Data Management Plan:

This project will compile and store a large number of coral specimens, collected field data, metadata, and photography, along with high-throughput sequence data on marine microbiomes. We will sub-sample all DNA samples for long-term storage and access at Vega's laboratory at OSU. Physical laboratory notebooks (and digital ones as described below) will be archived each of the PI's laboratories and will available for review by all interested scientific entities. These vouchers will be available to the research community upon any valid request.

Access and Sharing of Oceanographic/Ecological Data*:* Data (oceanographic and marine benthic data such as seawater temperature, nitrogen/phosphorus concentrations, coral abundance, and algal diversity data) and information management for this project will be done in close collaboration with the Moorea Coral Reef LTER program at UC Santa Barbara, which has agreed to provide data Archival, web access services to this project and linkage of project data to BCO-DMO. The MCR LTER Information Management System (IMS) will facilitate the archival cataloging of our data for long term preservation, enable the discovery of our fully documented data, and enhance their suitability for synthesis by us and others. In particular, data produced by this study will be curated by the MCR LTER IM using LTER Best Practices and made available through the MCR LTER local data catalog as well as the LTER Network Data Catalog. At BCO-DMO, a project page describing this study will point to the MCR local catalog, as is done with other projects funded by OCE. Additionally, when linkages become established in the near future, these datasets will appear in searches of the DataONE catalog alongside datasets from BCO-DMO. Data Quality Assurance and Metadata documentation will be done as part of this project by the PIs, who will work closely with the MCR Information Manager (M. Gastil-Buhl).

Data packages will conform to the most recent (August 2011) version of Best Practices for LTER dataset EML. Metadata features will include either embedded or online links to methods and protocols, full temporal, spatial, and taxonomic coverage, keywords from the MCR vocabulary, the NBII thesaurus, and/or the LTER Controlled Vocabulary, and units registered in the LTER Unit Dictionary. All EML will be version 2.1.1. All data tables will be congruent as far as the EML Congruency Checker will be able to check. Additionally, some datasets may provide explicit indexing keys and table-joining keys to facilitate cross-dataset synthesis.

Access and Sharing of Next-Generation Sequence Data**:** Sequence data not appropriate for the data management sites above (e.g., BCO-DMO) will be deposited in the appropriate major databases, such as the federal National Center for Biotechnology Information's (NCBI) GenBank/EMBL/SRA database. We will share nucleic acid sequences with wider research communities through deposition in publically available iPLANT Collaborative (http://www.iplantcollaborative.org/) and the Earth Microbiome Project and its global environmental sample database (http://www.earthmicrobiome.org/global-environmental-sample-database/) as we have done so in the past. We will also contribute data to the QIIME 16S database (http://www.microbio.me/qiime/index.psp). This will allow for cross-comparison with hundreds of studies and raw data download. Lastly, to ensure timely delivery of these open access data, since some websites require many months to make items truly available, we will also provide raw sequence reads on our own open access folder hosted at OSU's Center for Genome Research and Biotechnology (CGRB). Per guidelines under the Division of Ocean Sciences Sample and Data Policy, we will make all data publicly available within two years of its collection.

Data Entry/Management*:* One of the keys to a successful collaborative project is a process of data centralization that uses commonly available tools for data entry and sharing. In addition to relying on systems like GitHub for sharing data and program development, the labs will also maintain easy to use, widely available tools like DropBox and Google Docs. The Vega Thurber lab, and particularly the TDB bioinformatics postdoc and graduate student will oversee many of these specific procedures (technical implementation, documentation, training, etc.), as well as the general implementation of the data management plan. The graduate students in our labs will be trained to use these data management tools as part of their education.

Analysis of All Sequence Data*:* Our strategies for analyzing high-throughput sequencing data are described in the specific aims of this proposal. We will follow the latest directives from the Genomic

Standards Consortium (GSC) for the development of the minimal information checklists for any marker-gene amplicon datasets we generate. These datasets, “Minimal Information about a Marker Sequence” (MIMARKS), provide 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 annual meetings of the GSC (www.gensc.org). In addition, these standards are recognized by the INSDC and reported by a keyword (GSC) for compliant sequences. We will adhere to both standards for sequencing data generated using this proposal. All data will be made publicly available as soon as modeling and quality control are completed. This project aims to implement a truly open access data management plan. We will adhere to standards for spatially comprehensive environmental data generated using this proposal.

To ensure that analyses involving numerous steps on the commandline are replicable, we will use a system of “runnable lab notebooks”. For each analytical product, we generate a ‘procedure’ text file to document the exact steps of the analysis, starting from the shared raw data present in the CGRB/sequencing center repository. However, rather than being static text, the file will be a BASH script (with extensive additional comments explaining the results and reasoning of each step) *to allow large portions of the analysis to be regenerated in a single command-line step*. The Vega Thurber lab has found this system to be highly useful in documenting steps, ensuring that we can report all relevant parameters in methods documents, and reanalyzing data with slightly different parameters in response to reviewer requests. This simple system also provides an easy way to share procedures with collaborators or lab-members.

Data Backups: In addition to user backups, we use both an on-site cluster (see budget justification) with RAID storage at the CGRB and the commercial Dropbox software. All researchers also individually back up hard-drives approximately every two weeks. These layers of redundancy will be sufficient for internal analyses, paper manuscripts, etc. However, they do not suffice for data sharing with the broader community or permanent data storage. We will archive sequences in appropriate online repositories (see above).

Coding Practices: All software will be implemented as an open-source software package in the Python programming language (wrapping ecological modules from R, etc. as needed). This package will be developed openly through our GitHub site, which allows for good versioning practices and community input. It is our policy that all scientific software be accompanied by test code. Test code acts in a similar fashion to control experiments in molecular biology, and helps to ensure that code changes (to optimize speed, etc.) do not introduce biological errors. All code will follow a consistent, documented coding style (http://pycogent.org/coding_guidelines.html) and include substantial commentary. The Vega Thurber lab also manages and maintains a GitHub site that is publically available for all researchers to view and if so desired edit new versions of our scripts.

Data Publication and Presentation:
We aim to publish our data in peer-reviewed open access international scientific journals in a timely manner following the proposed timeframe in the project description, and to use the data in teaching undergraduate courses, a practice already routinely performed by the PIs. We have budgeted funds to make some of these publications ‘open access’ to allow for a broader community of researchers and the public to acquire these manuscripts. As state above, per guidelines under the Division of Ocean Sciences Sample and Data Policy, we will make all data publicly available within two years of its collection.