CORBATA

CORe micro**B**iome **A**nalysis **T**ools

Overview

This document describes the list of analyses that can be performed with Corbata.

All scripts will report their usage and necessary parameters when invoked without any options specified.

There are essentially two types of analyses that can be performed:

1.) Single cohort analyses
    a. The single cohort analyses are performed on a single group of samples. These include identifying the number of core in a cohort, variance-abundance (Var-Ab) plot, and the Ubiquity-Abundance (Ub-Ab) plots. By being single cohort analyses, they are performing any hypothesis testing. Rather, they are describing the cohort and helping to identify taxa of interest. The Var-Ab plot identifies core and minor core.
2.) Two cohort analyses
    a. The two cohort analyses are performed on a pair of cohorts. These include the Ubiquity-Ubiquity (UU) plot and computing the AWKS statistic. The UU plot helps you visualize the difference between two groups, and the AWKS statistic is a statistic that quantifies the differences seen in the UU plot so statistical significance can be established.

Scripts can be run by themselves by going into the lib directory and identifying the analysis script of interest, or they can be applied as a suite of analyses so that less typing is necessary. Its recommended to run the two scripts, Run_Two_SampleComparisons.pl or Run_Single_SampleAnalyses.pl, and exploring the resulting output, rather than trying to run the individual scripts unless you are having problems. Nonetheless, the individual scripts that were written are self-contained.

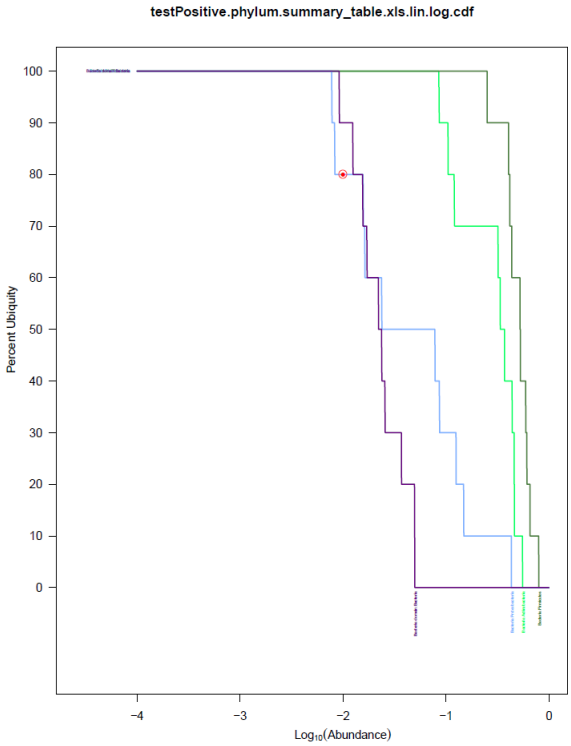
See the descriptions in the doc directory for more information.

Inputs

The various core analyses start with a user-generated tab-delimited summary file. The first and second column are reserved for recording the sample IDs and the total number of reads associated with the corresponding sample respectively. Additional meta-data in the summary file include the header (first row) which gives the taxonomy classification. The count data for the corresponding sample and taxonomy is used to fill the remaining portions of the table.

