DATA MANAGEMENT PLAN

I. Types of data

This project will generate nucleic acid sequence data, and data on the prevalence and load of viruses, bacteria and protozoa in sea star populations. The total amount of sequence information is estimated at 60 million sequence reads from 28 individual metagenomic libraries, representing 150 MBp of information. The sequence information will be captured by Illumina MiSeq sequencing, to be conducted by the Cornell Core Laboratory Center. Data on the prevalence and viral load of putative pathogens will be obtained by quantitative PCR. Data will originate from echinoderm samples collected in the Salish Sea, WA and Monterey Bay, CA.

Surveys of sea star populations in Washington will yield size and count data, and each individual counted will be assigned to a disease category (0-4). These data will be entered into the MARINe (Multi-Agency Rocky Intertidal Network) Access database, and integrated with similar data from other MARINe sites spanning the entire west coast of the United States.

II. Data and Metadata Standards

Nucleic acid sequence information will be stored in FASTA formatted files or SFF files which integrate sequence quality information with the sequence itself. The appropriate metadata to make the sequence information meaningful include accurate determination of host species identity, sex, size, and disease state. In addition, the latitude and longitude (measured by GPS), sampling depth, water temperature and conductivity (measured by CTD or YSI handheld meter, where available) will be measured and data saved as Excel files in the same location as the sequence data. These are consistent with genomics standards consortium (GSC) standards for metagenomic studies.

All sea star population data will be documented in accordance with the Data and Metadata Standards used by MARINe. This specifies the use of Ecological Metadata Language (EML), which was specifically developed by and for the ecology discipline. By adopting these standards, each data attribute is specifically defined, ensuring that the data are used and interpreted properly by the end user. For more information, please see the full Data and Metadata Standards for the North Central Coast MPA Baseline Program here: http://oceanspaces.org/sites/default/files/NCC_metadata_standards_Jul2011.pdf

III. Policies for access and sharing and provisions for appropriate protection/privacy

Nucleic acid sequence data and complementary metadata will be made available through two avenues. They will be submitted to the NCBI short read archive (SRA) which houses most metagenomic data obtained to date. Metadata is also submitted at the same time. We will also submit our data to the cyberinfrastructure for advanced marine microbiology research (CAMERA) database. All sea star population data will be stored in the MARINe Access database. Data will be released within 12 months of generation. These databases are accessible to the public. There will be no charge for access. There are no privacy issues regarding these data. The data will not be covered by a copyright. Data on viral prevalence and viral load will be available upon request 12 months after generation. There will be no charge to access the data, no privacy issues, and data will not be covered by copyright. All data will be published in peer-reviewed journals.

IV. Policies and provisions for re-use, re-distribution

There will be no permission restrictions needed for these data. The data may be of interest to biological oceanographers, invertebrate zoologists and ecologists at other institutions, and to disease ecologists and managers at government agencies. The intended and forseeable users of the data are oceanographers, modelers, ecologists, metagenomicists, and microbial ecologists within academia. It is anticipated that other scientists will compare their sequence information to nucleic acid sequences generated in this study via NCBI and/or CAMERA. There are no reasons not to share these data.

Sea star population data will be integrated into our comprehensive database and available in a summarized form on our publically accessible website (see Pacific Rocky Intertidal Monitoring: Trends and Synthesis, http://www.eeb.ucsc.edu/pacificrockyintertidal/index.html). Researchers requiring non-summarized data can contact the MARINe data manager via the website to make a data request.

V. Plans for archiving and Preservation of access

Initially, nucleic acid sequence data and data on prevalence will be archived on desktop computers in the Hewson Lab, and backed up onto the Cornell Microbiology Bioinformatics Cluster, which itself is backed up onto a remote server. Data will also be backed up onto terradrives in the laboratory. Data will be submitted to public databases (NCBI and CAMERA), where they will be permanently archived to preserve access to the public. A hard copy of all notes (i.e. lab notebooks) will be retained in the laboratory. All relevant metadata associated with genomic libraries will be submitted along with the nucleic acid sequences themselves. Research publications generated from this work will include all relevant data and refer readers to public databases where data is permanently archived.

Sea star population data will be integrated into the MARINe database and available in summarized form on our publically accessible website (see Pacific Rocky Intertidal Monitoring: Trends and Synthesis, http://www.eeb.ucsc.edu/pacificrockyintertidal/index.html). The MARINe database is updated frequently (6 month interval minimum). The database will be stored on a secure server at UC Santa Cruz, and will be backed up to a remote, secure server as well.