Guaymas Basin Data Management Plan

During the course of the proposed research we will produce a collection of sequencing data and associated environmental, sample processing and analytical metadata, which will be stored and manipulated using standard tools described here.
The environmental metadata associated with all field samples are pulled from shipboard CTD casts, *in situ* measurements by submersible, and cruise logs, and integrated to yield consistent spatiotemporal sampling sequences with consistent metadata. They include time of sampling, latitude, longitude, salinity, oxygen concentrations, temperature and water and sediment depth, plus linked biological observations, field notes and geochemical characteristics. Those of interest to the wider oceanographic community (for example, Guaymas Basin oxygen, salinity and T profiles) will be submitted to the BCODMO database in Woods Hole (bcodmo.org/home) which organizes biological and chemical oceanography metadata by cruise, marine area and PI. The PI has previously worked with BCO-DMO to document NSF-funded Gulf of Mexico Oil spill cruises in 2010.

Over three cruises, the proponents have developed a shipboard routine where every single sample and sediment core is immediately after retrieval catalogued, photographed, and recorded with full sampling context and time (using Alvin framegrabber images and data). Cores and samples enter the curation procedure after Submersible Alvin returns (ca. 5 pm) and are processed and recorded (digitally and in written lists and backup prints; paper has its place and is less fragile than most computers) before they are divided up for shipboard experiments, geochemistry, microbiology etc. in the daily science meeting after dinner.
Post-cruise, the sample data are organized and stored in the file sharing system that UNC Co-PI MacGregor has developed for highly complex field surveys in Guaymas Basin and the Gulf of Mexico. This system, on a dedicated file server in the Teske/MacGregor labs, now holds sample and site information for three major cruises (AT15-40 in December 2008; AT15-56 in November/December 2009, and AT18-02 in November/December 2010), and allows unambiguous retrieval of full sampling context, a digital photo of each core immediately after recovery, destination (the specific lab that has received and analyzed the core) and environmental *in situ* information (Temperature and geochemical gradients) for every individual sediment core. We are using this system extensively to keep track of samples for analysis, comparison and publication purposes, and enter and circulate updates of newly analyzed data and curated datafiles. This database has allowed us to provide full core information, geochemistry, sampling site metadata, and *in situ* and *ex situ* photographs on request, and to make this information quickly available for our collaborators and their current and future manuscripts.

Our sequence data documentation will follow the standards defined as Minimal Information about a Metagenomic Sequence (MIMS) and Minimal Information about a Marker Sequence (MIMARKS); these represent a curated standard format layer for the acquisition and display of information associated with sample acquisition, processing, handling, sequencing, and analysis. These are community standards, agreed using consensus and updated where necessary by routine annual meetings of the Genomic

Standards Consortium (www.gensc.org). In addition these standards are recognized by the INSDC and reported by a keyword (GSC) for compliant sequences. We will rigorously adhere to both standards for sequencing data generated using this proposal.