Running Header: Cohesion: Quantifying connectivity of communities

Title: Cohesion: A method for quantifying the connectivity of microbial communities

Authors: Cristina M. Herren[1*], Katherine D. McMahon[2]

Author affiliations:

[1]Freshwater and Marine Sciences Program, University of Wisconsin - Madison, Madison, Wisconsin, USA

[2]Departments of Bacteriology and Civil and Environmental Engineering, University of Wisconsin - Madison, Madison, Wisconsin, USA

* corresponding author

4440 Microbial Sciences Building, 1550 Linden Drive, Madison, WI, USA

cherren@wisc.edu

Word Count: 5127

Subject category: Microbial population and community ecology

Conflict of Interest: The authors declare no conflict of interest.

Abstract

The ability to predict microbial community dynamics lags behind the quantity of data available in these systems. Most predictive models use only environmental parameters, although a long history of ecological literature suggests that community complexity should also be an informative parameter. Thus, we hypothesize that incorporating information about a community's complexity might improve predictive power in microbial models. Here, we present a new metric, called community "cohesion," that quantifies the degree of connectivity of a microbial community. We validate our approach using long-term (10+ year) phytoplankton datasets, where absolute abundance counts are available. As a case study of our metrics' utility, we show that community cohesion is a strong predictor of Bray-Curtis dissimilarity ($R^2 = 0.47$) between phytoplankton communities in Lake Mendota, WI, USA. Our cohesion metrics outperform a model built using all available environmental data collected during a long-term sampling program. The result that cohesion corresponds strongly to Bray-Curtis dissimilarity is consistent across the five lakes analyzed here. Our cohesion metrics can be used as a predictor for many community-level properties, such as phylogenetic diversity, nutrient fluxes, or ecosystem services. We explain here the calculation of our cohesion metrics and their potential uses in microbial ecology.

Introduction

Most efforts to model microbial communities primarily use environmental drivers as predictors of community dynamics (Patterson 2009, Hambright et al. 2015). However, despite the vast quantities of data becoming available about microbial communities, predictive power in microbial models is often surprisingly poor (Blaser et al. 2016). Even in one of most well-studied microbial systems, the San Pedro Ocean Time Series (SPOT), there are sampling sites where none of the 33 environmental variables measured are highly significant ($P < 0.01$) predictors of community similarity (Cram et al. 2015). Thus, there may be room to improve predictive models by adding new parameters; ecological literature has long suggested that the degree of complexity in a community should inform community dynamics (MacArthur 1955, Cohen and Newman. 1985, Wootton and Stouffer 2016). We hypothesize that incorporating information about the complexity of microbial communities could improve predictive power in these communities.

Here, we present a workflow to generate metrics quantifying the connectivity of a microbial community, which we call "cohesion." We show how cohesion can be used to predict community dynamics (e.g. rate of compositional turnover, phylogenetic diversity, ecosystem services provided by the community). As an application of our metrics, we present a case study using our newly developed cohesion variables as predictors of the compositional turnover rate (a common response variable in microbial ecology) in phytoplankton communities.

Our cohesion metrics overcome two barriers that often preclude using information about community complexity in microbial analyses. First, the large number of taxa in microbial datasets makes it difficult to use information about all taxa in statistical analyses. Although methods exist to analyze microbial community interconnectedness (e.g. Local Similarity Analysis, artificial neural networks), this often involves constructing networks with many

62 parameters that are difficult to interpret. Second, microbial community data are often
63 "relativized" or "compositional" datasets, where the abundance of each taxon represents the
64 fraction of the community it comprises. This creates several problems in downstream analysis
65 (Weiss et al. 2016). For example, taxon correlation values are different in absolute versus
66 relative datasets (Faust and Raes 2012, Friedman and Alm 2012), and it is unclear how using
67 relative abundances influences the apparent population dynamics of individual taxa (Lovell et al.
68 2015). Thus, these two features (many taxa and relative abundance) have previously proven
69 problematic when analyzing microbial community connectivity.

70 Here, we describe and test a method to quantify one aspect of microbial community
71 complexity. Our resulting "cohesion" metrics quantify the connectivity of each sampled
72 community. Thus, our cohesion metrics integrate easily with other statistical analyses and can be
73 used by any microbial ecologist interested in asking whether community interconnectedness is
74 important in their study system. We demonstrate how to obtain these cohesion metrics from time
75 series data and, as a case study, show how cohesion relates to rates of compositional turnover in
76 long-term phytoplankton datasets. We develop this workflow with datasets where raw abundance
77 data are available and use these raw abundances to validate our methods when working with
78 relativized datasets. Thus, our approach was designed to overcome known challenges of
79 analyzing microbial datasets.

80

81 **Methods and Results**

82 *Description of datasets*

83 We restricted our search for datasets to those measured in absolute abundance that
84 spanned multiple years, as to cover a wide range of environmental conditions. We found several

85 datasets hosted on the North Temperate Lakes Long Term Ecological Research (NTL LTER)
86 database that were over 10 years in length and contained counts of phytoplankton in absolute
87 abundance. The datasets are from the following lakes in Wisconsin, USA: Lake Mendota (293
88 samples with 410 taxa over 19 years), Lake Monona (264 samples with 382 taxa over 19 years),
89 Paul Lake (197 samples with 209 taxa over 12 years), Peter Lake (197 samples with 237 taxa
90 over 12 years), and Tuesday Lake (115 samples with 121 taxa over 12 years). These lakes vary
91 in size, productivity, and food web structure. Peter Lake and Tuesday Lake were also subjected
92 to whole-lake food web manipulations during the sampling timeframe (detailed in Elser and
93 Carpenter 1988 and Cottingham et al. 1998). We present the workflow using results from the
94 Lake Mendota dataset, as it is the dataset of the longest duration with the largest number of taxa.
95 Details about datasets can be found at https://lter.limnology.wisc.edu/.

96 *Data curation*

97 Phytoplankton densities in Lake Mendota varied by more than 2 orders of magnitude
98 between sample dates. Densities of cells in these samples ranged from 956 cells/mL to 272 281
99 cells/mL. We removed individuals that were not identified at any level (e.g. categorized as
100 Miscellaneous). For each sample date, we converted the raw abundances to relative abundances,
101 such that all rows summed to 1. We removed taxa that were not present in at least 5% of
102 samples, as we were not confident that we could recover robust connectedness estimates for very
103 rare taxa. This cutoff retained an average of 98.9% of the identified cells in each sample. Results
104 of our analyses using other cutoff values can be found in the SOM.

105 *Overview*

106 The input of our workflow is the taxon relative abundance table, and the outputs are
107 measurements of the connectivity of each sampled community, which we call community

108 “cohesion” (Fig. 1). In the process, our workflow also produces metrics of the connectedness of
109 each taxon. Briefly, our workflow begins by calculating the pairwise correlation matrix between
110 taxa, using all samples. We use a null model to account for bias in these correlations due to the
111 skewed distribution of taxon abundances (i.e. many small values and a few large values) and
112 relativized nature of the dataset (i.e. all rows sum to 1). We subtract off these “expected”
113 correlations generated from the null model to obtain a matrix of corrected correlations. For each
114 taxon, the average positive corrected correlation and average negative corrected correlation are
115 recorded as the connectedness values. As previously noted, our goal was to create a metric of
116 connectivity for each community; thus, the next step in the workflow calculates cohesion values
117 for each sample. Cohesion is calculated by multiplying the abundance of each taxon in a sample
118 by its associated connectedness values, then summing the products of all taxa in a sample. There
119 are two metrics of cohesion, because we separately calculate metrics based on the positive and
120 negative relationships between taxa. Within each section (1, 2, and 3), we alternate between
121 presenting an analysis step and showing a validation of these techniques.
122
123 *1. CONNECTEDNESS METRIC*
124 Analysis
125 It is difficult to directly observe interactions within microbial communities, so
126 correlations are often used to infer relationships between taxa or between a taxon and the
127 environment. Thus, we used a correlation-based approach for determining the connectedness of
128 taxa. However, when using correlation-based approaches with relativized microbial datasets, it is
129 necessary to use a null model to evaluate how the features of the dataset (skewed abundances and
130 the fact that all rows sum to 1) contribute to correlations between taxa (Weiss et al. 2016). The

131 purpose of a null model is to assess the expected strengths of correlations when there are no true
132 relationships between taxa (Ulrich and Gotelli 2010).

