DATA MANAGEMENT PLAN

I. Types of data

This project will generate nucleic acid sequence data from P. californicus transcriptomes, amplicon libraries of flaviviruses, and 16S/18S rRNA amplicons from animal tissues and from overlying waters. Some biological parameters (metadata) from the holothurians will also be obtained, including the size, sex, color morph, gross pathology, histology, coelomic fluid hematology and chemical composition of their host (where applicable). The project will also collect data on dissolved oxygen concentrations, salinity, temperature, chlorophyll a by fluorometry and acetone-extracted spectrophotometry, and bacterial abundance as determined by epifluorescence microscopy. Data will originate from holothurians collected primarily from Southeast Alaska and at three other locations on the North American Pacific coast. Data will be obtained from experimental incubations of holothurians. The sequence information will be captured by Illumina HiSeq or MiSeq sequencing platforms, to be conducted by the Cornell University Biotechnology Resource Center (BRC). The total amount of sequence information is estimated at 135 million sequence reads representing ~162 GBp of information. In addition to sequence information, this project will generate data on the abundance and transcription of hypoxia inducible factors and hypoxia-sensitive host genes by qRT-PCR using the StepOne (Applied Biosystems) Instrument.

II. Data and Metadata Standards

Nucleic acid sequence information will be stored in FASTQ formatted files or SFF files which integrate sequence quality information with the sequence. The appropriate metadata to make the sequence information meaningful include accurate determination of host species identity, geographic location (latitude and longitude measured by GPS) and treatment, water temperature and salinity (measured by onset probes), and oxygen concentrations (measured by Onset HOBO loggers where appropriate) which will be measured and data saved as Excel files in the same location as the sequence data. These data are consistent with genomics standards consortium (GSC) standards for metagenomic studies including Minimum Information for Metagenomic Sequence (MIMS).

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 GenBank short read archive (SRA) which houses most metagenomic data obtained to date. Metadata is submitted at the same time. Completed flaviviral sequences will also be submitted to the non-redundant (nr) database. Data will be released within 18 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 or phylotype identity will be available upon request 12 months after generation, and will also be available via the project website as downloadable data files. Supporting biological oceanographic data will be deposited to the National Oceanographic Data Center (NODC) and BCO-DMO within 12 months of collection. 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, disease and microbial ecologists at other institutions. The intended and forseeable users of the data are oceanographers, modelers, disease 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. There are no reasons not to share these data.

V. Plans for archiving and preservation of access

Initially, nucleic acid sequence data and data on prevalence and viral abundance, and gene copy number of functional genes/transcription factors will be archived on desktop computers at Cornell University, 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 external hard drives in the laboratory. Data will be submitted to public databases (NCBI), 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.