CIRM Data Sharing and Management Plan (DSMP) for Omics / Flow Cytometry Data

Guidelines for Discovery Awards

DO NOT SUBMIT DSMP with APPLICATION
If funded, submit DSMP as Just in Time (JIT) material during pre-funding administrative review (PFAR)

Content:

1. Background

CIRM requires awardees to manage and preserve raw data, processed data and metadata, and make applicable data and metadata available to the broader scientific community. CIRM expects all applicable data generated under a CIRM award to be shared no later than the time of publication or by the end of the award, whichever comes first. Even data not used to support a publication, including null or negative findings, are considered data.

The Intellectual Property Policy for CIRM Awards **defines "Data"** as: Scientific, clinical, or technical recorded information derived during the Project Period of an Award, regardless of form or the media on which it may be recorded, but not any of the following: financial, administrative, management data, other information incidental to contract administration, preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. "Data" excludes physical objects (e.g., laboratory samples).

CIRM requires that funded DISC awardees develop and execute a **Data Sharing and Management Plan (DSMP)**. A DSMP should reflect the proposed approach to data management and sharing at the time it is prepared and can be updated, in consultation with CIRM, during the course of the award period to reflect any changes in the management and sharing of data (e.g., new scientific direction, new repository option, timeline revision). For some programs and data types, CIRM has developed specific data sharing expectations (e.g., data types to share, relevant standards, repository selection, timelines) that apply and should be reflected in a DSMP. When no specific CIRM data sharing expectations apply, researchers should propose their own approaches to data sharing and management in a DSMP.

To ensure data processing steps can be replicated and data can be reused by other researchers, CIRM requires that data management and sharing practices be consistent with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and reflective of practices within specific research communities.

All DISC awardees are expected to develop and execute a DSMP. If a project proposes to generate **omics and /or flow cytometry data**, a DSMP must be submitted to CIRM as Just in Time (JIT) material during pre-funding administrative review (PFAR), using DSMP for Omics and Flow Cytometry Data templates. For **data from other types of experiments** (e.g., imaging, electrophysiology, etc.), CIRM may work with the awardee to develop a DSMP and establish data sharing milestones prior to CIRM issuing a Notice of Award. **This document provides guidelines for completing the templates for the DSMP for Omics and Flow Cytometry Data.**

2. Data terminology

- **Data generation**: generation of raw data
- **Data processing**: all data processing steps (dry-lab processing) following generation of raw data
- **Data production**: overarching term, referring to both data generation and data processing
- **Data products**: the result of each data generation step and each data processing step (*Each data product should be listed in the DSMP Data Catalog*)
    - **Raw data**: data produced by an instrument (e.g., raw sequence data) or by other methods, such as measurements and surveys, or obtained from a data repository
    - **Processed data**: data produced from raw data and from subsequent processing steps (e.g., quantification files, alignment files, etc)
    - **Final processed data**: data produced from last processing step (e.g., aggregated quantification, etc), on which conclusions are based
- **Metadata:** data that provide additional information needed to make shared raw and processed data interpretable and reusable

  Types of metadata
    - Data production-based metadata refers to methods used for data generation (machine, instrument), data processing (software toolkits, pipelines) and data sharing (data repositories). This information is requested in the DSMP Data Catalog.
    - Tissue donor-based metadata includes clinical, phenotypic, demographic data.
    - Sample handling-based metadata includes quality of the sample, preparation of sample, and protocols used in the course of the research.

- Map of unique identifiers refers to a document that details the persistent unique identifiers or other standard indexing tools, assigned by data repositories and used to track projects and samples, enabling other researchers to find related data deposited in different repositories.

- **Data Standards:** guidelines or formal rules for producing, structuring, naming, and describing data.
  - CIRM expects that an awardee will apply data standards that are common to their field of study in the production of data and to metadata that are deposited in a Data Repository. Examples of data standards can be found at CDISC or LOINC.
  - Data Dictionary refers to a document that defines field names, such as male/female is represented by 0/1 or 1/2 or m/f etc. (only needed if not using an existing Data Standard, such as this LOINC code for sex at birth).