133 Our null model was used to calculate how strongly the features common to microbial
134 datasets contribute to taxon connectedness estimates, so that this structural effect could be
135 subtracted from the connectedness metrics. While testing various null models, it became clear
136 that a taxon's pairwise correlation values were strongly related to its mean abundance and
137 persistence (fraction of samples when present) across the dataset. Thus, these features (mean
138 relative abundance and persistence) were retained in the null model by only randomizing
139 abundance values within each taxon. Details about alternate null models investigated are given in
140 the SOM.

141 The objective of the null model was to calculate the strength of pairwise correlations that
142 would be observed if there were no true relationship between taxa. During each iteration, one
143 taxon was designated as the "focal taxon" (Fig. 2). For each taxon *besides the focal taxon*,
144 abundances in the null matrix were randomly sampled without replacement from their abundance
145 distribution across all samples. Then, we calculated Pearson correlations between the focal taxon
146 and the randomized other taxa. We iterated through this process of calculating pairwise
147 correlations between the focal taxon and all other taxa 200 times. The median correlations from
148 these 200 randomizations were called the "expected" correlations for the focal taxon. We
149 repeated this process for each taxon as the focal taxon, which resulted in a matrix of expected
150 taxon correlations. Finally, we subtracted the expected taxon correlations from their paired
151 observed taxon correlations, thereby producing a matrix where each value was an observed
152 minus expected correlation for the given pair of taxa.

153 We calculated taxon connectedness values from the corrected (observed minus expected)
154 correlation matrix. For each taxon, we separately averaged its positive and negative correlations
155 with other taxa to produce a value of positive connectedness and a value of negative
156 connectedness. We kept positive and negative values separate, because converting the dataset to
157 relative abundance should affect positive and negative correlations differently, since an increase
158 in the relative abundance in one taxon must result in the decrease in the relative abundances of
159 other taxa.

160 The averaging step in this workflow was intended to overcome the issue that individual
161 correlations between taxa can be influenced by many factors and may be spurious (Fisher and
162 Mehta 2014). However, assuming that correlations often (but not always) reflect complexity in a
163 community, the average of many correlations should be a more robust metric of complexity than
164 any single correlation. In other words, we assume only that highly connected taxa have stronger
165 correlations *on average*. Invoking the law of large numbers, these average correlations should be
166 increasingly accurate measures of a taxon's connectedness as the number of pairwise correlations
167 increases (i.e. as the number of taxa in the dataset increases).

168

169 Validation

170 As discussed previously, there are inherent limitations of using correlation-based
171 methods with relative abundance data instead of absolute counts (Fisher and Mehta 2014). Thus,
172 we examined whether a measure of connectedness based on absolute abundance would show the
173 same pattern observed using the relativized data. However, we needed a different approach for
174 calculating correlations in order to account for the following properties of count data: 1)
175 variance-mean scaling, which results in very large population variances of abundant taxa (Taylor

176 1961) and 2) the fact that individual population sizes are strongly related to overall community
177 densities, which causes positive correlations among all taxa (Doak et al. 1998). As noted
178 previously, phytoplankton densities in Lake Mendota samples varied by more than 2 orders of
179 magnitude among sample dates. Therefore, using correlations between raw abundances would
180 inflate the positive relationships between taxa as a result of changing overall community density.
181 Thus, we first detrended the count data to account for changing community density (on different
182 sampling dates) and drastically different variances of taxon populations (which are expected as a
183 result of mean-variance scaling).

184 We used a hierarchical linear model to estimate the effects of overall community density
185 and mean taxon abundance on individual taxon observations (sensu Jackson et al. 2012), so that
186 these effects could be removed when calculating correlations. We modeled the abundance of
187 each taxon at each time point as a function of sample date and taxon, assuming a quasipoisson
188 distribution (which accounts for increases in population variances when population means
189 increase). The residuals of this analysis represent the normalized (transformed) deviations of
190 taxon abundances after accounting for overall community density on the sample date and taxon
191 abundance/variance. We created a pairwise correlation matrix for the phtoplankton taxa using the
192 correlations between these residuals. We calculated connectedness metrics from the pairwise
193 correlation matrix using the same technique that we applied to the corrected correlation matrix
194 from the relativized data: we used the average positive and negative taxon correlations as their
195 connectedness values.

196 We validated our workflow for the relative abundance dataset using the estimates of
197 taxon connectedness obtained from the absolute abundance dataset. Comparing the
198 connectedness values from these two methods shows strong agreement between the two sets of

connectedness metrics (correlation for positive connectedness metrics = 0.820; correlation for negative connectedness metrics = 0.741, Fig. 3). Although two taxa deviate from the linear relationship between the negative connectedness metrics (appearing as outliers in Fig. 3B), both metrics rate these taxa as having strong connectedness arising from negative correlations. Thus, the two methods are qualitatively consistent for these two anomalous points.

2. COHESION METRIC

Analysis

Many researchers seek to detect differences in community connectivity across time, space, or treatments. Thus, it would be useful to have a metric that quantifies, for each community, the degree to which its component taxa are connected. We used a simple algorithm to collapse the connectedness values of individual taxa into two parameters representing the connectivity of the entire sampled community, termed "cohesion." Again, one metric of cohesion stems from positive correlations, and one metric stems from negative correlations. To calculate each cohesion metric, we multiplied the relative abundance of taxa in a sample by their associated connectedness values and summed these products. This cohesion index can be represented mathematically as the sum of the contribution of each of the *n* taxa in the community, after removing rare taxa (Eq. 1). Thus, communities with high abundances of strongly connected taxa would have a high score of community cohesion. We note that this index is bounded by -1 to 0 for negative cohesion or from 0 to 1 for positive cohesion.

$$cohesion = \sum_{i=1}^{n} abundance_i * connectedness_i \qquad \text{Eq. 1}$$

222

223 Validation

224 We had hypothesized that our cohesion metrics could be significant predictors of
225 microbial community dynamics. Thus, a natural question to ask was whether our metrics of
226 cohesion outperform environmental variables when analyzing the Lake Mendota data.
227 Fortunately, the NTL LTER program has collected paired environmental data for the Lake
228 Mendota phytoplankton samples. We obtained these environmental datasets to use as alternative
229 predictors of community dynamics in Lake Mendota. The environmental datasets available (11
230 variables) were: water temperature, air temperature, dissolved oxygen concentration, dissolved
231 oxygen saturation, Secchi depth, combined NO3 + NO2 concentrations, NH4 concentration, total
232 nitrogen concentration, dissolved reactive phosphorus concentration, total phosphorus
233 concentration, and dissolved silica concentrations. Protocols, data, and associated metadata can
234 be found at https://lter.limnology.wisc.edu/. We use these environmental data to build an
235 alternate model in our case study below.

