits” and  
212 ecological patterns of interest are more complicated and their ancestral state re-  
213 construction dubious. For instance, how can we identify the phylogenetic scales  
214 of changes in microbial community composition along a pH gradient, allow-  
215 ing possible non-linear associations that could be detected through generalized  
216 additive modeling? Answering such a question through ancestral state recon-  
217 struction requires conceiving and analyzing an evolutionary model of how the  
218 generalized additive models of pH association evolve along a tree. Phylofactor-  
219 ization aims to generalize the phylogenetic logic used for land/sea associations  
220 in order to identify phylogenetic scales for more complicated functional traits  
221 and ecological patterns, for which an evolutionary model would be dubious.  
222 Phylogenetic factorization generalizes the logic of land/sea associations through  
223 a graph partitioning algorithm iteratively identifying edges in the phylogeny  
224 along which meaningful differences arise (Figure 1c).

## 225 **General Algorithm**

226 Phylofactorization requires a set of phylogenies, rooted or unrooted graphs with  
227 no cycles, containing and connecting the units of interest in our data (the “units”  
228 can be species, genes or operons other evolving units of interest). Phylofactor-  
229 ization can be done using disjoint sub-graphs, such as viral phylogenies for  
230 which there are not clear common ancestors, and the sub-phylogenies can either  
231 be kept separate or joined at a polytomous root. The phylogeny may have an  
232 arbitrary number and degree of polytomies, and can even be a star graph.

233 Let  $[x]_{i,j}$  be the data for species  $i = 1, \dots, m$  in sample  $j = 1, \dots, n$ . Let  
234  $\mathbf{x}_{R,j}$  be the vector of a subset of species,  $R$ , in sample  $j$ . Let  $\mathbf{Z}$  be the  $n \times p$   
235 matrix containing  $p$  additional meta-data variables for each sample. Let  $\mathcal{T}$  be

236 the phylogenetic tree and let edge  $e$  partition the phylogeny into disjoint groups

237  $R$  and  $S$ . Phylofactorization requires:

238 • An aggregation function,  $A(\mathbf{x}_{R,j}, \mathcal{T})$  which aggregates any subset,  $R$ , of  
239 species

240 • A contrast function,  $C(A(\mathbf{x}_{R,j}, \mathcal{T}), A(\mathbf{x}_{S,j}, \mathcal{T}), \mathcal{T}, e)$  which contrasts the  
241 aggregates of two disjoint subsets of species,  $R$  and  $S$ , possibly using  
242 information from the tree  $\mathcal{T}$  and edge,  $e$ .

243 • An objective function,  $\omega(C, \mathbf{Z})$ .

244 With these operations, phylofactorization is defined iteratively as a special case  
245 of a graph partitioning algorithm (Figure 2). The steps of phylofactorization  
246 are:

247 1. For each edge,  $e$ , separating disjoint groups of species  $R_e$  and  $S_e$  within the  
248 sub-tree containing  $e$ , compute  $C_e = C(A(\mathbf{x}_{R_e,j}, \mathcal{T}), A(\mathbf{x}_{S_e,j}, \mathcal{T}), \mathcal{T}, e)$

249 2. compute edge objective  $\omega_e = \omega(C_e, \mathbf{Z})$  for each edge,  $e$

250 3. Select winning edge  $e^* = \operatorname{argmax}_e(\omega_e)$

251 4. Partition the sub-tree containing  $e^*$  along  $e^*$ , forming two disjoint sub-  
252 trees.

253 5. Repeat 1-5 until a stopping criterion is met.

254 Unlike more general graph-partitioning algorithms, phylofactorization does not  
255 impose a balance constraint - it does not require that the partitions have a simi-  
256 lar size or weight. Furthermore, phylofactorization, by working with phylogenies  
257 or graphs without cycles is centered around aggregation and contrast as princi-  
258 ple operations for defining scales and units of organization. Phylofactorization  
259 is limited to contrasts of non-overlapping groups. The incorporation of the tree,

260  $\mathcal{T}$ , in the contrast function encompasses a class of ancestral state reconstruction  
261 reconstruction methods. Ancestral state reconstruction with non-overlapping  
262 contrasts can be done with time-reversible models of evolution; in this case,  
263 phylofactorization contrasts the root ancestral states obtained in which the two  
264 nodes adjacent an edge are considered roots of the subtrees separated by an  
265 edge.

266 The edges,  $e^*$  and their contrasts,  $C_e$ , are interchangeably referred to as  
267 the “phylogenetic factors” due to their correspondence to hypothesized latent  
268 variables (traits) and their ability to construct basis elements that allow ma-  
269 trix factorization [51]. It’s possible to define objective functions through pure  
270 aggregation, but we limit our focus to contrast-based phylofactorizations which  
271 identify edges along which meaningful differences arose for reasons discussed  
272 later in the section on the “contrast basis”.

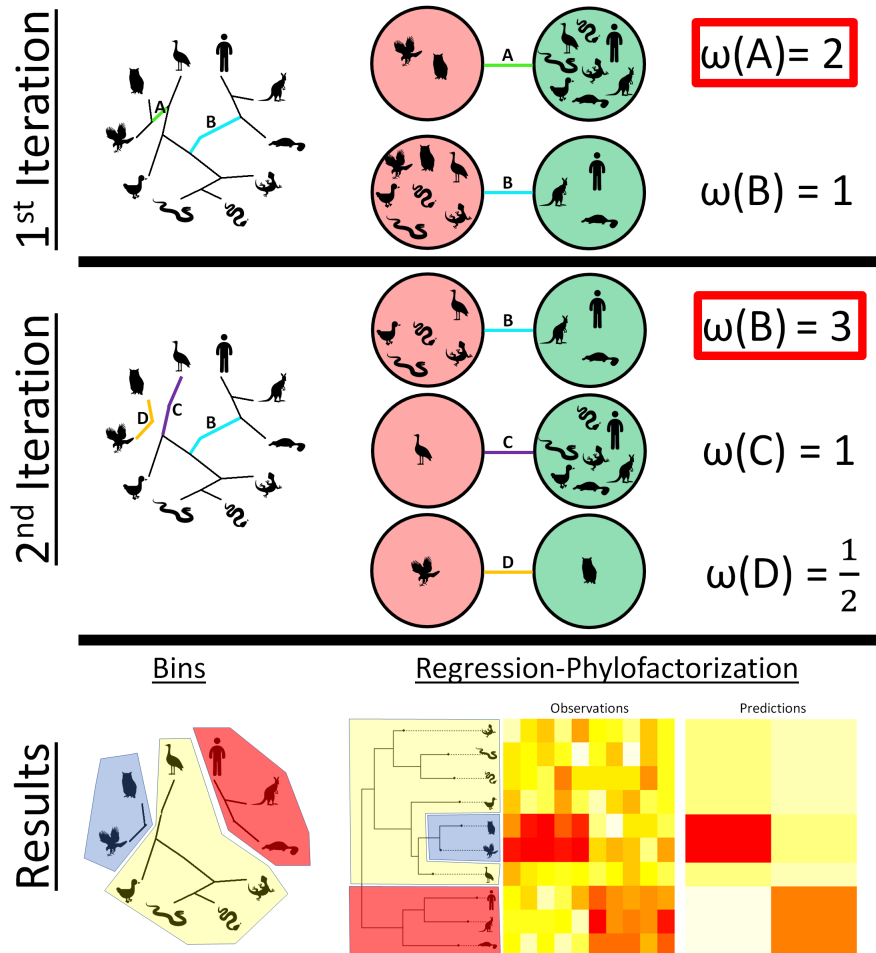


Figure 2: Phylofactorization is a graph partitioning algorithm. Defining an objective function,  $\omega$ , of a contrast of species separated by an edge allows one to iteratively partition the phylogeny along edges maximizing the objective function (1st iteration). After partitioning the phylogeny, the objective functions are re-computed to contrast species in the same sub-tree separated by an edge. Edge B in the first iteration contrasted mammals from non-mammals, but in the second iteration it contrasts mammals from non-mammals, excluding raptors (partitioned in the first iteration). The result of  $k$  iterations of phylofactorization is a set of  $k + 1$  bins of species with similar within-group behavior. A particularly useful case is “regression-phylofactorization”. Regression-phylofactorization is implemented by defining contrasts through the contrast basis (Figure 1c) and defining an objective function through regression on the component scores of each candidate contrast basis element. Regression-phylofactorization is a flexible way to search for clades with similar patterns of association with environmental meta-data while also obtaining low-rank, phylogenetically-interpretable representations of a data matrix.

273 The result of phylofactorization after  $t$  iterations is a set of  $t$  inferences on  
274 edges or links of edges. Links of edges occur following a previous partition,  
275 when two adjoining edges separate the same two groups in the resultant sub-  
276 tree. Partitioning the phylogeny along  $t$  edges results in  $t + 1$  bins of species,  
277 referred to as “binned phylogenetic units”. In general, the problem of maximizing  
278 some global objective function,  $\omega(e_1^*, \dots, e_t^*)$ , for a set of  $t$  edges,  $\{e_1^*, \dots, e_t^*\}$ , is  
279 NP hard [6]. However, stochastic searches of the space of possible partitions,  
280 via a stochastic computation of  $\omega_e$  in step 2 or a weighted draw of  $e^*$  in step 3,  
281 may better maximize a global objective function for general graph-partitioning  
282 algorithms such as phylofactorization [32, 20, 23].

283 Generalizing aggregation, contrast, and objective functions allows researchers  
284 several junctures to define and interpret meaningful quantities and outcomes  
285 from data analysis. Explicit decisions about aggregation formalize how a re-  
286 searcher would summarize data from an arbitrary set of species. Explicit de-  
287 cisions about contrast formalize how a researcher differentiates two arbitrary,  
288 disjoint groups of species - these common operations form an organizational  
289 framework for ecologists studying phylogenetic scales. Aggregation can be done  
290 through many operations, including but not limited to addition, multiplication,  
291 generalized means, and maximum likelihood estimation of ancestral states un-  
292 der models of trait diffusion away from the focal node. Likewise, examples of  
293 contrasts are differences, ratios, various two-sample tests, and more complicated  
294 metrics of dissimilarity such as the deviance of a factor contrast in a generalized  
295 additive model. Researchers must decide for themselves how best to aggregate  
296 information in groups of species, contrast two groups, and decide which group  
297 maximizes the objective for a research goal pertaining to a particular ecolog-  
298 ical pattern. Doing so allows objective, a priori definitions of what makes an  
299 informative phylogenetic scale, and the operations chosen are integrated into a

300 broader theoretical framework of phylofactorization.

301 Below, we run through several examples aimed to develop the generality  
302 and illustrate the results from phylofactorization. These examples were run  
303 using the R package “phylofactor”, using relevant functions for analyzing and  
304 visualizing phylogenies from the R packages ape [36], phangorn [43], phytools  
305 [40], and ggtree [52]. Scripts and datasets for every analysis are available in the  
306 supplemental materials.

### 307 **Example 1: two-sample tests and mammalian body-mass** 308 **phylofactorization**

309 If the data are a single vector of observations,  $\mathbf{x}$ , similar to the land/sea as-  
310 sociations of vertebrates, phylofactorization can be implemented through stan-  
311 dardized tests for differences of means or rate parameters in the two sets of  
312 species,  $R$  and  $S$ . Two-sample tests may bias away from the tips and towards  
313 the interior edges of the phylogeny due to increased power of two-sample tests  
314 of more equally-sized samples.

315 For example, a dataset of mammalian body mass from PanTHERIA [24]  
316 and the open tree of life using the R package “rotl” [33]. A single vector of data  
317 assumed to be log-normal can be factored based on a two-sample t-test (Figure  
318 3a). In this case,  $A(\mathbf{x}_R) = \overline{\log(\mathbf{x}_R)}$  is the arithmetic mean of the log-body-mass;  
319 we use the contrast operation

$$C = \frac{|A(\mathbf{x}_R) - A(\mathbf{x}_S)|}{\sqrt{\frac{1}{r} + \frac{1}{s}}} \quad (1)$$

320 and the objective function  $\omega_e = C_e$  - this is the two-sample t-test with the  
321 assumption of constant variance. Maximization of the objective function yields  
322 edges with the most significant difference in body mass of organisms on different



323 sizes of the tree.

324 The first five phylogenetic factors of mammalian body mass in these data are  
325 Euungulata, Ferae, Laurasiatheria (excluding Euungulata and Ferae), a clade  
326 of rodent sub-orders Myodonta, Anomaluromorpha, and Castorimorpha, and  
327 the simian parvorder Catarrhini. Five factors produce six binned phylogenetic  
328 units of species with different average body mass (Figure 3a). The most sig-  
329 nificant phylogenetic partition of mammalian body mass occurs along the edge  
330 basal to Euungulata, containing 296 species with significantly larger body mass  
331 than other mammals. The second partition corresponds to Ferae, containing 242  
332 species which have body masses larger than other mammals, excluding Euungu-  
333 lata. The third partition corresponds to 864 remaining species in Laurasiathe-  
334 ria, excluding Euungulata and Ferae, which contains Chiroptera, Erinaceomor-  
335 pha, and Soricomorpha. These mammals have lower body mass than non-  
336 Laurasiatherian mammals. The fourth partition identifies three rodent sub-  
337 orders comprising 926 species with lower body mass than non-Laurasiatherian  
338 mammals. Finally, 106 species comprising the Simian parvorder Catarrhini  
339 are factored as having higher body mass than the remaining mammals. These  
340 factors are fairly robust: 3000 replicates of stochastic Metropolis-Hasting phylo-  
341 factorization, drawing edges in proportion to  $C^\lambda$  with  $\lambda = 6$  (producing a 1/4  
342 probability of drawing the most dominant edge) could not improve upon these  
343 5 factors.

344 The first two phylogenetic factors of mammalian body size partition the  
345 mammalian tree at deep edges with ancestors near the K-Pg extinction event,  
346 corroborating evidence of ecological release [2, 3] and the exponential growth  
347 of maximum body sizes following the K-Pg extinction event [46] for these two  
348 dominant clades. The crown group of modern Euungulata are thought to have  
349 originated in the late Cretaceous [53] and its representatives may have expanded

350 into previously dinosaur-occupied niches during the rapid evolution of body  
351 size in mammals immediately after the K-Pg extinction event at the Creta-  
352 ceous/Paleogene boundary [45]. Cope’s rule posits that lineages tend to in-  
353 crease in body size over time, and a recent study [4] confirms Cope’s rule and  
354 found that mammals have, along all branch lengths in their phylogeny, tended  
355 to increase in size. The phylogenetic factors of mammalian body size discovered  
356 here illustrate an important feature of phylofactorization: correlated evolution  
357 within a clade, such as a consistently high body-size increase among lineages in  
358 a clade, can cause the edge basal to a clade to be an important partition for  
359 capturing variance in a trait. A more robust phylofactorization may be done  
360 through iterative ancestral-state reconstruction of the roots of subtrees parti-  
361 tioned by each edge (where the subtrees are re-rooted at the nodes adjacent  
362 the edge), but this unsupervised phylogenetic factorization body masses in 3374  
363 mammals takes 15 seconds on a laptops and yields partitions which simplify the  
364 story of mammalian body-mass variation to a set of 5 edges forming 6 binned  
365 phylogenetic units.