- **Data sharing:** make data available to the broader scientific community by deposit in a data repository accessible to other researchers

- **Applicable data**:
  CIRM expects all applicable data generated under a CIRM award to be shared no later than the time of publication or by the end of the award, whichever comes first. Even data not used to support a publication, including null or negative findings, are considered data.
  - CIRM requires that all data that are needed for another researcher to replicate results and to reuse data be deposited in a data repository. Minimally this includes raw data, final processed data and metadata.
  - CIRM does not anticipate that researchers will preserve and share all data produced in a study. Researchers should decide which data to preserve and share based on ethical, legal, and technical factors that may affect the extent to which data are preserved and shared. Provide the rationale for these decisions in the DSMP Questionnaire.
  - Data derived from living humans must be prepared to ensure privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures), in accordance with applicable federal, Tribal, state, and local laws and regulations.
  - Data submission rules of data repositories must be followed.

- **Replicate results:** another researcher uses shared data and same code/software as original researcher to obtain the same results

- **Reuse data:** another researcher uses shared data and different tools / software to obtain new results, or uses shared data in combination with their own data

3. DSMP for Omics and Flow Cytometry Data - Instructions

DO NOT SUBMIT DSMP with APPLICATION
If funded, submit DSMP as Just in Time (JIT) material during pre-funding administrative review (PFAR)

For all **omics and flow cytometry data** you propose to generate, please prepare a Data Sharing and Management Plan (DSMP) using the following 2 templates:

Part A - DSMP for Omics and Flow Cytometry Data - Data Catalog
('DSMP Data Catalog')
Part B – DSMP for Omics / Flow Cytometry DSMP - Questionnaire
('DSMP Questionnaire').

The 2 templates to complete the DSMP for omics / flow cytometry data can be found [here].

The DSMP Data Catalog intends to capture the methods used and the data files produced during the transition from wet-lab to data generation, and at each stage of subsequent dry-lab processing. The information collected in the DSMP Data Catalog and the DSMP Questionnaire collectively aims to provide (1) sufficient detail for another researcher to repeat the data processing stages (replicate results) and (2) sufficient context to use the data in new ways with confidence in their interpretation of the data and its provenance (reuse data). The expectation is that information captured in the DSMP will be included when data are deposited in a repository, with the goal of making the data findable, accessible, interoperable and reusable (FAIR). CIRM appreciates your careful attention to this matter and your support of these aims.

DSMP Data Catalog
The fields in the DSMP Data Catalog capture essential details of what data will be generated, how they will be processed, and how they will be shared. All fields in the Data Catalog are required. When filling out the Data Catalog, please consider

- the Example Answers included as a guide for the types of inputs we seek; and
- the Field Definitions (see section 4 below and pop ups in Data Catalog column headers)

DSMP Questionnaire

- Questions 3, 4, 6, 11 only need to be answered if applicable.
    - Questions 3, 4, 6 refer to entries in the DSMP Data Catalog
- Answering questions 1, 2, 5 & 7-10, 12 is required
- Answering question 13 is optional (but highly appreciated)

4. Field Definitions for Omics / Flow Cytometry Data Catalog

Please consult the below field definitions and the Example Answers provided in the Excel Data Catalog as a guide for the requested information in the Data Catalog. Values are suggested (Example Answers) to help with consistency. If there is no suggested value that is suitable, please provide your own descriptive term.

Organization of Data Catalog

As shown for the Example Answers in the Data Catalog, create a set of rows for each omics / flow cytometry Experiment Type you will generate, with each row representing a consecutive step in data processing as follows:

- First row: What is the raw data file coming from your instrument?
- Next rows: What are the intermediate files in your processing?
- Last row(s): What is your final file type(s)?