236

237 *3. CASE STUDY OF UTILITY*

238 Analysis

239 To demonstrate their utility, we applied our new metrics to the Lake Mendota
240 phytoplankton dataset. We tested whether community cohesion could predict compositional
241 turnover, a common response variable in microbial ecology. We used multiple regression to
242 model compositional turnover (Bray-Curtis dissimilarity between time points) as a function of
243 community cohesion at the initial time point. That is, Bray-Curis dissimilarity was the dependent
244 variable, while positive and negative cohesion were the independent variables. Because time

245 between samples influences Bray-Curtis dissimilarity (Nekola and White 1999, Shade et al.
246 2013), we only included pairs of samples taken within 36 to 48 days of each other. These criteria
247 included 186 paired communities across the 19 years. Cohesion values (both positive and
248 negative) were calculated at the first time point for each sample pair. We chose this timeframe
249 because it was sufficiently long for multiple phytoplankton generations to have occurred, and
250 because this timeframe was compatible with the sampling frequency.

251 Community cohesion was a strong predictor of compositional turnover (Fig. 4). The
252 regression using our cohesion metrics explained 46.5% of variability (adjusted $R^2 = 0.465$) in
253 Bray-Curtis dissimilarity. Cohesion arising from negative correlations was a highly significant
254 predictor, whereas cohesion arising from positive correlations was not significant (negative
255 cohesion: $F_{1, 183} = 6.81$, $p < 1* e^{-20}$; positive cohesion: $F_{1, 183} = 0.735$, $p = 0.405$).

256 For the purpose of model comparison, we used the associated environmental data to
257 model Bray-Curtis dissimilarity as a function of environmental drivers. We included as
258 predictors the 11 variables previously mentioned, as well as 11 additional predictors that
259 measured the change in each of these variables between the two sample dates. Finally, because
260 many chemical and biological processes are dependent on temperature (Brown et al. 2004), we
261 included first order interactions between water temperature and the 21 other variables. We first
262 included all 43 terms in the model, then used backward selection (which iteratively removes the
263 least significant term in the model, beginning with interaction terms) until all remaining terms in
264 the model were significant at $p < 0.1$, as to maximize the adjusted R^2 value. Although this
265 analysis does not represent an exhaustive list of possible environmental drivers, it includes all
266 available paired environmental data from the long-term monitoring program. Twenty-nine values
267 of Bray-Curtis dissimilarity were excluded from this analysis (leaving 157 of the 186 values),

268 because they lacked one or more associated environmental variables. Additional details about
269 this analysis can be found in the SOM.

270 In the final model after backward selection, 16 variables were retained as significant
271 predictors (see SOM). The adjusted R^2 of this model was 0.229. The *non-adjusted* R^2 value of
272 the full model (all 43 variables) was 0.393.

273 To address the generality of the relationship between cohesion and community turnover,
274 we calculated cohesion metrics and Bray-Curtis dissimilarity for the four other phytoplankton
275 datasets (Monona, Peter, Paul, and Tuesday lakes). Community cohesion was a significant
276 predictor of Bray-Curtis dissimilarity in all datasets. In each instance, stronger cohesion resulting
277 from negative correlations was related to lower compositional turnover. Table 1 presents the
278 results of these analyses and associated workflow parameters. Additional information about the
279 sensitivity of model performance to varying parameters can be found in the SOM.

280 Validation

281 Strong correlations between predictor variables are known to influence the results of
282 statistical analyses (Neter et al. 1996). Thus, we wondered whether strong correlations between
283 taxa would necessarily generate the observed relationship where greater cohesion is related to
284 lower compositional turnover. We conducted simulation studies to investigate whether our
285 significant results might be simply an artifact of strong inter-taxon correlations. We generated
286 datasets where taxa were highly correlated in abundance, as if they were synchronously
287 responding to exogenous forces. We calculated cohesion metrics and Bray-Curtis dissimilarities
288 for the simulated datasets to analyze whether strong taxon correlations was sufficient to produce
289 results similar to those we observed in the real data.

290 Here, we briefly describe the process used to simulate datasets, while additional details
291 can be found in SOM. First, we generated four autocorrelated vectors to represent exogenous
292 forces, such as environmental drivers. Taxa were artificially correlated to these external vectors,
293 thereby also producing strong correlations between taxa. We manipulated the taxon abundances
294 to mimic other important features of the microbial datasets, including skewed taxon mean
295 abundances and a large proportion of zeroes in the dataset. We calculated cohesion metrics and
296 Bray-Curtis dissimilarities for the simulated datasets, and we used a multiple regression to model
297 Bray-Curtis dissimilarity as a function of positive cohesion and negative cohesion. We recorded
298 the R^2 value and parameter estimates of this multiple regression. We repeated this simulation
299 process 500 times to generate distributions of these results.
300 Our cohesion metrics had a very low ability to explain compositional turnover (Bray-
301 Curtis dissimilarity) in the simulated datasets. The median model adjusted R^2 value was 0.022,
302 with 95% of adjusted R^2 values below 0.088 (Fig. 5). Although the community cohesion metrics
303 were highly significant predictors ($p < 0.001$) of community turnover more commonly than
304 would be expected by chance (1.0% of simulations for positive cohesion and 8.6% for negative
305 cohesion), the proportion of variance explained by these metrics was comparatively very low.
306 For comparison, across the five phytoplankton datasets from Wisconsin lakes, model adjusted R^2
307 values ranged from 0.36 to 0.50. Thus, there was comparatively little ability to explain
308 compositional turnover in the simulated datasets using our cohesion metrics.
309
310 **Discussion**
311 The ability to predict microbial community dynamics lags behind the amount of data
312 collected in these systems (Blaser et al. 2016). Here, we present new metrics, called "cohesion,"

313 which can be used as additional predictor variables in microbial models. The cohesion metrics
314 contain information about the connectivity of microbial communities, which has been previously
315 hypothesized to influence community dynamics (MacArthur 1955, May 1972, Nilsson and
316 McCann 2016). Our cohesion metrics are easily calculated from a relative abundance table (R
317 script provided online) and might be of interest to a variety of microbial ecologists and modelers.

318 In the Lake Mendota phytoplankton example, our two cohesion parameters outperformed
319 the available environmental data at predicting phytoplankton community changes. The two
320 cohesion parameters explained 46.5% of variability (adjusted $R^2 = 0.465$) in community turnover
321 over 19 years of phytoplankton sampling, in comparison to the final environmental model using
322 16 predictors, which explained 22.9% of community turnover (adjusted $R^2 = 0.229$). Although
323 there are almost certainly important predictors missing from the environmental model (e.g.
324 photosynthetically active radiation, three-way interactions), the environmental model represents
325 a commonly applied approach to explaining microbial compositional turnover (Tripathi et al.
326 2012, Chow et al. 2013) that uses all associated environmental data from a long-term sampling
327 program. Although we still strongly advocate for the collection of environmental data, we note
328 that cohesion was a much better predictor of compositional turnover than any available
329 environmental variable.

