Data Management Plan

(i) Data and Metadata Formats and Standards: The project will generate multiple types of data from discrete samples collected from bacterial cultures and field sampling (Table 1). We will follow the best management practices for metadata and data outlined by the Biological and Chemical Oceanography Data Management Office (BCO-DMO). We will register this project in BCO-DMO, work with the BCO-DMO staff to manage the data, and the relevant data generated by the project will be contributed to the BCO-DMO system and/or indexed in BCO-DMO for samples that are stored off-site (e.g. NCBI sequencing data). All data submitted to a public repository will be accompanied by as much metadata as possible, meeting or exceeding the appropriate MIxS / MIFlowCyt standard. Ontology terms will be developed from the raw data in accordance with the Open Biomedical Ontology and Library Foundry standards. We will use Python, shell scripts, and bash for pipeline development and data analysis on high-performance computers. We will use R and Python for quality assessment, data visualization, and statistical analyses. Any relevant analysis scripts or pipelines for interpreting the data will be open-source and available in GitHub.

Table 1. Description of the types of data to be generated during the project.

Type of samples	Brief description	Expected format	Repository
Culture / EV sample measurements	Cell/vesicle abundances, growth curves, enzyme activities, other discrete measurements	.csv/.tsv	Figshare
DNA/RNA sequences	Sequencing data from culture EV samples, 16S amplicon data	.fastq	NCBI SRA
Proteomics data	EV content	.raw/.xml	PRIDE
Imaging data	Cryo-EM images of EVs	.tif/.jpg	Figshare
Flow cytometry		.fcs	FlowRepository/ Figshare

(ii) Data availability and archives: All electronic data (both primary data files and derived work files) generated will be stored on multiple computer hard drives, shared with team members, and will be further backed up to cloud-based backup services provided courtesy of our institutions. All work products in the lab will be recorded in lab notebooks, which will be stored and maintained within the lab as per standard practices. For computational analyses, records of all processing steps and information about file contents will be maintained either in a physical or virtual lab notebook. Laboratory notebooks containing primary data will remain in the laboratories of the project's co-PIs. Original and unprocessed molecular sequence data will be shared in standard INSDC repositories (NCBI SRA/ENA) and made available within 2 years of generation or upon manuscript publication. Any computational scripts, pipelines or tools developed as part of this project will be placed into a public archive at Github (http://github.com) along with appropriate documentation. Other large-scale datasets will be stored in FigShare (https://figshare.com) or Zenodo (https://zenodo.org). The final accepted version of all peer-reviewed publications from this project will be submitted, as required, into NSF's Public Access Repository.

Educational materials: Curricula, lesson plans and exercises developed around the analysis and interpretation of these datasets will be made publicly available via peer-reviewed publications and/or through central hub repositories such as CURENet (https://serc.carleton.edu/curenet/index.html).

(iii) Policies for Data Sharing and Public Access
Access policies: All data will be available upon request, and made fully publicly available upon publication. No restrictions will be placed on the use or reuse of the datasets, as long as standard practices to properly cite the original source of the data are followed.

Code, analyses, and versioning: Any relevant analysis code will be made available via a GitHub repository to allow for adoption and refinement by the community. Where possible, bioinformatics protocols will be documented in protocols.io, where specific protocols can be assigned a DOI and referenced in associated publications.

Cultures: All living cultures used in this work will be cryopreserved and made available to qualified researchers upon request.

(iv) Roles and Responsibilities of all Parties with Respect to the Management of the Data
All members of the team will help coordinate the implementation of the data management plan. Adherence to this plan will be checked at least ninety days prior to the expiration of the award by Dr. Biller (PI) and Dr. Morris (Co-PI). Adherence checks will include a review of the data content and source code distribution.