Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1F327a6f3

Title: Germline transcription factor mutations and genetic mechanisms of disease development

Creator: Michael Drazer - **ORCID:** 0000-0002-8171-4106

Affiliation: University of Chicago (uchicago.edu)

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-21-038

Grant: https://grants.nih.gov/grants/guide/pa-files/PAR-23-145.html

Template: NIH-Default DMSP

Project abstract:

Germline mutations in transcription factors cause multiple syndromic disorders affecting varied organ systems. Our research aims to elucidate the genetic mechanisms underlying these mutations and their broad impact on human health. Specifically, we focus on GATA2, RUNX1, and ETV6, transcription factors whose germline mutations result in immunodeficiencies, bone marrow dysfunction, and a significantly increased risk of blood cancers. These mutations remain largely undruggable, necessitating innovative therapeutic approaches.

Over the next five years, our research will answer three pivotal questions: the influence of germline mutations on gene expression across various tissues, whether the effects of different mutations converge on shared molecular pathways, and the genetic impact of additional somatic mutations. This understanding may reveal treatment strategies applicable to patients with mutations in different transcription factors and identify therapeutic targets for synthetic lethality induced by specific mutation combinations.

Our recent work has defined the 'natural genomic history' in patients with these mutations, showing distinct disease development patterns and identifying high-risk genetic states for further medical complications. By developing high-fidelity models of these mutations, we can explore gene regulatory networks in the most genetically susceptible states.

Our approach combines advanced genomics techniques with patient-derived models, including induced pluripotent stem cells (iPSCs), to study gene regulatory disruptions caused by transcription factor mutations. In the short term, we aim to identify unique gene expression patterns resulting from these mutations. Long-term objectives include using iPSCs for genomic screens to discover synthetic lethality from combined germline and somatic mutations.

Our proposed research contributes to the understanding of fundamental genetic mechanisms and advances human health. Through this grant, we anticipate establishing innovative research pathways for addressing complex hereditary diseases and developing novel therapeutic strategies.

Start date: 12-01-2024

End date: 12-01-2029

Last modified: 07-08-2024

Copyright information:

Germline transcription factor mutations and genetic mechanisms of disease development

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/obtained from human cell lines derived from patients with germline transcription factor mutations. Data will be collected from approximately 36 research participants/specimens/experiments, generating 100 datasets totaling approximately 200 GB in size. The following data files will be used or produced in the course of the project: .fastq, .bam, and .vcf files. Raw data will be transformed by RNA sequencing pipelines (Genome Analysis Toolkit), and the subsequent processed dataset used for statistical analysis. To protect research participant identities, aggregated and summarized data will be made available for sharing.

Data collection will be performed at the University of Chicago.

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.

Based on legal considerations, only the following data produced in the course of the project will be preserved and shared: de-identified sequence data, transcriptomic data, epigenomic, and/or gene expression data. It includes both individual-level and aggregate-level data.

The final dataset will include clinical data from affected patients. We will share de-identified individual-participant level (IPD) data. Appropriate measures such as de-identification of date of birth will be used for data de-identification and sharing, and informed consent forms will reflect those plans.

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 the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets. The README file will include method description, instrument settings, RRIDs of resources such as antibodies, model organisms, cell lines, plasmids, and other tools (e.g., software, databases, services), and Protocol DOIs issued from protocols.io. The data dictionary will define and describe all variables in the dataset.

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.

All raw data will be analyzed via bioinformatics pipelines developed with the Genome Analysis Toolkit, which is available to the public. This will not require specialized tools to be accessed or manipulated. We will disseminate our code for these analyses using a GitHub repository that is available to the public.

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 .fastq, .bam, and .vcf files. Information needed to make use of this data, along with references to the sources of standardized names and metadata, 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.

Human data: Studies generating human genomic and any associated phenotypic data will use an NIH-designated data repository for submission. We will deposit de-identified human data to the dbGap, GEO, and SRA databases. These data will be accessible via the NIH Genomic Data Commons, which will require controlled access given the ability to track genomic data to individual patients.

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.

Dataset(s) resulting from this research will be deposited and shared via Knowledge@UChicago (https://knowledge.uchicago.edu/), which provides metadata, ensures long-term access, and registers a digital object identifier (DOI) for each dataset to facilitate discoverability and citation. Additionally, the dataset(s) will be openly licensed and made publicly available as soon as possible or at the time of associated publication. As the University of Chicago's institutional repository, Knowledge@UChicago is supported collaboratively by the University's Library and IT Services. It is built on a cloud-based platform maintained by a service provider named TIND. Knowledge@UChicago uses an open archival information system (OAIS) compliant approach to preservation, which is complemented by fixity checking, redundancy backup, and storage of archival packages on geographically separated servers.

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.

Human data will be available no later than six months after the completion of this project or at the time of acceptance of the first publication (whichever occurs first). The investigators will submit anonymous data to the designated repositories. The NIH-designated repositories will clean these data and disseminate without restrictions.

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.

Due to legal considerations, access/distribution/reuse of the resulting scientific data that is traceable to individual patients will be limited and approved/monitored by the primary investigators and the NIH. All patients on this study have provided informed consent to the 11-0014 protocol, which allows for the preservation and sharing of data. However, to protect patient identities, all patient data will be de-identified, and all data will be stored via NIH databases that require approved access.

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).

Raw data traceable to individual patients, such as gene expression data from patient-derived cell lines, will be available by controlled access only. To access data arising from this project, users must complete a data request form from the NIH and sign a Data Use Agreement (DUA), which limits subsequent use to the terms of the approved request and requires that users maintain data security, and refrain from any attempts to re-identify research participants or engage in any unauthorized uses of the data. To get access to the data, the user must submit a valid scientific question, include a statistical analysis plan, and complete all required fields on the data request form. The NIH will review the data request for completeness.

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).

In order to ensure participant consent for data sharing, IRB paperwork and informed consent documents will include language describing plans for data management and sharing of data, describing the motivation for sharing, and explaining that personal identifying information will be removed.

To protect participant privacy and confidentiality, shared data will be de-identified by removing any personal identifying information (date of birth, address, etc.).

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).

Michael Drazer, the lead PI, ORCID: 0000-0002-8171-4106, 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 as part of general campus stewardship, reporting, and compliance processes.

Planned Research Outputs

Physical object - "Germline transcription factor mutations and genetic mechanisms of disease development"

Manuscript detailing results of proposed studies.

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?
Germline transcription factor mutations and geneti ...	Physical object	2029-01-31	Open	None specified		Creative Commons Zero v1.0 Universal	None specified	No	No