Ubiquity-Abundance Plot

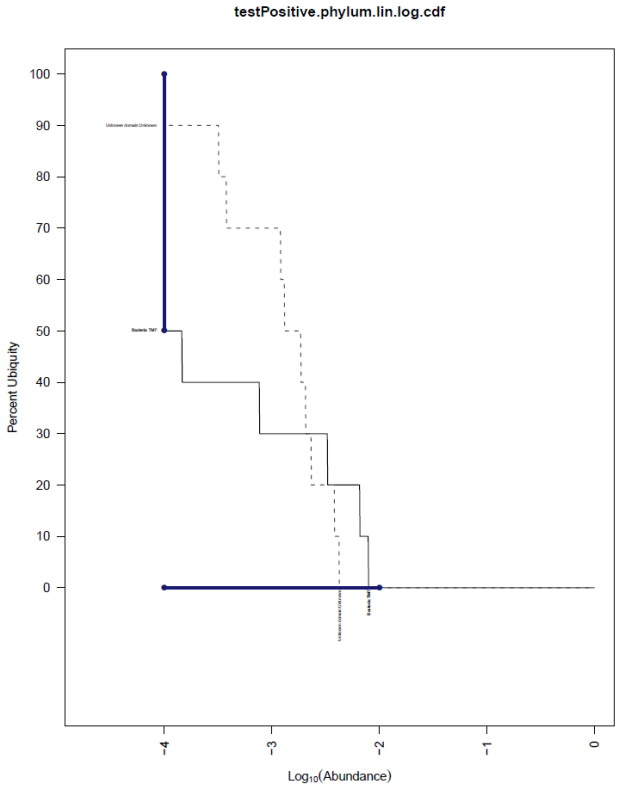
See **UbAb_Example.pdf** on how to generate this plot.



Plot above was generated based on the CDF values for each taxon with default cut-offs of ubiquity=0.8 and abundance=0.01. The y-axis represents the ubiquity percent (% of samples present) while the x-axis represents the abundance (log10 transformed). Each line is a representation of taxa that are above the specified cut-offs. The red “bulls’ eye” indicates where the cutoff is found.

Minor Core Ub-Ab Plot

The minor core analysis is also described in **OtherCore_Example.pdf**. This analysis highlights the minor-core using the previously described Ub-Ab plot, however with slightly different visualization highlights. The minor core is defined here as taxa with "Low Abundance but High Ubiquity". The default cut off values used are ubiquity=0.5 and abundance=0.01. Thus taxa which have ubiquity>0.5 and abundance<0.01 are considered part of the minor core.



In the plot above, y-axis represents the ubiquity percent (% of samples present) while x-axis represents the log transformed abundance. Each line is a representation of taxa which has ubiquity>0.5 and abundance<0.01 across the cohort. The blue bars represent the where the taxonomic curves most originate and end in order to be considered part of the minor core.

Number of Core

See **NumCore_Example.pdf** on how to generate these statistics. This analysis estimates the number of core found in the cohort based on the user-specified cutoff. Default values are generated at ubiquity=0.8 and abundance=0.01. The output looks like:

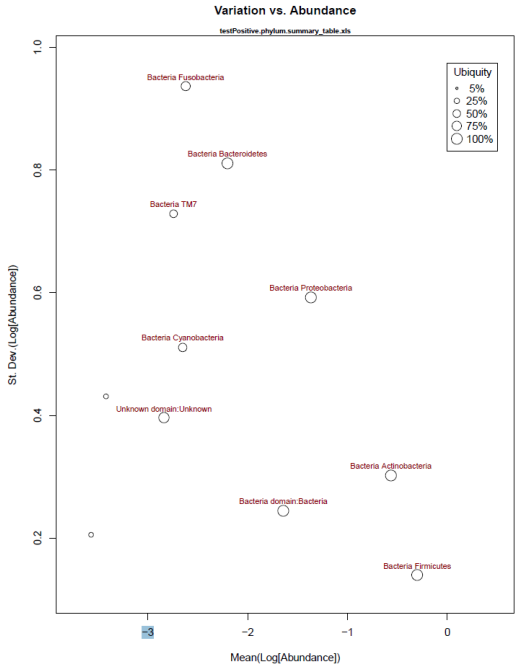
```
FilenameRoot, NumCore, Median, LowerBound, UpperBound, alpha

testPositive.phylum.summary_table.xls, 3, 3, 2, 4, 0.05
```

In this example, for the given file, there were 3 core taxa observed. The median (3), lower (2) and upper bound (4) estimates for the number of core taxa are based on 100x(1-alpha) where alpha=0.05, ie. 95% confidence interval.

