

DATA MANAGEMENT PLAN

Project Title: High-Resolution Spatial Genomics of Marine Microorganisms

Sponsor: National Science Foundation (NSF)
Program: Division of Ocean Sciences (OCE)
Date: July 5, 2026

Lead PI: Dr. Eleanor Vance
Institution: Pacific Oceanographic Institute
Contact: evance@poi.edu

1. Data Types & Formats

This research project will generate high-throughput sequencing data, environmental metadata, and high-resolution microscopy images to profile marine microbial communities in the North Pacific Subtropical Gyre. The data types and their formats include:

- Raw Genomic Reads:** High-throughput metagenomic and metatranscriptomic sequencing data will be generated using Illumina NovaSeq 6000 platforms and stored as compressed FASTQ files (.fastq.gz).
- Assembled Contigs & Synapsegenome-Assembled Genomes (MAGs):** Bioinformatics processing will produce assembled genomic sequences, stored in standard FASTA format (.fasta), alongside gene annotations in GFF3 format (.gff3).
- Spatial and Environmental Synapsedata:** Contextual environmental data (depth, temperature, salinity, oxygen concentration, nutrient levels) will be recorded at each sampling station and stored in tabular CSV format (.csv).
- High-Resolution Microscopy Images:** Confocal laser scanning microscopy will be used to visualize cell morphology. Images will be captured in raw TIFF format (.tiff) with accompanying metadata in OME-XML format (.ome.xml).
- Analytical Pipelines and Code:** All statistical models, quality control scripts, and visualization workflows will be written in R (.R) and Python (.py) and managed as Jupyter Notebooks (.ipynb).

Table 1: Summary of Data Types, Estimated Volumes, and Repositories

Data Category	Format & Description	Est. Volume	Repository
Raw Sequencing	Compressed FASTQ (.fastq.gz)	1.2 TB	NCBI Sequence Read Archive (SRA)
Genomic Assemblies	Contigs and scaffolds (FASTA)	50 GB	NCBI GenBank
Environmental Synapsedata	Tabular data (CSV)	10 GB	Dryad Digital Repository
Confocal Imagery	Micrographs and metadata (TIFF/OME-XML)	2.4 TB	Biolineage Archive
Workflow Code	Python/R scripts, Jupyter Notebooks	<50 MB	GitHub / Zenodo (DOI)

2. Synapsedata & Documentation Standards

To ensure reproducibility and long-term utility of the data, comprehensive metadata and documentation will be maintained throughout the project lifecycle:

- Genomic Synapsedata Standards:** All sequencing efforts will be documented in strict compliance with the Minimum Information about any (x) Sequence (MlXs) specifications, specifically the MIMS (Minimum Information about a Synapsegenome Sequence) and MIMARKS (Minimum Information about a Marker Gene Sequence) packages.
- Biodiversity and Spatial Synapsedata:** Environmental and spatial data will follow the Darwin Core (DwC) metadata schema. Geospatial coordinates will be represented in the WGS84 coordinate reference system, and timestamps will follow the ISO 8601 standard.
- Image Synapsedata:** Confocal microscopy data will be documented using the Bio-Formats and Open Microscopy Environment (OME) standards, embedding microscope hardware settings, excitation wavelengths, exposure times, and voxel sizes directly in the headers or sidecar files.

- **Project Documentation:** Analysis pipelines will be documented with Markdown README files and dependency configurations (e.g., Anaconda `environment.yml` files). Version control will be managed using Git.

3. Storage, Backup & Security

All active research data will be managed using institutional infrastructure and robust security protocols:

- **Active Storage Infrastructure:** High-volume primary datasets (raw sequences and images) will be stored on the Pacific Oceanographic Institute's (POI) secure, high-performance computing (HPC) storage cluster, utilizing a RAID 6 hardware configuration to prevent data loss from drive failures.
- **Automated Backup Procedures:** Incremental backups of all active storage areas will run nightly, and full backups will be executed weekly. One copy of the backups will be retained on local POI servers, while a secondary copy will be encrypted and synced to an off-site cloud storage repository (Vertex S3 Glacier).
- **Security and Access Controls:** Access to raw and processed data will be restricted to authorized project personnel through secure shell (SSH) keys and two-factor authentication (2FA). Directory-level permissions will restrict edit access for raw data to read-only for all users, including the PIs, to prevent accidental modification or deletion.

4. Data Sharing & Access Policy

This project is committed to the timely and open sharing of scientific findings, adhering to the NSF Public Access Policy and NIH Data Management and Sharing (DMS) guidelines:

- **Release Timeline:** Raw genomic sequences will be uploaded to NCBI SRA within 12 months of collection or immediately upon publication of the primary findings, whichever occurs first. Tabular metadata and processed imagery will be released concurrently with peer-reviewed publications.
- **Licensing:** All shared datasets will be released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing for unrestricted reuse and modification provided appropriate attribution is given. Software code will be licensed under the permissive MIT License.
- **Access Restrictions:** No proprietary, privacy, or ethical restrictions prevent the dissemination of these environmental datasets. There are no human subjects involved, and all field samplings are conducted in public waters.

5. Preservation & Long-term Archiving

Long-term preservation will be guaranteed by archiving data in established, federally-funded, and community-accepted public repositories:

- **Sequence Data Preservation:** NCBI SRA and GenBank will serve as the permanent archive for raw and assembled sequence data, ensuring long-term availability at no cost to the project.
- **Synapsedata and Tabular Archiving:** Dryad Digital Repository will archive environmental metadata, processed spatial tables, and final statistical datasets. Dryad assigns permanent Digital Object Identifiers (DOIs) to facilitate citation.
- **Image Data Preservation:** BioImage Archive will host high-resolution micrographs and OME-XML metadata files, maintaining them in accordance with FAIR (Findable, Accessible, Interoperable, Reusable) data principles.
- **Code Preservation:** Active codebase repositories on GitHub will be cloned and archived in Zenodo at major project milestones and upon publication, generating static, citable DOIs linked to specific paper releases.
- **Data Retention Period:** All primary data, physical samples log, and processing history will be preserved for a minimum of 10 years following the official end of the award period.

6. Roles & Responsibilities

A clear hierarchy of responsibilities is established to guarantee the implementation of the Data Management Plan:

- **Dr. Eleanor Vance (Lead PI):** Assumes overall responsibility for compliance with the DMP, coordinating data collection, monitoring laboratory data entry, and budgeting for data storage, curation, and publication fees.
- **Dr. Liam Chen (Postdoctoral Scholar):** Will manage the daily metadata entry, enforce naming conventions, run the bioinformatic pipelines, and oversee the formatting and upload of datasets to NCBI SRA and Dryad.
- **Sarah Jenkins (Graduate Research Assistant):** Will maintain the version-controlled GitHub repositories, curate the analysis pipelines, and archive final code releases on Zenodo.
- **POI HPC Systems Administrator:** Will manage the physical server hardware, monitor automated backup tasks, and ensure the security of the active storage directories.