330 Our workflow overcomes many challenges associated with using correlation-based
331 techniques in microbial datasets. The validations we conducted indicated that our connectedness
332 metrics are robust for relativized datasets, because connectedness metrics from relative and
333 absolute datasets showed strong correspondence. Additionally, our cohesion metrics address a
334 common problem of techniques describing community complexity (such as network analyses),
335 which is that they do not quantify the connectivity of individual communities. For instance, the

336 “hairball” generated from a network analysis has no parameters that are associated with each
337 sample, and therefore the network cannot be used as a predictor variable. Thus, existing methods
338 to quantify connectivity do not pair easily with other analyses. Our cohesion metrics quantify
339 sample connectivity using only two parameters, which can be used as predictors in a variety of
340 further analyses (linear regression, ordinations, time series, etc.). Finally, our simulations showed
341 that strong inter-taxon correlations were not sufficient to reproduce the observed result that
342 cohesion was a strong predictor of Bray-Curtis dissimilarity. In the simulations, cohesion had
343 low explanatory power, even though taxa were highly correlated. From this result, we infer that
344 correlations between taxa in real communities are an important aspect of complexity that is
345 captured by our cohesion metrics.

346 Our cohesion metrics explain a significant amount of compositional change in all five
347 phytoplankton datasets. Yet, it is not immediately clear what cohesion is measuring. There are
348 two broad factors that could cause correlations between taxa: biotic interactions and
349 environmental drivers. Thus, at least one of these two factors must underlie our connectedness
350 and cohesion metrics. Here, we discuss the evidence supporting either of these interpretations:

351 *Cohesion as a Measure of Biotic Interactions*

352 Even if shared responses to environmental drivers underlie most pairwise taxon
353 correlations, cohesion could still indicate biotic interaction strength in a community. This would
354 occur if taxa were influenced to the same degree by environmental drivers, but differentially
355 influenced by species interactions. In this case, averaging over all correlations would give larger
356 connectedness values for strong interactors and smaller connectedness values for weak
357 interactors. Many studies have indicated that microbial taxa have differential interaction
358 strengths. For example, some microbial communities contain keystone taxa, which have

359 disproportionate effects on community dynamics through their strong taxon interactions (Trosvik
360 and de Muinck 2015, Banerjee et al. 2016). Similarly, recent work suggests that many taxa
361 within candidate phyla are obligate symbionts, meaning they must interact strongly with other
362 taxa for their survival and reproduction (Kantor et al. 2013, Hug et al. 2016). Conversely, there
363 are many taxa that can be modeled adequately as a function of environmental drivers; this is true
364 for some bloom forming cyanobacteria, which are known to respond strongly to nutrient
365 concentrations and temperature (McQueen and Lean 1987, Beaulieu et al. 2013). Taken together,
366 these studies suggest that there is a wide spectrum of how strongly taxa interact with one
367 another. These differences in interaction strength would be detected by our connectedness metric
368 due to averaging over the large number of pairwise correlations. Thus, it is plausible that
369 connectedness and cohesion are reflecting biotic interactions in communities.

370 We now examine our phytoplankton analysis results under the assumption that cohesion
371 measures biotic interactions. The Bray-Curtis dissimilarity regression results would mean that
372 communities with many strong interactors have lower rates of change, especially when the
373 interactions create negative correlations between taxon abundances. This finding is in line with
374 prior work showing that biotic interactions affect microbial community stability (Coyte et al.
375 2016). Thus, the interpretation that stronger biotic interactions lead to lower compositional
376 turnover is a plausible explanation for our observed results. However, we specifically refrain
377 from interpreting positive or negative connectedness values as indications of specific biotic
378 interactions, such as predation, competition, or mutualism. For example, a positive correlation
379 between two taxa could be the result of a mutualism between the taxa, or it could be the result of
380 a shared predator declining in abundance. Further work, both empirical and theoretical, is

381 necessary to identify what these positive and negative correlations signify in the context of the
382 ecology of these organisms.

383 *Cohesion as a Measure of Environmental Synchrony*

384 We now consider the possibility that connectedness and cohesion are simply detecting
385 environmental synchrony. If a subset of taxa respond to a changing environmental driver, then
386 these taxa will have strong pairwise correlations. For example, correlations between
387 phytoplankton species of the same genus (and, therefore, with similar niches) can be upwards of
388 0.9, indicating strong similarity in abundance patterns. In this case, connectedness would
389 measure the degree of environmentally-driven population synchrony that a taxon has with other
390 taxa. A high cohesion value would indicate that a community has many taxa that respond
391 simultaneously to external forces; then, cohesion would quantify overall community
392 responsiveness to one or more environmental drivers. Under this assumption, cohesion should
393 correlate with environmental drivers (e.g. cohesion is high because many taxa are positively
394 correlated to warm temperatures, but cohesion drops when it gets colder and these taxa senesce).
395 We tested this prediction with 22 variables from the environmental model (11 for the
396 environmental variables and 11 for the changes in environmental variables) and found that
397 negative cohesion in the Lake Mendota phytoplankton dataset generally had weak correlations
398 with these predictors (absolute correlations < 0.25, SOM). We also looked for a seasonal trend in
399 cohesion, but found no significant correlation between negative cohesion and Julian Day, or a
400 quadratic term for Julian Day. Thus, we do not find any evidence that cohesion is simply
401 reproducing the information contained in environmental data. Finally, our simulations show one
402 example where taxon abundances could be driven exclusively by external factors (such as the
403 environment), but this does not necessarily lead to strong predictive power of compositional

turnover. However, our simulations omitted many features of real ecological communities, and so we cannot completely rule out the possibility that environmental drivers contributed to our cohesion metrics in the phytoplankton datasets.

Under the assumption that cohesion measures environmentally driven population synchrony, we examine our result that stronger cohesion was related to lower Bray-Curtis dissimilarity. In this scenario, communities that have strong cohesion contain high abundances of taxa that respond simultaneously to environmental forces. Then, communities with many synchronous taxa would turn over more slowly than communities with taxa whose abundances are independent of the environment. This conclusion is counterintuitive, but possible. This pattern could occur if taxa that are strongly influenced by the environment have lower variability than taxa that are weakly influenced by the environment; in that case, highly correlated taxa would have their abundances more tightly regulated than other taxa. Although plausible, this explanation disagrees with many studies that have found that environmental gradients regulate which taxa can persist in communities (Fierer and Jackson 2006, Walter and Ley 2011, Freedman and Zak 2015).

Comparing the two possible signals that cohesion might be detecting, we believe the evidence points to biotic interaction as the larger contributor. However, we expect that environmental synchrony is captured to some extent, with the relative importance of environmental factors depending on the particular communities and ecosystem. Regardless of the ecological force measured by cohesion, there is a clear result in the five datasets analyzed that stronger negative cohesion is related to lower compositional turnover. This result suggests that negative correlations between taxa are arranged non-randomly to counteract one another, thereby stabilizing community composition. In other words, relationships between taxa appear to form

negative feedback loops that buffer changes in community composition. This result agrees with prior theoretical models that propose that feedback loops are integral to modulating food web stability (Neutel et al. 2007, Brose 2008). The finding that negative cohesion was stabilizing was not easily replicated in our simulations, where positive and negative correlations were interspersed with random magnitude throughout the dataset. Thus, the arrangement of correlations between taxa in the dataset appears to be an important feature of real communities that may contribute to their stability (Worm and Duffy 2003).

Guidelines for Using Our Metrics

Although we used long-term time series datasets for the analyses presented here, our cohesion metrics can be used to predict community dynamics in a variety of datasets. For example, cohesion could be used with a spatially explicit dataset, where samples were collected from different locations across a landscape. In the context of phytoplankton samples, this could be a dataset consisting of samples from different locations in a lake or watershed. Then, the cohesion metrics could be used to predict community composition change at one location over time, or to predict differences in community composition between locations. It would also be interesting to investigate how cohesion is affected by experimental perturbations. Finally, cohesion could be used as a predictor in of many response variables. Additional applications of the cohesion metrics could include identifying communities susceptible to major compositional change (e.g. cyanobacterial blooms, infection in the human microbiome), relating community cohesion to spatial structure (e.g. how taxon connectedness relates to the dispersal abilities of different microbial taxa), and investigating how disturbance influences cohesion (e.g. how illness influences the cohesion of communities in a host-associated microbiome, how oil spills affect cohesion of marine microbial communities).