366 Two-sample tests can be used for phylogenetic factorization of any vector of  
367 trait data. For another example, Bernoulli trait data, such as presence/absence  
368 of a trait, can be factored using Fisher’s exact test that there is the same  
369 proportion of presences in two groups,  $R$  and  $S$ . In this case, the aggregation  
370 operation  $A(\mathbf{x}_R) = \sum_{i \in R} x_i$  counts the number of successes in group  $R$ , the  
371 contrast operation is the computation of the P-value using Fisher’s exact test  
372 and the contingency table

Successes	Failures	Total
$A(\mathbf{x}_R)$	$r - A(\mathbf{x}_R)$	$r$
$A(\mathbf{x}_S)$	$s - A(\mathbf{x}_S)$	$s$
$A(\mathbf{x}_R) + A(\mathbf{x}_S)$	$r + s - (A(\mathbf{x}_R) + A(\mathbf{x}_S))$	$r + s$

373 and an objective function can be defined as the inverse of the P-value from

374 Fisher's exact test,  $\omega_e = |C_e^{-1}|$ . The phylofactorization of vertebrates by  
375 land/water association in Figure 1, using an ad-hoc selection of vertebrates for  
376 illustration, was performed using Fisher's exact test, and the factors obtained  
377 correspond to Tetrapods, Cetaceans, and Pinnipeds. Unlike the phylofactoriza-  
378 tion of mammalian body mass, all three factors obtained from phylofactorization  
379 of vertebrate land/water association correspond to a set of traits. Tetrapods  
380 evolved lungs and limbs which allowed them to live on land. Cetaceans evolved  
381 fins and blowholes, and Pinnipeds evolved fins, all traits adaptive to life in the  
382 water.

383 Two-sample tests are used when partitioning a vector of traits. Phylofac-  
384 torization of body mass and land/water associations illustrate two potential  
385 evolutionary models under which edges are important: correlated evolution of  
386 members of a clade and punctuated equilibria. More complicated methods for  
387 phylofactorization can keep these cases in mind when interpreting the edges  
388 identified: they may correspond to traits, or they may correspond to ancient  
389 (and possibly ongoing) evolutionary processes common within a clade, such  
390 as ecological release or niche partitioning. When the objective function from  
391 two-sample tests has a well-defined null distribution, the uniformity of the dis-  
392 tribution of P-values from two-sample tests can be used to define a stopping criteria  
393 as discussed later (see: "stopping criteria").

## 394 **Example 2: Contrast basis and phylogenetic components** 395 **analysis**

396 For datasets with multiple samples of the same feature, such as abundance data  
397 for a set of species across a range of habitats, the phylogeny provides a natu-  
398 ral scaffold for low-rank, phylogenetically interpretable approximations of the  
399 data. One reliable algorithm for producing phylogenetically-interpretable low-

400 rank approximations of data is to construct basis elements through aggregation  
 401 and contrast vectors (Figure 1c). An aggregation basis element for a group  
 402  $Q = R \cup S$  can be constructed through a vector whose  $i$ th element is

$$\mathbf{v}_{A_Q,i} = \begin{cases} a & i \in Q \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

403 and such aggregation basis elements can be subsequently partitioned with a  
 404 contrast vector

$$\mathbf{v}_{C_{R|S},i} = \begin{cases} b & i \in R \\ -c & i \in S \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where  $b > 0$  and  $c > 0$ . By meeting the criteria

$$rb - sc = 0 \quad (4)$$

$$rb^2 + sc^2 = 1 \quad (5)$$

, one can ensure that  $\mathbf{v}_{A_Q}$  and  $\mathbf{v}_{C_Q}$  are orthogonal and with unit norm. These criteria are satisfied by

$$b = \sqrt{\frac{s}{r(r+s)}} \quad (6)$$

$$c = \sqrt{\frac{r}{s(r+s)}}. \quad (7)$$

In this case, the aggregation and contrast operations for sample  $j$  are

$$\begin{aligned} A(\mathbf{x}_{R,j}) &= \bar{\mathbf{x}}_{R,j} \\ C(A(\mathbf{x}_{R,j}), A(\mathbf{x}_{S,j})) &= \sqrt{\frac{rs}{r+s}} (\bar{\mathbf{x}}_{R,j} - \bar{\mathbf{x}}_{S,j}). \end{aligned} \quad (8)$$

405 where  $\bar{x}_{R,j}$  is the sample mean of species in group  $R$  and sample  $j$ . Projecting  
406 a dataset onto  $\mathbf{v}_{C_{R|S}}$  yields coordinates which are a standardized difference of  
407 means: the absolute value of the projection of a single multi-species sample  
408 onto a contrast vector yields the two-sample t-statistic from equation (1). The  
409 contrast vector is comprised of two sub-aggregations of opposite sign, one for  
410 group  $R$  and the other for group  $S$ . By ensuring criterion (4), the groups aggre-  
411 gated within a contrast vector can be subsequently partitioned with additional,  
412 orthogonal contrast vectors splitting each group  $R$  and  $S$ . Maintaining criterion  
413 (5), the aggregation and contrast vectors defined here can be used to construct  
414 an orthonormal basis for describing data containing our species,  $\mathbf{x}_j \in \mathbb{R}^m$ , by  
415 defining a set of  $q \leq m$  orthogonal aggregation vectors corresponding to disjoint  
416 sets of species  $Q_l$  such that the entire set of aggregations,  $\bigcup_{l=1}^q Q_l = \{1, \dots, n\}$ ,  
417 covers the entire set of  $m$  species. Then,  $m - q$  contrast vectors partitioning  
418 the aggregations and the sub-aggregations within contrast vectors can complete  
419 the basis (Figure 1c). Of note is that, as defined in equations (2) and (3), the  
420 span of any aggregate and its contrast is equal to the span of the contrasts'  
421 sub-aggregates, i.e. for  $R \cup S = Q$ ,

$$\text{span}(\mathbf{v}_{A_Q}, \mathbf{v}_{C_{R|S}}) = \text{span}(\mathbf{v}_{A_R}, \mathbf{v}_{A_S}) \quad (9)$$

422 (Figure 1c) and the two natural ways of changing variables with the phylogeny,  
423 an aggregate of species and its orthogonal contrast (grouping species and parti-  
424 tioning the group) or two orthogonal aggregates (two disjoint groups of species),  
425 are rotations of one-another. Aggregation and contrast vectors translate the no-  
426 tion of phylogenetic scale and group-differences into a basis that can be used to  
427 analyze community ecological data.

428 Pure aggregation vectors as defined in equation (2) can be defined a priori  
429 based on traits or clades of species thought to be important for the question

430 at hand (e.g. aggregate “terrestrial” and “aquatic” animals), or defined by the  
431 data through myriad clustering algorithms or phylofactorization based purely on  
432 aggregation by converting steps (1) and (2) in the phylofactorization algorithm  
433 into a single step: maximizing an objective function of the aggregate of a clade.  
434 A special case occurs when data are compositional [1], in which case the sum  
435 of any sample across all species in the community will equal 1 and thus the  
436 data are constrained by an aggregation element - the aggregate of all species  
437 - which can only be subsequently contrasted. Phylofactorization via contrasts  
438 of log-relative abundance data allows one to construct an isometric log-ratio  
439 transform, a commonly used and well-behaved transform for the analysis of  
440 compositional data [10, 9, 44]. Since the span of an aggregate and its contrast  
441 is equal to the span of the contrasts’ two aggregates (equation 9), we simplify  
442 construction of the basis by considering, from here on out, only the “contrast  
443 basis” in which the an initial aggregate of all species is then partitioned with a  
444 series of contrasts.

445 An orthonormal basis, including one constructed via aggregation and con-  
446 trast vectors, enables researchers to partition the variance captured by each of  
447 a set of orthogonal directions corresponding to discrete, identifiable features in  
448 the phylogeny. Using the phylofactorization algorithm, a dataset  $\mathbf{X} = [x]_{i,j}$  can  
449 be summarized by defining the objective function

$$\omega_e = \text{Var} [\mathbf{v}_{C_e}^T \mathbf{X}] \quad (10)$$

450 where  $\mathbf{v}_{C_e}$  is the contrast vector from (3) corresponding to the sets of species,  
451  $R$  and  $S$ , split by edge  $e$ . Phylofactorization by variance-maximization yields a  
452 phylogenetic decomposition of variance, referred to as “phylogenetic components  
453 analysis” or PhyCA. PhyCA is a constrained version of principal components  
454 analysis, allowing researchers to focus only on the loadings,  $\mathbf{v}_{C_e}$ , corresponding

455 to contrasts of species separated by an edge.

456 The variance of component scores,  $\mathbf{y}_e = \mathbf{v}_{C_e}^T \mathbf{X}$ , can be easily understood if  
457 the data  $[x_{i,j}]$  are assumed to be Gaussian. The component score for sample  $j$ ,  
458  $\mathbf{y}_{e,j}$ , can be written as

$$\mathbf{y}_{e,j} = \sqrt{\frac{rs}{r+s}} (\bar{\mathbf{x}}_{R,j} - \bar{\mathbf{x}}_{S,j}) \quad (11)$$

459 where  $\bar{\mathbf{x}}_{R,j}$  is the sample mean of  $x_{i,j}$  for  $i \in R$  and  $\bar{\mathbf{x}}_{S,j}$  is the sample mean of  
460  $x_{i,j}$  for  $i \in S$ . The variance of the component score across all samples  $j = 1, \dots, n$   
461 is

$$\text{Var}[\mathbf{y}_e] = \frac{rs}{r+s} (\text{Var}[\bar{\mathbf{x}}_R] + \text{Var}[\bar{\mathbf{x}}_S] - 2\text{Cov}[\bar{\mathbf{x}}_R, \bar{\mathbf{x}}_S]). \quad (12)$$

462 The variance of  $\mathbf{y}_e$  increases through a combination of variances in aggregations  
463 of groups  $R$  and  $S$  across samples ( $\bar{\mathbf{x}}_R$  and  $\bar{\mathbf{x}}_S$ , respectively) and a high negative  
464 covariance between aggregations for groups  $R$  and  $S$  across samples. Species  
465 with a negative covariance may be competitively excluding one-another or may  
466 be differentiated due to a trait which arose along edge  $e$  which causes different  
467 habitat associations or responses to treatments. Edges extracted from PhyCA  
468 are edges along which putative functional ecological traits arose differentiating  
469 the species in  $R$  and  $S$  in the dataset of interest.

470 **Phylogenetic Components of the American Gut** To illustrate, we per-  
471 form PhyCA to identify 10 factors from a sub-sample of the American Gut  
472 dataset and the greengenes phylogeny [8] containing  $m = 1991$  species and  $n =$   
473 788 samples from human feces (Figure 3b). The American Gut dataset was fil-  
474 tered to only fecal samples with over 50,000 sequence counts and, for those sam-  
475 ples, otus with an average of more than one sequence count per sample. After  
476 performing PhyCA, each identified resulting component score,  $\mathbf{y}_{e^*}$ , is assessed  
477 for a linear association with seven explanatory variables: `types_of_plants` (a

478 question asking participants how many types of plants they've eaten in the past  
479 week), age, bmi, alcohol consumption frequency, sex, antibiotic use (ABX), and  
480 inflammatory bowel disease (subset\_ibd) (Figure 3b). The raw P-values are  
481 presented below, but for a reference, the P-value threshold for a 5% family-wise  
482 error rate is  $7.1 \times 10^{-4}$ .

483 The first factor splits 1229 species of Firmicutes from the remainder of mi-  
484 crobes. The component score for the first factor,  $y_{e_1^*}$ , is strongly associated with  
485 antibiotic use ( $P=3.6 \times 10^{-4}$ ), with dramatic decreases in relative abundance  
486 in patients who have taken antibiotics in the past week or month. The second  
487 factor identifies 217 species of several genera of Lachnospiraceae, a clade con-  
488 tained within the Firmicutes, strongly associated with age ( $P=1.2 \times 10^{-15}$ ) and  
489 bmi ( $P=3.2 \times 10^{-6}$ ) and alcohol ( $P=6.4 \times 10^{-3}$ ). The third factor is a clade of  
490 81 Bacteroides most strongly associated with types\_of\_plants ( $P=2 \times 10^{-9}$ ).  
491 By identifying a clade of Bacteroides as a major axis of variation, factors 1  
492 and 3 refine the Firmicutes to Bacteroidetes ratio commonly used to describe  
493 variation in the gut microbiome and found associated with obesity and other  
494 disease states [28, 7]. It's been found that the Firmicutes/Bacteroidetes ratio  
495 changes with age [31], but the picture from phylofactorization is more nuanced:  
496 the large clade of Firmicutes in the first factor does not change with age, but  
497 the Lachnospiraceae within that clade decrease strongly with age relative to  
498 the remaining Firmicutes, while the Bacteroides show only a moderate decrease  
499 with age. The strong decrease with age in Lachnospiraceae is found in a few  
500 other clades within the Firmicutes: the 4th factor identified a clade of Firmi-  
501 cutes of the family Ruminococcaceae strongly associated with types of plants  
502 ( $P=3.6 \times 10^{-5}$ ), sex ( $P=5.9 \times 10^{-4}$ ) and decreasing with age ( $P=9.2 \times 10^{-4}$ ),  
503 and the 5th factor identified a group of Firmicutes of the family Tissierellaceae  
504 that decrease strongly with age ( $P=1.9 \times 10^{-5}$ ).



