Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1CW4V

Title: Rhode Island IDeA Network of Biomedical Research Excellence

Creator: Christopher Hemme - **ORCID:** 0000-0002-4092-211X

Affiliation: University of Rhode Island (ww2.uri.edu)

Data Manager: Christopher L. Hemme

Project Administrator: Bongsup Cho, Brett Pellock, Samantha Meenach

Contributor: Laura Bellavia, Ang Cai, Janet Atoyan

Funder: National Institutes of Health (nih.gov)

Grant: P20GM103430

Template: NIH-GEN DMSP (Forthcoming 2023)

Start date: 04-01-2024

End date: 03-31-2029

Last modified: 01-19-2024

Rhode Island IDeA Network of Biomedical Research Excellence

Because of the nature of the RI-INBRE program, the large number of investigators, and the yearly turnover of projects, we do not know exactly from year to year what projects will be funded and what type of data will be generated. However, we can predict based on current and past projects, RI-INBRE focus areas (cancer, neuroscience, and environmental health sciences) and on the current resources of the program. Data generated by the program will likely include:

- Non-human genomic data from environmental samples (e.g., metagenomics)

- Non-human genomic data from model organisms (e.g., mouse, rat, Drosophia, C. elegans, zebrafish, etc.)

- Proteomics and metabolomics data

- Microscopy and imaging data (microscopes, Cyrostat, etc.)

- Additional instrument data (e.g. HPLC, NMR, etc.)

- Data from psychological studies (human, rat)

- Data from chemical synthesis/drug development

- Program metrics data (e.g., number of punlications/presentations, student tracking, workforce development, etc.)

- Software code (e.g., bioinformatics workflows, algorithms, educational modules, virtual reality applications, etc.)

- Additional data generated by individual laboratories

RI-INBRE manages two primary core facilities, the Centralized Research Core Facility (CRCF) and the Molecular Informatics Core (MIC) as well as local satellite facilities at network institutions. The CRCF maintains an Illumina MiSeq machine for next-gen sequencing, Sanger sequencing, multiple mass spectrometers for proteomics/metabolomics/biologics work, HPLC, microscopes, and other instruments. Analysis of data generated by the CRCF can be conducted externally by the researcher or by a third party, or it may be analyzed internally by the MIC. Other sources of data such as higher throughput NGS are generated externally but may be analyzed internally by the MIC. Whenever possible, the CRCF and MIC will develop internal workflows for the storage, processing and analysis of omics data generated by the CRCF.

The MIC maintains software and hardware for the analysis of data. This includes software for molecular modeling, virtual reality (VR) application development, VR demonstrations, and analysis on long-read sequencing data from Oxford Nanopore MinION devices. The MIC also makes use of URI's various high-performance computing (HPC) systems operated by URI Information Technology Services (ITS) Department of Research Computing. For VR app development, the MIC collaborates with the URI ITS Student Technology Assistants (STA) program.

The amount of data generated depends on the number of researchers, the types of research, and the instruments used. The CRCF sees extensive use of the sequencing (~350 Gbase/year) and mass spec instruments, resulting in a steady stream of omics and chemical data. In most cases, the data is not based on clinical patient data and thus does not require additional data security procedures such as anonymization. Data types will include but are not limited to: raw spectrum data (e.g. mass spec, HPLC, NMR), processed proteomics/metabolomics data (MaxQuant, Spectronaut), sequence read files (fastq), image stacks (e.g. tiff), count matrices for omics data (.csv file), plain text files (.txt). As much as possible, we will encourage the use of standardized data formats (e.g. fastq).

In addition to scientific data, the MIC also maintains a database of program metrics that includes projects, personnel, publications, presentations, etc. This includes tracking of current and former students for the purposes of measuring workforce development. When such data is made public (e.g., presentation on websites or promotional materials) the data will be anonymized to protect student privacy.

For scientific data, the responsibility for preserving and sharing data will fall on the individual researchers. All researchers

are expected to make every reasonable effort to make their data publicly available at the earliest opportunity. RI-INBRE will require specific policies for data generated by the RI-INBRE core facilities. Human subject data will be properly anonymized before release of data to the public. Internal program metrics data will only be made available as anonymized versions or as data visualizations.

To facilitate interpretation of data, particularly for omics data, relevant information about the data will be released. This includes metadata (environmental, clinical, etc.), protocols, code, statistical models. Documentation and support materials related to clinical information will be compatible with the clinicaltrials.gov Protocol Registration Data Elements.

Code generated by the MIC will be stored on the MIC GitHub account. Bioinformatics workflows are typically coded in Snakemake (Python) and R using Anaconda and containers. All code generated by the MIC will be open source and available to the public. The MIC also employs workflows generated by other laboratories. MIC workflows are typically deployed on URI HPC resources. Individual researchers operating under the RI-INBRE program will be expected to follow similar procedures.

Sequencing data generated by the Illumina MiSeq is automatically transferred to the Illumina BaseSpace system where it can then be transferred to the user or to the URI HPC systems. The CRCF, MIC and URI College of Pharmacy will develop workflows for proteomics and metabolomics data analysis. Future workflows may include data generated for single cell or spatial omics methods.

Commercial software purchased by RI-INBRE will be made available to network researchers but may require user fees for expensive software.

