DATA-SHARING PLAN FOR MOORE FOUNDATION
Coral resilience investigated in the field and via a sea anemone model system

(Arthur Grossman, Steve Palumbi, and John Pringle)

GENERAL PHILOSOPHY

The three Principal Investigators subscribe fully to the GBMF philosophy both as a matter of general principle and because we think that the immediacy of the worldwide threat to coral reefs demands a particularly high level of cooperation among investigators. Thus, we intend both to publish our findings promptly in journals cooperating with open-access policies and to make our data publicly accessible at or before the time of publication (as appropriate for the particular type of data).

1. DATA DESCRIPTION

A. Data to be collected include (i) annotated transcriptome assemblies for the anemone *Aiptasia,* several strains of the dinoflagellate *Symbiodiniuim,* and one species of coral; (ii) gene-expression (*i.e.,* transcript-abundance) data for the same organisms, collected under various environmental conditions (temperature, light intensity, possibly other variables); (iii) physiological data, mainly on photosynthetic function and optimal growth conditions, for various *Symbiodinium* strains, both while free-living and while living in a compatible host, and under various environmental conditions; (iv) detailed environmental data (including temperature, light levels, and pH over time) for the lagoon pools constituting the project's field site at Ofu, American Samoa.

B. Each of the four types of data listed in A will require a distinct format. (i) Transcriptome data will consist of annotated assemblies and raw sequence reads (short, paired-end reads from Illumina sequencing). (ii) Gene-expression data will be built into the transcriptome assembly and coupled to an SQL database of terms (tissue, environmental conditions, etc.) to allow searching for transcript-abundance changes correlated with specific parameters of interest. (iii) Physiological data will be provided as tabulations of the various measured responses as a function of the conditions applied and the identity of the strain tested. (iv) Environmental data will be provided in graphical (*e.g.,* temperature *vs.* time) form and as downloadable spreadsheets.

C. Data will be collected throughout the project, starting essentially immediately (because some transcriptome and gene-expression data are already available). Transcriptome and gene-expression data will be entered into electronic databases as soon as assembly and quality-control tests are completed. Physiological data will be entered into a database not later than the time at which they have been analyzed and organized for publication. At least initially, the *Aiptasia* transcriptome data (without expression data) will be made available through AiptasiaBase (http://aiptasia.cs.vassar.edu/AiptasiaBase/index.php), an existing sequence database set up in a collaboration involving the Pringle laboratory and currently hosted by Dr. Jodi Schwarz and her colleagues at Vassar College. Transcriptome data (containing both gene annotations and expression data), physiological data, and environmental data will also be available through databases hosted by our laboratories as part of this project and available for others to access through a project website that we will develop. As the pool of gene-expression data expands, it may prove preferable to have it housed by the Stanford Microarray Database.

D. The only pre-existing data that we envision utilizing are those previously collected by our own groups or in earlier collaborative efforts in which our groups were involved. No special data-sharing arrangements are required.

E. We do not anticipate developing new analytical tools as part of this project. The sharing of gene-expression, physiological, and environmental data will involve the development of new databases that we will host on local servers and make available through the project website and our individual laboratory websites.

F. Metadata will include (i) for transcriptome data, detailed descriptions (as for publication) of the environmental conditions pertaining at the times of mRNA collection, and of the procedures used for processing, sequencing, assembly, and annotation; (ii and iii) for gene-expression and physiological data, the precise environmental conditions and stresses that applied prior to and at the time of data collection; and (iv) for environmental data, the precise geographical coordinates and depths of the recording devices and the specifications of those devices. Note also that we intend to distribute freely both our *Aiptasia* clone (this has already been supplied to about a dozen other laboratories) and the various *Symbiodinium* strains that we will isolate and characterize; such distributions will always be accompanied by information on whatever we know about the properties of the organisms (*e.g.,* precise identification to clade and subclade and possible special nutritional requirements of particular *Symbiodinium* strains).

G. For the purposes of quality assurance, the data will be owned by the three collaborating Principal Investigators. However, the data will be placed in the public domain, with no restrictions on access or use, as soon as processing and quality control are complete (transcriptome sequences and gene-expression analyses) or organization and analysis (as for publication) are complete (physiological and environmental data).

