DATA MANAGEMENT PLAN

I. Types of data

This project will generate nucleic acid sequence data, data on the prevalence and load of circoviruses amongst amphipod populations and in environmental reservoirs, and data on amphipod circoviral prevalence in incubations. The total amount of sequence information is estimated at 15 million sequence reads from 4 individual metaviromic libraries, representing 100 MBp of information. The sequence information will be captured by Illumina sequencing, to be conducted by the Columbia Genome Center Facility. Data on the prevalence and viral load of individual circoviral genotypes will be obtained by quantitative PCR. Data will also be obtained using several microbial ecological techniques outlined in the table below:

Supporting Data Measurement	Method	Reference
Salinity, Temperature, DO, pH, Depth	CTD	-
Water Column NH_4^+, NO_3^-, PO_4^{3-}	Autoanalyzer	(Parsons et al. 1985)
Sediment porewater NH_4^+, NO_3^-, PO_4^{3-}	Autoanalyzer and Porewater Squeeze	(Parsons et al. 1985)
Dissolved Organic Carbon (DOC)	Oxidation and IMS	(Hansell et al. 1993)
Sediment Total Organic Carbon (TOC)	Loss on Ignition	-
Water and Sediment bacterial production	Tritiated Leucine Incorporation	(Kirchman et al. 1985)
Sed Cellulase and □-Galactosidase Activity	Methylumbelliferone-labeled substrate uptake	(Hewson et al. 2007)
Bacterial and Viral Abundance	SYBR Green I Epifluorescence Microscopy	(Noble & Fuhrman 1998)

Data will originate from field samples collected in Lake Michigan and Santa Catalina Island, and from experimental incubations of amphipods.

II. Data and Metadata Standards

Nucleic acid sequence information will be stored in FASTA formatted files or SFF files which integrate sequence quality information with the sequence itself. The appropriate metadata to make the sequence information meaningful include accurate determination of host species identity. In addition, the latitude and longitude (measured by GPS), sampling depth (estimated by CTD), water temperature and conductivity (measured by CTD) will be measured and data saved as Excel files in the same location as the sequence data. These are consistent with genomics standards consortium (GSC) standards for metagenomic studies.

III. Policies for access and sharing and provisions for appropriate protection/privacy

Nucleic acid sequence data and complementary metadata will be made available through two avenues. They will be submitted to the NCBI short read archive (SRA) which houses most metagenomic data obtained to date. Metadata is also submitted at the same time. We will also submit our data to the cyberinfrastructure for advanced marine microbiology research (CAMERA) database. Data will be released within 12 months of generation. These databases are accessible to the public. There will be no charge for access. There are no privacy issues regarding these data. The data will not be covered by a copyright. Data on viral prevalence and viral load will be available upon request 12 months after generation, and will also be available via the project website as downloadable data files. Supporting biological oceanographic data will be deposited to the National Oceanographic Data Center (NODC) within 6 months of collection.

There will be no charge to access the data, no privacy issues, and data will not be covered by copyright. All data will be published in peer-reviewed journals.

IV. Policies and provisions for re-use, re-distribution

There will be no permission restrictions needed for these data. The data may be of interest to biological oceanographers, invertebrate zoologists, ecologists and limnologists at other institutions. The intended and forseeable users of the data are oceanographers, modelers, ecologists, metagenomicists, and microbial ecologists within academia. It is anticipated that other scientists will compare their sequence information to nucleic acid sequences generated in this study via NCBI and/or CAMERA. There are no reasons not to share these data.

V. Plans for archiving and Preservation of access

Initially, nucleic acid sequence data and data on prevalence, mortality, and viral abundance in environments will be archived on desktop computers in the Hewson Lab, and backed up onto the Cornell Microbiology Bioinformatics Cluster, which itself is backed up onto a remote server. Data will also be backed up onto terradrives in the laboratory. Data will be submitted to public databases (NCBI and CAMERA), where they will be permanently archived to preserve access to the public. A hard copy of all notes (i.e. lab notebooks) will be retained in the laboratory. All relevant metadata associated with genomic libraries will be submitted along with the nucleic acid sequences themselves. Research publications generated from this work will include all relevant data and refer readers to public databases where data is permanently archived.