Sample Acquisition and Processing
cruises KM0703, X0804, KM1309 (2007, 2008, 2013)

Trichodesmium sequence accessions: https://www.bco-dmo.org/dataset/716817

Project: Dissolved Phosphorus Processing by *Trichodesmium* Consortia: Quantitative Partitioning, Role of Microbial Coordination, and Impact on Nitrogen Fixation (P Processing by Tricho)
Principal Investigator: Dr Sonya T. Dyhrman (Lamont-Doherty Earth Observatory, LDEO)
BioProject PRJNA314461
2017-10-27

Sampling:
Trichodesmium colonies were collected from the near surface (approximately within the top 25 m) using a handheld 130 µm net. Single colonies were picked and transferred into 0.2 µm filtered local surface water collected at 5 m with a Rosette sampling device. To avoid potential contamination, and to assure that only organisms tightly associated with *Trichodesmium* were sequenced, colonies were rinsed in fresh 0.2 µm filtered local surface water three more times. Then, colonies were separated into two morphologies, "rafts" (colonies with a parallel organization of trichomes), and "puffs" (colonies with a radial organization of the trichomes). Between 10-20 washed *Trichodesmium* puffs or rafts were filtered onto 47 mm, 10 µm polycarbonate filters, which were then placed in 2 mL cryovials, snap-frozen, and stored in liquid nitrogen until DNA extraction was performed.

Processing:

DNA from each sample was sent to Argonne National Laboratory (Lemont, IL, USA) for paired-end sequencing (2x150bp) of partial 16S rDNA genes using the Illumina Miseq platform. Genomic DNA was amplified using the Earth Microbiome Project barcoded primer set, adapted for the MiSeq platform by adding nine extra bases in the adapter region of the forward amplification primer that support paired-end sequencing. As suggested by Kozich et al. (2013) and Mizrahi-Man et al. (2013) for short-read sequencing strategies, the V4 region of the 16S rDNA gene (515F-806R) was amplified, using region-specific primers that included the Illumina flowcell adapter sequences. The reverse amplification primer also contained a twelve base barcode sequence for the later distinction of individual sample sequences (Caporaso et al., 2011, 2012). Each 25 µl PCR reaction contained 12 µl of MoBio PCR Water (Certified DNA-Free), 10 µl of 5 Prime HotMasterMix (1x), 1 µl of Forward Primer (5 µM concentration, 200 pM final), 1 µl Golay Barcode Tagged Reverse Primer (5 µM concentration, 200 pM final), and 1µl of template DNA. The conditions for PCR were as follows: 94°C for 3 minutes to denature the DNA, with 35 cycles at 94 °C for 45 s, 50 °C for 60 s, and 72 °C for 90 s; with a final extension of 10 min at 72 °C to ensure complete amplification. The PCR amplicons were quantified using PicoGreen (Invitrogen, Carlsbad, CA, USA). Once quantified, different volumes of each of the products were pooled into a single tube so that each amplicon was represented equally. This pool was then cleaned up using UltraClean® PCR Clean-Up Kit (MoBIO), and then quantified using a Qubit (Invitrogen). After quantification, the molarity of the pool was determined and diluted down to 2 nM, denatured, and then diluted to a final concentration of 6.75 pM with a 10% PhiX spike for sequencing on the Illumina MiSeq. Sequencing was performed as described previously (Caporaso et al., 2012).

rDNA 16S sequences were processed using modules implemented in the Mothur v.1.34.0 software following Kozich et al. (2013). Briefly, any sequences with ambiguous bases or homopolymers longer than 8 bases were removed from the data set and sequences were trimmed to a uniform length of 253 bp. Sequences were aligned using the SILVA-compatible alignment database available within Mothur and chimeric sequences were removed using Uchime (Edgar et al., 2011). Unique sequences were then classified by a Bayesian approach using the Mothur-formatted version of the RDP training set (v.9) with an 80% cutoff. Sequences unique to *Trichodesmium* sp. (624,599 sequences, 67.7%) were extracted and assigned to three different *Trichodesmium* clades (Clade I, III and IV) by mapping to rDNA 16S sequences obtained from Hynes et al. (2012), after Janson et al. (1999) and Lundgren et al. (2005) (Table S2), using the classify.seqs command implemented in Mothur. Herein, Clade I is represented by sequences mapping to T. thiebautii and T. tenue Z-1, Clade III is represented by sequences mapping to T. havanum F34-5, T. erythraeum K-02#2 and T. erythraeum 21-75, and Clade IV is represented by sequences mapping to T. contortum and T. tenue (accession numbers AF013028 and AF013029, respectively). A bacterial-epibiont-only dataset (298,174 sequences, 31.8%) was created by removing any sequences classified as "chloroplast" (which included *Trichodesmium* sp. sequences and other photosynthesizers), "mitochondrial," "archaeal," or "unknown." Bacterial-epibiont-only sequences were clustered into operational taxonomic units (OTUs) based on 97% sequencing identity. Depth coverage was assessed using Good's coverage estimator calculated in Mothur. Venn diagrams were created within Mothur to visualize the number of OTUs shared between ocean basins or colony morphology. To analyze the effects of morphology and ocean basin on both epibiont community composition (species richness) and community structure (species relative abundances) a two-way Permanova was performed using a Bray-Curtis dissimilary distance matrix (Anderson, 2001). Permanova tests were also performed separately with both morphology and ocean basin as fixed effects in the *Trichodesmium* and the epibiont dataset and a Benjamini-Hochberg correction for multiple testing was applied, limiting the overall false discovery rate to 5% (Benjamini and Hochberg, 1995). Variance in epibiont community structure described by Bray-Curtis dissimilarities was visualized using principal coordinate analysis in R. Mantel tests testing for correlations between community dissimilarity and geographic distance were conducted for each colony morphology. All statistical tests were performed in R.

Metabolic predictions of the epibiont community were carried out on OTUs at the 97% similarity level, using PICRUSt (Langille et al., 2013) through the Galaxy server (Goecks et al., 2010). OTUs were assigned with Mothur against the Greengenes database (v5). The mean Nearest Sequenced Taxon Index, calculated by PICRUSt to provide a reliability estimation of the metagenome predictions, was in a normal range for all samples (0.099±0.04), according to Langille et al. (2013). The Linear Discriminant Analysis (LDA) Effect Size (LEfSe) algorithm (Segata et al., 2011) was used through the Galaxy server to identify significant differentially abundant microbial relevant pathways enriched in the puff or raft morphology within each ocean basin. KEGG pathways were considered differentially represented if their LDA score was higher than 2 (Segata et al., 2011). Raw sequences for each sample have been submitted to NCBI Sequence Read Archive, BIOPROJECT PRJNA314461 and SRA accession ID SRP072053.