DATA MANAGEMENT PLAN

P.I. Jonathan Zehr, Co-P.I. Kevin Arrigo (Stanford)

August 8, 2015

Project Name: **Collaborative Research: Biogeochemical significance of the abundant, uncultivated symbiotic cyanobacteria UCYN-A**

Description of the expected data:

1. **Samples:**
   - *N_2 and C fixation rate samples:* Photic zone water samples will be collected three times per year at the Scripps Institute of Oceanography (SIO Pier) and at a site near the Virginia Institute for Marine Sciences (VIMS) in years 1 and 2, and incubated with either $^{15}N_2$ or acetylene for the determination of N_2 fixation rates and $^{13}C\text{-}HCO_3^-$ for the determination of carbon fixation (Zehr/Arrigo/Mills). Two process cruises per year (one at each site) will be conducted in Y1 and Y2 making the same measurements with the objective of understanding the dynamics of the UCYN-A associations over larger spatial scales.
   - *Cell-specific N_2 and C fixation rate samples (HISH-nanoSIMS):* Photic zone water samples will be collected from the above incubations in years 1 and 2, and the picoeukaryote fraction will be immediately fixed, when parallel acetylene reduction assays indicate active N_2 fixation, UCYN-A-specific cellular N_2 and C fixation rates will be measured using HISH and nanoSIMS analyses (Arrigo/Mills).
   - *DNA and RNA samples:* Photic zone water samples will be collected gently filtered 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^{o}C at UCSC. RNA extracts will be used to generate complementary DNA (cDNA), which will be stored at -80^{o}C. DNA extracts and cDNA will be used in PCR amplification and in qPCR assays (Zehr).
   - *Picoeukaryote enrichments for novel UCYN-A subclades screens:* Seawater samples will be analyzed immediately upon collection. Non-phycoerythrin containing photosynthetic picoeukaryote populations will be sorted into samples of 5,000 cells into nuclease-free water then frozen by flash freezing in liquid nitrogen and stored at -80^{o}C until use. Concentrates and single cell-sorts will be generated to complement these FACS-based isolations, and will also be stored at -80^{o}C until use (Zehr).
   - *Single picoeukaryote cells*: Single picoeukaryote events will be sorted into individual wells of 96-well or 384-well plates and stored at -80^{o}C until use (Zehr).
   - *CARD-FISH:* Samples for CARD-FISH/HISH will be collected from surface water and experimental manipulations and will be immediately fixed and gently filtered onto 0.6 μm polycarbonate filters, which will be dried then stored at -80^{o}C until processing.
   - *Chemical and Biological samples:* Seawater samples will be generated for characterizing the autotrophic community, measuring primary productivity, POC,PON, chl a, O_2, ammonium, nitrite+nitrate, silicate, phosphate during proposed cruises.

2. **Data:**
   - Each of the listed sample type 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 (Zehr).
   - The following types of measurements will generate digital data from experiments during and after sampling at SIO: N_2 fixation rate measurements (Zehr/Arrigo/Mills), chemical and biological measurements including primary production, POC/PON, chl, O_2, and nutrients (Zehr/Arrigo/Mills), nanoSIMS data (Arrigo/Mills), quantitative PCR targeting UCYN-A (Zehr), and raw and processed sequence data from Illumina MiSeq runs. The raw data received from the listed measurements will be processed following a standard procedure for each type of measurement, and stored in a mySQL database (Zehr).

- Image data will be obtained from FACS and FISH analyses (Zehr). Images will be converted into digital data.
- Digital data from UCYN-A qPCR analysis will be generated and processed at UCSC (Zehr) following a standard protocol and stored in a mySQL database at UCSC.
- Nucleotide sequences will be obtained from the sequencing of PCR amplified genes 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 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 cells obtained from ^{15}N nanoSIMS analyses will be archived for future HIMS-nanoSIMS method development and stored in cryovials or on gold and palladium coated membrane filters at -80^{o}C at Stanford University. N_2 fixation rate samples and nutrient samples will be processed immediately..
2. **Digital Data:** Raw data, processed data, and metadata 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 and acetylene reduction) 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.