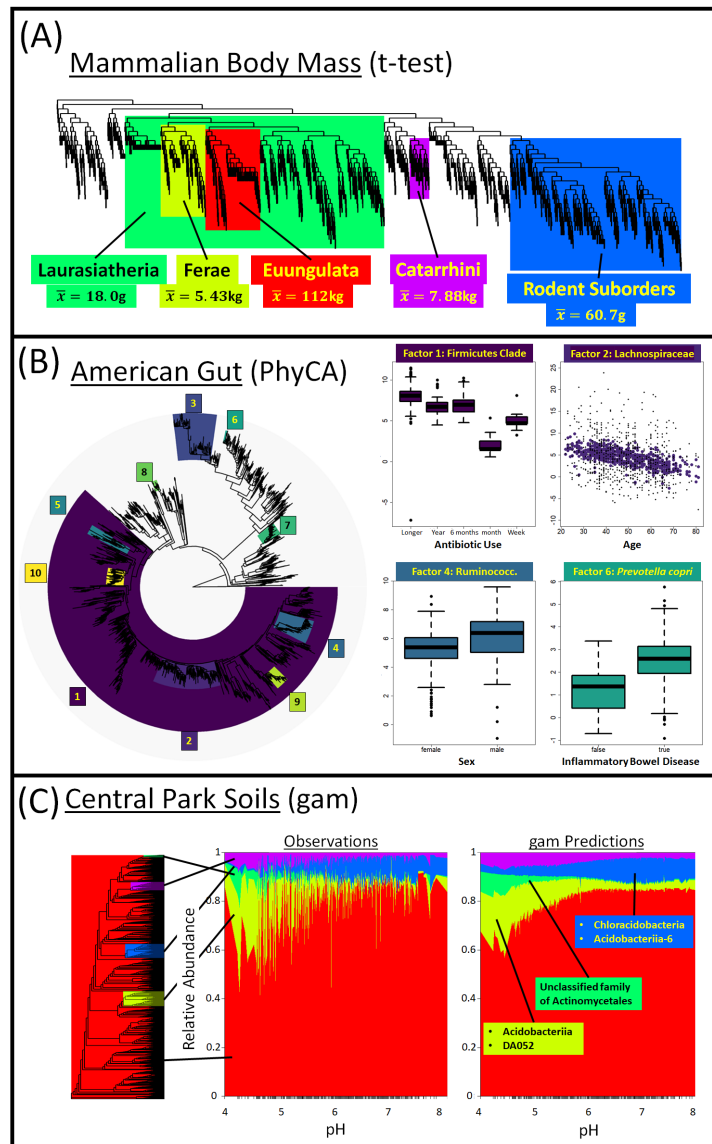


Figure 3: Phylofactorization with contrast basis. (A) The contrast basis defines variables similar to t-statistics, and maximizing the projection of data onto the contrast basis can identify phylogenetic factors. Five iterations of phylofactorization on a dataset of mammalian log-body mass yields five clades with very different body masses. (B) Maximizing the variance of component scores,  $y_e$ , of log-relative abundance data produces a “phylogenetic components analysis” (PhyCA) of the American Gut dataset. The most variable clades cover a range of phylogenetic scales. Downstream analysis of component scores tested associations with meta-data - plotted are linear predictors against relevant meta-data; the plot of Lachnospiraceae includes the raw data as black dots. (C) More complicated methods can be used, such as generalized additive modeling with  $y_e$ . Using the central park soils dataset,  $y_e$  of log-relative abundances, the model  $y_e \sim s(\log(\text{Carbon})) + s(\log(\text{Nitrogen})) + s(\text{pH})$ , and the objective of maximizing the explained variance, we obtained the same 4 factors obtained using generalized linear modeling in the original data, including the misnomer group of Chloracidobacteria that don't thrive in low pH environments. The relative importance of pH in the generalized additive models and exact clades with a high amount of variance explained by pH allows a projection of 3000 species into 5 BPUs for clear visualization of a dominant feature of how soil bacterial communities change along a key environmental gradient.

505 The sixth factor is a small group of 5 OTUs of *Prevotella copri* strongly as-  
506 sociated with types\_of\_plants ( $P=2.8 \times 10^{-4}$ ) and inflammatory bowel disease  
507 ( $P=2.5 \times 10^{-3}$ ). Previous studies have found that *Prevotella copri* abundances  
508 are correlated with rheumatoid arthritis in people and inoculation of *Prevotella*  
509 *copri* exacerbates colitis in mice. Consequently, *Prevotella copri* is hypothesized  
510 to increase inflammation in the mammalian gut [42], and the discovery of *Pre-*  
511 *vetella copri* as one of the dominant phylogenetic factors of the American Gut, as  
512 well as the discovery of its association with IBD, corroborates the hypothesized  
513 relationship between *Prevotella copri* and inflammation. Likewise, the seventh  
514 factor is a clade of 41 Gammaproteobacteria of the order Enterobacteriales also  
515 associated with types\_of\_plants ( $P=6.7 \times 10^{-8}$ ) and weakly associated with  
516 inflammatory bowel disease ( $P=0.022$ ). Gammaproteobacteria were used as  
517 biomarkers of Crohn’s disease in a recent study [49] and their associations with  
518 IBD in the American Gut project corroborates the possible use of Gammapro-  
519 teobacterial abundances for detection of IBD from stool samples. Summaries of  
520 the models for all factors’ component scores are in the supplemental information.

### 521 **Example 3: Compositional, log-normal and Gaussian regression-** 522 **phylofactorization**

523 Phylogenetic contrast vectors can be used to define more complicated objective  
524 functions for data assumed to be Gaussian or easily mapped to Gaussian, such  
525 as logistic-normal compositional data or log-normal data. Conversion of the  
526 data to an assumed-Gaussian form can then allow one to perform least-squares  
527 regression using  $\mathbf{y}_e$  as either an independent or dependent variable. Rather  
528 than performing PhyCA and subsequent regression, one can choose phylogenetic  
529 factors based on their associations with meta-data of interest.

530 Maximizing the explained variance from regression identifies clades through

531 the product of a high contrast-variance from equation (10) and the percent  
532 of explained-variance from regression - such clades can capture large blocks  
533 of explained variance in the dataset. Another common objective function is  
534 the deviance or  $F$ -statistic from regression which identifies clades with more  
535 predictable responses - such clades can be seen as bioindicators or particularly  
536 sensitive clades, even if they are not particularly large or variable clades in  
537 the data. Regression-phylofactorization can use the component scores as an  
538 independent variable, as was used in the phylofactorization-based classification  
539 of Crohn's disease [49]. For multiple regression, one can use the explanatory  
540 power of the entire model, or a more nuanced objective function of a subset of  
541 the model. More complicated regression models can be considered, including  
542 generalized additive models.

543 To illustrate the flexibility of regression phylofactorization to identify phy-  
544 logenetic scales corresponding to nonlinear patterns of abundance-habitat asso-  
545 ciations, we perform a generalized additive model analysis of the Central Park  
546 soils dataset [39] analyzed previously using a generalized linear model. To iden-  
547 tify non-linear associations between clades and pH, Carbon and Nitrogen, we  
548 perform a generalized additive model of the form

$$\mathbf{y}_e \sim s(\text{pH}) + s(\text{Carbon}) + s(\text{Nitrogen}) \quad (13)$$

549 and maximize the explained variance (Figure 3c). The resultant phylofactor-  
550 izations identifies the same 4 factors as the generalized linear model, but allows  
551 nonlinear and multivariate analysis of how community composition changes over  
552 environmental meta-data. Combining the high relative-importance of pH with  
553 the identified 4 factors, splitting over 3,000 species 5 binned phylogenetic units,  
554 allows clear and simple visualization of otherwise complex behavior of how a  
555 community of several thousand microbes changes across several hundred soil

556 samples. As with the original analysis, the generalized additive modeling phylo-  
557 factorization identifies a clade of Acidobacteria - the Chloracidobacteria - which  
558 have highest relative abundances in more neutral soils.

#### 559 **Example 4: Phylofactorization through generalized linear** 560 **models**

561 Many ecological data are not Gaussian. Presence-absence data or count data  
562 with many zeros cannot be easily transformed to yield approximately Gaus-  
563 sian random variables. However, the graph-partitioning algorithm we describe  
564 provides a framework for implementing phylofactorization with the appropriate  
565 choice of aggregation and contrast operations defined through more complex re-  
566 gression models. Data assumed to be exponential family random variables can  
567 be analyzed with regression-phylofactorization by adapting generalized linear  
568 models through shared coefficients and assumptions of within-group homogene-  
569 ity that allow algebraic group operations for aggregation within the exponential  
570 family. We present three options for aggregation and contrast in generalized  
571 linear models, intended to be an illustrative, but not exhaustive, account of the  
572 application of phylofactorization in the context of generalized linear and addi-  
573 tive models. These options correspond to the contrast basis, either explicitly  
574 using the contrast basis to approximate the coefficient matrix in multivariate  
575 generalized linear models, or performing shared-coefficient or factor-contrasts  
576 in generalized linear modeling which, we'll show later, have a similar graph-  
577 topological behavior as the contrast basis.

578 The first method is to perform multivariate generalized modeling of one  
579 generalized linear model or generalized additive model using the same formula  
580 for each species and subsequently use contrast basis elements,  $\mathbf{v}_{C_e}$ , to change the  
581 basis for regression parameters of interest - such expansions of the maximum-

582 likelihood estimates of regression coefficients are maximum likelihood estimates  
583 of the expansion by the invariance of maximum likelihood estimates. To be  
584 precise, given an  $m \times p$  matrix,  $B$ , of coefficients used in regression on species-  
585 specific data. In particular, generalized linear models will model the predictors,  
586  $\boldsymbol{\eta} \in \mathbb{R}^s$ , for each species through a linear model

$$\boldsymbol{\eta} \sim \mathbf{B}\mathbf{Z}. \quad (14)$$

587 Instead of using the exhaustive  $s \times p$  list of coefficients, one can represent the  
588 coefficient matrix  $B$  through contrast basis elements and their component scores

$$\mathbf{B} = \mathbf{1}\mathbf{w}_0^T + \mathbf{V}\mathbf{W} + \boldsymbol{\epsilon} \quad (15)$$

589 where  $\mathbf{1} \in \mathbb{R}^s$  is the one vector,  $\mathbf{w}_0 \in \mathbb{R}^p$  contains the sum of the regression  
590 coefficients for each of the  $p$  predictors,  $\mathbf{V} \in \mathbb{R}^{s \times K_t}$  is a matrix whose columns  
591 are contrast basis elements and  $\mathbf{W} \in \mathbb{R}^{K_t \times p}$  is a matrix whose rows are the  
592 component scores for each contrast basis element. One example of an objective  
593 function guiding the choice of contrast basis elements can be the norm

$$\omega_e = \|\mathbf{v}_{C_e}^T \mathbf{B}\| \quad (16)$$

594 which captures the extent to which coefficients in  $B$  are different between the  
595 sets of species partitioned by the edge  $e$ . Another option for an objective func-  
596 tion is the deviance of a reduced model with shared coefficients.

597 Other options for aggregation and contrast exploit the factor-contrasts built  
598 into generalized linear and additive modeling machinery. Factor contrasts, such  
599 as a variable  $g \in \{R, S\}$  indicating which group a species is in, can capture  
600 the assumption of shared coefficients within-groups and different coefficients

601 between-groups in multivariate generalized linear modeling across all species. A  
602 third option is to assume within-group homogeneity and aggregate exponential  
603 family random variables to a “marginally stable” exponential family random  
604 variable used for analysis. Marginal stability, to the best of our knowledge, has  
605 not been explicitly defined elsewhere, and thus we introduce the term here by  
606 loosening the definition of stable distributions [41].

607 **Stable distribution** A distribution with parameters  $\theta$ ,  $\mathcal{F}(\theta)$ , is said to be  
608 stable if a linear combination of two independent random variables from  $\mathcal{F}(\theta)$   
609 is also in  $\mathcal{F}(\theta)$ , up to location and scale parameters.

610 **Marginally stable distribution** A distribution with parameters  $\{\theta_1, \theta_2\}, \mathcal{F}(\theta_1, \theta_2)$ ,  
611 is said to be marginally stable on  $\theta_1$  if  $\mathcal{F}(\theta_1, \theta_2)$  is it is stable conditioned on  $\theta_1$   
612 being fixed.

613

614 For example, the Gaussian distribution is stable: the sum of two Gaus-  
615 sian random variables is also Gaussian. Meanwhile, binomial random variables  
616  $Binom(\rho, N)$  are marginally stable on  $\rho$ ; random variables  $x_i \sim Binom(\rho, N_i)$   
617 can be summed to yield  $A(\mathbf{x}) \sim Binom(\rho, \sum N_i)$ . The marginal stability can  
618 also be used with transformations that connect the assumed distribution of the  
619 data to a marginally stable distribution. Log-normal random variables can be  
620 converted to Gaussians through exponentiation; chi random variables can be  
621 converted to chi-squared through squaring - random variables from many dis-  
622 tributions may be analyzed by transformation to a stable or marginally stable  
623 family of distributions. Such transformation-based analyses implicitly define  
624 aggregation through a generalized  $f$ -mean

$$A_f(\mathbf{x}_R) = f^{-1} \left( \sum_{i \in R} f(x_i) \right) \quad (17)$$

625 where  $f(x) = \log(x)$  for log-normal random variables,  $f(x) = x^2$  for Chi ran-  
626 dom variables, etc. The goal of such aggregation, whether through exploiting  
627 marginal stability or generalized  $f$ -means or other group operations in the ex-  
628 ponential family, is to produce summary statistics for each group,  $R$  and  $S$ , in a  
629 manner that permits generalized linear modeling of the summary statistics. By  
630 ensuring summary statistics are also exponential-family random variables, one  
631 can perform a factor-contrast style analysis as described above but only on the  
632 two summary statistics and not on all  $s$  species. Doing so can greatly reduce  
633 the computational load of phylofactorizing large datasets and, as we show be-  
634 low, can increase the power of edge-identification even when the within-group  
635 homogeneity assumption does not hold. Marginal stability, for the purposes of  
636 phylofactorization, must be on the parameter of interest in generalized linear  
637 modeling (Figure 3a).

638 Marginal stability opens up more distributions to stable aggregation. Pres-  
639 ence absence data, for instance, can be assumed to be Bernoulli random vari-  
640 ables. The assumption of within-group homogeneity for the probability of pres-  
641 ence,  $\rho$ , allows addition of Bernoulli random variables within each group,  $R$   
642 and  $S$ , to yield a respective binomial random variable,  $x_R$  and  $x_S$ . Likewise,  
643 the addition of a group of binomial random variables with the same proba-  
644 bility of success,  $\rho$ , yields an aggregate binomial random variable. A homo-  
645 geneous group of exponential random variables with the same rate parameter,  
646  $\lambda$ , can be added to form a gamma random variable. Gamma random vari-  
647 ables,  $x_i \sim \text{Gamma}(\kappa_i, \theta)$ , parameterized by their shape,  $\kappa_i$ , and scale,  $\theta$ , are  
648 marginally  $A$ -stable on  $\theta$ . Addition of geometric random variables with the  
649 same rate parameter forms a negative binomial, and the addition of a group  
650 of negative binomial random variables,  $x_i \sim \text{NB}(\pi_i, \rho)$ , with the same proba-  
651 bility of success  $\rho$  but different numbers of failures,  $\pi_i$ , can be aggregated into

652  $x_R = \sum_{i \in R} x_i$  where  $x_R \sim NB(\sum_{i \in R} \pi_i, \rho)$ . All of these distributions are not  
653 stable, but they are marginally stable.

654 For a practical example of regression phylofactorization of an exponential  
655 family random variable, we consider a presence/absence dataset  $\mathbf{X}$ , whose en-  
656 tries  $x_{i,j}$  are assumed to be Bernoulli random variables with some probability  
657 dependent upon meta-data,  $\rho_{i,j}(\mathbf{Z})$ , modeled naturally through the canonical  
658 link function,  $\eta$ . Phylofactorization can identify edges which separate species  
659 based on their response to a set of environmental variables,  $\{z_k\}$ . For exam-  
660 ple researchers sequencing microbial 16S sequences in the soil may have data  
661 on the presence/absence of microbes across a range of biomass, pH, and nitro-  
662 gen concentrations and be interested in identifying the edges that best separate  
663 microbes based on differential probability of presence in response to nitrogen,  
664 controlling for common responses to biomass and group-specific responses to pH.  
665 Such questions can be addressed through appropriate choice of factor contrasts  
666 in a generalized linear model, with the optional use of within-group homogeneity  
667 to allow aggregation of presence/absences to binomial random variables.

