DATA MANAGEMENT PLAN

Our goal is to have all raw project data in the public domain a maximum of 6 months past the end date of the project – most likely much sooner (e.g. as it is published or submitted to major databases for analysis). Reasonable requests from interested parties for early data distribution will be discussed by PIs, and, if deemed appropriate, honored.

Chemical and biological oceanographic data

Oceanographic data will be collected in conjunction with the SPOT program. Some data will be collected during sampling; additional chemical assays will be performed later, in the laboratory. This aspect of the project is a continuation of the existing time series work, and protocols for data handling are well developed. The data are quality-checked, provided to cruise participants when ready, and incorporated into the SPOT website (https://dornsife.usc.edu/spot) for local distribution and eventually public release. Biological assays on the samples will performed in the Caron and Fuhrman labs. We will work with the SPOT program to coordinate these data and submit to the Biological and Chemical Oceanography Data Management Office's (BCO-DMO) Biological and Chemical Oceanography Database.

in situ oceanographic measurements

pressure, temperature, salinity (sensor), oxygen (sensor), PAR, chlorophyll fluorescence (sensor), CDOM fluorescence (sensor)

chemical and nutrient measurements

salinity, oxygen, phosphate, nitrate, nitrite, silicate, chlorophyll, ammonium

microbial community counts

viruses, bacteria&archaea, phototrophic picoeukaryotes, nanoplankton, microplankton

productivity

primary productivity (Profiling Natural Fluorometer-based estimate)
Phytoplankton growth and grazing estimates (dilution-based)
secondary productivity (^{3}H-thymidine and -leucine incorporation)

Sequence data

Raw short-read sequence data (16S or 18S tags and also shotgun metagenomes and metatranscriptomes) generated for the project will be submitted to international nucleotide data repositories, either NCBI (U.S.) or ENA (EMBL in Europe). We have found it is currently more straightforward to submit to ENA, and it then gets shared with NCBI from there. Metagenomic and metatranscriptomic sequences may also be submitted for processing to MG-RAST and/or IMG databases, where they automatically become public within 6 months. Metadata describing the samples and sample sites will be provided in a format compliant with the Genome Standards Consortium MIMS/MIENS standard (Field et al., 2008).

Assembled sequence reads ("contigs") and contig assemblies into genomes or "genomic bins" (collections of contigs presumed to be part of a single genome) are not actually raw data (their assemblies require decision-making and therefore they are interpretations of the data), but will be released as they are used for publications, or in many cases earlier as part of MGRAST or IMG submissions (contigs submitted to them for annotation are automatically made public within a year).

Data storage, distribution, and presentation

After quality checks are complete, data are put into local databases and protected spreadsheets (to prevent accidental corruption). Data in each lab are stored locally on servers and local backup drives, and for additional prevention of loss they are automatically backed up offsite.

A project-specific public website will be maintained as another mechanism for data access and distribution. Project data will be available for viewing and download (directly or by linking to appropriate files at SPOT or national databases). Graphical representations of the data will be integrated into the site. OTU groupings and taxonomic identifications will be available as tab-delimited text files. Precalculated trees and graphs of correlation and diversity analysis will be presented.

Publications will also include links to raw data and to the relevant intermediate processed data and the detailed means by which they were processed (example in Needham and Fuhrman, Nature Microbiology in press).

REFERENCES

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P. *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**: 541-547