2. DATA MANAGEMENT

A. Data will be stored at Vassar (AiptasiaBase – see above) and on local servers at Stanford.

B. Data will be hosted as noted above and will be searchable or downloadable remotely using conventional software. For extremely large data sets (*e.g.,* raw sequence data from Illumina sequencing or raw gene-expression data from RNA-seq), external hard-drives with the data will be made available for shipment to other laboratories.

C. The data archive will use the same location(s) and software as the initial data entry.

D. Metadata will be stored in the same databases as the primary data and will be accessed by links within the websites. See note under 1.B about use of SQL databases.

E. Data will be entered and maintained by AiptasiaBase staff and by our project personnel (including the dedicated bioinformatician) at Stanford. Our intent is to maintain archives indefinitely or until they are judged no longer useful.

F. Data quality will be assured by (i) tests whose nature and outcome will be available along with the data themselves and (ii) peer-review of publications that are based on the data.

G. Databases will be populated with the data obtained by our project personnel.

H. No proprietary data will be used.

3. DATA SHARING

A. Potential data users include all investigators working with the *Aiptasia* model system (a number that is expected to grow with time) or with cultured *Symbiodinium,* as well as many investigators working directly with reef-building corals.

B. Some data will be released for general access as soon as the dataset is complete and quality is assured. For example, the transcriptome assembly will be released as soon as the data are fully assembled, checked adequately for quality, and sufficiently annotated to be widely useful. Before general release, the adequacy of data storage and access procedures will be tested first by project personnel, then by selected colleagues external to the project. A publication(s) describing the data collected and conclusions drawn from them would be submitted soon thereafter. Other data will more appropriately be made generally available at the time publications reporting on them are accepted.

C. Archived data will be made available initially as just described and are intended to be available indefinitely or until judged no longer useful.

D. Any data pertaining to this project that we control, which have resulted from non-GBMF projects (prior, ongoing, or future), will be made available according to the same principles as data resulting from the GBMF-supported studies.

E. Data will be accessed by other users as described above.

F. Raw data will be made available to interested users. In general, however, the data will be made available after processing such as quality control, assembly, and annotation (for transcriptome sequence assembly), organization by mapping RNA-seq reads onto the assembled transcriptome (for gene-expression studies), etc.

G. We intend that the data be available indefinitely beyond the term of the grant. The databases will be hosted by ourselves or potentially by consortia established in the field.

H. At this point, we cannot foresee the future hardware needs sufficiently clearly to include them in the budget for the proposed Grant.

I. We do not anticipate the development of data-analysis tools under the auspices of this Grant. If such tools should be developed, we will make them available through the project website with sufficient tutorial associated to allow future users to use the tools without undue difficulty.

J. No data-sharing agreement with outside vendors will be needed.

K. We intend to make all data freely and fully available to other investigators without the need for a Creative Commons-type license.

L. For the data that we provide, we will expect appropriate acknowledgment (of our publications and/or databases) by other users in their own publications or other uses of the data, in accord with well established scientific norms.

M. As noted above, we anticipate the prompt publication of papers in open-access journals that will be based on our data and make its availability widely known to other potential users.

DATA SHARING PHILOSOPHY

DATA SHARING PHILOSOPHY

The Gordon and Betty Moore Foundation's goals of scientific advancement, environmental conservation and health care improvement will best be served through a culture of open access to data. It is our philosophy that:

- All data used in or developed in whole or in part by foundation-funded projects (and that can be shared in a manner consistent with applicable laws) will be made widely available and freely shared as soon as possible[1]. If data used in foundation-funded projects are owned by an additional party other than the grantee, we do not require it to be released, but the grantee will use its best efforts to encourage the data owners to make it openly and freely available.
- Data are shared with full and proper attribution to the data provider.
- Data developed in whole or in part by foundation grant funding are the property of the grantee unless otherwise specified. The grantee may protect its property through patent, copyright and/or other intellectual property protection instruments, except that it may not impede the effective access and use of the data by the public.
- The foundation is not responsible for any liabilities associated with errors in the data or misrepresentations or misinterpretations of publicly available data.
- The foundation supports grant funding for costs associated with data sharing and open access publication of scientific findings, where appropriate.
- The foundation and prospective grantees will jointly develop a Data Management and Sharing Plan prior to the finalization of a grant agreement.