668 A more general formula for phylofactorization based on predictors in regres-  
669 sion models for exponential family random variables can be made by partitioning  
670 the independent variables,  $\{z_k\}_{k=1}^P$ , into three disjoint sets: a set U of universal  
671 effects assumed to have a common effect across species, a set B of group-specific  
672 effects one wishes to control for, and a set P of group-specific effects one wishes  
673 to use for phylofactorization. Instead of a species-specific, multivariate general-  
674 ized linear model of the predictor for each species  $i$ ,  $\eta_i$ ,

$$\eta_i = \beta_{i,0} + \beta_{i,1}z_1 + \dots + \beta_{i,p}z_p, \quad (18)$$

675 one can define a factor,  $g \in \{R, S\}$ , which indicates which group a species is  
676 in (or, for aggregated data  $A(\mathbf{x}_R)$ , which group the aggregate corresponds to),



677 and construct a generalized linear model

$$\eta = \sum_{l \in U} \beta_l z_l + g \times \sum_{j \in B} \beta_j z_j + g \times \sum_{k \in P} \beta_k z_k, \quad (19)$$

678 where  $g \times z_j$  indicates an interaction term between the group factor and the  
 679 independent variable  $z_j$ . For example, a data frame contrasting the counts of  
 680 “birds” from “non-birds” can be constructed as follows

Site	Species	Abundance	$z_1$	$z_2$	$g$
1	Sparrow	10	1	.5	$R$
1	Dove	8	1	.5	$R$
1	Lizard	1	1	.5	$S$
1	Mouse	3	1	.5	$S$
1	Cat	1	1	.5	$S$
2	Sparrow	2	0	-2	$R$
2	Dove	1	0	-2	$R$
2	Lizard	10	0	-2	$S$
2	Mouse	4	0	-2	$S$
2	Cat	3	0	-2	$S$
...	...	...	...	...	...

681 and a generalized linear model for a count family (e.g. Poisson, binomial, or  
 682 negative binomial) with the formula

$$\text{Abundance} \sim z_1 + g \times z_2$$

683 can be used for maximum likelihood estimation of  $g$ , the factor which contrasts  
 684 birds from non-birds whose coefficient or deviance can be used as the objective  
 685 function.

686 The contrast function is defined through the factor-contrast, and one exam-

687 ple of an objective function is an omnibus test for all interaction terms between  
688 g and predictors for phylofactorization, P, relative to the model containing only  
689 the terms from U and B. Another example of an objective function can be  
690 the  $L_2$  norm of the coefficients  $\vec{\beta}_P$  of interest for phylofactorization. For the  
691 example with carbon, pH, and nitrogen, one can perform phylofactorization to  
692 identify edges differentiating microbial presence-absences (or negative binomial  
693 sequence-counts) through factor-contrasts in the model

$$\eta = \text{Carbon} + g \times \text{pH} + g \times \text{Nitrogen}. \quad (20)$$

694 The same principle of optional aggregation to marginally stable distributions  
695 followed by factor-contrasts can be applied to perform phylofactorization of  
696 exponential family random variables through generalized additive models.

697 These two approaches are by no means an exhaustive list of how to integrate  
698 generalized linear modeling into phylofactorization. For instance, it may be  
699 possible to perform phylofactorization by representing a vector of canonical link  
700 functions for two groups or multiple species,  $\eta$ , in terms of “canonical contrasts”  
701 using an aggregation-contrast basis defined above. The examples included are  
702 intended to illustrate the feasibility and creative options for robust and statis-  
703 tically well-calibrated phylofactorization of datasets comprised of non-Gaussian  
704 random variables.

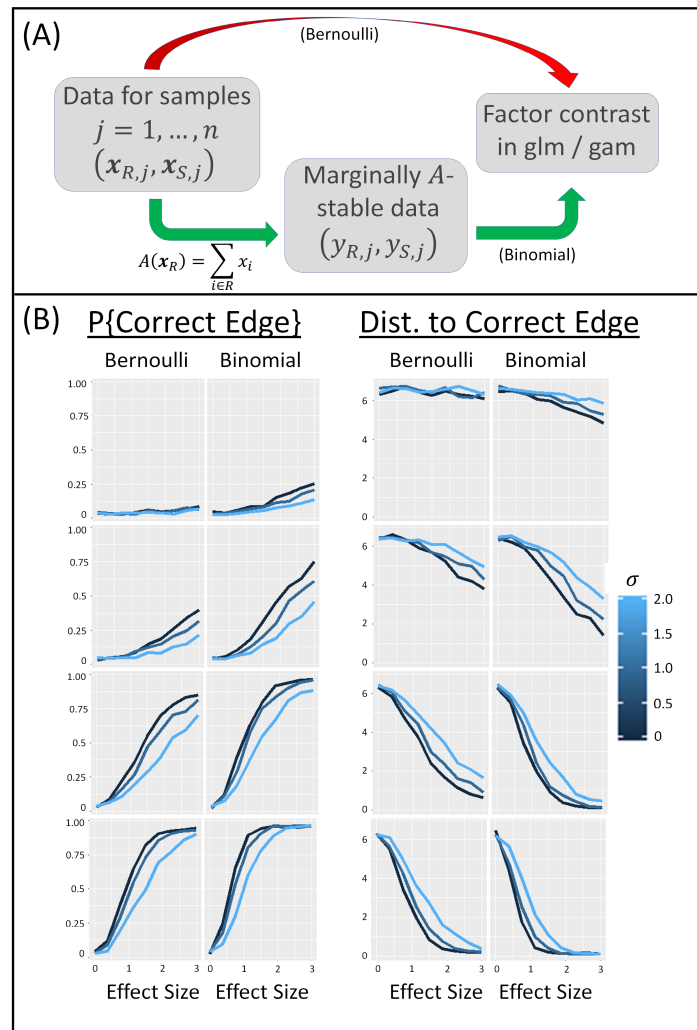


Figure 4: Factor contrasts can be used to define objective functions for phylofactorization of exponential family random variables. (A) Each edge separates the species in a sample into two groups. These groups can be used as factors directly in a generalized linear model as in equation 13. Alternatively, a within-group homogeneity assumption can be used to aggregate data of many exponential family random variables to a marginally stable distribution, such as addition of Bernoulli random variables with the same probability of success to obtain a binomial random variable, or addition of exponential random variables with the same rate parameter to obtain a gamma random variable. Regression on marginally stable random variables may dramatically reduce computational costs and improve accuracy. (B) Simulations of Bernoulli presence/absence data of 30 species with a random phylogeny suggest that aggregation to binomial improves power across a range of effect sizes,  $\delta$ , (x-axis), sample sizes,  $n$  (rows), and within-group heterogeneity,  $\sigma$  (see supplemental info for more details on the simulations). In all cases considered here, aggregation of presence-absence data to binomial random variables for subsequent factor-contrasts outperformed the raw factor contrast of Bernoulli presence/absence data, suggesting it is at least a viable tool for large datasets, but the generality of improved power of regression on surrogate, marginally stable aggregates remains to be seen.

To test and compare the viability of the two proposed methods - raw factor contrasts and aggregation to a marginally stable distribution - we simulated 700 replicates of effects of the form in equation (13) for the probability of presence on random edges, with varying effect sizes, sample sizes and within group homogeneity. For 700 replicates for each combination of sample size  $n \in \{5, 10, 30, 60\}$ , effect size  $\delta \in \{0, 0.375, 0.75, 1.125, 1.5, 1.875, 2.25, 2.625, 3\}$ , and within-group variance  $\sigma \in \{0, 1, 2\}$ , we simulated three explanatory variables  $\{z_1, z_2, z_3\}$  as independent, identically distributed  $n$ -vectors of standard normal random variables. The log-odds of presence for individual  $i$  in group  $R$  or group  $S$  was modeled as

$$\begin{aligned}\eta_{R,i} &= z_1 + z_2 + \left(0.1 + \frac{\delta}{2}\right) z_3 + z_{4,i} \\ \eta_{S,i} &= z_1 - z_2 + \left(0.1 - \frac{\delta}{2}\right) z_3 + z_{4,i}\end{aligned}\tag{21}$$

705 where  $z_{4,i} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  are independent Gaussian random variables particular  
706 to the individual and sample. The data were either kept as Bernoulli random  
707 variables or aggregated via summation to binomial random variables and then  
708 analyzed using factor contrasts in a generalized linear model of the form

$$\eta = z_1 + \mathbf{g} \times z_2 + \mathbf{g} \times z_3.\tag{22}$$

709 The objective function was the deviance from the final term,  $\mathbf{g} \times z_3$ . The proba-  
710 bility of identifying the correct edge and the distance between the identified and  
711 correct edge (in the number of nodes separating the two edges) are plotted in  
712 Figure 4b. The method of factor-contrasts for glm-phylofactorization asymptot-  
713 ically approaches perfect edge-identification, both in the probability of detecting  
714 the correct edge and in distance from the correct edge, as the sample sizes and  
715 effect sizes increase. Aggregation to binomial and subsequent factor-contrast of

716 the aggregates slightly improved the power of edge-identification in these sim-  
717 ulations. While the performance of the Bernoulli to binomial aggregation may  
718 decrease with differences in within-group means as opposed to an addition of  
719 individual within-group variance through  $z_{4,i}$ , our purpose here is to illustrate  
720 that there exist methods of aggregation and contrast which permit maximum-  
721 likelihood regression-phylofactorization of exponential family random variables.  
722 Marginally-stable aggregation and stepwise construction of factor contrasts are  
723 but one viable way to extend regression-phylofactorization to exponential family  
724 random variables.

## 725 **Phylogenetic factors of space and time**

726 So far, we've demonstrated phylofactorization through examples of cross-sectional  
727 data, either through two-sample tests of cross-sections of species or through anal-  
728 yses of contrast-basis projections or factor contrasts in communities sampled  
729 across a range of meta-data. Phylofactorization can also be used in conjunction  
730 with many analyses of spatial and temporal patterns. Samples of a commu-  
731 nity over space can be projected onto contrast basis elements and the resulting  
732 component scores,  $y_e$ , can be analyzed much like PhyCA to identify the phy-  
733 logenetic partitions of community composition over space. Spatial samples can  
734 also be analyzed using factor contrasts as defined for generalized linear mod-  
735 els. Multivariate Autoregressive Integrated Moving Average (ARIMA) models  
736 can be constructed either as ARIMA models of the component scores,  $y_e$ , or as  
737 multivariate ARIMA models with factor contrasts as used in generalized linear  
738 models perform phylogenetic partitions based on differences in drift, volatility,  
739 and other features of interest.

740 Marginal stable aggregation in spatial and temporal data requires a more  
741 complex consideration of the marginal stability of spatially explicit random vari-

742 able and stochastic processes. Stability”, for spatially and temporally explicit  
743 random variables, must preserve the underlying model for the spatial or tempo-  
744 ral process being used for analysis. An example of marginally stable aggregation  
745 and analysis of time-series data is the stability of neutral drift (sensu Hubbell  
746 [22]) to grouping and the use of a constant volatility transformation for neutral-  
747 ity testing.

748 Neutral communities fluctuate, and those fluctuations have a drift and volatil-  
749 ity unique to neutral drift. Neutral drift can also be defined either by discrete,  
750 finite-community size urn processes or stochastic differential equations for the  
751 continuous approximations of finite but large communities. Recently, Wash-  
752 burne et al. [50] articulated the importance of a feature of neutral drift which  
753 enables time-series neutrality tests: its invariance to grouping of species. If a  
754 stochastic process of relative abundances,  $\mathbf{X}_t$ , obeys the probability law de-  
755 fined by neutral drift (either for discrete, finite communities or their continu-  
756 ous approximations, referred collectively as “neutral process”), then any disjoint  
757 groupings of  $\mathbf{X}_t$  is also a neutral process. Thus, neutral processes are stable  
758 to aggregation by grouping or summation of relative abundances. Collapsing  
759 all species into two disjoint groups,  $R$  and  $S$ , yields a two-dimensional neu-  
760 tral drift well-define neutrality test for time-series data. Specifically, if  $\mathbf{X}_t$  is a  
761 Wright Fisher process and  $R$  and  $S$  are disjoint groups whose union is the entire  
762 community, the quantity

$$\nu_t = \arcsin \left( \left( \sum_{i \in R} X_{i,t} \right) - \left( \sum_{j \in S} X_{j,t} \right) \right) \quad (23)$$

has a constant volatility which serves as a neutrality test for time-series data.  
Thus, phylofactorization can be done to partition edges across which the dy-  
namics appear to be the least neutral. For the test developed by Washburne

et al., the aggregation operation is the  $L_1$  norm and the contrast operation is subtraction:

$$\begin{aligned} A(\mathbf{x}_R) &= |\mathbf{x}_R| \\ C(A(\mathbf{x}_R), A(\mathbf{x}_S)) &= A(\mathbf{x}_R) - A(\mathbf{x}_S) \end{aligned} \quad (24)$$

763 and the objective function,  $\omega$ , for edge  $e$  is the test-statistic of a homoskedasticity  
764 test of  $\arcsin(C_e)$ . Neutrality is a relative measure - biological units are neutral  
765 relative to one-another - and thus the use of aggregation of species into a unit  
766 and a contrast of two units is a natural connection between the theory and  
767 operations of phylogenetic factorization and the concept of neutrality.

768 Whether the data are cross-sectional or spatially/temporally explicit, phylo-  
769 factorization can be implemented through analysis of data projected onto the  
770 contrast basis, factor contrasts in autoregression, or model-specific marginally-  
771 stable aggregation and contrast such as that demonstrated for neutrality testing  
772 of time-series data.

## 773 **Statistical Challenges**

774 We present a unifying algorithm which partition organisms into functional groups  
775 by identifying meaningful differences or contrasts along edges in the phylogeny.  
776 Phylogenetic factorization is formally defined as a graph-partitioning algorithm. How-  
777 ever, maximizing the variance of the data projected onto contrast basis elements  
778 corresponding to edges in the phylogeny is a constrained principal components  
779 analysis. The use of regression-based objective functions and the iterative con-  
780 struction of a low-rank approximation of a data matrix is similar to factor anal-  
781 ysis. The discovery of a sequence of orthogonal factor contrasts in generalized  
782 linear models is a form of stepwise or hierarchical regression. The maximization  
783 of the objective function at each iteration is a greedy algorithm. Each of these

784 connections between phylofactorization and other classes of methods produces a  
785 body of literature from related methods which could inform phylofactorization.  
786 The relation of phylofactorization to pre-existing methods presents a suite of  
787 opportunities for rapid development of this exploratory tool into a more robust,  
788 inferential one.

789     There are statistical challenges common across many methods for phylofac-  
790 torization. In this section, we enumerate some of the statistical challenges and  
791 discuss work that has been done so far. First, as with any method using the phy-  
792 logeny as a scaffold for creating variables or making inferences, the uncertainty  
793 of the phylogeny and the common use of multiple equally likely phylogenies war-  
794 rant consideration and further method development. Other challenges discussed  
795 here are: understanding the propagation of error; development of Metropolis al-  
796 gorithms to better arrive at global maxima; the appropriateness, and error rates,  
797 of phylofactorization under various evolutionary models underlying the effects  
798 (e.g. trait differences, habitat associations, etc.) and residuals in our data;  
799 understanding graph-topological biases and confidence regions; cross-validating  
800 the partitions and inferences from phylofactorization; determining the appropri-  
801 ate number of factors and stopping criteria to stop a running phylofactorization  
802 algorithm; and understanding the null distribution of test-statistics when objec-  
803 tive functions being maximized are themselves test-statistics from a well-known  
804 distribution. Any exploratory data analysis tool can be made into an inferential  
805 tool with appropriate understanding of its behavior under a null hypothesis,  
806 and the connections of phylofactorization to related methods can accelerate the  
807 development of well-calibrated statistical tests for phylogenetic factors.

