Data Management Plan
P.I. Kendra Turk-Kubo (UCSC), Co-P.I. Kevin Arrigo (Stanford)
February 3, 2020

Project Name: ***Collaborative Research: Quantifying N_2 fixation rates of non-cyanobacterial diazotrophs and environmental controls on their activity***

Description of the expected data:

1. **Samples:**
   - *N_2 and C fixation rate samples:* Seawater samples from surface waters and from experimental manipulations will be collected from multiple field sites (San Diego, CA, and North Pacific Ocean) using $^{15}N_2$ and $^{13}C\text{-}HCO_3^-$ stable isotopes.
   - *MIMS samples:* Samples to measure the $^{15}N_2$ enrichment of seawater used for $^{15}N_2$ rate incubations will be collected in exetainers from each batch generated and stored at room temperature until processed by Sam Wilson (see letter, UH-Manoa).
   - *GeneFISH and CARD-FISH:* Samples for gene/CARD-FISH will be collected from surface water and experimental manipulations, and will be immediately fixed with formaldehyde, then and gently filtered onto >0.6 µm, >5 µm and > 20 µm polycarbonate filters, which will be dried then stored at -80°C until processing.
   - *Cell-specific N_2 and C fixation rate samples (nanoSIMS):* Samples for nanoSIMS will be chosen from gene/CARD-FISH samples. Gene/CARD-FISH samples will be visualized, then transferred onto gridded Si wafers, mapped and measured at Stanford University's nanoSIMS facility.
   - *DNA and RNA samples:* Samples will be gently filtered onto 0.2, 5, and 20 µm filters and immediately flash frozen. DNA and RNA will be extracted using a modified bead-beating protocol, and the quality of the extracts will be evaluated using a Bioanalyzer. DNA and RNA extracts will be archived at -80°C at UCSC. RNA extracts will be used to generate complementary DNA (cDNA), which will be stored at -80°C. DNA extracts and cDNA will be used in PCR amplification and in qPCR/ddPCR assays. Large volume samples from the 5-20 µm size fraction will be taken from stations with high abundances of gamma A for metagenome sequencing and construction of a gamma A metagenome assembled genome (MAG).
   - *Chemical and Biological samples:* Hydrographic and biogeochemical samples from surface waters and from experimental manipulations will be subsampled for characterizing the autotrophic community via flow cytometry, measuring primary productivity, particulate concentrations (POC/PON and Chl *a*), O_2 concentrations, DOM inventories, and nutrient concentrations (nitrite+nitrate, silicate, phosphate).
   - *NCD Culture isolates:* Enrichments will be maintained in the light or dark at *in situ* temperatures. Strains will be isolated on agar plates or tubes. Isolates confirmed as NCDs will be archived in gycerol and stored at -80°C.

2. **Data:**
   - Each of the listed sample types will receive a unique identifier and the associated metadata (e.g. location and date of sampling, volume filtered, etc.) will be stored in a mySQL sample database maintained at UCSC.
     - The following types of measurements will generate digital data from station samples and experiments during and after sampling: N_2 and C fixation rate measurements, chemical and biological measurements including primary production, measuring primary productivity, particulate concentrations (POC/PON and Chl *a*), O_2 concentrations, nutrient concentrations (nitrite+nitrate,

silicate, phosphate), DOM inventories, nanoSIMS data, quantitative PCR or digital droplet PCR targeting NCDs, and raw and processed sequence data from Illumina MiSeq runs for both amplicons and metagomes. The raw data received from the listed measurements will be processed following a standard procedure for each type of measurement.

- Image data will be obtained from FISH analyses as well as nanoSIMS analyses. Images will be converted into digital data.
- Sequence data will be obtained from the sequencing of PCR amplified genes and size-fractionated metagenomes using Illumina MiSeq technology. Raw data will be processed into formatted nucleotide sequences, counts and metadata. All sequence data will be submitted to the NCBI's Sequence Read Archive, and metadata for nucleotide sequences will be created manually following NCBI's standard.
- Processed data will be in csv, txt, tiff and fasta file formats.

3. **Publications:** The results of this study will be shared through presentations at scientific meetings and in peer reviewed publications.

Plans for data storage and preservation:

1. **Physical Samples:** DNA extracts, RNA extracts, as well as cDNA and amplified nucleic acid samples will be archived at -80^{o}C at UCSC for a minimum of 5 years, after which their integrity is questionable. Fixed filters for gene-/CARD-FISH analyses will be archived for future nanoSIMS method development in cryovials at -80^{o}C at UCSC. Glycerol-stabilized NCD isolates will be stored at -80^{o}C at UCSC. N_2 fixation rate samples, FCM, MIMS, and nutrient samples will be processed immediately.
2. **Digital Data:** All data will be stored in relational databases on servers at UCSC, and data servers are backed up using RAID 4 set-up and applications are backed up using a Time Machine (Apple). For long term storage, the data will be converted to stable file forms such as pdf, tiff, and ascii. At the termination of this research, long-term identifiers will be obtained using the UC3EZID system. Archives will be created for all raw and processed data and stored at the Merritt repository service at the University of California Curation Center.

Plans for Data Sharing:

Prior to publication, samples and data will be shared with other researchers upon request.

- Nucleotide sequences obtained from Illumina sequencing will be submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive in accordance with their protocols, by the time of publication.
- All field data and relevant genetic (e.g. gene expression data and qPCR data) and rate ($^{15}N_2$ assimilation, $^{15}NO_3^-$, $^{15}NH_4^+$ uptake rates) will be submitted to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) using their formats and standards and will be available online (http://www.bco-dmo.org) within one year of collection. Dataset documentation will include PI and sample analysts, references to analytical methods, calibration and blank corrections, and estimated accuracy and precision.