Plan Overview

A Data Management Plan created using DMP Tool

Title: Automated and Segmentation-Optional Approaches for Extracting and Characterizing Clinically-Predictive Patterns of Immune Cell Organization from Multi-Sample Imaging Mass Cytometry Datasets

Creator: First Last

Affiliation: University of North Carolina at Chapel Hill (UNC-CH) (unc.edu)

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-20-084

Template: NIH-Default DMSP

Project abstract:

Imaging mass cytometry (e.g. IMC or imaging CyTOF) is rapidly advancing for uncovering clinically-predictive spatial patterns of immune cell-types. As such technology is still in its infancy, agnostic and efficient bioinformatics approaches for deriving compact and predictive representations of the overall sample-specific cellular landscapes are lacking. Existing manual or bioinformatics approaches for linking particular cell-types and their spatial patterns to clinical outcomes typically involve a sequential process of 1) segmenting cells, 2) identifying cellular microenvironment, and 3) applying machine learning techniques for prioritizing microenvironments with prognostic power. To readily and efficiently prime multiple profiled IMC samples for downstream predictive modeling tasks that could inspire therapeutics, vaccines, or diagnostic tests, we propose to develop segmentation-optional bioinformatics techniques for automated analysis of IMC data. Such methods will computationally derive rich, mathematically abstract summaries of cellular heterogeneity and spatial organizational patterns without concern for time-consuming pre-processing steps like segmentation (Aim 1). Furthermore, we will develop machine learning techniques for prioritizing particular immune cells or their co-occurence patterns that are likely driving a particular clinical phenotype, while simultaneously correcting for additional potentially confounding clinical covariates (Aim 2). Our proposed methodology makes significant strides toward enabling fully automated analysis of clinical, multi-sample IMC datasets. Our proposed methodology strengthens and expands the techniques developed for traditional suspension CyTOF to accommodate the nuanced, clinically-predictive spatial organizational patterns that can be highlighted through IMC.

Start date: 09-01-2023

End date: 09-01-2025

Last modified: 07-08-2024

Copyright information:

Automated and Segmentation-Optional Approaches for Extracting and Characterizing Clinically-Predictive Patterns of Immune Cell Organization from Multi-Sample Imaging Mass Cytometry Datasets

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 proposed research is for developing bioinformatics approaches for automated bioinformatics analysis of imaging cytometry, or IMC data. We will develop and test our methodology with three already publicly available IMC datasets. The details of the three publicly available datasets can be summarized as follows:

1) Maternal-Fetal Interface (Krop et al., iScience. 2022): Samples were acquired from the decidua basalis at various gestational ages. N=3 samples were from women in the first trimester, N=5 samples were from women in the second trimester, and N=5 samples were from women at term. The ultimate goal of applying IMC to these tissues is to understand the interplay between immune cells and trophoblasts. These data have an antibody panel specific for 42 proteins. We will formulate a binary classification problem to classify term samples from non-term samples (e.g. first and second trimester). The raw IMC images (TIFF files) are available in a Mendeley data repository.

2) Diabetes Progression (Damond et al. Cell Metabolism. 2019): Tissue samples were extracted from the pancreases of 12 patients to understand the pathogenesis of type-1 diabetes. N=4 samples were from non-diabetic patients, N=4 samples were from patients during diabetes onset, and N=4 samples from patients with diabetes over a long duration. These data have an antibody panel specific for 35 proteins. Using these data, we will formulate a multiclass classification problem seeking to classify non-diabetes, from onset samples from long duration samples. The raw IMC images (TIFF files) are available in a Mendeley data repository.

3) Tuberculosis (TB) Granuloma (McCaffrey et al. Nature Immunology. 2022): TB-infected tissues were extracted from 15 total pulmonary (lung) and extrapulmonary tissues. N=8 of the samples were extracted from lung and N=7 were extracted from extrapulmonary tissue. These data have an antibody panel specific for 37 proteins. Using these data, we will formulate a binary classification problem to predict whether a sample was taken from a pulmonary or extrapulmonary tissue. The raw images (TIFF files) are available in a mendeley data repository.

For all three datasets, each sample has corresponding clinical labels (outlined above), which we will incorporate in our analyses.