Use of artificial intelligence, machine learning, and deep learning (AI/ML/DL) tools in all funded projects will be comprehensively documented. Creative endeavors including research projects are expected to be substantively original unless the use of AI/ML/DL is integral to the project. All use of AI/ML/DL tools, including prompts for generative AI algorithms, should be documented and made available through publication or by other means (e.g. GitHub). Researchers should indicate in the acknowledgements section of all publications the use of AI/ML/DL tools and the degree of their use.

For all data generated by the core facilities or by RI-INBRE researchers, FAIR (**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse) principles for data will be followed including the use of open file formats and persistent unique identifiers.

Omics data generated by the core facilities or by individual researchers will follow common data standards for the process. Users may use workflows developed by the MIC or a third party. NGS workflows will be designed to use standard omics data formats (.fastq, .gff., .gtf., .sam, .bam, .bed). Proteomics and metabolomics workflows will be designed to use standard omics data formats (). Count data from omics workflows will be stored as .csv files and relevant metadata will be stored as plain text or csv files (.txt, .csv).

For other types of experiments, data will as much as possible follow conventional data standards for the instruments/workflows in question.

Sequencing and proteomics/metabolomics data and related protocols and metadata will be required to be deposited in public repositories, specifically Genbank, Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), and Proteomics Identification Database (PRIDE). The MIC will post all relevant data (e.g. code) to the MIC GitHub account, and use of Github to share code, protocols, and small non-standardized datasets will be encouraged for individual RI-INBRE researchers.

RI-INBRE will use Persistent Unique Identifiers (PIDs) to improve data findability . 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. Data placed in public repositories will use the PIDs assigned as by those repositories (e.g. PubMed ID, accession numbers, BioProject ID, etc.). RI-INBRE will keep records of these PIDs as part of our program metrics tracking.

The core facilities will assist users on depositing data in the relevant repositories. However, responsibility for making data generated from individual projects will be the responsibility of the individual users. All data generated using RI-INBRE funding will be subject to all relevant federal data sharing policies (e.g. 2023 NIH Data Management and Sharing policies, 2025 White House mandate, etc.). It is expected that all such data will be available at the time of publication of the data. Users are also expected to properly acknowledge the RI-INBRE grant in all publications and presentations and to comply with NIH Public Access Data policies (i.e., deposition of the manuscript in PubMed Central).

RI-INBRE will follow all relevant data privacy laws and regulations (i.e., HIPAA, FERPA, IRB policies). In the event of research generating clinical and/or human subject data, all efforts will be made to protect the privacy of the subjects including but not limited to anonymization of data and use of certificates of confidentiality. All such research will follow federal inclusion policies to ensure the research benefits individuals of all sexes/genders, races, ethnicities, and ages.

Research conducted on vertebrate animal subjects will be conducted by the standards of Public Health Service (PHS) Policy on Humane Care and Use of Laboratory Animals and the Animal Welfare Act.

All human subject data funded by RI-INBRE will follow federal policies on the use of human subjects and is ultimately the responsibility of the individual researcher. Consent of participants on data sharing and preservation of data will be required. Anonymization and managed access procedures will be employed to protect participant privacy. All relevant regulations and laws (e.g., HIPAA) will be followed.

All work on human subject data will follow standard IRB protocols for the investigator's institution and HIPAA regulations which includes informed consent documentation, plans for data management and sharing, and anonymization of data. Researchers will individually chose the proper methods to deanonymize human subject data.

The Director of RI-INBRE MIC will oversee RI-INBRE data management and sharing policies, will be responsible for disseminating relevant policies to network participants, and will manage program metrics tracking with RI-INBRE administrators. The MIC Director will be responsible for data generated by the MIC. The CRCF Director and CRCF Manager will be responsible for data generated by CRCF until it transferred to the individual researcher. The individual researchers will ultimately be responsible for their own data.

Planned Research Outputs

Software - "Virtual/Augmented Reality Applications"

In cooperation with the URI Information Technology Services (ITS) Student Technology Assistants (STA) Program, the RI-INBRE Molecular Informatics Core will develop virtual/augmented reality applications for use in STEM education.

Software - "Bioinformatics Workflows and Omics Data"

General bioinformatics workflows for generation and analysis of omics data, including next-gen sequencing and mass spec omics methods, and resulting raw and processed omics data.

Dataset - "Other Research Data"

Generic research data from instruments (non-omics), experiments, field studies, etc. Data will be deposited in the appropriate repository, GitHub, or personal/institutional websites. Data will also conform to appropriate metadata standards.

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?
Virtual/Augmented Reality Applications	Software	Unspecified	Open	GitHub		Creative Commons Attribution Non Commercial Share Alike 4.0 International	None specified	No	No
Bioinformatics Workflows and Omics Data	Software	Unspecified	Open	GitHub NCBI NCBI Assembly NCBI Nucleotide NCBI Genome NCBI Gene NCBI dbVar NCBI Protein NCBI Reference Sequence Database Gene Expression Omnibus Biosamples PRoteomics IDEntifications Database BioProject protocols.io		Creative Commons Attribution Non Commercial Share Alike 4.0 International	None specified	Yes	Yes
Other Research Data	Dataset	Unspecified	Open	GitHub		Creative Commons Attribution Non Commercial Share Alike 4.0 International	MIBBI (Minimum Information for Biological and Biomedical Investigations)	Yes	Yes