DATA MANAGEMENT PLAN

The proposed work will generate several different types of data including environmental data (temperature, salinity, etc.), Sanger sequence data and pyrosequencing reads. The PIs are committed to proper curation of resulting data and deposition in public databases in a timely manner. Further, we will meet standards for annotation by following guidelines such as those generated by the Genomic Standards Consortium (GSC: http://gensc.org/).

Environmental data: All environmental data (T, S, chlorophyll-a, etc.) will be deposited in BCODMO, in accordance with NSF policy (*Division of Ocean Sciences Data and Sample Policy*, NSF 04-004). This database will also accept species counts for marine organisms, and we will deposit any microscope-based abundance data there as well. PI McManus will be responsible for ensuring the timely deposition of the oceanographic data.

Photodocumentation: We will rely on light microscopy for several aspects of this proposal including: identifying morphospecies for mesocosms work and fluorescence in situ hybridization. Given the uncertainties around identifications of some of our study organisms, we recognize the need to provide photodocumentation of cells, which can then be returned to in future studies. Both the McManus and Katz labs have extensive experience identifying and documenting microbial morphospecies (*e.g*. McManus_lab_images). For this project, we will store our images on the publically accessible micro*scope website, which is linked to the encyclopedia_of_life. The Katz lab is currently working on creating a collection of images for a separate project that focuses on the diversity of amoebae in a local bog environment (hawley_bog_project).

Attributes for molecular data: To increase the utility of all molecular data, a standardized set of attributes will be collected with every sample, and then deposited with sequences on GenBank. These attributes will also be designed to meet the Genomic Standards Consortium standards, MIMS and MIMARKS, for annotation of the metagenomic and marker gene surveys respectively: these standards are available at gensc.org and are the subject of a forthcoming Nature Biotechnology publication.

Attributes will include:

- Sample location – GPS coordinates
- Sample depth
- Sampling date
- Temperature
- Salinity
- Sample size
- Sampling container type
- Type of sequencing (*e.g.* Sanger, 454)
- Target gene and primers
- Additional abiotic features including chlorophyll, physical aspects of the local environment where the sample was taken (fronts, pycnoclines, etc. as deemed appropriate)
- For assemblies representing multiple reads, we will also track read number

Sanger sequencing**:** As with our past collaborations, sequences generated by Sanger sequencing will be deposited on GenBank when manuscripts are submitted. Sanger sequencing is performed at Smith College on an ABI3100, or at Penn State sequencing center, and raw reads are assembled using Lasergene software (DNAstar Inc). These contigs are scanned by

eye for quality and all polymorphisms are confirmed before being submitted to GenBank. We will submit the data with the attributes listed above to allow meaningful comparison of our data to other available oligotrich and choreotrich data.

***Pyrosequencing data**:* We will submit pyrosequencing data in two forms: 1) raw reads and 2) assemblies edited for quality and used in analyses. Raw read sff files will be deposited into something like GenBank's Sequence Read Archive (http://trace.ncbi.nlm.nih.gov/Traces/sra/),. We will also explore posting these files on other metagenomics resources such as MG-RAST and CAMERA, both of which are currently accepting data from the community.

We will submit curated data in the form of alignments that will be used both to generate phylogenies and explore community assemblages using tools like QIIME (http://qiime.sourceforge.net). Again, these data will be tagged according to GSC standards that users can access the attributes listed above. Curated alignments will then be deposited in GenBank to increase the availability of uncultured choreotrich and oligotrich data for comparison with future studies.