The Data Sharing Philosophy applies to all activities that are financially supported in whole or in part by the foundation that include, but are not limited to:

> Data collection and analyses, data, meta-analyses and information derived from pre-existing datasets, and database development

Data sharing includes, but is not limited to, data contained within the following:

> Publications, databases, derived data products, mathematical models and model code, metadata (defined as appropriate documentation describing the data, relevant specifics of their collection and the data format) and statistical and other forms of data reduction and analysis

[1] Examples of when data should be released: For data created for scientific and environmental conservation purposes, public release should occur not more than six months from the "date of collection" (defined as the date when data enters an electronic database), unless otherwise specified in the grant agreement between the grantee and the foundation; for DNA sequence data, "public release" (i.e. submission to an appropriate public database), should occur not more than six months after "completion" of the DNA sequence determination (as defined in the grant agreement between the grantee and the foundation).

DATA SHARING AND MANAGEMENT PLAN

As part of the foundation grant development process, potential grantees are required to develop a Data Management and Sharing Plan with their foundation grant team. In these cases, before submitting a final grant proposal to the foundation for approval, both the potential grantee and the foundation grant team must approve a final version of the plan that is consistent with this Data Sharing Philosophy. Any exceptions to the Data Sharing Philosophy must be clearly articulated in the plan and approved by the grant team. Funds needed for data sharing and management may be requested as part of the proposal. Once finalized, the plan will be referenced in the grant agreement for the approved grant.

The plan should address the following three topics and any other topics identified by the foundation and/or grantee:

1. **Data description**. Questions to consider as appropriate:

   - What data will be collected during this project?
   - How many different data formats are anticipated? Please list formats.
   - When will the data be collected, when will they be entered into electronic databases and what databases will harbor the data?
   - Does this project involve organization or analysis of pre-existing data, and what are the data sharing arrangements for these data?
   - What are the anticipated data products (e.g., databases, analyses, tools)?
   - What kinds of metadata will be associated with the data?
   - Who is the owner of the data?

2. **Data management**. Questions to consider as appropriate:

   - Where (physically) will the data be stored?
   - What type of data access or data distribution mechanism and software will be used?
   - Will the location or software for initial data entry differ from the data archive?
   - How will metadata be stored, and what provisions will be made to enable metadata searching capability?
   - Who will be responsible for entering and maintaining data archives, and over what period of time will archives be maintained?
   - What data quality controls and assurances will be provided?
   - Who will contribute to the database?
   - Will proprietary data be used? If so, describe the permissions obtained to use the data.

3. **Data Sharing**. Questions to consider as appropriate:

   - Who are the potential data users?
   - What is the appropriate timing for release of data to the public or relevant users, and why?
   - When will archived data be openly available to other users?
   - If data from non-foundation-supported or previous projects are integral to the successful completion of the Grant Purposes, will the non-foundation-supported and/or pre-existing data also be made freely available?
   - How will other users (i.e., beyond the grantee and the foundation) access data and metadata?
   - Are the publicly available data in raw form? If not, what treatments have been

applied to the data prior to their being released to the public?
- How long beyond the grant term will the data be maintained and by whom?
- Does the proposed grant include provisions for future hardware upgrades in the event that data is to be stored and maintained well beyond the project period of the grant?
- If data analysis tools are to be created as a consequence of the grant, will a tutorial be available for training of future users of the data, and if so, how can it be accessed?
- Will a data sharing agreement be required between outside vendors? If so, a brief description of the agreement needs to be provided in the grant proposal.
- Is a Creative Commons type-license appropriate for sharing the data? Why or why not?
- How will appropriate attribution to the data provider be provided?
- Do you anticipate publishing a “Data Release Paper” for referencing and sharing the data?