450 The critical step in the cohesion workflow is calculating reliable correlations between
451 taxa. Thus, some datasets will be more suitable for our cohesion metric than others. For example,
452 a dataset consisting of 20 samples from five lakes over multiple years might be a poor candidate
453 for the cohesion metrics. In this case, correlations between taxa might be driven mainly by
454 environmental differences or location, and the sample number would be too low to calculate
455 robust correlations. Based on the phytoplankton datasets analyzed here, we suggest a lower limit
456 of 40-50 samples when calculating cohesion metrics, with more samples necessary with more
457 heterogeneous datasets. We also suggest including environmental variables as covariates when
458 analyzing heterogeneous datasets. Finally, the persistence cutoff for including taxa should be
459 adjusted based on the dataset being analyzed. For example, in datasets obtained by DNA
460 sequencing, the sequencing depth affects taxon persistence (Smith and Peay 2014). Thus, it may
461 be more effective to implement a cutoff by mean abundance.
462
463 **Conclusion**
464 Our cohesion metrics provide a method to incorporate information about microbial
465 community complexity into predictive models. These metrics are easy to calculate, needing only
466 a relative abundance table. Furthermore, across all datasets analyzed in this study, cohesion was
467 strongly related to compositional turnover. In systems where cohesion is a significant predictor
468 of community properties (e.g. nutrient flux, rates of photosynthesis), this result could guide
469 further investigation into the effects of microbial interactions in mediating community function.
470 In this case, researchers might focus their efforts on understanding the role of highly connected
471 taxa, which are identified in our workflow. We aim to eventually determine the features that

472 distinguish systems in which cohesion is important versus systems in which cohesion does not
473 predict community properties.
474
475 **Acknowledgements**
476 We thank the North Temperate Lakes LTER program for the use of their publicly available data
477 on Lake Mendota and Lake Monona. We also thank the Cascade research group for the use of
478 their data from Peter, Paul, and Tuesday lakes, which is hosted on the LTER website. This
479 manuscript has been much improved as a result of comments from the McMahon lab. This work
480 was funded by a United States National Science Foundation (NSF) GRFP award to CMH (DGE-
481 1256259). KDM acknowledges funding from the NSF Long Term Ecological Research program
482 (NTL-LTER DEB-1440297) and an INSPIRE award (DEB-1344254).
483
484 Supplementary material is available at ISME Journal's website.
485

References

Banerjee S, Kirkby CA, Schmutter D, Bissett A, Kirkegaard JA, Richardson AE. (2016). Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* **97**: 188–198.

Beaulieu M, Pick F, Gregory-Eaves I. (2013). Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. *Limnol Oceanogr* **58**: 1736–1746.

Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL, *et al*. (2016). Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *mBio* **7**: e00714–16.

Brose U. (2008). Complex food webs prevent competitive exclusion among producer species. *Proceedings of the Royal Society of London B: Biological Sciences* **275**: 2507–2514.

Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. (2004). Toward a Metabolic Theory of Ecology. *Ecology* **85**: 1771–1789.

Chow C-ET, Sachdeva R, Cram JA, Steele JA, Needham DM, Patel A, *et al*. (2013). Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *ISME J* **7**: 2259–2273.

Cohen JE, Newman CM. (1985). When will a large complex system be stable? *Journal of Theoretical Biology* **113**: 153–156.

Cottingham KL, Carpenter SR, Amand ALS. (1998). Responses of epilimnetic phytoplankton to experimental nutrient enrichment in three small seepage lakes. *J Plankton Res* **20**: 1889–1914.

Coyte KZ, Schluter J, Foster KR. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science* **350**: 663–666.

Cram JA, Chow C-ET, Sachdeva R, Needham DM, Parada AE, Steele JA, *et al*. (2015). Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J* **9**: 563–580.

David Hambright K, Beyer JE, Easton JD, Zamor RM, Easton AC, Hallidayschult TC. (2015). The niche of an invasive marine microbe in a subtropical freshwater impoundment. *ISME J* **9**: 256–264.

Doak DF, Bigger D, Harding EK, Marvier MA, O'Malley RE, Thomson D. (1998). The Statistical Inevitability of Stability‐Diversity Relationships in Community Ecology. *The American Naturalist* **151**: 264–276.

Elser JJ, Carpenter SR. Predation-driven dynamics of zooplankton and phytoplankton communities in a whole-lake experiment. *Oecologia* **76**: 148–154.

Faust K, Raes J. (2012). Microbial interactions: from networks to models. *Nat Rev Micro* **10**: 538–550.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *PNAS* **103**: 626–631.

Fisher CK, Mehta P. (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLOS ONE* **9**: e102451.

Freedman Z, Zak DR. (2015). Soil bacterial communities are shaped by temporal and environmental filtering: evidence from a long-term chronosequence. *Environ Microbiol* **17**: 3208–3218.

Friedman J, Alm EJ. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology* **8**: 1–11.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, *et al*. (2016). A new view of the tree of life. *Nature Microbiology* **1**: 16048.

Jackson MM, Turner MG, Pearson SM, Ives AR. (2012). Seeing the forest and the trees: multilevel models reveal both species and community patterns. *Ecosphere* **3**: 1–16.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, *et al*. (2013). Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio* **4**: e00708–13.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Comput Biol* **11**: e1004075.

MacArthur R. (1955). Fluctuations of Animal Populations and a Measure of Community Stability. *Ecology* **36**: 533–536.

May RM. (1972). Will a Large Complex System be Stable? *Nature* **238**: 413–414.

McQueen DJ, Lean DRS. (1987). Influence of Water Temperature and Nitrogen to Phosphorus Ratios on the Dominance of Blue-Green Algae in Lake St. George, Ontario. *Can J Fish Aquat Sci* **44**: 598–604.

Nekola JC, White PS. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography* **26**: 867–878.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. (1996). *Applied linear statistical models* (Vol. 4). Chicago: Irwin.

Neutel A-M, Heesterbeek JAP, van de Koppel J, Hoenderboom G, Vos A, Kaldeway C, *et al*. (2007). Reconciling complexity with stability in naturally assembling food webs. *Nature* **449**: 599–602.

Nilsson KA, McCann KS. (2016). Interaction strength revisited—clarifying the role of energy flux for food web stability. *Theor Ecol* **9**: 59–71.

Patterson DJ. (2009). Seeing the Big Picture on Microbe Distribution. *Science* **325**: 1506–1507.

Shade A, Gregory Caporaso J, Handelsman J, Knight R, Fierer N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J* **7**: 1493–1506.

Smith DP, Peay KG. (2014). Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing. *PLOS ONE* **9**: e90234.

Taylor LR. (1961). Aggregation, Variance and the Mean. *Nature* **189**: 732–735.

Tripathi BM, Kim M, Singh D, Lee-Cruz L, Lai-Hoe A, Ainuddin AN, *et al*. (2012). Tropical Soil Bacterial Communities in Malaysia: pH Dominates in the Equatorial Tropics Too. *Microb Ecol* **64**: 474–484.

Trosvik P, de Muinck EJ. (2015). Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. *Microbiome* **3**: 44.

