Data Management Plan:

This project will compile and store a large amount of physical specimens, collected metadata, and next generation sequence data. Each lab will maintain vouchers of all coral specimens and any novel microbial cultures for storage redundancy. We will provide subsamples of DNA samples for archives at the other laboratories for long-term storage and access. Physical laboratory notebooks (and digital ones as described below) will be archived in each of the PI's laboratory and 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. To properly store this information, database software (MySQL) 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/) 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, corals, and other invertebrates will be submitted to the Ocean Biogeographic Information System (OBIS) (http://www.iobis.org).

Site Partner Data Entry/Management*:* One of the keys to a successful experimental network is a process of data centralization that uses commonly available tools for data entry and sharing. For sharing data within the network, we will rely on 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 easily be edited to enter data by the site partners. These templates can then be easily quality controlled by the PI's.

Analysis of All Sequence Data*:* Our strategies for analyzing high-throughput sequencing data are described in the specific aims. 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 called Minimal Information about a Metagenomic Sequence (MIMS) and Minimal Information about a Marker Sequence (MIMARKS) provided 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 publically 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.

We would like to emphasize that our strategy of sequencing artificially constructed mock communities of known composition on 1-2 barcodes of each sequencing run will serve as a powerful internal control ensuring that bioinformatic analyses recapture known patterns. To ensure replicability of analyses that involve numerous steps on the commandline, we 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 Dropbox/CGRB/sequencing center repository. However, rather than being static text, the file is 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*. We have 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 (the Center For Genome Research and Biocomputing on the Oregon State University Campus) and the commercial Dropbox software. All researchers also individually back up hard-drives approximately every two weeks. We have found these layers of redundancy 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. Additionally, they do not suffice for maintaining and releasing versioned software for which we will use the open-access GitHub repository.

Contribution of a coral holobiont sequence resource to the community: Based on our own experiences using published data in meta-analyses, we find that there are two successful strategies for making data available to the larger community: 1) a simple, but well documented web page that allows for

the download of all raw data and metadata 2) a collaborative database spanning hundreds of studies that includes built-in methods for cross-comparison, ideally maintained by multiple full-time developers. Unsuccessful approaches do not allow full raw data download (e.g. allowing only search results to be downloaded), or implement inferior or duplicative versions of already available tools. Based on these experiences, our strategy will be to provide the data in the simplest, most accessible form (strategy 1); and simultaneously to contribute to existing standards-based comparative platforms (strategy 2).

Data generated in this project will be deposited in the major databases, such as the federal National Center for Biotechnology Information's (NCBI) GenBank/EMBL database. We will share nucleic acid sequences with wider research communities through deposition in publically available databases: the Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (http://metagenomics.nmpdr.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. Moreover these resources have both adopted the MIxS metadata standard and the BIOM observation x sample matrix standard, enabling data interchange. All policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, and other rights or requirements will be adhered to. We will also use built-in tools in the QIIME database to export sequence reads to the EBI and NCBI's SRA. *Symbiodinium* ITS data and metadata will be contributed to GeoSymBio (https://sites.google.com/site/geosymbio/contact), while Fungal ITS sequences will be uploaded to the UNITE short read repository (http://unite.ut.ee/repository.php).

To ensure that the resource we provide to the community is useful for comparative analysis (an essential goal of the project, and one that is very important to us), we will use the same resources that we release publically to exchange data internally. This approach results in much more useful data products, since inconveniences in format or difficulties with data access are noticed and corrected by team members in the course of normal research. This approach ensures that off-site backups of data are regularly enforced.

Coding Practices: All software (including the predictive models described in Aim 4) 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. Reviewers can already examine numerous examples of test code from our PICRUSt project on GitHub (picrust.github.com). All code follows a consistent, documented coding style (http://pycogent.org/coding_guidelines.html) and includes substantial commentary. Finally, we encourage team coding, which ensures that scripts are interpretable by multiple coders, promoting good practices.

User support: This key aspect of data/software management is often neglected, but essential if data or software are to be used by the broader community. Our practice is to establish a help forum/e-mail list (a simple Google group) for each major software product, and to respond rapidly to user queries. This approach allows user questions to be archived, such that a large trove of solutions and best practices is built up over time. Users will use the list to support one another in a collaborative fashion. We have found this practice to provide extensive benefits in ensuring the reliability and practical utility of software. For example, our PICRUSt software has already been cited 13 times and has more than 180 unique users despite being less than a year old. The QIIME forum has more than 1660 unique users and has been cited 1320 times since 2010. User feedback from these lists is invaluable in driving development of new features, refining output formats, improving tutorials, etc. It also provides broad exposure, training, and networking opportunities for early-career scientists as they assist other groups.

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 use the data in teaching undergraduate courses, a practice already routinely performed by the PIs. We have budgeted funds to make these publications 'open access' to allow for a broader community of researchers and the public to acquire these manuscripts.