Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1C058

Title: Dartmouth Cancer Center - Genomics and Molecular Biology Shared Resource

Creator: Fred Kolling - **ORCID:** 0000-0002-6178-9901

Affiliation: Dartmouth College (dartmouth.edu)

Principal Investigator: Fred Kolling, PhD

Data Manager: Fred Kolling, PhD, Christian Lytle

Project Administrator: Christian Lytle

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-20-043

Template: NIH-Default DMSP

Project abstract:

The Genomics and Molecular Biology Shared Resource (GMBSR) provides services for the isolation and analysis of DNA and RNA from bulk tissues and in single cells using microarray, Nanostring and next generation sequencing technologies. While the PI funding the project is ultimately responsible for data storage and distribution, the GMBSR provides support for the activities as outlined in this document.

Start date: 12-01-2024

End date: 11-30-2029

Last modified: 01-19-2024

Copyright information:

relationship to, your project or proposal

Dartmouth Cancer Center - Genomics and Molecular Biology Shared Resource

The GMBSR performs a number of genomics assays including bulk RNA-seq, DNA-seq and SNP genotyping as well as single cell RNA and ATAC-seq and spatial transcriptomics. The GMBSR maintains 50Tb of data stored on Dartmouth's DartFS secure, redundant and networked storage system. Data types include Illumina BCL and FASTQ files, Oxford Nanopore FASTQ, FAST5 and POD5 files, Illumina iScan IDAT files, and TIFF image files produced by several imaging platforms for spatial transcriptomics applications.

All raw data generated in the GMBSR is preserved and shared with the sponsoring PI and their designated affiliates for a period of 5 years. Data types include FASTQ, FAST5, POD5, IDAT and TIFF files.

Metadata for all projects is maintained in the GMBSR's Laboratory Information Management System (LIMS), hosted on a Virtual Machine (VM) managed by Dartmouth's Research Computing group. While these data are not freely accessible to users, they are available upon request.

The DartFS file system can be directly mounted onto PC, Mac or Linux machines, and is also accessible via the command line and Dartmouth's Andes, Polaris and Discovery High Performance Computing (HPC) clusters. Instructions for accessing DartFS by these means are available here: https://rc.dartmouth.edu/index.php/dartfs-access-guide/

All data generated in the GMBSR uses file formats established by the instrument vendor and/or are considered standards in the field. Examples include FASTQ, FAST5, POD5, TIFF and IDAT file formats.

Raw FASTQ, FAST5, IDAT and TIFF are retained in a read-only format for 5 years from the date of acquisition. Data are placed in PI-specific directories, protected and credentialed using Dartmouth-assigned NetIDs with access determined by the PI sponsoring the project. After 4 years and 11 months, an email alert is sent to the sponsoring PI indicating the data will be removed in 1 month after which time they will be responsible for the maintenance of the data. The GMBSR expects that 5 years will be sufficient time for the PI to publish the data should they choose to do so, and will submit the data to the appropriate public repository in accordance with the requirements of the funding source supporting the project.

Data is stored in a directory with the naming convention "[PI-Last-Name]". In the case that two PIs have the same last name, the naming convention "[PI Last Name][PI First Initial]" is used. Within this directory, each project is named as [Assay Type]_[Date]. The raw data files within the project directory are named using the order number associated with the project, assigned by Dartmouth's RaDar ordering system. A metadata file is provided to map RaDar order numbers to the sample names provided by the user when submitting the project.

Data are shared with the sponsoring PI within 1 week of generation and are available in a read-only format for 5 years.

Question not answered.

Question not answered.

All human subjects material provided to the GMBSR is de-identified and is untraceable by GMBSR personnel to the person of origin. It is the responsibility of the sponsoring PI to ensure proper de-identification of human subjects material and the GMBSR refuses to accept submissions that do not meet this criteria.

Fred Kolling, PhD serves as director of the GMBSR and is ultimately responsible for the management and sharing of data generated within the facility. Dr. Kolling is listed as the owner of the storage partition on DartFS and manages the user groups and associated permissions controlling access to data. GMBSR staff perform day-to-day data management tasks including data generation and migration of data to appropriate locations within DartFS.

Planned Research Outputs

Dataset - "FASTQ Files"

Sequencing data generated from Illumina and Oxford Nanopore platforms. Contains basecalled sequences, quality data and instrument and run metadata associated with the sample.

Image - "TIFF Files"

Raw image file format generated for spatial transcriptomics applications. In the GMBSR, these images contain immunofluoresence or hematoxylin and eosin-stained tissues.

Dataset - "FAST5 Files"

Data format established for raw Oxford Nanopore sequencing data.

Dataset - "POD5 Files"

Raw format for Oxford Nanopore sequencing data.

Dataset - "IDAT Files"

File format generated by Illumina iScan instruments containing position and intensity information for Illumina beadchip arrays.

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?
FASTQ Files	Dataset	Unspecified	Restricted	Gene Expression Omnibus dbSNP SRA - Reads	5 GB	None specified	None specified	No	No
TIFF Files	Image	Unspecified	Restricted	None specified		None specified	None specified	No	No
FAST5 Files	Dataset	Unspecified	Restricted	Gene Expression Omnibus dbSNP SRA - Reads		None specified	None specified	No	No
POD5 Files	Dataset	Unspecified	Restricted	dbSNP Gene Expression Omnibus SRA - Reads		None specified	None specified	No	No
IDAT Files	Dataset	Unspecified	Restricted	Gene Expression Omnibus		None specified	None specified	No	No