Ulrich W, Gotelli NJ. (2010). Null model analysis of species associations using abundance data. *Ecology* **91**: 3384–3397.

Walter J. and Ley R (2011). The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. *Annual Review of Microbiology* **65**: 411–429.

Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, *et al*. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* **10**: 1669–1681.

Wootton KL, Stouffer DB. (2015). Many weak interactions and few strong; food-web feasibility depends on the combination of the strength of species' interactions and their correct arrangement. *Theor Ecol* **9**: 185–195.

Worm B, Duffy JE. (2003). Biodiversity, productivity and stability in real food webs. *Trends in Ecology & Evolution* **18**: 628–632.

585 **Figure 1:** This diagram shows an overview of how our cohesion metrics are calculated,
586 beginning with the relative abundance table and ending with the cohesion values. The relative
587 abundance table shows six samples (S1 indicating "Sample 1", etc.) and a subset of taxa (A, B,
588 C, and Z). First, pairwise correlations are calculated between all taxa, which are entered into the
589 correlation matrix. We then used a null model to account for how the features of microbial
590 datasets might affect correlations, and we subtracted off these values (null model detailed in Fig.
591 2). For each taxon, we averaged the positive and negative corrected correlations separately and
592 recorded these values as the positive and negative connectedness values. Cohesion values were
593 obtained by multiplying the relative abundance table by the connectedness values. Thus, there
594 are two metrics of cohesion, corresponding to positive and negative values.
595
596 **Figure 2:** Microbial data are in the form of relative abundance, and some taxa are much more
597 abundant than others, which are factors that may cause taxa to be spuriously correlated. Thus, we
598 devised a null model to account for the bias that these data features introduce into our metrics.
599 We repeated this process with each taxon as the "focal taxon," which is A in this figure. For each
600 of 200 iterations, we randomized all taxon abundances *besides the focal taxon*. We then
601 calculated correlations between the focal taxon and all other taxa. We recorded the median value
602 of the 200 correlations calculated for each pair of taxa in the median correlation matrix.
603
604 **Figure 3:** Comparing the metrics of connectedness obtained from the absolute abundance dataset
605 (x-axes) and the relative abundance dataset (y-axes) shows agreement between the two methods
606 of generating these metrics. Correlations between the metrics are 0.810 (panel A) and 0.741
607 (panel B). We used separate variables for positive and negative metrics because relativizing the

608 dataset is expected to differentially affect positive and negative correlations. Solid lines show the
609 fit of linear models.
610
611 **Figure 4:** We used our metrics of community cohesion as predictors of the rate of compositional
612 turnover (Bray-Curtis dissimilarity) in the Mendota phytoplankton communities. Negative
613 cohesion was a significant predictor ($p < 1*e^{-20}$) of Bray-Curtis dissimilarity, and the regression
614 explained 46.5% of variation in compositional turnover.
615
616 **Figure 5:** We simulated datasets where correlations between taxa were artificially produced by
617 forced correlation to external factors. We calculated cohesion values for the simulated
618 communities to test whether cohesion and Bray-Curtis dissimilarity were strongly related in
619 simulated datasets. The histogram of model adjusted R^2 values from our simulations shows that
620 the median adjusted R^2 was 0.022 (dashed line), with 95% of values falling below 0.088. For
621 comparison, observed adjusted R^2 values ranged from 0.36 to 0.50.
622
623

624 **Table 1: Cohesion predicts community turnover in five phytoplankton datasets**

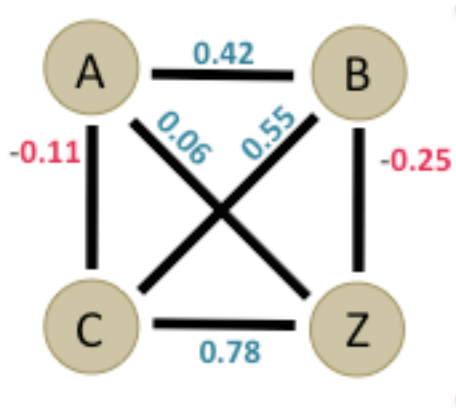
Lake	*Taxon Pers. Cutoff* *	*Model Adjusted R^2*	*Positive Cohesion P value*	*Negative Cohesion P value*	*Positive Cohesion Direction* +	*Negative Cohesion Direction* +	*Days Between Samples*	*Number of Samples*
Mendota	5%	0.465	0.405	$< 1*e^{-20}$	n.s.	Stronger is stabilizing	36-48	186
Monona	5%	0.355	0.413	$< 1*e^{-15}$	n.s.	Stronger is stabilizing	36-48	166
Peter	10%	0.357	0.062	$< 1*e^{-3}$	n.s.	Stronger is stabilizing	39-45	121
Paul	10%	0.500	$< 1*e^{-11}$	$< 1*e^{-19}$	Weaker is stabilizing	Stronger is stabilizing	39-45	125
Tuesday	10%	0.374	0.355	$< 1*e^{-8}$	n.s.	Stronger is stabilizing	39-45	72

625 * Stands for "taxon persistence cutoff," which was the minimum proportion of presences across
626 the dataset that we used as a cutoff for including taxa in the connectedness and cohesion metrics.
627 Other cutoffs may give higher model adjusted R^2 values (see SOM), but we wanted to use the
628 same cutoff for datasets collected within the same sampling program.
629 + These columns indicate the direction of a significant relationship between cohesion and Bray-
630 Curtis dissimilarity. For example, "stronger is stabilizing" means that greater cohesion is related
631 to lower Bray-Curtis dissimilarity. Non-significant relationships are denoted "n.s.".
632

Relative Abundance Table

	A	B	C	...	Z
S1	0.06	0.20	0.02		0.11
S2	0.03	0.43	0.00		0.13
S3	0.61	0.00	0.04		0.08
S4	0.12	0.33	0.05		0.07
S5	0.04	0.25	0.00		0.15
S50	0.04	0.28	0.05		0.10

Taxon Correlations



Correlation Matrix

	A	B	C	...	Z
A	1	0.42	-0.11		0.06
B	0.42	1	0.55		-0.25
C	-0.11	0.55	1		0.78
...					
Z	0.06	-0.25	0.78		1

subtract null model

Observed – Expected Correlations

	A	B	C	...	Z
A	1	0.35	-0.10		0.03
B	0.35	1	0.46		-0.18
C	-0.10	0.46	1		0.70
...					
Z	0.03	-0.18	0.70		1

single taxon

Taxon Correlations

	A
A	0
B	0.35
C	-0.10
...	
Z	0.03

Average Correlations

A: (+) = mean(0.35 + 0.03 + ...)
= 0.17

A: (-) = mean(-0.10 + ...)
= -0.08

Connectedness Metrics

	+	-
A	0.17	-0.08
B		
C		
...		
Z		

repeat by taxon

Connectedness Metrics

	+	-
A	0.17	-0.08
B	0.26	-0.13
C	0.38	-0.03
...		
Z	0.09	-0.21

Multiply Sample Table by Connectedness

	A	B	C	...	Z
S1	0.06	0.20	0.02		0.11
S2	0.03	0.43	0.00		0.13
S3	0.61	0.00	0.04		0.08
S4	0.12	0.33	0.05		0.07
S5	0.04	0.25	0.00		0.15
S50	0.04	0.28	0.05		0.10

X

	+
A	0.17
B	0.26
C	0.38
...	
Z	0.09

=

	+
S1	0.06*0.17 + 0.20*0.26+ 0.02*0.38 + ... + 0.11*0.09
S2	0.03*0.17 + 0.43*0.26 + 0*0.38 + ... + 0.13*0.09
S3	0.61*0.17 + 0*0.26 + 0.04*0.38 + ... + 0.08*0.09
...	

=

	+
S1	0.035
S2	0.052
S3	0.019
S4	0.063
S5	0.044
S50	0.022

	A	B	C	...	Z
S1	0.06	0.20	0.02		0.11
S2	0.03	0.43	0.00		0.13
S3	0.61	0.00	0.04		0.08
S4	0.12	0.33	0.05		0.07
S5	0.04	0.25	0.00		0.15
S50	0.04	0.28	0.05		0.10

X

	-
A	-0.08
B	-0.13
C	-0.03
...	
Z	-0.21

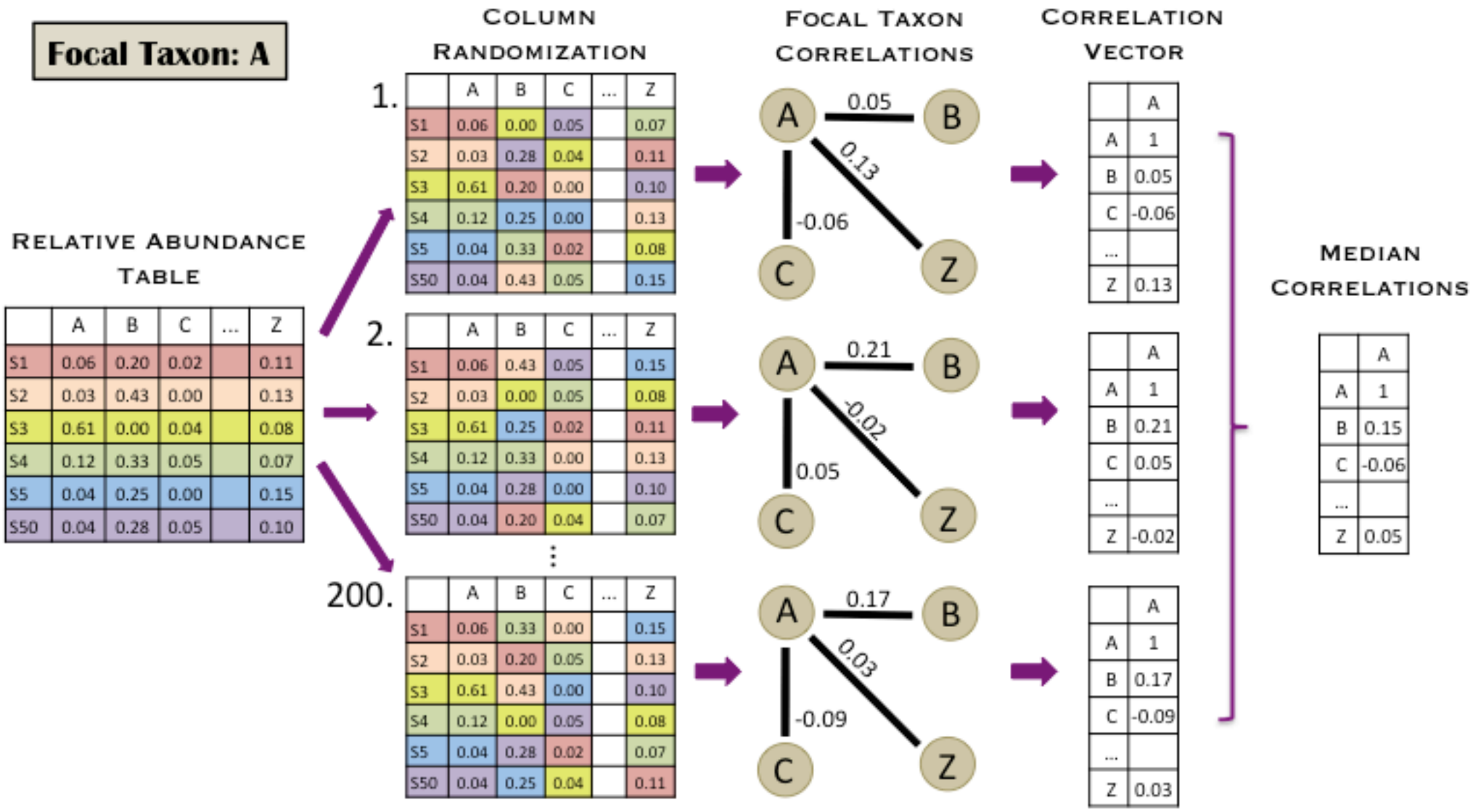
=

	-
S1	0.06*-0.08 + 0.20*-0.13+ 0.02*-0.03 + ... + 0.11*-0.21
S2	0.03*-0.08 + 0.43*-0.13 + 0*-0.03 + ... + 0.13*0-0.21
S3	0.61*-0.08 + 0*-0.13 + 0.04*-0.03 + ... + 0.08*-0.21
...	