Taxonomic Variation

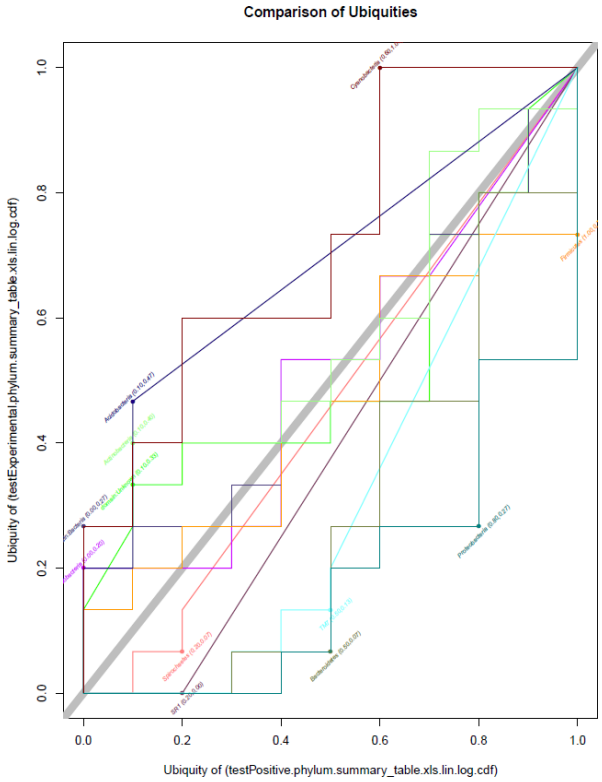
This analyses looks at the relationship between variation and abundance for each taxa. See **VarAb_Example.pdf** for examples on how to generate this plot.



The taxonomic variation plot is a scatter plot that separates each taxon by its average abundance and variation across the cohort. Since the majority of taxonomic abundances for complex environments are less than 1% and appear to decrease following a power distribution, a logarithm base-10 transform is first applied to the abundances to encourage the distribution of abundances to be more normally distributed. After this is done, the standard deviation of the log(abundances) is computed to estimate the amount of variation for each taxa across the cohort. Following sampling theory, the amount of variation in samples from a fixed proportion should exhibit depends on the average proportion and number of reads, (p*(1-p))/(n-1). As a result, one should expect the amount of variation of a taxon to increase as abundance increases. Interestingly, when a taxon is core, it will tend to be found on the bottom right of the variation plot, where it has a greater abundance and less variation across the cohort. The variation plot helps to identify core and also members in the cohort with unusually high variation as well. The ubiquity, from 5-100%, of the taxa across the cohort is also represented based on the size of each glyph.

Ubiquity – Ubiquity Plot

Ubiquity-Ubiquity plot compares the microbiomes between two cohorts. Two summary files are needed to generate the plot. See **UU_Example.pdf** for additional details on how to run this script.



The plot above compares the ubiquity values for two datasets at matching abundance levels. The thick grey reference line (with a slope of 1), represents matching ubiquities between the two cohorts at all abundances for a taxon of interest. If a taxon's classification line deviates above the grey line, then that taxon is more ubiquitous (present in more samples) for the dataset represented along the y-axis. Contrariwise, if a taxon line deviates below the grey reference line, then it is more ubiquitous for the dataset along the x-axis. The highlighted dots for each classification e.g. ***Acidobacteria (0.10, 0.47)*** indicates the maximum difference of ubiquities between the two datasets. For Acidobacteria, the maximum difference of ubiquity along matching abundances was 0.47 (testPositive) - 0.10 (testExperimental)= 0.37.

Abundance-Weighted Kolmogorov-Smirnov (AWKS) Statistic

AWKS statistic complements the Ubiquity-Ubiquity plot by providing a value which represents the magnitude of the difference between the two cohorts. See **AWKS_example.pdf** for additional details on how to run this script.

The key output is the AWKS test statistic and the estimated p-value, based on bootstrapping.

In this example, the two cohorts were not statistically significantly different from each other.

Output example: