The probiotic potential of symbiotic partnerships: Can contact between aquatic animals enhance the anti-pathogen potential of protective microbiomes?

Data Management Plan

Description of Data Used and/or Generated
The types of data we anticipate generating include experimental metadata such as the daily logging of aquarium parameters and fish health, Sangar sequence data in .fasta format, and meta-omic sequence data in .fastq format. All project related data will be made publicly available through BCO-DMO: https://www.bco-dmo.org/. We will also generate bacterial pure cultures, stocks of which will be frozen in sterile sea water and glycerol at -80°C.

- *Metadata.* Metadata and observations will be archived after each collection in a single laboratory notebook as well as an online notebook accessible to all research participants.

- *Sanger sequence data.* Sanger-based sequences will be submitted to NCBI's GenBank and annotated with associated domain characterizations and metadata.

- *Meta-'omic data.* All 'omic data will be filtered using default protocols associated with the sequencing platform. Additional quality filtering will be imposed using protocols established in the Stewart lab, or via downstream analytical processing (e.g., chimera checks). Following automatic quality assurance filtering on the Illumina system, demultiplexed raw sequencing data with combined quality scores (FASTQ format) will be archived and stored on servers at Georgia Tech and Montana State and for public access in the Sequence Read Archive (SRA) at the NCBI (Co-I Stewart is an adjunct professor at Georgia Tech; both the Co-I and PI retain access to Georgia Tech servers and other institutional resources). Our proposed amplicon sequencing involves multiple runs on Illumina MiSeq instruments, ultimately generating hundreds of gigabases of sequence. Rapid dissemination of these data to the broader community will be a priority - all sequence data will be made publicly accessible within one year of generation. We will fully abide to the Minimum Information about a Genome Sequence and Metagenomic Sequence standards that have been recently established by the scientific community (MIGS and MIMS, respectively). Following automatic quality assurance filtering on the Illumina system, demultiplexed raw sequencing data with combined quality scores (FASTQ format) will be archived and stored on servers at Georgia Tech and for public access in the Sequence Read Archive (SRA) at the NCBI. SRA data will be assigned a single BioProject identifier with linked metadata.

- *Bacterial cultures.* Subcultures and supporting descriptions of growth conditions will be made available to colleagues upon request.

Accountability
All individuals involved in the research will log metadata and observations in the same notebook, which will be uploaded to an online journal in OneDrive that is accessible by these individuals. Sanger sequencing data, meta-'omic data, and bacterial cultures will be maintained by PI Pratte and shared with all individuals involved. The PIs, or senior members of the PIs' labs (trained in archiving protocols), will share archiving responsibilities and cultures as needed if one of PIs leaves the project.

Data Sharing
- *Dissemination of data between partner groups*. Regular meetings will be held between PIs, undergraduate researchers, and other participants to discuss on-going research and

provide opportunities for integrative discussions. In addition to presenting research results at lab meetings, PIs, students and other participants involved in this project will be expected to present at national and international meetings and to submit research manuscripts to peer-reviewed journals. Team members will exchange data using OneDrive. Montana State University has a campus-wide license for OneDrive, providing cloud storage space for each faculty member and his/her immediate collaborators.

- *Dissemination of datasets to publicly accessible data repositories.* Rapid dissemination of sequence (omic) data and associated metadata will be a priority in this project- all sequence data will be made publicly accessible within one year of generation. We will fully abide to the Minimum Information about a Genome Sequence and Metagenomic Sequence standards that have been recently established by the scientific community (MIGS and MIMS, respectively). SRA data will be assigned a single BioProject identifier with linked metadata. Our submissions will be annotated with detailed descriptions of the sampled environment or experimental treatments (project description, lat/long, date, habitat type, hydrography, chemical measurements, etc), including brief summaries of any associated variables. Additionally, .pdf copies of all protocols used in the generation of sequence data (if not prohibited by manufacturer copyright restrictions) will be linked to the data submissions, either directly or via instructions for accessing copies on the PIs website.

Protection of Data: Security and Integrity
Metadata, and sequencing data shared within OneDrive will only be accessible to those individuals with whom the file link is shared.

Data Preservation
All analysis results will be stored using internal data formats within the utilized software packages and common formats such as TIFF, ASCII, and Excel. Phylogenetic trees will be stored in Newick format, and any multivariate analysis data will be encoded both as Excel files and as R/Matlab binary images. All accession numbers will also be archived via BCO-DMO, which will serve as a central site for identifying omic datasets generated by this project. Digital data copies will be kept in the labs of the PIs for 10 years past the lifetime of the project. The anticipated data occupies a relatively small amount of space relative to our computing capabilities and storage resources; thus, long-term preservation will be easily accomplished by keeping several copies of the data on local computers at Montana State and Georgia Tech (Co-I Stewart is an adjunct professor at Georgia Tech; both the Co-I and PI retain access to Georgia Tech servers and other institutional resources).