Data Management Plan:

This project will compile and store a large amount of coral specimens, collected metadata, photography, viral production, SIP, and high-throughput sequence data. We will sub-sample all DNA samples for long-term storage and access at each of the laboratories. Physical laboratory notebooks (and digital ones as described below) will be archived in the PI's laboratory 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*:* This proposal will generate a large amount of oceanographic metadata, species abundance/distribution data, and sequence data on the viromes and microbiomes of the samples. To properly store this information, database software (Microsoft Access & SQL) will be used to manage the data and facilitate statistical analyses. All oceanographic metadata (e.g., seawater temperature, irradiance, nitrogen/phosphorus data) will be submitted to the National Oceanographic Data Center (NDOC) (http://www.nodc.noaa.gov/) and to the Biological and Chemical Oceanography Data Management (BCO-DMO) office (http://www.bco-dmo.org) in compliance with the guidelines for the Division of Ocean Sciences Data and Sample Policy. Further, all species abundance, distribution, and diversity data for algae, fishes, and corals and other invertebrates will be submitted to the Ocean Biogeographic Information System (OBIS) (http://www.iobis.org). Image data will also be stored on iPLANT for any user to access. We will also store the data on our freely accessible folder at OSU's Center for Genome Research and Biotechnology (CGRB).

***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 the Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (http://metagenomics.nmpdr.org/), 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/). We will contribute data to the QIIME database (http://www.microbio.me/qiime/index.psp), and MG-RAST (http://metagenomics.anl.gov/). 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 assess folder at the Center for Genome Research and Biotechnology (CGRB) at OSU.

Site Partner 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. Likewise, for data that will be entered directly by the site partners (e.g., collection coordinates), we will have a series of Google Forms templates on Google Docs that will be easily edited by site partners for data entry. These templates can then be easily quality controlled by the PI's. The Vega Thurber lab 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 genomes, metagenomes, and 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 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 GitHub, 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.

Data Publication and Presentation:
We aim to publish our data in peer-reviewed 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.