**README- Statistical Analyses of LOSCO Taxonomic Profiles**

Summary:

Several statistical analyses have been completed using the taxonomic profiles from the entire sample data set generated by 16S rRNA gene sequencing of the V1V3 and V3V5 regions.

1. Genera based results from classification of reads to a reference 16S rRNA gene sequence database (taxonomy dependent) for V1V3 and V3V5
2. Operational Taxonomic Units (OTUs) generated from the clustering of sequences based on sequence similarity (taxonomy independent) for V1V3 and V3V5

The three key analyses and their descriptions are given below:

1. **Permutational Multivariate ANOVA (PERMANOVA)**

PERMANOVA is an analysis of variance (ANOVA) method that uses bootstrapping to estimate the statistical significance of the effects of various factors on the distances between samples. The variance being analyzed is based on the distance metrics (i.e. Euclidean distance) which is computed between all pairs of samples. Each of these distances represents a positive scalar value that is a collapsed down representation of the magnitude of difference between two multidimensional profiles. Because these distance values are not independently, nor identically normally distributed, bootstrapping is necessary to perform this ANOVA. A linear model uses the computed distances as the response variable, and the factors as the predictors. P-values are estimated with bootstrapping by resampling with replacement the samples included in the analysis.

The output files are organized in a manner as outlined below:

PCA and MDS Ordination plots: (P.1) These two ordination plots visualize the relative differences between samples by positioning them in 2-D space. The principal components analysis (PCA) plot positions each sample according to the two largest independent components (largest eigenvalues) of variance across the samples. The multidimensional scaling (MDS) plot positions samples according to a reduced dimensional (2D) representation of the distances between samples. In both the PCA and MDS plots, the samples are color-coded based on the factor levels where each page focuses on a separate factor.

MDS plots: (P.2) The MDS plot on the right hand side plots the centroid of the samples for each factor level. These plots are rotated so that the centroid of the first factor level is on the left hand side of the second factor level. The MDS plot on the left hand side have the sample names labeled according to the rotation/repositioning of the factor level centroids.

Contributions to variation: (P.3) This bar plot represents the contributions to variation (residuals) of each sample. The samples are ordered from the largest contributors to the smallest contributors. When the linear model is fit, samples that are further away from their predicted distances contribute more to the variation in the model.

1. **Multivariate Linear Regression**

The effect of factors across specific taxon can also be estimated by performing a multivariate linear regression. Because the profile of each sample is represented as a composition (proportion) of taxa, to perform a multivariate regression, the compositional representation must first be transformed into independent and normally distributed values. This disrupts the artificial negative correlation among taxa due to the requirement of the taxonomic proportions necessity to sum to 1 and produces a distribution that is significantly more Gaussian. The transformation that is applied is called the "additive log transform" (ALR). First, a subset of the most abundant taxa are selected, e.g. the top p = 20 taxa may represent 71% of the average cumulative abundance across all the samples. Next, for each of the selected taxa (p), their abundances are divided by the sum of the remaining abundances not utilized. This ratio of individual taxa abundance to unused taxonomic abundance is then logarithmically transformed, yielding values that are closer to normally distributed. The multivariate linear regression is then performed on these transformed values.

The output files are all organized in a similar manner as outlined below:

Factor Correlations (P.1): Factors that are continuous or categorical, but ordinal, are pairwise analyzed for correlation. Factors that are highly correlated need to be removed from the model because they contribute to multicollinearity. This prevents a stable estimate of both the predictor coefficients and standard errors.

Reference Factor Levels (P.2): Here the reference factor levels are summarized. In this case, the reference factors to which all others are compared to are given. In this case, LEM and SCR are modeled relative SKM (“healthy” swabbed skin), male (M) is modeled relative to female (F), all fish species are modelled relative to Red Snapper, and all stations are modelled relative to East Louisiana Offshore.

Multivariate Analysis of Variance Table (P.3) Multivariate ANOVA table for factors tested using Pillai's trace statistic. P-values are uncorrected for multiple factors. Pillai's trace, as well as other MANOVA test statistics, take the variance-covariance among the multivariate response variable (the ALR transformed abundances) into account when determining the significance of the factors effect across all taxa.

Taxonomic Correlations (P.4) A heat map is shown which summarizes the correlations among the top 20 taxa based on ALR transformed abundances. The cumulative abundance that these 20 taxa represent is given on the top of P.3. Positive or negative correlations that are large in magnitude suggest cooperation or competition, respectively. Taxa with correlations consistently close to zero may suggest taxa with minimal interaction with other organisms.

Univariate ANOVA Summary (P.5) A heat map is shown which summarizes the p-values for each of the factors against each taxa. Low values indicate higher statistical significance. These are not corrected for multiple testing nor do they take into account correlation among the taxa that the multivariate ANOVA statistics provide. These values are provided as a diagnostic for relative statistical significance of factors on individual taxa and should not be considered accurate estimates of actual significance.