808 **Phylogenetic inference** So far we have assumed that the phylogeny is known  
809 and error free, but the true evolutionary history is not known - it is estimated.  
810 Consequently, phylofactorizations are making inferences on an uncertain scaf-



811 fold - the more certain the scaffold, the more certain our inferences about a  
812 clade. Two challenges remain for dealing with phylofactorization on an uncer-  
813 tain phylogeny. For a consensus tree, there is the question of what statistics of  
814 the consensus are most easily integrated for precise statements of uncertainty  
815 in phylofactorization inferences. Bootstrapped confidence limits for monophyly  
816 [12] are the most commonly used statement of uncertainty for a consensus tree,  
817 but there may be others as well. Different organisms will have different lever-  
818 ages in regression or two-sample test phylofactorization, and thus monophyly  
819 is only part of the picture: leverage is another. For a set of equally likely  
820 bootstrapped trees, there is a need to integrate phylofactorization across trees.  
821 Phylofactorization of bundles of phylogenies has not yet been done, but may be  
822 a fruitful avenue for future research. One last option for researchers with trees  
823 containing clades with low bootstrap monophyly is to lower the resolution of the  
824 tree. Phylofactorization can still be performed on a tree with polytomies - the  
825 mammalian phylogeny used above contained many - and reducing the number  
826 of edges considered at each iteration can focus statistical effort (and chances of  
827 false-discovery) on clades about which the researcher is more certain.

828 **Propagation of error** Phylofactorization is a greedy algorithm. Like any  
829 greedy algorithm, the deterministic application of phylofactorization is non-  
830 recoverable. Choosing the incorrect edge at one iteration can cause error to  
831 propagate, potentially leading to decreased reliability of downstream edges. Lit-  
832 tle research has been done towards managing the propagation of error in phylo-  
833 factorization, but recognizing the method as a greedy algorithm suggests options  
834 for improving performance. Stochastic-optimization schemes, such as replicate  
835 phylofactorizations using Metropolis algorithms and stochastic sampling as im-  
836 plemented in the mammalian tree phylofactorization (sampling of edges with  
837 probabilities increasing monotonically with  $\omega_e$  and picking the phylofactor ob-

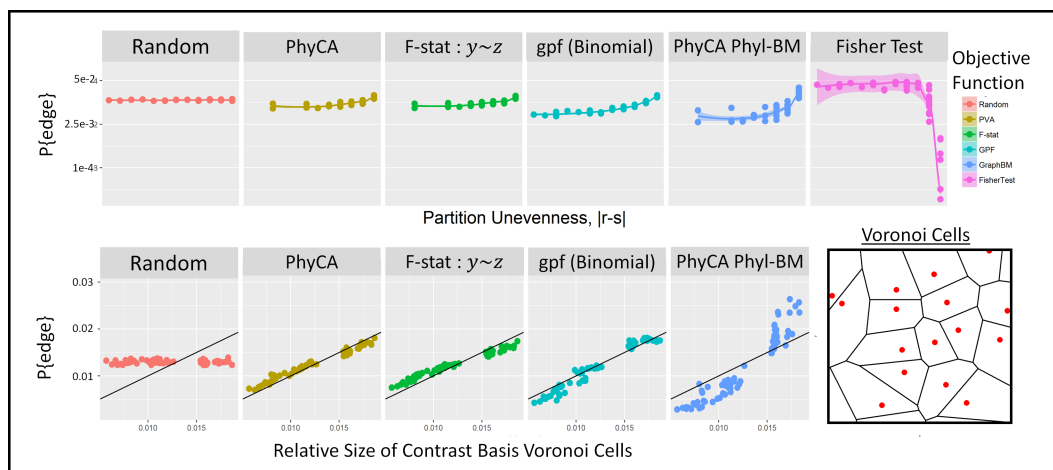
838 ject which maximizes a global objective function), may reduce the risk of error  
839 cascades in phylofactorization [20]. We leave this important problem, and the  
840 construction of suitable algorithms, to future research.

841 **Behavior under various evolutionary models** Phylofactorization is hy-  
842 pothesized to work well under a punctuated-equilibrium model of evolution or  
843 jump-diffusion processes [15, 26] in which jumps are infrequent and large, such as  
844 the evolution of vertebrates to land or water. If few edges have large changes in  
845 functional ecological traits underlying the pattern of interest, phylofactorization  
846 is hypothesized to work well. Phylofactorization may also work well when in-  
847 frequent life-history traits arise or evolutionary events occur (such as ecological  
848 release) along edges and don't yield an obvious trait but instead yield a cor-  
849 related, directional evolution in descendants. Phylofactorization of mammalian  
850 body sizes yielded a scenario hypothesized to be in this category. In this case the  
851 exact trait may not have arisen along the edge identified, but a precursor trait,  
852 or a chance event such as extinctions or the emergence of novel niches, may  
853 precipitate downstream evolution of the traits underlying phylofactorization.  
854 Both aggregation and contrast functions can incorporate phylogenetic structure  
855 and edge lengths to partition the tree based on likelihoods of such evolutionary  
856 models. The sensitivity of phylofactorization to alternative models, such as the  
857 myriad Brownian motion and Ornstein-Uhlenbeck models commonly used in  
858 phylogenetic comparative methods [13, 19], remains to be tested and will likely  
859 vary depending on the particular method used.

860 **Basal/distal biases** Researchers may be interested in the distribution of fac-  
861 tored edges in the tree. If a dataset of microbial abundances in response to  
862 antibiotics is analyzed by regression-phylofactorization and results in many tips  
863 being selected, a researcher may be interested in quantifying the probability of

864 drawing a certain number of tips given  $t$  iterations of phylofactorization. Alter-  
865 natively, if several edges are drawn in close proximity researchers may wonder  
866 the probability of drawing such clustered edges under a null model of phylofac-  
867 torization. For another example, researchers may wonder if the number of im-  
868 portant functional ecological traits arose in a particular historical time window  
869 (e.g. due to some hypothesis of important evolutionary event or environmental  
870 change), and thus want to test the probability of drawing as many or more  
871 edges than observed under a null model of phylofactorization. All of these tests  
872 would require an accurate understanding of the probability of drawing edges in  
873 different locations of the tree.

874 All methods described here, save the Fisher exact test, have a bias for tips  
875 in the phylogeny (Figure 5a). Such biases affect the calibration of statistical  
876 tests of the location of phylogenetic factors, such as a test of whether/not there  
877 is an unusually large number of differentiating edges in mammalian body mass  
878 during or after the K-Pg extinction event.



**Figure 5: Graph topological bias in null data and the relative size of Voronoi cells of contrast basis elements.** The method and the null distribution determine graph-topological bias of phylofactorization, but many methods share a common source of bias. A random draw of edges does not discriminate against edges based on the relative sizes of two groups contrasted by the edge, but 16,000 replicate phylofactorizations of null data reveal that contrast-basis methods are slightly biased towards uneven splits (e.g. tips of the phylogeny). Standard Gaussian null data were used for PhyCA, F-statistics from regression on contrast basis elements ( $y_e \sim z$ ), and binomial null data was used for generalized phylofactorization (gpf) through marginally-stable aggregation. Other methods, such as Fisher’s exact test of a vector of Bernoulli random variables, have opposite biases. The tip-bias of contrast-basis analysis is amplified for marginal-stable aggregation in generalized phylofactorization, and amplified even more if the null data have residual structure from a Brownian motion diffusion along the phylogeny (Phyl-BM). The common bias when using contrast bases across a range of objective functions is related to the uneven relative sizes of Voronoi cells produced by the bases, simulated here by equation (26).

879 Phylofactorization using the contrast basis is biased towards the tips of  
 880 the tree. Some progress can be made towards understanding the source of  
 881 basal/distal biases in phylofactorization via the contrast-basis. The biases from  
 882 analyses of contrast basis coordinates,  $\mathbf{y}_e$ , stem from a common feature of the  
 883 set of  $K_t$  candidate basis elements  $\{\mathbf{v}_{C_e}\}_{e=1}^{K_t}$  considered at iteration  $t$  of phylo-  
 884 factorization. For the example of the t-test phylofactorization of a vector of  
 885 data,  $\mathbf{x}$ , the winning edge  $e^*$  is

$$e^* = \operatorname{argmax}_e |\mathbf{v}_{C_e}^T \mathbf{x}|. \quad (25)$$

886 If all basis elements have unit norm, which they do under equation (5), then  
887 each basis element being considered corresponds to a point on an  $m$ -dimensional  
888 unit hypersphere. If the data,  $\mathbf{x}$ , are drawn at random, such that no direction  
889 is favored over another, the probability that a particular edge  $e$  is the winning  
890 edge is proportional to the relative size of its Voronoi cell on the surface of the  
891 unit  $m$ -hypersphere. Thus, the basal/distal biases for contrast-basis analyses  
892 with null data assumed to be drawn from a random direction can be boiled  
893 down to calculating or computing the relative sizes of Voronoi cells. For our  
894 simulation, we estimated the size of Voronoi cells through matrix multiplication

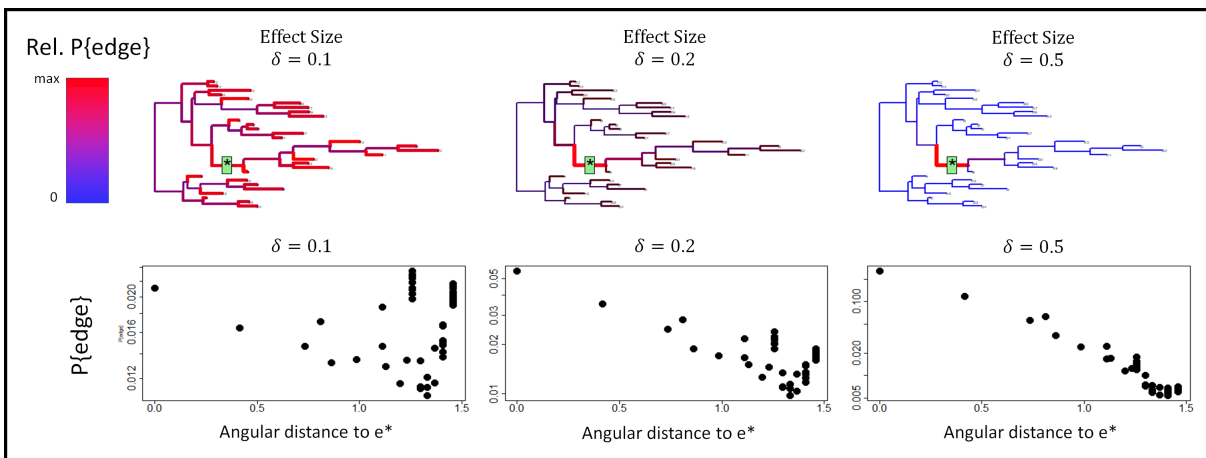
$$\mathbf{Y}_{null} = \mathbf{V}^T \mathbf{X}_{null} \quad (26)$$

895 where  $\mathbf{V}$  is a matrix whose columns  $j$  is the contrast basis elements for edge  
896  $e_j$  being considered and  $\mathbf{X}_{null}$  is the dataset simulated under the null model  
897 of choice whose columns are independent samples  $\mathbf{x}_j$ . Each column of  $\mathbf{Y}_{null}$   
898 contains the projections of a single random vector - the element of each column  
899 with the largest absolute value is the edge closest to that random vector.

900 **Graph-topology and confidence regions** As a graph-partitioning algo-  
901 rithm, phylofactorization also invites a novel description of confidence regions  
902 over the phylogeny. The graph-topology of our inferences - edges, and their  
903 proximity to other edges, both on the phylogeny and in the  $m$ -dimensional hy-  
904 persphere discussed above - can be used to refine our statements of uncertainty.  
905 95% Confidence intervals for an estimate, e.g. the sample mean, give bounds  
906 within which the true value is likely to fall 95% of the time in random draws of  
907 the estimate. Confidence regions are multi-dimensional extensions of confidence  
908 intervals. Conceptually, it's possible to make similar statements regarding phy-  
909 logenetic factors - confidence regions on a graph indicating the regions in which

910 the true, differentiating edge is likely to be.

911       Extending the concept of confidence regions to the graph-topological infer-  
912 ences from phylofactorization requires useful notions of distance and “regions” in  
913 graphs. One example of such a distance between two edges is a walking distance:  
914 the number of nodes one crosses along the geodesic path between two edges. Al-  
915 ternatively, one could define regions in terms of years or branch-lengths. The  
916 issue of confidence regions on graphs is conceptually possible and may prove im-  
917 portant for statements of certainty in phylogenetic factorization; it is an area of  
918 fruitful, future research. Defining confidence regions in phylofactorization must  
919 combine the uneven Voronoi cell sizes as well as the geometry of the contrast  
920 basis. For low effect sizes, confidence regions extend generously to edges whose  
921 contrast basis have a large relative Voronoi cell size (e.g. the tips). As the effect  
922 sizes increase, confidence regions over the graph can be described in terms of  
923 angular distances between the contrast basis elements and that of the winning  
924 edge,  $e^*$  (Figure 6).



**Figure 6: Graph-topological confidence regions for phylofactorization.** It may be possible to describe confidence regions around inferred edges by defining distances relevant to the method and graph topology. A tree with 30 species was given a fixed effect about edge  $e^*$  in their mean values as a function of meta-data  $z \sim Gsn(\pm\delta/2, 1)$ .  $7 \times 10^5$  iterations of phylofactorization were run and the relative probability of drawing each edge was visualized through both the color and width of the edge. The relationship between the angular distance of an edge's contrast basis element to that of  $e^*$  and the probability of drawing the edge suggests that for low effects, confidence intervals must incorporate a mix of tip-bias and angular distance, but larger effect sizes, in which the edge drawn is correctly in the neighborhood of  $e^*$ , the angular distance may provide a tractable method for defining confidence regions around the location of inferred phylogenetic factors.

925 **Cross-validation** How do we compare phylofactorization across datasets to  
 926 cross-validate our results? If a researcher observes a pattern in the ratio of  
 927 squamates to mammalian abundances in North America, say a decrease in the  
 928 ratio of lizard and snake to mammal abundance with increasing altitude, they  
 929 may wish to cross-validate their findings in other regions, including regions with  
 930 few or none of the same species in the original study. Researchers replicating  
 931 the study in Australia and New Zealand would have to grapple with whether  
 932 or not to include monotremes in their grouping of “mammals” and whether or  
 933 not to include the tuatara, a close relative of squamates, in their grouping of  
 934 “squamates” - such branches were basal to the squamate & mammalian clades  
 935 contrasted in the hypothetical North American study.

