1. Types of data, samples, physical collections, software, and other materials to be produced in the course of the project.

Information, materials and data that will be produced in the course of this project include: 1) Tissue samples, DNA extracts from sponge vouchers, and ARMS community samples; 2) targeted COI and 28S rRNA amplicon loci sequences in FASTA format from sponge vouchers which will include hundreds of unique species; 3) primer sequences optimized for targeting 28S rRNA sequences from sponges; 4) paired-end COI and 28S rRNA metabarcoding sequences in BAM format; 5) high resolution images (in *TIF* or *JPG* format) depicting diagnostic characters of museum vouchers such as histological sections, light or scanning electron micrographs (SEM) of spicule and different cell types; 6) microscopic measurements in excel spreadsheet format of diagnostic characters of vouchers such as cell diameter, spicule length and width, collagen fiber width, and measurements of the skeletal arrangement of these microstructures; 7) permanent slide preparations of spicules and histological sections embedded in Permount; 8) vouchers, with relevant metadata and pertinent observational data included in the description of new species; 9) percent cover data of sponge species in excel spreadsheet format; and 10) R scripts for data manipulation or statistical analyses, and bioinformatic pipelines required for repeatable processing of Illumina sequences.

2. The standards to be used for data and metadata format and content

Data will be stored in both the original data format as well as non-proprietary formats in accordance with GBIF/TDWG standards (following protocols and formats used for the GBIF Integrated Publishing Toolkit; IPT). Metadata are crucial to repeatable science and interpreting data but often undervalued in Data Management strategies. Metadata associated with all DNA sequences such as voucher number, location (GPS coordinates), depth, collector name, date, and primer sequences will be uploaded to Genomic Observatories MetaDatabase (GEOME, Deck et al. 2017), a new genomic standards compliant georeferenced database developed as part of the DIPnet RCN initiative (Toonen et al. NSF-DEB 14-57848, see **Results from Prior NSF Support**). Through GEOME, raw sequence data are stored in the appropriate Federated database, but permanently linked to the metadata through persistent identifiers generated by EZID (California Digital Library). The BOLD database will be used to upload COI sequence data linked to images associated with the voucher specimen. 28S rRNA sequences will be added to the SILVA database. Metadata associated with sequencing including information on library preparation protocols and details of sequencing events will be meet recommended standards of the DIPnet consortium and formatted according to the Genomic Standards Consortium recommendations (MIMS, MIMARKS, and MINSEQE). Long-term archiving and curation of all genetic metadata will occur through GEOME and other records associated with specimen occurrence and identification records through InvertEBase.

Material sample data and all metadata associated with vouchers and ARMS samples, such as high resolution images, percent cover, location of slide preparations of spicules, and observational data included in the description of new species will follow the data standards of, and be archived with, the Florida Museum of Natural History and the Smithsonian Natural History Museum.

3. Policies for access and sharing, re-use and re-distribution of data

For short term storage and organization during the project, data will be collected and added to the ToBo lab redundant personal cloud storage (see **Facilities: Computing & Data Storage Capacity**) and mirrored on personal computers of the individuals who are working with specific datasets, as well as their personal GitHub accounts. Additional backups on portable USB hard drives will be stored

remotely for individual workflows. All members of the team will always have access to all data, and will share the responsibility for access sharing, re-use and re-distribution.

None of the data produced will require provisions for protection of privacy, confidentiality, intellectual property, or other rights. We are committed to the public sharing of all scientific data and information resulting from this project. We will work with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to archive all raw data they will accept and make those data publicly available through NSF. Individual data sets will also be archived to GitHub and shared through more specialized databases that maximize their chances of discovery during searches. For example, genetic data and metadata will be made publicly available and searchable online through GEOME within 2 years of the date of collection and served to both GBIF and OBIS in addition to the appropriate federated sequence archive (BOLD, SILVA, GenBank and/or SRA) as is customary with GEOME submissions.

Tissue backups and DNA extractions will be stored in the Toonen lab and archived for future use. At the end of the project, remaining samples will be posted to *Otlet.io*, a searchable global open-access platform to share and source biological samples and archived with one of our museum collaborators as appropriate.

In keeping with NSF policy, all data and metadata will be made publicly available within 2 years of collection, and all non-electronic data products or samples accumulated through this project (such as images, histological slides, tissue samples, measurements and image analysis of % cover data) will be made publicly available upon request at no more than the cost of reproduction or shipping & handling.

4. Plans for archiving data, samples, and other products, and for preservation of access to them.

In compliance with university and national standards, all digital data will be made publicly available upon publication in a peer-reviewed journal or within 2 years of collection, whichever comes first, and maintained in identifiable and retrievable format for a minimum of 5 years post publication.

As outlined above, all digital data will be backed-up locally on our lab redundant personal cloud server, and also archived long-term via individual GitHub accounts and through BCO-DMO. Sequence data will be archived via NCBI GenBank/SRA, SILVA (for 28S rRNA) and BOLD (for COI) databases, and permanently linked to the associated metadata through persistent identifiers generated by EZID (California Digital Library) through GEOME.

Images associated with the description of sponge voucher species will be served publicly by each collaborator museum with which the sample is associated, uploaded to the InvertEBase website data portal, and any from Moʻorea will also be uploaded to the Moʻorea Biocode Databases. Individual data sets (individual genotypes with full metadata, including accurate georeferencing, habitat notes, etc.) will be deposited either as data publications (such as Figshare, DataONE, re3data, or DASH), open access publications that include the data as part of the peer-reviewed product, or as openly accessible files at the Dryad Digital Repository (if neither of the first two options are available).

New species described in scientific publications will be officially registered on the ZooBank website which assigns a new species registration number and automatically uploads onto The World Porifera Database and World Register of Marine Species websites.