Data Management Plan

The proposed research will include cell culture, microfluidic, and mass spectrometry studies to elucidate the role of viral-induced chemotaxis in cross-trophic microbial interactions and carbon flux. The proposed work will be managed by PI Sheri Floge at The Ohio State University/ Wake Forest University (pending transfer), PI Jeffrey Guasto at Tufts University, and PI Trent Northen at Lawrence Berkeley National Lab (LBNL). In accordance with NSF data management policy, the PIs will disseminate results through publications, conference presentations, invited talks, and group and online data management websites, making them available to all interested parties no later than the publication date.

1. **Types of Data**
   The primary (raw) experimental data produced by PI Floge's team during the course of the project will be flow cytometry, isotope ratio mass spectrometry (IRMS), and dissolved nutrient data utilized to characterize viral impacts on cell cultures and carbon and nutrient cycling in model laboratory-based experimental systems. Flow cytometry data will be stored as raw .fcs files, and IRMS and dissolved nutrient data will be stored in MS Excel (.xlsx) format and .csv files. Flow cytometry analyses will be conducted with FlowJo® data analysis software and all data processing workflows will be recorded and saved as .wsp files. Raw data generated by the Tufts team will primarily take the form of video microscopy and image data, including but not limited to cell motility in multispecies communities and in microfluidic devices for chemotaxis assays. Imaging modalities may include phase contrast, dark field, and epifluorescence microscopy. Using quantitative image analysis techniques, the Guasto lab has generated extensive custom particle/cell tracking velocimetry (PTV) software, which is used to post-process raw video data for statistical and fluid dynamic analyses. Lastly, the Northen group will generate mass spectrometry data, which will be processed and archived on the NERSC super computing facility at LBNL. The PIs will work with interested parties to make sure that datasets are intelligible, such as by communicating relevant metadata, such as experimental protocols, when and how the experimental data are collected and organized, descriptions of materials and equipment used, computer codes, analyzed data, and relevant algorithms for data processing and analyses.

2. **Data and Metadata Standards**
   Cell and virus isolate maintenance and experimental growth condition metadata will be recorded in text (.txt) and Excel (.xlsx) or ASCII-formatted files. Microbial trophodynamics will be captured via flow cytometry (.fcs). The BCO-DMO metadata authoring tool will be used prepare information for submission to the archive, and the core reporting guidelines set forth by the Metabolomics Standards Initiative will be used to ensure adequate metadata collection during culture-based experimentation and subsequent metabolomics sample preparation, and analysis. Cell motility records will be stored in the form of video (image sequence) recordings using lossless tiff images (.tif) and captured using optical (light) microscopes equipped with a variety of digital, scientific cameras. Corresponding experimental conditions are logged in text files (.txt). Tracking and analysis codes are written in the Matlab (Mathworks, Natick, MA) scientific computing language (.m), which is a worldwide standard for image processing and analyses. The resulting statistical quantification of microbial motility data will be stored in Matlab data format (.mat). Although no formal standards exist for metadata requirements for proteomic and metabolomic data, MIGS/MIMS/MIENS standards for metadata will be captured where possible. Metabolomics

data handing programs will be made available via GitHub. IPython Notebooks used for metabolomics analysis will be made publicly available through the GitHub repositories upon publication of results. Publications will provide information about the configurations and uses of these tools. Software tools reported in publications will be made available to manuscript reviewers. Participating students and postdocs will document all their progress (including data) in a timely manner, and prepare detailed guidelines regarding experimental procedures for interested parties.

3. Data Access and Sharing

The PIs will provide or facilitate access to the primary data and other supporting materials within a reasonable time. The mechanisms for dissemination of the data will include submission to the BCO-DMO data management service and scientific publications and meetings, such as journal papers, conference proceedings, and technical talks. We will continue to promptly publish findings in peer-reviewed journals, and to present at international conferences across a wide variety of disciplines. In addition, metabolomics data will be submitted to The Metabolomics Consortium Data Repository and Coordinating Center (DRCC) via the Metbolomics Workbench. Upon request, the PIs will share with other researchers the primary data, cell and virus isolates, available experimental samples, physical collections and other supporting materials created or gathered in the course of work in a timely fashion. Processed data and metadata will be available to the general public no later than the publication of major findings. The PIs' research group websites will include text and links designed to aid in the broad dissemination of data gathered during this project, including material such as technical papers, presentations, curriculum materials, as well as information about experimental procedures, and equipment and source code for image processing. Select video-recorded behaviors will also be posted on YouTube and hosted on PI Guasto's laboratory website, where such visual records can reach a broad audience, be freely available to the public, and be effectively used as teaching materials.

4. Policies for Re-use

All unpublished materials will be freely provided upon request for reuse.

5. Plan for Archiving Data

The PIs will also ask all the participants of this project to back up their raw and processed data, and experimental metadata, in designated folders for long-term archiving. Data generated in the Floge lab will be automatically archived daily to both external hard drives and cloud storage for long-term data preservation. At Tufts, since certain datasets, such as videos (image sequences), are relatively large (> 1 gigabytes per video), we will employ an internal archiving system using a network-attached storage unit configured for RAID 5 for long- term storage. Over the course of the three years of the proposed project, up to 30 terabytes (TB) of raw video data are expected to be produced, along with the resulting statistical quantification of cell motility and flow field data, all of which will be archived on magnetic hard drives for at least 5 years after the completion of the project. Offsite and cloud storage through Tufts will also be used for added security and long-term data retention of the raw and processed data. Tracking and analysis codes written in Matlab will be permanently archived on servers at Tufts. PI Guasto will be responsible for overseeing archiving of all raw and processed data at Tufts. All Northen lab data metabolomics data is stored and analyzed using the NERSC super computer at LBNL, and as such is automatically tape archived in the zipped mzML format.