936 Phylofactorization formalizes the issues arising with such phylogenetic cross-  
937 validation (Figure 7). If all species in the training/testing datasets can be  
938 located on a universal phylogeny, phylofactorization of a training set of species  
939 and data identifies edges or links of edges in the training phylogeny which are  
940 guaranteed to correspond to edges or links of edges in the universal phylogeny.  
941 The testing set of species may introduce new edges to the phylogeny which  
942 interrupt the links of edges in the universal phylogeny along which training  
943 contrasts were conducted. In the example above, the tuatara and monotremes  
944 all interrupt the link of edges separating North American mammals from North  
945 American reptiles on the universal phylogeny.

946 Robust cross-validation for phylofactorization requires directly addressing  
947 the issues arising from the interruptions of edges produced by novel species.  
948 Interruptions may be either ignored, or used to refine the inference. Returning  
949 to the previous example, one can use the presence of monotremes and tuatara to  
950 refine the definition of North American mammals to mean “all mammals” and  
951 “all placental and marsupial mammals”, and likewise one can optionally refine  
952 the definition of “squamates” to the broader “Lepidosauria” clade.



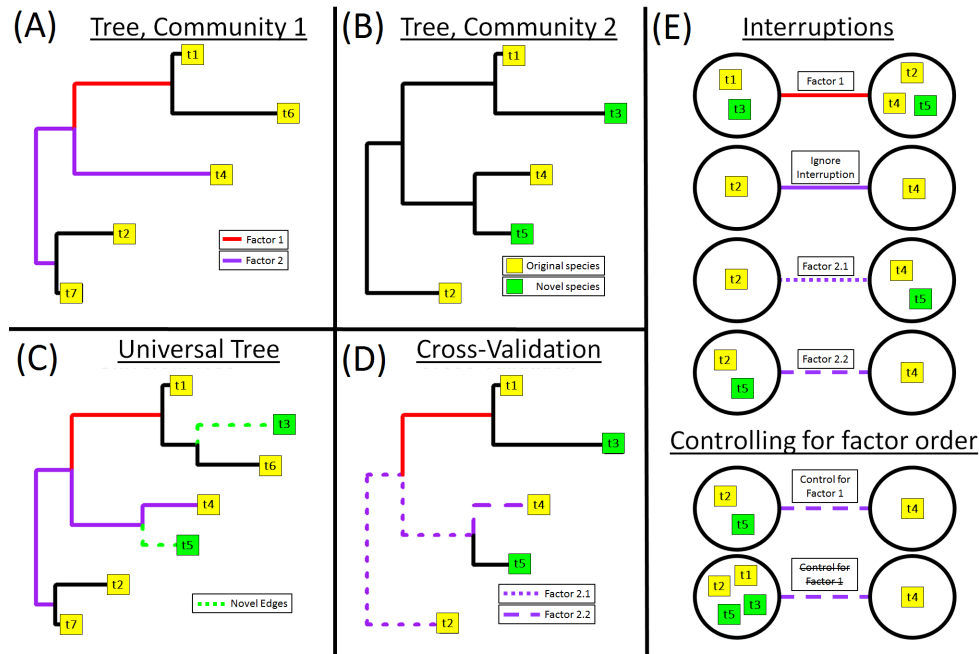
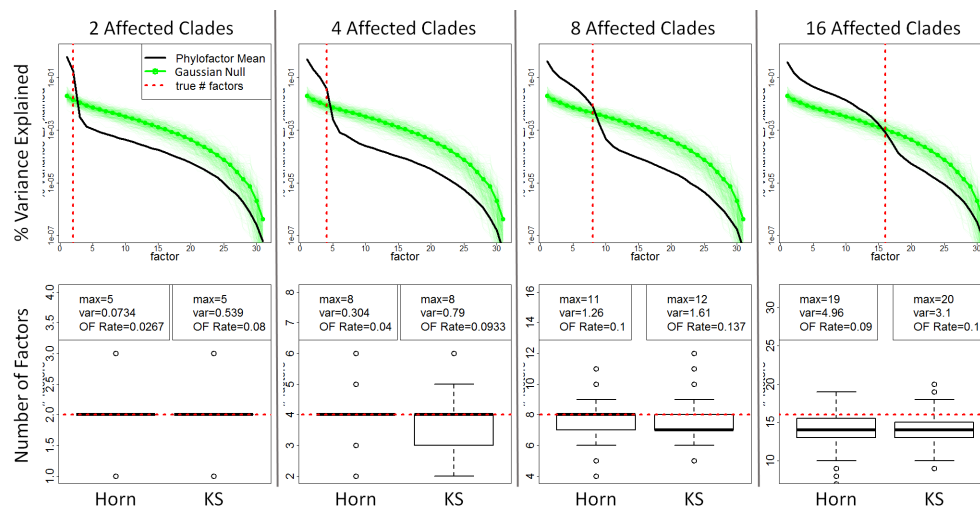


Figure 7: Graph-topological considerations with cross-validation. (A) The training community has 5 species (yellow boxes) split into two factors. The second factor forms a partition separating  $t_4$  from  $\{t_2, t_7\}$ . The second factor does not correspond to a single edge, but instead a chain of two edges. (B) A second, testing community is missing species  $t_6$  and  $t_7$  and contains novel species  $t_3$  and  $t_5$  (green boxes). (C) All factors can be mapped to chains of edges on a universal phylogeny. Novel species “interrupt” edges in the original tree; cross-validation requires deciding what to do with novel species and interrupted edges. Species  $t_3$  does not interrupt a factored edge, and so  $t_3$  can be reliably grouped with  $t_1$  in factor 1. However, species  $t_5$  interrupts one of the edges in the edge-path of factor 2. (D-E) Interruptions can be ignored, or they can be used to refine the location of important edges (illustrated in Factor 2.1 and Factor 2.2). Another topological and statistical question is whether/not to control for factor order. For instance, controlling for factor order with Factor 2.2 would partition  $t_4$  from  $\{t_2, t_5\}$ . Not controlling for factor order would partition  $t_4$  from  $\{t_1, t_2, t_3, t_5\}$ .

953 **Stopping Criteria** With appropriately defined aggregation and contrast func-  
 954 tions, phylofactorization can be iterated until every species is split and the graph  
 955 is fully partitioned. However, such full partitioning is rarely desired. Rather,  
 956 researchers may often want a minimal set of partitions for prioritization of find-  
 957 ings, simplicity of summarizing the data, and certainty in the inferences made.  
 958 There are two broad options for stopping phylofactorization: a stopping func-

959 tion demonstrated to be sufficiently conservative, and null simulations allowing  
960 quantile-based cutoffs (e.g. stop phylofactorization when the percent variance  
961 explained by PhyCA is within the 95% quantile of null phylofactorizations).  
962 Null simulations may allow statistical statements stemming from a clear null  
963 model, but stopping criteria can be far more computationally efficient and can  
964 be constructed to be conservative.

965 Washburne et al. [51] proposed a stopping criterion for regression-phylofactorization  
966 which extends to all methods of phylofactorization using an objective function  
967 that is a test-statistic whose null-distribution is known. The original stopping  
968 criterion is based on the fact that, if the null hypothesis is true, the distribu-  
969 tion of P-values from multiple hypothesis tests is uniform. Phylofactorization  
970 performs multiple hypothesis tests at each iteration. At each iteration, one  
971 can perform a one-tailed KS test on the uniformity of the distribution of the  
972 P-values from the test-statistics on each edge; if the KS-test is non-significant,  
973 stop phylofactorization. KS-test stopping criteria can conservatively stop simu-  
974 lations at the appropriate number of factors when there is a discrete subset of  
975 edges with effects. Such a method performs similarly to Horn's stopping crite-  
976 rion for factor analysis [21], whereby one stops factorization when the scree plot  
977 from the data crosses that expected from null data (figure 8). It's also possible  
978 to first use a stopping criterion and subsequently run null simulations to under-  
979 stand the likelihood of observed results under a null model of the researcher's  
980 choice (figure 8).



**Figure 8: Null simulations and stopping criteria.** A challenge of phylofactorization is determining the number of factors,  $K$ , to include in an analysis. Null simulations can be used to construct quantile-based cutoffs such as those in Horn’s parallel analysis from factor analysis. Stopping criteria aim to stop the computationally intensive iteration of phylofactorization without using null simulations but instead using features available during phylofactorization of the observed data. Abundances of  $m = 32$  species across  $n = 10$  samples were simulated as i.i.d. standard Gaussian random variables. To simulate effects, a set of  $u$  clades were associated with environmental meta-data,  $\mathbf{z}$ , where  $z_j \stackrel{i.i.d.}{\sim} N(0, 1)$ . Regression-phylofactorization on the contrast-basis scores  $y_e$  was performed on 300 datasets for each  $u \in \{2, 4, 8, 16\}$  and on data with and without effects, with objective function being the variance explained by regression  $y_e \sim \mathbf{z}$ . (**top row**) The percent explained variance (EV) decreases with factor,  $k$ , and the mean EV curve for data with  $u$  affected clades intersects the mean EV curve for null data near where  $k = u$ , motivating a stopping criterion (Horn) based on phylofactorization of null datasets to be evaluated and compared to the KS-based stopping criterion proposed by [51]. (**bottom row**) The Horn stopping criterion has a lower over-factorization (OF) rate than the standard KS stopping criterion (where OF rate is the fraction of the 300 phylofactorizations of data with simulated effects in which  $K > u$ ). Both criteria can be modified to be made more conservative (e.g. the P-value threshold for the KS stopping criterion can be lowered, or the Horn criterion can be modified to stop the simulation at different quantiles of null simulations). The KS stopping criterion, however, is far less computationally intensive for large datasets as it requires running phylofactorization only once. Null simulations, however, can allow inferential statistical statements regarding the null distribution of test statistics in phylofactorization.

**Calibrating Statistical Tests for  $\omega_{e^*}$**  Often, the objective function for the winning edge in phylofactorization,  $\omega_{e^*}$ , corresponds directly to a common test-statistic such as an  $F$ -statistic. Applying a standard test for the resul-

tant test-statistic, however, will lead to a high false-positive rate and an over-estimation of the significance of an effect, as the statistic was drawn as the best of many. Even when using a test-statistic not equal to the objective function, researchers should be cautious of dependence between their test-statistic and the objective function as a possible source of high false-positive rates. Two non-exhaustive avenues for calibrating, or making conservative, statistical tests of  $\omega_{e^*}$  are multiple-comparisons corrections to control a family-wise error rate (or other multiple-hypothesis-test methods) or conservative bounds on the distribution of the maximum of many independent, identically distributed statistics. For example, if each edge of  $K_t$  edges resulted in an independent  $F$ -statistic,  $F_e$ , then the distribution of the maximum  $F$ -statistics,  $F_{e^*}$ , is

$$\begin{aligned} P \{F_{e^*} > F\} &= P \{F_{e_1} > F \cap F_{e_2} > F \cap \dots \cap F_{e_K}\} \\ &= P \{F_e > F\}^{K_t}. \end{aligned} \tag{27}$$

981 Such an approximation may be used to yield conservative estimates, but the  
982  $F$ -statistics are not independent and thus more nuanced analyses are needed for  
983 well-calibrated statistical tests.

984 **Summary of limitations** Phylofactorization can be a reliable statistical tool  
985 with a careful understanding of the statistical challenges inherent in the method  
986 and shared with related methods such as graph-partitioning, greedy algorithms,  
987 factor analysis, and the use of a constrained, biased basis for matrix factoriza-  
988 tion. Phylofactorization can first and easiest be an exploratory tool, but all  
989 exploratory tools can be made inferential with suitable understanding of their  
990 behavior under an appropriate null model. For example, principal components  
991 analysis was and still is primarily an exploratory tool, but the discovery of the  
992 Marcenko-Pastur distribution [30] has improved the calibration of statistical

993 tests on principal components for standardized, mean-centered data. Improved  
994 understanding of how uncertainties in phylogenetic inference translate to uncer-  
995 tainties in phylofactorization, conservative stopping criteria, null distributions  
996 of test-statistics for winning edges, propagation of error and stochastic sampling  
997 algorithms to avoid deterministic ruts, graph-topological biases and confidence  
998 regions on a graph, can all improve the reliability of phylofactorization as an  
999 inferential tool.

1000 While phylofactorization was built with an evolutionary model of punctuated  
1001 equilibria in mind, it may also work well under other evolutionary models such  
1002 as niches leading to correlated evolution of descendants. There may also be  
1003 many evolutionary models under which phylofactorization does not perform  
1004 well. For instance the graph-topological biases of PhyCA are increased under  
1005 a Brownian motion model of evolution. All statistical tools operate well under  
1006 appropriate assumptions, and understanding the assumptions, as well as the  
1007 known limitations, are necessary for responsible and academically fruitful use  
1008 of statistical tools like phylofactorization.

## 1009 Discussion

1010 Functional ecological traits underlie many observed patterns in ecology, includ-  
1011 ing species abundances, presence/absence of species, and responses of traits  
1012 or abundances to experimental conditions or along environmental gradients.  
1013 Where the ecological pattern of interest is associated with heritable traits, the  
1014 phylogeny provides a scaffold for the discovery of functional groupings of clades  
1015 underlying the ecological pattern of interest. Traits arise along edges, and con-  
1016 trasting taxa on opposing sides of an edge allows one to uncover edges best  
1017 separating species with different functional associations or links to the ecologi-

1018 cal pattern. By noting that each edge partitions the phylogeny into two disjoint  
1019 sets of species, by generalizing the operations of “grouping” - aggregating and  
1020 contrasting disjoint sets of species - and by defining the objective function of  
1021 interest (the pattern), we have proposed a universal method for identifying rel-  
1022 evant phylogenetic scales in arbitrary datasets.

1023     Phylofactorization is a graph-partitioning algorithm intended to separate the  
1024 phylogeny into binned phylogenetic units with a combination of high within-  
1025 group similarity and high between-group differences. Two-sample tests are a  
1026 natural method for making such partitions in vectors of data, although such  
1027 partitions can also be made with ancestral state reconstruction. The idea be-  
1028 hind two-sample tests, however, can be extended to larger, real-valued datasets  
1029 by analysis of a contrast basis. Objective functions for choosing the appropriate  
1030 contrast basis include maximizing variance - a phylogenetic analog of principal  
1031 components analysis - maximizing explained variance from regression, maximiz-  
1032 ing F-statistics from regression, and more. We’ve illustrated that two-sample  
1033 tests can partition a dataset of mammalian body mass into groups with very  
1034 different average body masses. We’ve demonstrated that maximizing variance  
1035 of data projected onto a contrast basis can identify major clades of bacteria in  
1036 human feces that have been known, at a coarser resolution, to be highly variable  
1037 and determined that one of the top phylogenetic factors in the American Gut  
1038 dataset is a clade of Gammaproteobacteria associated with IBD and used re-  
1039 cently in an effort to diagnose patients with Crohn’s disease. We’ve shown that  
1040 such analysis of contrast bases can couple with non-linear regression, and within  
1041 minutes of analysis on a laptop found a natural way put over 3,000 species into 5  
1042 binned phylogenetic units, sort them along an axis of the dominant explanatory  
1043 variable, and produce a simplified story of the dominant phylogenetic scales of  
1044 explained variation in Central Park soil. One can also perform phylofactoriza-

1045 tion when doing maximum-likelihood regression of exponential family random  
1046 variables. Factor contrasts are a natural, built-in method for extending the con-  
1047 cepts of aggregation and contrast to generalized linear models and generalized  
1048 additive models. One can either perform the factor contrasts on the raw data,  
1049 or, for many exponential family random variables, one can aggregate the data  
1050 from each group to a marginally stable distribution for more computationally  
1051 efficient and powerful factor contrasts. These methods can be implemented in  
1052 the R package “phylofactor”, and scripts for running each analysis are available  
1053 in the supplemental materials.

