Overview

This collection of scripts and metadata was used to perform analysis on the pristine mangrove tidal zone set in CO Santana et al. 2020. References to Supplemental Files and Figures are in reference to that article.

Data Accessibility

The original reads are accessible at NCBI under BioProject PRJNA608697: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA608697

Read Trimming

Trimmomatic was used to filter and trim demultiplexed sequences (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:100).

A minimum average read quality score of 15 was required for inclusion while the sliding window cuts any read at the point where the median quality score over a 4 nucleotide window is less than 15.

Read-pairing

This section summarizes the "Read_preprocessing.sh" (SF6 Supplemental file 6, Read_preprocessing.sh) and "jantar.py" (SF6 Supplemental file 6, jantar.py) files.

QIIME2 uses an implementation of vsearch for read pair joining ('vsearch join-pairs') which has a lower criteria of a minimum of 10 base-pairs of overlap. To extend vsearch with accurate read-pair joining below this cut-off we used a custom script (jantar.py) that would attempt to combine read-pairs that had failed to combined with vsearch. In order to be combined a read pair must have a "perfect" (starting from tails of reads, with no mismatches, in the correct orientation) 7bp overlap or 10bp with a single mismatch.

QIIME2

This section summarizes the "QIIME_run.sh" file (SF6 Supplemental file 6, QIIME_run.sh).

Reads were denoised using 'dada2 denoise-single', default settings using '--p-trim-left 3 and --p-trunc-len 0'. Phylogeny was determined using 'phylogeny align-to-tree-mafft-fasttree' with default settings. Minimum (9800) and maximum (40153) read depths were determined using 'feature_table.qzv'. Alpha-rarefaction was then calculated using the rooted MAFFT tree and the maximum read depth.

Alpha and Beta diversities were calculated using 'diversity core-metrics-phylogenetic' using the minimum read depth ('--p-sampling-depth') and the rooted MAFFT tree.

Taxonomy was calculated using 'feature-classifier classify-sklearn' and plotted using 'taxa barplot'.

Finally, gneiss was used to calculate the hierarchical correlation-clustering 'gneiss correlation-clustering' using a pseudo-count of 1. This was plotted using 'dendrogram-heatmap' for the meta-data column 'Site'.

PICRUSt2

This section summarizes the 'PICRUSt2_run.sh' file. (SF6 Supplemental file 6, PICRUSt2_run.sh)

PICRUSt2 was run twice using '--stratified' and '--per_sequence_contrib' options. One run was set to limit NSTI thresholds to 0.15, while the other (and data set subsequently used) was ran with the default NSTI threshold of 2.0.

Taxonomic abundance analysis

This section summarizes the use of the custom python script 'novembro.py' and 'Novembro_run.sh' (SF6 Supplemental file 6, novembro.py)

Novembro.py normalizes the ASV abundances generated by QIIME2 into their corresponding taxa.This requires feature-table.tsv and taxonomy.tsv files. While the taxonomy.tsv file an be found by unzipping the taxonomy.qza file the feature-table.tsv needs to be generated using a biom conversion command (eg. 'biom convert -i feature-table.biom -o feature-table.biom.txt --to-tsv')

Novembro.py iterates from the lowest to the highest taxonomic level combining ASV abundances into their corresponding taxa. Replicate normalizations are downsampled to match the lowest ASV abundant replicate. We use novembro.py to calculate zone specific taxonomic enrichment. This is accomplished by first using chi2 to compare the replicates of each zone against the others. To be deemed significant a zone must have an greater than a 5% effect size and p-value less than, or equal to, 0.05 relative to both other sites.

Taxa that are found to be significantly enriched are saved to a tab-delimited file along with their log10 transformed abundances.

Zone specific differential KO abundances

This sections summarizes the use of ALDEx2, an R package, that uses GLM to identify differential expression abundances across sites using replicates (SF6 Supplemental file 6, sigilo.py).

First we generate Monte Carlo samples of each KO distribution of the unstratified metagenome predictions (eg. 'pred_metagenome_unstrat.tsv') using ALDEx2's centered log-transformed ('aldex.clr') module. We then use ALDEx2's GLM ANOVA test ('aldex.kw'), taking those KOs with an expected p-value ('glm.ep') of less than, or equal to, 0.05 to be significant.

Potential functional abundance analysis

This section summarizes the use of the custom python script 'sigilo.py' and 'Sigilo_run.sh'(SF6 Supplemental file 6, sigilo.py)

Sigilo.py performs several functions to aid in the visualization and analysis of PICRUSt2 and ALDEx2 generated KO data.

Generate heatmap: KO enrichment heatmaps can be generated using '--generate_heatmap', this function combines the predicted functional abundances of KOs identified by ALDEx2 as significant into their corresponding KEGG Ortholog pathways, performs a log10 transform and plots them as heatmaps.

Correlate ASV with Functional Abundance: Using the '--asv2fa' command we first combine all ASVs into their corresponding taxa at a given level (for Family, level = 4) and then calculate the total functional abundance for those combined ASVs to derive the predicted functional abundance of a given taxa.

Correlate ASV with NSTI: Using the '--asv2nsti' command we first combine all ASVs into their corresponding taxa at a given level (for Family, level = 4) and then combine the associated NSTI values for those ASVs for each taxa. We can then use these to identify the mean, median, and standard deviation of NSTI scores for each given taxa.