=

	-
S1	-0.041
S2	-0.024
S3	-0.012
S4	-0.039
S5	-0.050
S50	-0.062

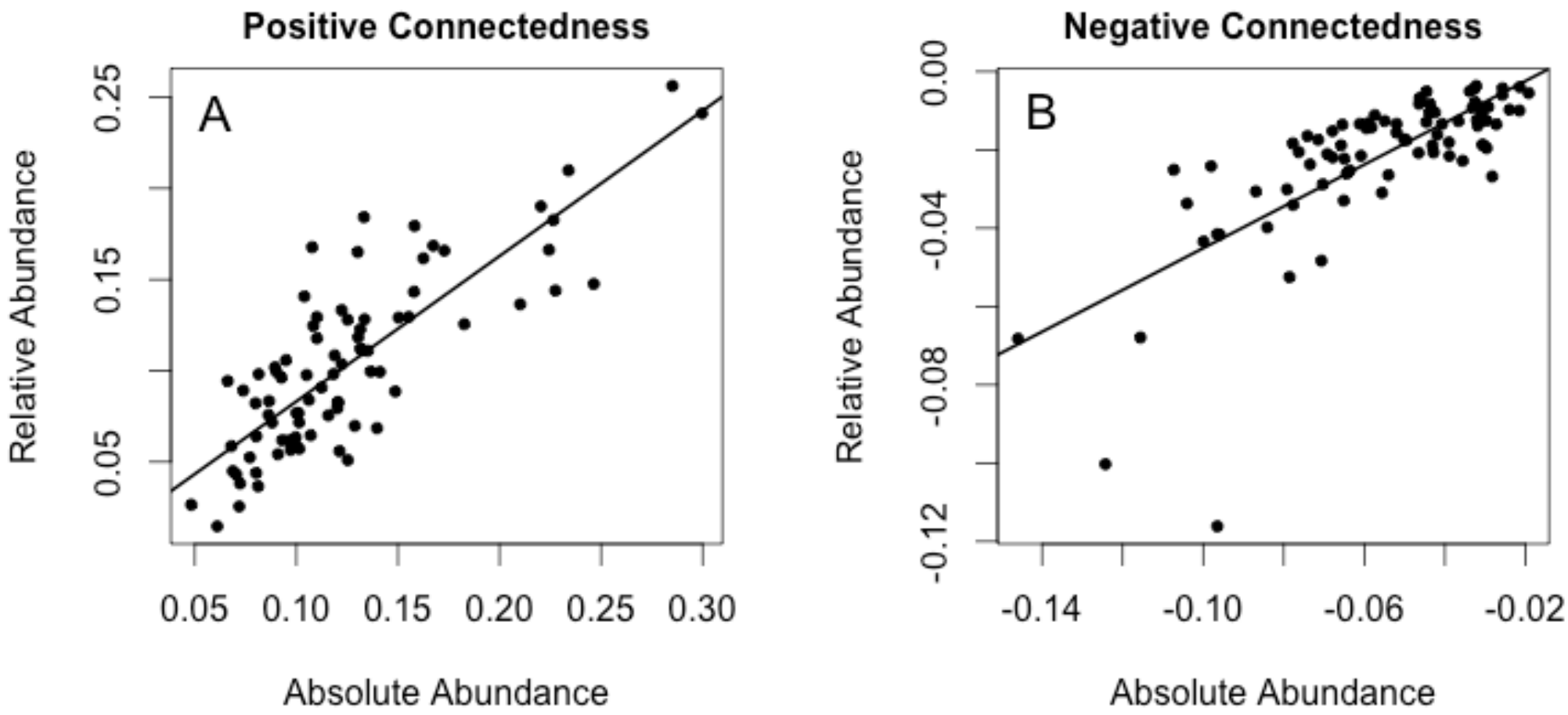
Cohesion Metrics

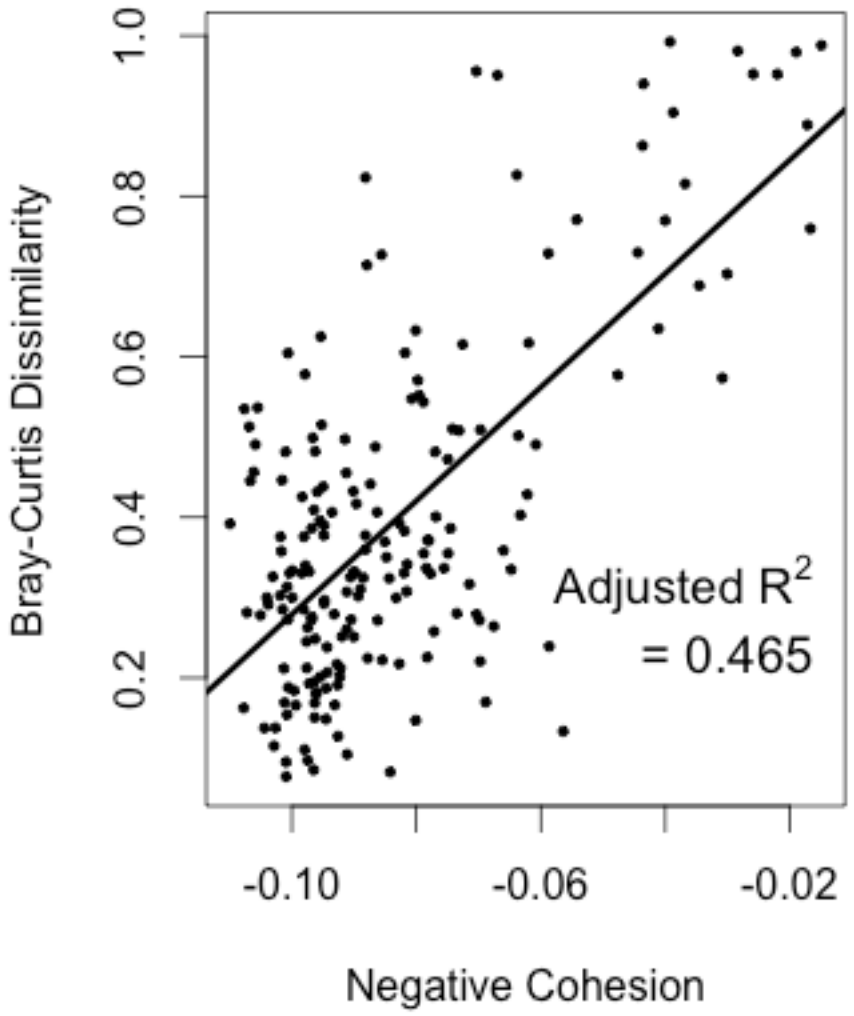


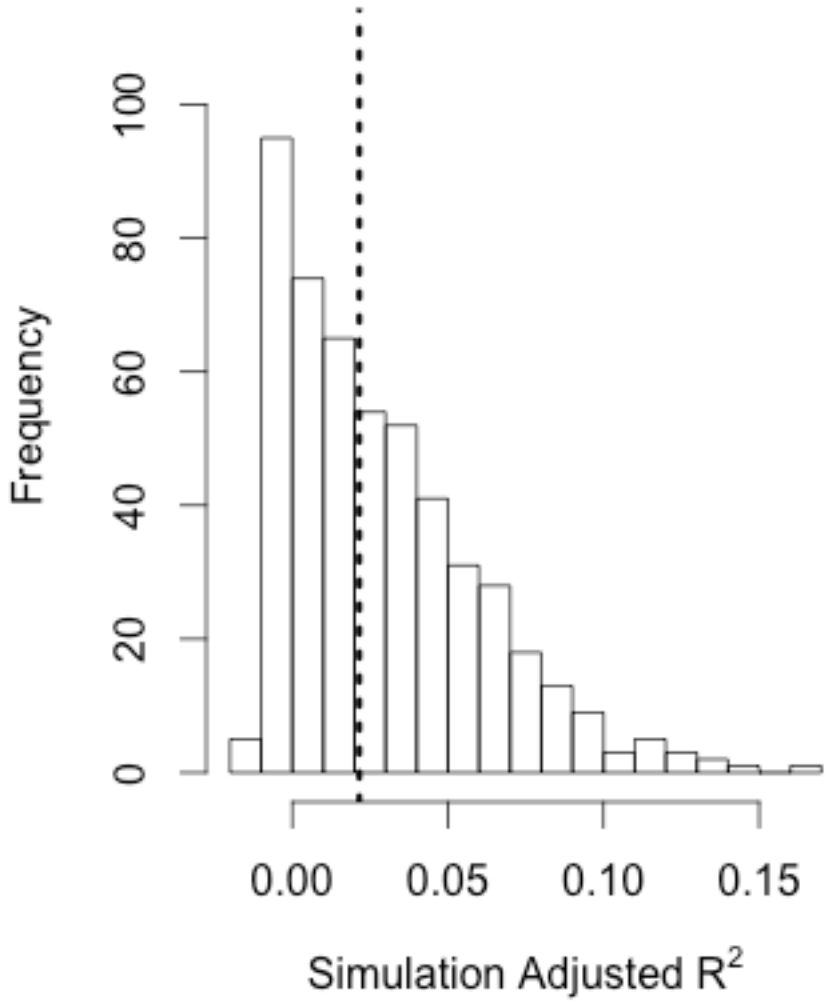
Median Correlation Matrix

Repeat above for each taxon

	A	B	C	...	Z
A	1	0.12	-0.02		0.06
B	0.15	1	0.01		-0.03
C	-0.06	0.03	1		0.05
...					
Z	0.05	-0.04	0.03		1


Positive Connectedness
A
Relative Abundance
0.05
0.15
0.25
0.05
0.10
0.15
0.20
0.25
0.30
Absolute Abundance
Negative Connectedness
B
Relative Abundance
0.00
-0.04
-0.08
-0.12
-0.14
-0.10
-0.06
-0.02
Absolute Abundance


Bray-Curtis Dissimilarity
0.2
0.4
0.6
0.8
1.0
Adjusted R^2
= 0.465
-0.10
-0.06
-0.02
Negative Cohesion


Frequency
0
20
40
60
80
100
0.00
0.05
0.10
0.15
Simulation Adjusted R^2