readme for Cohesion R Script:

script maintainer: Cristina Herren, cristina.herren@gmail.com

*While developing the cohesion analysis, we tested our workflow on many different datasets. These are some suggested best practices and diagnostics for using the cohesion metrics.*

Suggestions for setting abundance and persistence thresholds:

* The cohesion R script includes a parameter (pers.cutoff) to exclude taxa from the analysis if they are present in fewer than a specified proportion of samples. The purpose of this parameter is to exclude taxa for which reliable correlations cannot be calculated. As a guideline, taxa should be present in at least 5-10 samples to be included in the analysis. Thus, if your dataset has 50 samples, the pers.cutoff parameter might initially be set to 0.1, as to exclude taxa present in fewer than 5 samples.
* For samples with deep sequencing, rare taxa (taxa with very low abundances) are present in a large proportion of samples. In this case, it may be optimal to cut out taxa from the analysis based on mean abundance. As a guideline, the abundance threshold might be initially set to exclude, on average, 5-10% of the community. In our analyses, this approach dramatically reduced the number of taxa included (thereby also making the script run faster) while retaining the vast majority of the community.
* When using cohesion as a predictor variable in analyses, we found that analysis results were often qualitatively similar across a range of abundance and persistence thresholds (see sensitivity analysis SOM). Thus, we suggest trying various thresholds and selecting final parameter values within the range of values where results are stable.

Suggestions for correcting or importing the correlation matrix:

* While creating the cohesion workflow, we tested dozens of null models for correcting correlations between taxa. We selected a final version based on what worked well for a variety of different datasets. However, we imagine that the default null models included in the script might not be ideal for every dataset, because microbial datasets vary in richness and evenness. You can test the influence of the null model on the cohesion metrics by importing the true (uncorrected) correlation matrix as a custom correlation matrix. This will bypass the null model and calculate connectedness on the uncorrected correlation matrix.
* We found that the row shuffle null model worked better as sample evenness increased. Conversely, the column shuffle null model worked better when there were taxa that consistently comprised > 20% of the community. In these cases, it was important to maintain taxon mean abundances in the null model, which only occurs in the column shuffle null model.
* If importing a custom correlation matrix, we urge users to consider the aim of the analysis used to generate the custom matrix. Correlation methods that are intended to determine significance may not yield appropriate correlation matrices. For example, in Local Similarity Analysis (LSA), LS scores do not equate to significance; a lower LS score can be more significant than a higher LS score. This is not a deficit of the Local Similarity method, because the aim of LSA is to identify significant pairwise correlations. However, the differing objectives of these two analyses (LSA and cohesion) mean that the matrix produced from LSA may not be an appropriate custom matrix.

Types of datasets appropriate for the cohesion metrics

* Time series datasets from a single location
* Time series datasets from multiple similar locations (e.g. water samples from across a watershed or from many groundwater wells)
* Spatial datasets (e.g. soil samples across a landscape)
* Host-associated microbiome samples from a single individual/organism (e.g. stool samples from a single human subject)
* Host-associated microbiome samples from multiple individuals/organisms within the same population (e.g. stool samples from many human subjects)
* Experimental datasets (e.g. calculating cohesion metrics or the relationship between cohesion and a response variable in two or more experimental treatments)

Types of datasets that may be problematic for the cohesion metrics

* Time series collected at high frequencies (such as a daily time scale)
  + An implicit assumption of this analysis is that there is sufficient time between samples for populations to change enough in relative abundance such that the real population change can be distinguished from background/methodological noise in the data. This assumption may not be met at a daily sampling interval.
  + Another problem that may arise in time series at high frequency is autocorrelation. Autocorrelation may inflate correlations between taxa. One possible way to account for autocorrelation would be to calculate connectedness using the first differences of the dataset. Correlations on the first differences would indicate whether changes in two populations are synchronous.
* Datasets containing samples from different sites where most OTUs are not shared.
  + OTUs being present at one site but absent at others may generate spurious correlations between taxa.

Diagnostic tests when cohesion is used as a predictor in a regression

* Test for autocorrelation in residuals. As a simple diagnostic, plot sequential residuals against one another. Test for significant correlation between sequential residuals.
* If analyzing a time series, test for trends in residuals over time.
* If analyzing a time series, split the dataset into two parts corresponding the first half of the time series and the second half of the time series. Fit separate regressions for the two halves, and test whether the slope estimates for cohesion are significantly different.
* Cross validation: randomly split dataset into two halves many times. Use one half to train the model, and the other half to calculate the error in predicted values from the linear model. Compare the residual error in the fitted model to the error in the predicted values.
* Test the significance and model R2 value of cohesion metrics when cohesion metrics are used to predict dynamics of random samples. To do this, randomize the order of your samples. Conduct the same analysis (such as a regression of cohesion predicting Bray-Curtis dissimilarity) using randomly ordered samples. Repeat many times to get a distribution of results. Cohesion should be more significant when used to predict dynamics in ordered samples, versus randomized samples.
* The most abundant taxon in microbial communities can comprise more than 10% of the entire community. To test how strongly the most abundant taxon contributes to cohesion values, manually set the connectedness values of the most abundant taxon to zero. Re-calculate cohesion, and re-run the analysis using these new cohesion metrics.

We suggest not rarefying datasets before calculating cohesion

* We found that rarefying decreases the strength of pairwise correlations. We tested the effects of rarefying on pairwise correlations using a dataset of bacterial samples from bog lakes in northern Wisconsin (available at the Earth Microbiome Project, study ID 1288). This dataset contained replicate samples at many time points. We calculated correlations between the same OTU in replicated samples in the non-rarefied dataset and in a rarefied dataset. These correlations should be close to 1. We found that rarefying consistently decreased the magnitude of these strong correlations. We also found that rarefying increased the standard deviation of OTUs due to the stochasticity of rarefying. In the equation for Pearson correlations, the denominator is the product of the standard deviations of the two populations. Thus, the mathematical reason for the decreased correlations is the increase in the standard deviations of the OTUs without a compensating increase in the covariance of the two populations.

Miscellaneous

* The null models implemented in the cohesion R script are stochastic due to random sampling. To generate reproducible results, manually set a seed from a predetermined vector before each randomization.
* When running the script on a datasets where correlations for 500 taxa were calculated, the script took 30-60 minutes to run. The run time is directly proportional to the number of iterations (iter). We found that using iter = 40 was sufficient during parameter optimization (i.e. while determining the persistence cutoff and type of null model to use), but that a larger value (iter = 200) was best to generate final results.
* If using DNA sequencing data, try clustering the sequences based on different percent similarity cutoffs.