1054 As with any method, there are limitations to be aware of. First, the gen-  
1055 eral problem of separating species into  $k$  bins that maximize some global ob-  
1056 jective function of high within-group similarity and high between-group dif-  
1057 ferences is NP hard. Second, like any greedy algorithm, purely deterministic  
1058 phylofactorization may fall into ruts and errors in one step might propagate  
1059 into downstream inferences. Third, the null distribution of test-statistics re-  
1060 sulting from phylofactorization is not known and is biased towards extreme  
1061 values due to the algorithm choosing species which maximize objective func-  
1062 tions. We propose null simulations, conservative stopping functions, and/or  
1063 extremely stringent multiple comparisons corrections for users attempting to  
1064 make inferences through phylofactorization while maintaining a certain family-  
1065 wise error or false-discovery rate. When the objective function being maximized  
1066 is also a test-statistic with a well-defined null distribution, one-sided KS-tests  
1067 of the P-values from the test-statistic can serve as a computationally efficient  
1068 and conservative stopping function. Fourth, the contrast basis is biased towards  
1069 the tips due, we hypothesize, to the unequal relative sizes of the Voronoi cells  
1070 of the contrast basis elements in the unit hypersphere in which they lie. Such  
1071 topological bias is exacerbated by data produced through Brownian motion dif-

1072 fusion along the phylogeny, and reversed for Fisher's exact test of a vector of  
1073 binary trait data. Understanding the graph-topology of errors can assist the  
1074 description of graph-topological confidence regions for each inference. Finally,  
1075 phylofactorization formalizes the logic of cross-validating ecological comparisons  
1076 even when the training and testing sets of species are completely disjoint, but  
1077 such cross-validation must address the issues of interrupting edges and whether  
1078 or not to control for factor order in cross-validation. Many of these limita-  
1079 tions may be resolved with future work, allowing the general algorithm and its  
1080 common implementations to become a reliable, well-calibrated inferential tool.

1081 Phylofactorization is its ability to objectively identify phylogenetic scales for  
1082 ecological big-data and instantly produce avenues for future research to elucidate  
1083 mechanisms that underlie patterns in big-data. By iteratively identifying clades,  
1084 phylofactorization provides a sequence of low-rank approximations of a dataset,  
1085 such as that visualized in figure 3c, which correspond to groups of species with  
1086 a shared evolutionary history. What traits characterize the Chloracidobacteria  
1087 which don't like acidic soils? What traits characterize the monophyletic clade  
1088 of Gammaproteobacteria that are associated with IBD? What traits underlie  
1089 the Clostridia/Erysipelotrichi being such variable species in the American gut?  
1090 Phylofactorization has identified clades from big-data, and produced questions  
1091 that can be subsequently answered by comparative genomics, microbial physio-  
1092 logical studies, and other clear avenues of future research.

1093 **Relation to other phylogenetic methods** Phylofactorization is proposed  
1094 amidst an explosion of literature in phylogenetic comparative methods and vari-  
1095 ous other phylogenetic methods for analyzing ecological datasets [29, 38, 14], and  
1096 some careful thinking is beneficial to clarify the distinctions between the myriad  
1097 methods. First, phylogenetically independent contrasts [13] produces variables  
1098 corresponding to contrasts of descendants from each node, whereas phylofactor-



1099 ization uses contrasts of species separated by an edge, picks out the best edge,  
1100 splits the tree, and repeats. Phylogenetic generalized least squares [16] aims to  
1101 control for residual structure in the response variable expected under a model of  
1102 trait evolution, and is thus used when performing regression on a trait, whereas  
1103 phylofactorization aims to partition observed trait values or abundances into  
1104 groups with different means or associations with meta-data along edges along  
1105 which differences most likely arose. Thus, while methods of phylogenetic sig-  
1106 nal, such as Pagel's  $\lambda$  [35] or Blomberg's  $\kappa$  [5], summarize global patterns of  
1107 phylogenetic signal by parameterizing the extent to which a particular model of  
1108 evolution can be assumed to underlie the residual structure of observed traits  
1109 (often for downstream use in PGLS), phylofactorization iteratively identifies  
1110 precise locations of putative changes and precise locations partitioning phyloge-  
1111 netic signal or structure. Phylofactorization can be implemented by a contrast  
1112 of ancestral state reconstructions of nodes separated by edges, for example by  
1113 looking for edges with nodes whose reconstructed ancestral states are most dif-  
1114 ferent, but is limited by disallowing the descendant clade of an edge to impact  
1115 the ancestral state of the edge's basal node - a proper non-overlapping contrast  
1116 would separate the groups of species being used to reconstruct each node, and  
1117 thus phylofactorization can be implemented with ancestral state reconstruction  
1118 under the assumption of time-reversible evolutionary models. Phylofactoriza-  
1119 tion develops a set of variables and an orthonormal basis to describe ecological  
1120 data, but limits itself to bases interpretable as non-overlapping contrasts along  
1121 edges; eigenvectors of phylogenetic distances matrices or covariance matrices  
1122 under diffusion models of traits [35], are not encompassed in phylofactorization  
1123 as they do not construct non-overlapping contrasts along edges. Such eigenvec-  
1124 tor methods construct quantities whose evolutionary interpretation is less clear.  
1125 Unlike many modern methods for re-defining distances, such as UniFrac dis-

1126 tances [29] or phylogenetically-defined inner products [38], phylofactorization is  
1127 principally about discovering phylogenetically-interpretable directions - vectors  
1128 which characterize primary axes of variation in the community and represented  
1129 through the contrast basis, a multilevel-factor developed from stepwise selection  
1130 of factor contrasts, or a basis made of aggregations of the binned phylogenetic  
1131 units.

1132 **Phylofactorization as a species concept** There is great debate about what  
1133 constitutes a species in microbes, let alone all organisms. There is a need for  
1134 objectivity and universality in the definition of “species” and other units in  
1135 ecology and evolution. The biological species concept is complicated by asexual  
1136 reproduction. Genetic species concepts are limited by the subjectivity of a  
1137 sequence-similarity cutoff, such as the 97% sequence similarity commonly used  
1138 in defining operational taxonomic units or OTUs, which is additionally compli-  
1139 cated by the fact that functional ecological similarity may not be uniform at  
1140 a given sequence-similarity cutoff. Ecological species concepts are often useful  
1141 once researchers have a clear sense of the functional ecological groups, but it is  
1142 difficult to objectively define what constitutes an important functional ecologi-  
1143 cal group, especially for taxa whose life histories are unknown. Species concepts  
1144 coarse-grain the diversity of life in a way that connects our coarse-grained units  
1145 to biological, ecological, and evolutionary theory. To that end, phylofactoriza-  
1146 tion can be seen as defining a species concept.

1147 Species concepts are fundamental to biology as they partition the diversity of  
1148 life into units between which we define ecological interaction and within which  
1149 we define evolution and natural selection. At the heart of species concepts are  
1150 the operations fundamental to phylofactorization: aggregation, contrast, and an  
1151 objective function. Species are aggregations of finer units of diversity: individual  
1152 subpopulations of individual organisms and their individual cells and the cells’

1153 individual genes are all aggregated to define a “population”. Aggregation in a  
1154 species concept defines a clear partition for later “within-species” contrasts (evo-  
1155 lution) and “between-species” interactions (competition & ecological interactions  
1156 among populations or aggregates of species). A species concept must meaning-  
1157 fully contrast the units of diversity - the biological species concept contrasts  
1158 species based on reproductive isolation, the genetic species concept contrasts  
1159 species based on genetic dissimilarity, and ecological species concepts contrast  
1160 species based on distinct functional ecological traits. The objective function in  
1161 phylofactorization is the theoretical placeholder for a researcher’s “meaningful  
1162 contrast”. The units for aggregation and contrast must be done in light of some  
1163 objective, such as a common fitness or pattern of relative abundance within units  
1164 over time, space, across environmental gradients and/or between experimental  
1165 treatments. A full theoretical consideration of phylofactorization as a species  
1166 concept, as it relates to evolutionary and ecological theory, is saved for future  
1167 research. For the time being, we note that phylofactorization partitions diver-  
1168 sity and yields notions of a “species” which can be aggregated and contrasted  
1169 with other “species”.

1170 Phylofactorization is a flexible species concept, a hybrid of the phylogeny-  
1171 based phylogenetic species concept [34] and the character-based ecological species  
1172 concept [48]. After  $k$  iterations of phylofactorization, the phylogeny is par-  
1173 titioned into  $k + 1$  bins of species referred to as “binned phylogenetic units”  
1174 (BPUs). BPUs are aggregations of the phylogeny which, up to a certain level  
1175 of partitioning, are more similar to one-another with respect to the aggrega-  
1176 tion, contrast and objective function, than they are to other groups. BPUs are  
1177 a coarse-grained way to cluster entities into “units” of organization with com-  
1178 mon behavior with respect to the ecological pattern defined in the objective  
1179 function. Phylofactorization defines functional groups based on phylogenetic

1180 partitions and a similar association with some ecological pattern of interest.  
1181 Consequently, phylofactorization can be seen as an ecological species concept  
1182 constrained to a phylogenetic scaffold. Whereas the phylogenetic species con-  
1183 cept is character-based and pattern oriented, phylofactorization is pattern-based  
1184 and phylogenetically-constrained. A textbook example of a phylofactorization-  
1185 derived species are “land-dwelling tetrapods”, a group which can be obtained  
1186 objectively through phylofactorization and which defines a scale for aggregating  
1187 and summarizing the pattern of vertebrate species-abundances on Earth.

1188 Phylofactorization permits optional fine-graining and coarse-graining of our  
1189 patterns of diversity. Phylofactorization provides an algorithm for identifying  
1190 relevant units, and those units may be at different taxonomic or phylogenetic  
1191 depths but will have shared evolutionary history and similar associations with  
1192 the ecological pattern of interest. For microorganisms, for which the biological  
1193 species concept doesn’t apply, the genetic species concept appears too detached  
1194 from ecology, and the ecological species concept is unavailable due to lack of  
1195 life history detail, phylofactorization serves as a way to organize diversity for  
1196 focused between-species interactions and within-species comparisons.

1197 **R package: phylofactor** An R package is in development and, prior to its  
1198 stable release to CRAN, publicly available at <https://github.com/reptalex/phylofactor>.  
1199 The R package contains detailed help functions and supports flexible definition  
1200 of two-sample tests (the function `twoSampleFactor`), contrast-basis analyses with  
1201 the function `PhyloFactor`, and generalized phylofactorization of exponential fam-  
1202 ily random variables with the function `gpf`. Phylofactorization is highly paral-  
1203 lelizable, and the R package functions have built-in parallelization. The R pack-  
1204 age in development also works with phylogenies containing polytomies, allowing  
1205 researchers to collapse clades with low bootstrap support to make more robust  
1206 inferences. The output from each of the three phylofactorization functions is

1207 a “phylofactor” object one can input into various functions which summarize,  
1208 plot, cross-validate, run null simulations, and parse out the information from  
1209 phylofactorization. Future releases aim to simplify this into a single function:  
1210 phylofactor. Researchers are invited to contact the corresponding author for  
1211 assistance with the package and how to produce their own customized phylo-  
1212 factorizations - such feedback will be essential for a user-friendly stable release  
1213 to CRAN.

1214       Until then, the supplemental information contains the data and scripts used  
1215 for all analyses done in this manuscript in an effort to accelerate method devel-  
1216 opment in this field.

1217 **“Everything makes sense in light of evolution”** Phylogenetic factoriza-  
1218 tion is a new paradigm for analyzing a large class of biological data. Ecological  
1219 big-data, as Thomas Dhobzansky noted about biology in general, makes sense  
1220 “in light of evolution”. Phylofactorization extends a broad category of data anal-  
1221 yses - two sample tests, generalized linear modelling, factor analysis and PCA,  
1222 and analysis of spatial and temporal patterns - to incorporate a natural set of  
1223 variables and operations defined by the phylogeny. Phylofactorization localizes  
1224 inferences in big data to particular edges or chains of edges on the phylogeny  
1225 and, in so doing, can accelerate our understanding of the phylogenetic scales  
1226 underlying ecological patterns of interest. The problem of pattern and scale is  
1227 central to biology, and phylofactorization uses the pattern to objectively uncover  
1228 the relevant phylogenetic scales in ecological datasets.

## 1229 Acknowledgments

1230 This work is published in loving memory of Diana Nemergut. This research was  
1231 developed with funding from the Defense Advanced Research Projects Agency  
1232 (DARPA; D16AP00113).

