Data Management Plan

The proposed research will produce numerous types of raw and processed data that will be disseminated to the scientific community. All sequence data will be made available at the time such data is published in the scientific literature. Intended routes of distribution are detailed below and will act as minimum modes of dissemination. Best effort will be made to utilize the available dissemination methods and best scientific practices prevailing at the time of release to promote effective sharing of data.

The Viral Resource for Metagenome Exploration (VIROME; http://virome.diagcomputing.org) web tool produced and maintained by Wommack and Polson [1], along with the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA; http://camera.calit2.net) resource maintained by their collaborators at the University of California, San Diego [2], will be heavily utilized for data dissemination purposes. VIROME provides a viral metagenome annotation pipeline and web-based metagenome exploration interface. It also acts as a repository for metadata, annotations, and processed sequences. VIROME's analysis pipeline will feed all shotgun metagenome data into a back-end MySQL database that will be used to drive the websites' export and exploration features. CAMERA will also function as a repository for these data and will further allow users to perform computational tasks utilizing the deposited data such as BLAST, sequence clustering, and hmm searches.

Appropriate metadata will be collected for any environmental samples or single cell genomes used in the proposed research. Metadata will comply with and exceed the minimum requirements of, the Genome Standards Consortium's (GSC) Minimum Information about Metagenome Sequences (MIMS), Minimum Information about a Genome Sequence (MIGS) [3] and proposed Minimum Informational about Environmental Sequences (MIENS) [4] standards. Information collected will include: date, location, site description, physiochemical sample characteristics, and collection / laboratory processing methodology. This information will be disseminated through both the CAMERA and VIROME project web sites, both of which participate in the GSC and are committed to providing MIMS/MIENS compliant metadata collection and reporting.

At present, the National Center for Biotechnology Information (NCBI) will be supporting the Sequence Read Archive (SRA) through Oct. 1, 2011 [5] and are seeking continued support in the face of budgetary constraints. As metagenomic and 16S rRNA gene library data do not fall into the scope of NCBI's Gene Expression Omnibus (GEO) database to which next generation sequence data for gene expression datasets flows, there is no apparent home for such raw data in the current NCBI organization. There are indications that the European (EMBL/EBI) and Japanese (DDBJ) SRA repositories will continue functioning for the time being [6, 7]. We will continue to monitor this situation as it develops and will make every effort to make raw sequence data available according to the prevailing best scientific practices of the time.

Processed sequence data will be made available as an integrated metagenome project at NCBI, encompassing metadata, contigs, predicted gene and proteins, and 16S rRNA gene libraries. Processed sequence data will also be made available through both the CAMERA and VIROME databases. The VIROME web tool in particular will provide an integrated method for viewing metagenomic data, complete with annotations and the underlying evidence. The VIROME project is committed to public dissemination of all annotation data. VIROME will continue to expand these offerings looking to implement any future standards in such export formats.

Single cell genomics data will be handled by the Bigelow Single Cell Genomics Center's Geneus-based (Genologics) laboratory information management system (LIMS). This LIMS has been developed to store and track project and sample metadata (e.g. environmental and field sample

handling) as well as all the analytical data, such as FACS records, real-time WGA and PCR results, DNA sequences, etc. During the first 1.5 year in operation, SCGC processed over 150,000 individual cells, and we anticipate this number to grow to over 200,000 per year, each generating diverse types of data described above. The LIMS and single cell genomics-related bioinformatics analyses are hosted on a computer cluster, mirrored in two separate buildings for secure, uninterrupted service, with database backups performed on a daily basis. All DNA and protein sequences generated by this project will be deposited in GenBank for public access immediately upon their curation and the submission of corresponding manuscripts for publication.

The collective project data will be shared among the collaborating parties through a secure ftp site and backed up daily. A redundant copy will be stored in the SCGC LIMS. Bigelow Laboratory has established a robust institutional data management policy, which defines general protocols for data use, access, share, formats, security, backups and long-term archival, in the light of scientific needs, legal responsibilities and ethical considerations.

References

1. Wommack, K. E., S. Srinivasiah, M. Liles, J. Bhavsar, S. Bench, K. E. Williamson, and S. W. Polson. 2011. Metagenomic contrasts of viruses in soil and aquatic environments. In F. J. d. Bruijn (ed.), Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats. John E. Wiley & Sons, New York.
2. Seshadri, R., S.A. Kravitz, L. Smarr, P. Gilna, and M. Frazier. CAMERA: a community resource for metagenomics. PLoS biology, 2007. **5**(3): p. e75. PMCID:PMC1821059.
3. Field, D., G. Garrity, T. Gray, *et al*. The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology, 2008. **26**(5): p. 541-547.
4. Yilmaz, P., R. Kottmann, D. Field, R. Knight, *et al*. The Minimum Information about an ENvironmental Sequence (MIENS) specification. Nature Preceedings, 2010. (http://dx.doi.org/10.1038/npre.2010.5252.2).
5. *SRA Archive is still in service*. 9 May 2011; Available from: http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement.
6. *DDBJ will continue Sequence Raw Data Archiving*. 22 February 2011; Available from: http://www.ddbj.nig.ac.jp/whatsnew/2011/DRA20110222.html.
7. *EMBL-EBI will continue to support the Sequence Read Archive for raw data*. 20 February 2011; Available from: http://www.ebi.ac.uk/ena/sites/ebi.ac.uk.ena/files/documents/sra_announcement_feb_2011.pdf.