Data Catalog Fields

Sample Data Description

Data Product: Identify each row by entering a name for the data product that will be produced at each stage of processing.

Experiment Type: The omics / flow cytometry data modality or data type represented by these Data Products.

Example Answers

- Bulk WGS
- scWGS
- Bulk RNA-seq
- scRNA-seq
- Total RNA-seq
- miRNA-seq
- HiC-seq
- snATAC-seq
- iPONDS
- Untargeted Proteomics
- FACS
- etc

File Type: The file extension or common name for the primary files produced at this stage.

Number of Data Samples: The number of samples for this Experiment Type that will be produced for this stage of processing. This is typically the number of files of the Primary File Type.

Total GB: The anticipated data volume in aggregate for all samples of this File Type. Supporting files that are not identified as separate Data Products (rows in the Data Catalog) should be included as part of the Total GB for the Data Product to which they most pertain. If less than 1, enter 1.

Specimen Details

Number of participants, animals, or cell lines: The number of participants (sample / tissue donors), animals or cell lines analyzed in this study. This is often less than the number of data samples, as the same File Type (e.g. fastq, bam, etc) is often produced from cell lines at different time points following perturbations, or from participants or animals at different collection times in a longitudinal study.

Species: Enter species of subjects or cell lines used in this study, such as Homo Sapiens, Mus Musculus.

Research Context - biological question or disease addressed

Example Answers

Biological question
- Lung development
- Effects of stress on motor neurons
- Fetal brain cell atlas
- etc.

Disease
- Parkinson's Disease
- Polycystic Kidney Disease
- Liver failure
- etc

Biological Material Classification: The type of biological material used in the experiment, to be further refined in the field "field "Type of Primary Tissue / Differentiated Product".

Example Answers

- Primary tissue
- Tissue stem cell
- Tissue stem cell derivative
- iPSC
- iPSC derivative
- hESC
- hESC derivative

Type of Primary Tissue / Differentiated Product: The type of primary tissue or differentiated stem cell derivative used in the experiment. (This is NOT a question about the iPSC source cell type.) In cases where tissue types are nuanced between two Data Products, please specify (e.g. atrial vs ventricular cardiomyocytes)

Example Answers

If primary tissues studied
- Fibroblast
- Placenta
- Whole blood
- Plasma
- CSF
- etc.

If stem cell derivatives studied
- Dopaminergic neurons
- Cardiomyocytes
- Hepatocytes
- etc.

Data Generation and Processing Method

Machine, Instrument, Software Toolkit, or Pipeline: The data generation/processing method. At the transition from wet-lab to dry-lab, provide a machine or instrument name used to generate the data. At the transition from one dry-lab stage of processing to another, please provide the name of the primary software toolkit or pipeline used.

Machine Model or Software Version: The machine model or software version used for this stage of processing.

Model / Version Links: It is incredibly helpful for reproducibility and for data reuse to identify links or references to machine models and software versions used at each stage of your processing.

Is Software Open Source? This field applies to software that is used in the data production of this stage of processing and to software that is required for a researcher to reuse the resulting file types. This field is expected to be either “yes” (i.e. open source) or “no” (i.e. not open source).

A description or justification should be included outside of this DSMP Data Catalog, in the DSMP Questionnaire, for any processing software or resulting file types that are novel (developed as part of your work) or otherwise require proprietary (not open source) software to replicate or to reuse resulting files produced at each stage of processing.

Data Sharing Method

Data Repository: The repository where the data will be deposited and shared for this stage of processing.

CIRM expects that data will be deposited in established data repositories when possible. If you propose a data repository that is not covered in the CIRM guidance for data repositories, such as a repository developed as part of your work, or on-premise or institutional hosting solutions, please provide a description of the repository and a justification for its use in the DSMP Questionnaire.

Certain intermediate stage data processing files may not need to be shared, if they can be reproduced with free open-source software available to users. You may indicate n/a if that is the case for a certain Data Product.