1233 Table of mathematical notation

Term	Description	Terms	Description
$A(\cdot)$	Aggregation operator	$r, s$	Numbers of species in groups $R, S$ respectively
$C(\cdot, \cdot)$	Contrast operator	$s(\cdot)$	Smoothing spline notation for term in generalized additive model
$\mathcal{F}(\theta)$	Distribution parameterized by $\theta$	$t$	Iteration of phylofactorization
$F_e$	F-statistic for edge $e$	$x_{i,j}$	The $i, j$ th element of data matrix $\mathbf{X}$
$K_t$	Number of edges considered in iteration $t$ of phylofactorization	$x_{R,j}$	Aggregate, $A(\mathbf{x}_j)$ of group $R$ for sample $j$ . If $j$ is missing then sample is arbitrary.
$N$	Size of a binomial random variable	$x_{S,j}$	See $x_{R,j}$ above.
$Q$	A group $Q = R \cup S$ aggregated at a current or previous iteration	$x_i$	A random variable (assumed to be a single species $i$ for arbitrary sample)
$R, S$	Two groups contrasted containing $r$ and $s$ species, respectively	$[x]_{i,j}$	$i, j$ th entry of data matrix, $\mathbf{X}$
$U, B, P$	Meta-data subsets for phylofactorization	$z_i$	Column of meta-data matrix, $\mathbf{Z}$
$\mathcal{T}$	Phylogenetic tree	$v_{Q,i}$	$i$ th element of aggregation basis for set $Q$
$\mathbf{V}$	$m \times K_t$ matrix of contrast basis elements considered at iteration $t$	$v_{C_{R S}}$	Contrast vector splitting groups $R$ and $S$
$\mathbf{X}$	$m \times n$ data matrix used for phylofactorization	$v_{C_e}$	Contrast vector for edge $e$ (which splits sub-tree into two disjoint groups)
$\mathbf{Y}$	$K \times n$ matrix of component scores, one for each edge considered	$\mathbf{x}_{R,j}$	$r$ -vector containing only the species in group $R$ for sample $j$
$\mathbf{Z}$	$n \times p$ matrix of meta-data used in regression-phylofactorization	$\mathbf{x}_{S,j}$	See $\mathbf{x}_{R,j}$ above.
$a$	Coefficient in aggregation vector	$\mathbf{x}$	$m$ -vector of species' data for an arbitrary sample
$b, c$	Coefficients in a contrast vector	$\bar{\mathbf{x}}$	Sample mean of vector $\mathbf{x}$
$e_k$	Edge $k$	$\mathbf{y}_e$	$n$ -vector of component scores for edge $e$
$e^*$	Winning edge	$\mathbf{z}_k$	Vector of meta-data of type $k$ .
$e_t^*$	Winning edge at iteration $t$	$\beta_i$	Coefficients for linear model
$f(\cdot)$	Transformation in generalized $f$ -mean	$\eta$	Natural parameter for exponential-family random variable
$g$	Factor containing two levels, $\{R, S\}$	$\kappa$	Scale parameter for Gamma distribution
$i, j, k, l$	Indexes. Often, $i$ is the index for species and $j$ for samples.	$\pi$	Number of failures parameter for Negative Binomial distribution.
$m$	Number of species	$\rho$	Probability of success for Bernoulli, Binomial, Negative Binomial distributions
$n$	Number of samples	$\sigma$	Standard deviation for Gaussian random variable
$p$	Number of meta-data types for each sample	$\theta$	Arbitrary parameters for probability distribution
$q$	Number of pure aggregates in a basis for $\mathbb{R}^m$		

## 1234 References

- 1235 [1] J. AITCHISON, *The statistical analysis of compositional data*, (1986).
- 1236 [2] J. ALROY, *Cope's rule and the dynamics of body mass evolution in north*  
1237 *american fossil mammals*, *Science*, 280 (1998), pp. 731–734.
- 1238 [3] ———, *The fossil record of north american mammals: evidence for a paleo-*  
1239 *ocene evolutionary radiation*, *Systematic Biology*, 48 (1999), pp. 107–118.
- 1240 [4] J. BAKER, A. MEADE, M. PAGEL, AND C. VENDITTI, *Adaptive evolution*  
1241 *toward larger size in mammals*, *Proceedings of the National Academy of*  
1242 *Sciences*, 112 (2015), pp. 5093–5098.
- 1243 [5] S. P. BLOMBERG, T. GARLAND JR, A. R. IVES, AND B. CRESPI, *Test-*  
1244 *ing for phylogenetic signal in comparative data: behavioral traits are more*  
1245 *labile*, *Evolution*, 57 (2003), pp. 717–745.
- 1246 [6] A. BULUÇ, H. MEYERHENKE, I. SAFRO, P. SANDERS, AND C. SCHULZ,  
1247 *Recent advances in graph partitioning*, in *Algorithm Engineering*, Springer,  
1248 2016, pp. 117–158.
- 1249 [7] J. C. CLEMENTE, L. K. URSELL, L. W. PARFREY, AND R. KNIGHT, *The*  
1250 *impact of the gut microbiota on human health: an integrative view*, *Cell*,  
1251 148 (2012), pp. 1258–1270.
- 1252 [8] T. Z. DESANTIS, P. HUGENHOLTZ, N. LARSEN, M. ROJAS, E. L.  
1253 BRODIE, K. KELLER, T. HUBER, D. DALEVI, P. HU, AND G. L. ANDER-  
1254 SEN, *Greengenes, a chimera-checked 16s rrna gene database and workbench*  
1255 *compatible with arb*, *Applied and environmental microbiology*, 72 (2006),  
1256 pp. 5069–5072.



- 1257 [9] J. J. EGOZCUE AND V. PAWLOWSKY-GLAHN, *Groups of parts and their*  
1258 *balances in compositional data analysis*, *Mathematical Geology*, 37 (2005),  
1259 pp. 795–828.
- 1260 [10] J. J. EGOZCUE, V. PAWLOWSKY-GLAHN, G. MATEU-FIGUERAS, AND  
1261 C. BARCELO-VIDAL, *Isometric logratio transformations for compositional*  
1262 *data analysis*, *Mathematical Geology*, 35 (2003), pp. 279–300.
- 1263 [11] C. E. FARRIOR, R. DYBZINSKI, S. A. LEVIN, AND S. W. PACALA, *Com-*  
1264 *petition for water and light in closed-canopy forests: a tractable model of*  
1265 *carbon allocation with implications for carbon sinks*, *The American Natu-*  
1266 *ralist*, 181 (2013), pp. 314–330.
- 1267 [12] J. FELSENSTEIN, *Confidence limits on phylogenies: an approach using the*  
1268 *bootstrap*, *Evolution*, (1985), pp. 783–791.
- 1269 [13] ———, *Phylogenies and the comparative method*, *The American Naturalist*,  
1270 125 (1985), pp. 1–15.
- 1271 [14] L. Z. GARAMSZEGI, *Modern phylogenetic comparative methods and their*  
1272 *application in evolutionary biology*, *Concepts and Practice*. London, UK:  
1273 Springer, (2014).
- 1274 [15] N. E.-S. J. GOULD, *Punctuated equilibria: an alternative to phyletic grad-*  
1275 *ualism*, (1972).
- 1276 [16] A. GRAFEN, *The phylogenetic regression*, *Philosophical Transactions of the*  
1277 *Royal Society of London. Series B, Biological Sciences*, 326 (1989), pp. 119–  
1278 157.
- 1279 [17] C. H. GRAHAM, D. STORCH, AND A. MACHAC, *Phylogenetic scale in*  
1280 *ecology and evolution*, *bioRxiv*, (2017).

- 1281 [18] B. G. HALL AND M. BARLOW, *Evolution of the serine  $\beta$ -lactamases: past,*  
1282 *present and future*, Drug Resistance Updates, 7 (2004), pp. 111–123.
- 1283 [19] T. F. HANSEN, *Stabilizing selection and the comparative analysis of adap-*  
1284 *tation*, Evolution, 51 (1997), pp. 1341–1351.
- 1285 [20] W. K. HASTINGS, *Monte carlo sampling methods using markov chains and*  
1286 *their applications*, Biometrika, 57 (1970), pp. 97–109.
- 1287 [21] J. L. HORN, *A rationale and test for the number of factors in factor anal-*  
1288 *ysis*, Psychometrika, 30 (1965), pp. 179–185.
- 1289 [22] S. P. HUBBELL, *The Unified Neutral Theory of Biodiversity and Bio-*  
1290 *geography (MPB-32)*, Princeton University Press, 2001.
- 1291 [23] M. JERRUM AND G. B. SORKIN, *The metropolis algorithm for graph bi-*  
1292 *section*, Discrete Applied Mathematics, 82 (1998), pp. 155–175.
- 1293 [24] K. E. JONES, J. BIELBY, M. CARDILLO, S. A. FRITZ, J. O’DELL,  
1294 C. D. L. ORME, K. SAFI, W. SECHREST, E. H. BOAKES, C. CAR-  
1295 BONE, ET AL., *Pantheria: a species-level database of life history, ecology,*  
1296 *and geography of extant and recently extinct mammals*, Ecology, 90 (2009),  
1297 pp. 2648–2648.
- 1298 [25] Y. KATZ, K. TUNSTRØM, C. C. IOANNOU, C. HUEPE, AND I. D. COUZIN,  
1299 *Inferring the structure and dynamics of interactions in schooling fish*, Pro-  
1300 ceedings of the National Academy of Sciences, 108 (2011), pp. 18720–18725.
- 1301 [26] M. J. LANDIS, J. G. SCHRAIBER, AND M. LIANG, *Phylogenetic analysis*  
1302 *using lévy processes: finding jumps in the evolution of continuous traits*,  
1303 Systematic biology, 62 (2012), pp. 193–204.
- 1304 [27] S. A. LEVIN, *The problem of pattern and scale in ecology: the robert h.*  
1305 *macarthur award lecture*, Ecology, 73 (1992), pp. 1943–1967.

- 1306 [28] R. E. LEY, P. J. TURNBAUGH, S. KLEIN, AND J. I. GORDON, *Microbial*  
1307 *ecology: human gut microbes associated with obesity*, *Nature*, 444 (2006),  
1308 pp. 1022–1023.
- 1309 [29] C. LOZUPONE AND R. KNIGHT, *Unifrac: a new phylogenetic method for*  
1310 *comparing microbial communities*, *Applied and environmental microbiol-*  
1311 *ogy*, 71 (2005), pp. 8228–8235.
- 1312 [30] V. A. MARČENKO AND L. A. PASTUR, *Distribution of eigenvalues for*  
1313 *some sets of random matrices*, *Mathematics of the USSR-Sbornik*, 1 (1967),  
1314 p. 457.
- 1315 [31] D. MARIAT, O. FIRMESSE, F. LEVENEZ, V. GUIMARÃES, H. SOKOL,  
1316 J. DORÉ, G. CORTHER, AND J. FURET, *The firmicutes/bacteroidetes ra-*  
1317 *tio of the human microbiota changes with age*, *BMC microbiology*, 9 (2009),  
1318 p. 123.
- 1319 [32] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H.  
1320 TELLER, AND E. TELLER, *Equation of state calculations by fast computing*  
1321 *machines*, *The journal of chemical physics*, 21 (1953), pp. 1087–1092.
- 1322 [33] F. MICHONNEAU, J. W. BROWN, AND D. J. WINTER, *rotl: an r package to*  
1323 *interact with the open tree of life data*, *Methods in Ecology and Evolution*,  
1324 7 (2016), pp. 1476–1481.
- 1325 [34] K. C. NIXON AND Q. D. WHEELER, *An amplification of the phylogenetic*  
1326 *species concept*, *Cladistics*, 6 (1990), pp. 211–223.
- 1327 [35] M. PAGEL, *Inferring the historical patterns of biological evolution*, *Nature*,  
1328 401 (1999), pp. 877–884.
- 1329 [36] E. PARADIS, J. CLAUDE, AND K. STRIMMER, *Ape: analyses of phyloge-*  
1330 *netics and evolution in r language*, *Bioinformatics*, 20 (2004), pp. 289–290.

- 1331 [37] R. K. PLOWRIGHT, C. R. PARRISH, H. MCCALLUM, P. J. HUDSON,  
1332 A. I. KO, A. L. GRAHAM, AND J. O. LLOYD-SMITH, *Pathways to zoonotic*  
1333 *spillover*, Nature Reviews Microbiology, (2017).
- 1334 [38] E. PURDOM, *Analysis of a data matrix and a graph: Metagenomic data and*  
1335 *the phylogenetic tree*, The Annals of Applied Statistics, (2011), pp. 2326–  
1336 2358.
- 1337 [39] K. S. RAMIREZ, J. W. LEFF, A. BARBERÁN, S. T. BATES, J. BET-  
1338 LEY, T. W. CROWTHER, E. F. KELLY, E. E. OLDFIELD, E. A. SHAW,  
1339 C. STEENBOCK, ET AL., *Biogeographic patterns in below-ground diversity*  
1340 *in new york city’s central park are similar to those observed globally*, in  
1341 Proc. R. Soc. B, vol. 281, The Royal Society, 2014, p. 20141988.
- 1342 [40] L. J. REVELL, *phytools: an r package for phylogenetic comparative biology*  
1343 *(and other things)*, Methods in Ecology and Evolution, 3 (2012), pp. 217–  
1344 223.
- 1345 [41] K.-I. SATO, *Lévy processes and infinitely divisible distributions*, Cambridge  
1346 university press, 1999.
- 1347 [42] J. U. SCHER, A. SZESNAK, R. S. LONGMAN, N. SEGATA, C. UBEDA,  
1348 C. BIELSKI, T. ROSTRON, V. CERUNDOLO, E. G. PAMER, S. B. ABRAM-  
1349 SON, ET AL., *Expansion of intestinal prevotella copri correlates with en-*  
1350 *hanced susceptibility to arthritis*, Elife, 2 (2013), p. e01202.
- 1351 [43] K. P. SCHLIEP, *phangorn: phylogenetic analysis in r*, Bioinformatics, 27  
1352 (2011), pp. 592–593.
- 1353 [44] J. D. SILVERMAN, A. D. WASHBURNE, S. MUKHERJEE, AND L. A.  
1354 DAVID, *A phylogenetic transform enhances analysis of compositional mi-*  
1355 *crobiota data*, Elife, 6 (2017), p. e21887.

- 1356 [45] F. A. SMITH, A. G. BOYER, J. H. BROWN, D. P. COSTA, T. DAYAN,  
1357 S. M. ERNEST, A. R. EVANS, M. FORTELIUS, J. L. GITTLEMAN, M. J.  
1358 HAMILTON, ET AL., *The evolution of maximum body size of terrestrial*  
1359 *mammals*, science, 330 (2010), pp. 1216–1219.
- 1360 [46] F. A. SMITH AND S. K. LYONS, *How big should a mammal be? a macroe-*  
1361 *cological look at mammalian body size over space and time*, Philosophical  
1362 Transactions of the Royal Society of London B: Biological Sciences, 366  
1363 (2011), pp. 2364–2378.
- 1364 [47] P. J. TURNBAUGH, R. E. LEY, M. A. MAHOWALD, V. MAGRINI, E. R.  
1365 MARDIS, AND J. I. GORDON, *An obesity-associated gut microbiome with*  
1366 *increased capacity for energy harvest*, nature, 444 (2006), pp. 1027–131.
- 1367 [48] L. VAN VALEN, *Ecological species, multispecies, and oaks*, Taxon, (1976),  
1368 pp. 233–239.
- 1369 [49] Y. VÁZQUEZ-BAEZA, A. GONZALEZ, Z. Z. XU, A. WASHBURNE, H. H.  
1370 HERFARTH, R. B. SARTOR, AND R. KNIGHT, *Guiding longitudinal sam-*  
1371 *pling in ibd cohorts*, Gut, (2017), pp. gutjnl–2017.
- 1372 [50] A. D. WASHBURNE, J. W. BURBY, AND D. LACKER, *Novel covariance-*  
1373 *based neutrality test of time-series data reveals asymmetries in ecological*  
1374 *and economic systems*, PLoS computational biology, 12 (2016), p. e1005124.
- 1375 [51] A. D. WASHBURNE, J. D. SILVERMAN, J. W. LEFF, D. J. BENNETT,  
1376 J. L. DARCY, S. MUKHERJEE, N. FIERER, AND L. A. DAVID, *Phyloge-*  
1377 *netic factorization of compositional data yields lineage-level associations in*  
1378 *microbiome datasets*, PeerJ, 5 (2017), p. e2969.
- 1379 [52] G. YU, D. K. SMITH, H. ZHU, Y. GUAN, AND T. T.-Y. LAM, *ggtree: an*  
1380 *r package for visualization and annotation of phylogenetic trees with their*

- 1381 *covariates and other associated data*, *Methods in Ecology and Evolution*, 8  
1382 (2017), pp. 28–36.
- 1383 [53] X. ZHOU, S. XU, J. XU, B. CHEN, K. ZHOU, AND G. YANG, *Phy-*  
1384 *logenomic analysis resolves the interordinal relationships and rapid diver-*  
1385 *sification of the laurasiatherian mammals*, *Systematic biology*, 61 (2011),  
1386 pp. 150–164.