Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1S92Z

Title: cdG_surface_sensing

Creator: George O'toole - **ORCID:** 0000-0002-2861-4392

Affiliation: Dartmouth College (dartmouth.edu)

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PA-20-185

Template: NIH-Default DMSP

Project abstract:

cdG and surface sensing

Start date: 01-01-2024

End date: 12-31-2029

Last modified: 07-08-2024

Copyright information:

cdG_surface_sensing

Data Type

Types and amount of scientific data expected to be generated in the project:
Summarize the types and estimated amount of scientific data expected to be generated in the project.

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

The following datatypes will be generated.

1. Bacterial strains. All strains are recorded in a digital file on a secured server at Dartmouth, backed up daily. A redundant paper copy is kept, as well as copies on local laptops (password protected). All published strains are listed in the appropriate published manuscript.

2. Phenotypic data. Raw images (JPEG), raw data (XL) and quantified phenotypic data (XL). All original datasets are saved on a secured server in a folder labeled with the corresponding figure in the paper, and published as supplementary data with the manuscripts whenever possible. All statistical analysis is listed in the raw data file or associated PRISM file, as well as described in the manuscript wherein the data are published.

3. Western blot data. Raw images (JPEG), raw data (XL) and quantified phenotypic data (XL). All original datasets are saved on a secured server in a folder labeled with the corresponding figure in the paper, and published as supplementary data with the manuscripts whenever possible. All statistical analysis is listed in the raw data file or associated PRISM file, as well as described in the manuscript wherein the data are published.

4. Our microscopy workflow generates a high volume of binary files, archived in common Tagged Image File Format (TIFF) for wide support via data handling software, both closed- and open-source, to ensure long-term accessibility. We plan to eventually perform automatic data archival onto duplicate LTO tapes each month, a set of which will be stored offsite. To track history and file changes, we will explore open-source version control tools such as Subversion (SVN) and GIT. Since significant expertise is required for extracting and interpreting cell tracking data, data sharing will be done on an individual basis. We will also explore offsite sharing and repository services with university libraries (e.g. Merritt at California Digital Library, and Chronopolis at UCSD).

5. Code used to generate and/or analyze the various data sets will be placed on github (GeiselBiofilm, Wong-Lab)

Scientific data that will be preserved and shared, and the rationale for doing so:
Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

All data will be preserved and shared upon request.

Metadata, other relevant data, and associated documentation: Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

N/A. All metadata associated with datasets is described above for each data set.

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

N/A

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

Formal standards for the data sets generated have not yet been widely adopted. However, our data and other materials will be structured and described according to best practices which are as follows: all raw data will be archived with associated metadata. All original datasets are saved on a secured server in a folder labeled with the corresponding figure in the paper, and published as supplementary data with the manuscripts.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository)

The UCLA facility uses a primary data server with redundant storage. Raw data is routinely backed up to external hard drives (cold storage), with multiple data servers at different locations to mirror the data.

At Dartmouth, all original datasets are saved on a secured server (DartFS).

For UCLA and Dartmouth, all original datasets and the resulting figures/tables are placed in a folder labeled with the corresponding figure in the paper (i.e., using the format recommended by the journal eLife), and published as supplementary data with the manuscripts. The secured server at Dartmouth is backed up daily, off site. Original datasets are also kept on password-secured laptops and deposited, whenever possible on journal databases.

Bulk cyclic nucleotide measurements are performed at the Michigan State University Metabolomics Core (Text provided by the MSU Core Facility). All instrument raw data is identified by date, investigator, experiment info and sample number and backed up to a Sharepoint server at MSU. Processed data/tabulated results are likewise identified and backed up to the MSU Sharepoint server. Processed/tabulated data are shared with the investigator and raw data files are also shared with investigators upon request (these data files are handled as described here). Raw data files are also converted to open-source (vendor independent) formats and uploaded with corresponding experimental metadata to public data repositories such as the Massive DB or Metabolomics Workbench upon request.

Microscopy images. The primary experimental data in this category will consist of microscopy data from UCLA and Dartmouth, and the associated analysis at UCLA. Microscopy data will be in the form of image sequences and movies from both our microscopes and those at UCLA CNSI. Both of these will be stored on our group PC computers and regularly backed up onto hard disk drives. In addition, detailed procedures for the construction of new equipment and for sample preparation will be recorded in laboratory notebooks and reported in publications. The PIs will be responsible for ensuring that all datasets, publications, software, strains and reagents arising from this research are managed according to this data sharing plan.

Strains and plasmid vectors will be available upon request and stored in a -80C freezer with backup power. We typical send strains 1-3 days after the request.

Laboratory notebooks and PCs will always remain the property of the institutions, with copies given to graduating investigators to facilitate future interactions.

All data will be also deposited into Zenodo, using the account geiselbiofilm.

The code for processing and analyzing data are stored in repositories with Github.

There are no sensitive data that cannot be shared.

How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

We will use Persistent Unique Identifiers (PIDs) to improve data findability across all dissemination outputs. PIDs used will include ORCID iDs for people, DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. We will also use indexed metadata, such as MeSH terms with a unique URL to make scientific data easily findable. We will keep our ORCID Records up to date with DOIs for our datasets and publications, ROR, and funder IDs to increase findability.

All data will be deposited into into Zenodo, which generated a unique DOI for each project, using the account geiselbiofilm.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

Shared data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 5 years after the end of the funding period.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

There are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated by the proposal.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

Controlled access will not be used. The data that is shared will be shared by unrestricted download.

Protections for privacy, rights, and confidentiality of human research participants: If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

N/A

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The Lead PI George O'Toole, PhD, ORCID: 0000-0002-2861-4392, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing at Dartmouth. Broader issues of DM Plan compliance oversight and reporting will be handled by the PI and Co-I team as part of general campus(es) stewardship, reporting, and compliance processes.

Planned Research Outputs

Data paper - "published manuscripts"

We strive to publish all data as soon as possible, as indicated by our publication track record. All data will be posted to the appropriate repositories as outlined elsewhere in the DMP.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
published manuscripts	Data paper	Unspecified	Open	None specified		None specified	None specified	No	No