Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1J935

Title: Unraveling glioblastoma heterogeneity through developmentally partitioned proliferative compartments

Creator: Christopher Plaisier - **ORCID:** 0000-0003-3273-5717

Affiliation: Arizona State University (asu.edu)

Principal Investigator: Patrick Paddison

Funder: National Institutes of Health (nih.gov)

Template: NIH-Default DMSP

Project abstract:

Single cell RNA-seq studies of the deadly brain tumor glioblastoma (GBM) have helped establish that intratumoral heterogeneity is likely an important contributor to tumor growth, homeostasis, and therapeutic responses. From these data, GBMs are complex, yet maligned, neuro-developmental ecosystems, harboring diverse tumor cell types, including cells resembling astrocytes, neural progenitors, oligodendrocyte progenitor cells, mesenchymal cells and radial glial cells. While illuminating, these same datasets have, unfortunately, failed to produce general models for how GBM cells transition in and out of specific transcriptional states in tumors to maintain tumor homeostasis and/or cause recurrence. Our failure to account for their behavior and molecular features during tumor growth and regrowth represents a critical knowledge gap. This grant bridges this knowledge gap by creating a new paradigm for characterizing tumor heterogeneity. We propose that GBM tumors are comprised of developmentally partitioned proliferative compartments (DPPCs), which can be computationally defined, isolated, and analyzed from single cell genomics data. DPPCs consist of division capable tumor cells that exist in compartmentalized developmental states, trajectories, and cell cycles. Critically, our DPPC computational classification approach appears sufficient to account for the heterogeneity of cell states in GBM tumors. From preliminary data, we find that human GBM tumors are admixtures of 2 to 6 DPPCs (out of likely >20 DPPCs) that drive tumor growth, which can be captured in experimental models of primary tumors. The data is consistent with tumors having both interconvertible DPPCs, which can transition to other DPPCs, and, also, non-interconvertible DPPCs, which have clonally evolved away from other DPPCs. Importantly, this approach can be tailored to individual tumors for more personalized tracking of tumor

subpopulations. The purpose of this grant is to understand DPPCs through applications of novel machine learning tools and wet-lab experimentation in patient samples to resolve their cell cycle states and identify key intrinsic and extrinsic regulators of DPPC identity. In general, we will test the hypothesis that DPPCs, rather than developmental hierarchies, are the primary generators of human GBM intratumor heterogeneity. DPPCs may arise, be maintained, and evolve through a number of mechanisms: tumor cell plasticity, cell-cell/TME interactions, cell stress (hypoxia), and divergent clonal evolution of tumor cells.

Last modified: 08-09-2023

Unraveling glioblastoma heterogeneity through developmentally partitioned proliferative compartments

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)

This project will produce sequencing data generated from RNA-seq, scRNA-seq, and sgRNA screens, and CROP-seq. We plan to collect scRNA-seq samples from 40 GBM grade IV IDH-WT tumors.

The following data files will be used or produced in the course of the project: fastq, BAM, and cellranger output files.

Raw data will be transformed by sequence alignment, and the subsequent processed dataset used for statistical analysis.

To protect research participant identities, data will be made available for sharing through dbGAP through controlled access.

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 sequencing data and non-identifiable meta-data will be preserved and shared.

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.

To facilitate interpretation of the data, metadata will be created, shared, and associated with the relevant datasets.

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.

No specialized tools will be required to access or manipulate the shared data.

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

Data will be stored in common and open formats, such as fast or BAM for our sequencing data. Information needed to make use of this data [e.g., the meaning of variable names, codes, information about missing data, other metadata, etc.] along with references to the sources of those standardized names and metadata items will be included wherever applicable.

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.

All dataset(s) that can be shared will be deposited in GEO, SRA, dbGAP, or flowrepository.org.

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.

These repositories provide metadata, persistent identifiers (DOI), and long-term access. These repositories are supported by NIH and dataset(s) are available under a free access license.

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 10 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.

There are no anticipated factors or limitations that will affect the access, distribution or reuse of the processed scientific data generated by the proposal. However, raw data will be access controlled to ensure the privacy of the individuals whose tumors were sequenced.

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).

Given the sensitive nature of the dataset, data will be made available in dbGAP data repository, which restricts access to the data to raw sequencing data to qualified investigators with an appropriate research question and approved data use agreement (DUA).

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).

To protect participant privacy and confidentiality, shared data will be de-identified using the standard methods and ensuring access control for raw data that may have personally identifiable genetic information.

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).

Lead PI Dr. Christopher Plaisier, ORCID: 0000-0003-3273-5717, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI , Co-PI and Co-I team as part of general Arizona State University, Fred Hutch Cancer Research Center, and Duke University stewardship, reporting, and compliance processes.

Planned Research Outputs

Software - "DPPCGBM classifier"

A classifier that attributes Developmentally Partitioned Proliferative Compartment (DPPC) states to cells from scRNA-seq studies.

Software - "Probabilistic Boolean Network model of DPPCs"

A probabilistic Boolean network (PBN) model for each DPPC that predicts interventions that change DPPC abundance.

Dataset - "scRNA-seq study of DPPCs"

A dataset of 40 scRNA-seq from ~8,200 patient tumor cells.

Dataset - "Functional genomics of DPPCs"

Functional genomics screening results from DPPCs.

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?
DPPCGBM classifier	Software	Unspecified	Open	GitHub		GNU General Public License v3.0 or later	None specified	No	No
Probabilistic Boolean Network model of DPPCs	Software	Unspecified	Open	GitHub		GNU General Public License v3.0 or later	None specified	No	No
scRNA-seq study of DPPCs	Dataset	Unspecified	Open	Gene Expression Omnibus SRA - Reads NCBI dbGaP		GNU General Public License v3.0 or later	MIBBI (Minimum Information for Biological and Biomedical Investigations)	No	No
Functional genomics of DPPCs	Dataset	Unspecified	Open	Gene Expression Omnibus SRA - Reads NCBI dbGaP		GNU General Public License v3.0 or later	MIBBI (Minimum Information for Biological and Biomedical Investigations)	No	No