Univariate R2 Summary (P.6) The R2 and adjusted R2 (corrected for the number of factors/predictors in the model) for each taxa. Higher values, towards 1.0, indicate a better model fit suggesting that a larger proportion of the variance can be attributed to the predictors supplied in the model. Because a high predictor-to-sample count ratio may contribute to "overfitting", the adjusted R2 calculation uses the number of predictors to penalize/reduce the R2 value. The adjusted R2 may become negative in particularly poor models.

Univariate analyses(P.7- end) These remaining reports provide a breakdown of the univariate regression fits for each of the top selected taxa, ordered decreasingly starting with the taxon with the greatest average abundance.

a.) Univariate ANOVA

i.) Taxon name

ii.) Mean abundance of taxon across all samples

iii.) R2, and adjusted R2

iv.) ANOVA table

b.) Univariate regression results:

i.) For factors with non-continuous or ordered levels, a coefficient (Estimate) is computed for each factor level with respect to the reference factor level. (Page 2)

ii.) If the factors are continuous or ordered, then only a single coefficient needs to be computed.

iii.) These p-values are testing whether the coefficients are non-zero.

c.) Marginal model and model fit plots

i.) Marginal Model Plots: When factor levels are continuous, a plot of the factor levels versus the ALR transformed abundance is provided. The blue line is a LOWESS fit, essentially a smoothed fit of the observed data. The red line represents what the linear model predicted. The closer the LOWESS fit is to the model fit, the better the factor contributes to the linear model.

ii) Model Fit Plot: The last plot (or the only plot if there were no factors with ordinal levels) is a plot of the observed ALR transformed abundance versus the predicted or fitted ALR transformed values. The interpretation is similar to the marginal model plots in that the blue line represents the observed data, and the red line represents the linear model. The more similar the red and blue lines, the better the fit of the model.

Note that even when the observed (blue) and model (red) fit curves overlap, the model may still be weak (with a low R2). The overlapping of the two lines only provides evidence that the observed data can be modeled with a linear model. To acquire a strong R2, the points also need to fall on the red line. If the red and blue lines appear to diverge, for example if the blue line is an arc and the red line is straight (it is always straight in a linear model), then the data may need to be further transformed to improve the model's fit. If the lines in the marginal model plot diverge, then the factor may need to be transformed. If the lines in the model fit plot diverge, then the response (i.e. observed ALR transformed abundance), may need to be further transformed. When the R2 is low due to points falling off the red line in the model fit, then this may only be improved by identifying additional predictors that may explain this variation.

1. **Univariate Regression of Taxonomic Diversity:**

The taxonomic diversity can be represented with several alpha diversity indices. Those being analyzed here include: Tail, Shannon, Simpson, Evenness, and Simpson's Reciprocal. Each diversity index has its own mathematical interpretation, however in general, they tend to be highly correlated, as one should expect. Because each diversity measure is implemented with an alternative calculation for representing diversity, some of which may contribute to a better model fit, all will be analyzed. Note that the concept of diversity focuses on the distribution of taxa in a sample. Thus, two samples with very different compositions of taxa may have the same measured diversity if the shape of their rank abundance curve is identical.

In these analyses, a univariate model is fit for each of the alpha diversity indices based on the factors that have been specified in the model. The output will look similar to the MLR output format. The goal here is to determine if the factors have had any effect on the taxonomic diversity across the samples.

A key step that needs to be performed before running the regression analysis between diversity indices and the predictors is the transformation of the diversity measures. This is necessary because the residuals from fitting a linear model for these diversity indices may not be normally distributed. After the diversity indices have been transformed, the subsequent p-values for the effect of each of the predictors will be more accurately calculated. The transformation that is applied is called the Box-Cox transformation. In short, Y (the response) is transformed into Y', with the formula Y^lambda. Lambda is identified computationally by iterating over a range of candidate values, and then identifying the value of lambda which minimizes the sum of the residuals of the model. This is performed independently on each of the alpha diversity measures.

Factor correlations: (P.1) Used to identify potential collinearities across the factors.

Diversity Index Histograms: (P.2) These plots represent the distribution of each of the diversity indices across the input data set. Here, you can see how the range of each diversity index is compressed or expanded, given the same domain of samples profiles. When a range is expanded, it allows the regression to be more sensitive towards samples across those regions.

Identification of the Box-Cox Lambda: (P.3) These plots exhibit the range of lambda's that were tried in order to determine its optimal value for each diversity index. The lambda that maximizes the log-likelihood is the lambda is that selected for that diversity index. The closer lambda is to 1, the smaller the impact of the Box-Cox transformation because the residuals of the model fit for the untransformed diversity indices are already close to normally distributed.

Transformed Diversity Index Histogram: (P.4) These plots demonstrate the distribution of the diversity indices after they have been transformed. The transformed indices will frequently become more symmetric after their transformation.

Summary of factor levels: (P.5) The reference factor levels and sample counts across the factor levels are summarized here.

Univariate Regression Results and Marginal model plots: (Beginning on P.6) For each of the diversity indices, the regression results and marginal model plots are generated.

Summary of the R2 values: (Last page) A summary of the R^2 values are provide for each of the diversity indices.