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.*

A strength of our proposed project is to provide computationally extracted summarizations or featurizations capturing the diversity of the cellular landscapes. Such representations of each sample encode complicated patterns in the images and particular cellular co-occurence patterns that can be used for downstream prediction tasks. In each of the three datasets, we intend to generate data matrices in annData (a python library which interfaces well with the single-cell specific scanpy library) format which contain per-sample, computationally derived features. Further, the annData format will allow for any associated clinical metadata to be stored in the same data structure as the computed featurizations. This will enable reproducibility of results, and for other machine learning researchers to explore the most meaningful ways to predict clinical outcomes from IMC data.

1) Matrices of Spatially-Informed Kernel Mean Embeddings: One of our proposed featurizations is to represent each sample in terms of a spatially-informed Kernel Mean Embedding. This is, each sample will be represented as an approximately 256-length vector. Such representation will be computed across all samples and datasets.

2) Matrices of Frequencies Across Determined Microenvironments: Cellular microenvironments in each sample will be determined by partitioning the graph of super-pixels computed for each image. Assuming super-pixels are partitioned into *m* different microenvironments, then frequencies or the proportion of super-pixels in each microenvironment can be used to featurize each sample.

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.

The clinical metadata for each sample in each dataset will be extracted from the associated Mendeley data repositories where the raw IMC data were downloaded. These data repositories and clinical metadata for each IMC sample can be accessed at the following DOIs:

1) Maternal-Fetal Interface (Krop et al., iScience. 2022): DOI of DOI:10.17632/gs2bj33r6f.1.

2) Diabetes Progression (Damond et al. Cell Metabolism. 2019): DOI of DOI:10.17632/cydmwsfztj.1.

3) Tuberculosis (TB) Granuloma (McCaffrey et al. Nature Immunology. 2022): DOI of DOI:10.17632/dr5fkgtrb6.4.

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.

Our proposed methodology will certainly rely on existing software, and Python libraries to pre-process, visualize, and assist with the implementation of new methodology.

1) General Single-Cell Analysis Tools: We will rely on the following Python libraries to access many existing tools for single-cell data analysis.

- Scanpy (general single-cell analysis). This is an open-source Python library.
- AnnData (creating required data structures). This is an open-source Python library.
- SquidPy (spatial single-cell analysis). This is an open-source Python library.

2) Image Analysis Tools: We will use the following software and libraries in order to process, visualize, and work with our images.

- Fiji (general image processing package). This is an open-source software package.
- Mesmer (python library for cell segmentation). This is an open-source Python library.
- scikit-image (python library for general image processing, and defining super-pixels). This is an open-source Python library.

3) General Python Libraries for Developing Methodology: We will use the following libraries to help us with implementing the software that we will develop.

- Numpy (general linear algebra, matrix computations). This is an open-source Python library.
- Networkx (general graph computations). This is an open-source Python library.
- PyGSP (graph signaling processing). This is an open-source Python library.
- matplotlib (plotting and data visualization). This is an open-source Python library.

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

The computationally-derived features and associated clinical metadata will be stored in annData format and saved as .h5ad files for easy access and future experiments by ourselves or other scientists.

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; see Selecting a Data Repository)

All of our computationally-derived IMC features and sample-specific metadata will be made available in a Zenodo repository. Processed data files with features and clinical metadata will be distributed to the Zenodo repository as .h5ad files in annData format.

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.

Processed IMC features distributed in Zenodo repositories will be findable and identifiable through hyperlinks to the associated and specific Zenodo repository webpage and through a DOI.

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.

Data and associated code to reproduce the results will be made available when the associated manuscripts are submitted for publication. Data will be available forever.

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. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

Since we are simply providing mathematically-abstract representations for individual samples of already publicly-available IMC data (in Mendeley data repositories), we do not anticipate any limitations related to access, distribution, or reuse of scientific data.

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).

Controlled access will not be used. The data that are shared will be shared by unrestricted download.

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).

Our study does not involve human research participants, hence, this question is not applicable.

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).

The lead PI Natalie Stanley 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 and research team as part of general stewardship, reporting, and compliance processes.
