Computational Pipeline for Cell Type Annotation of HuBMAP Spatial-omics Data

Zachary Deutsch, Yang Miao, John Hickey
Department of Biomedical Engineering, Duke University, Durham, NC

Website

Hickey Lab

Duke Biomedical Engineering

Introduction

Open data facilitates scientific collaboration by allowing researchers to leverage, share, and combine data. The Human BioMolecular Atlas Program (HuBMAP) is a research consortium focused on collecting single-cell datasets of healthy organs of the human body.

- However, many HuBMAP Co-detection by indexing (CODEX) datasets lack cell type annotations.
- Our group is using our background in spatial-omics data analysis to annotate these datasets.
- We have created an end-to-end Jupyter Notebook that can be used with the HuBMAP virtual workspace to annotate these datasets within the HuBMAP Data Portal.

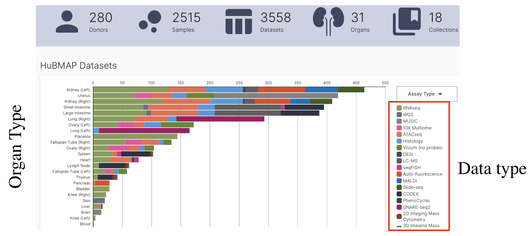


Fig. 1. HuBMAP Data Portal

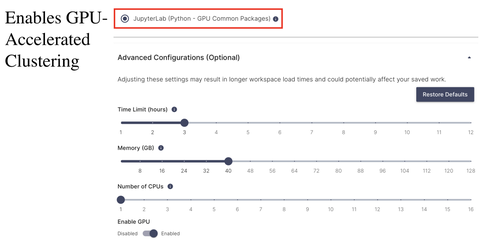


Fig. 2. HubMAP Virtual Workspace

Methods: HuBMAP Dataset Extraction

- For the purpose of our pipeline, we choose processed CODEX from the small and large intestine.
- We retrieve 8 datasets each from 8 donors, or 64 total datasets.
- The marker expression profile for each single cell, as well as its spatial coordinate information, is then extracted for downstream analysis.

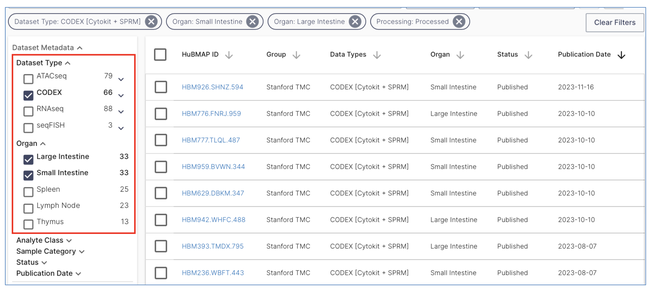


Fig. 3. Small and Large Intestine Codex Datasets

Methods: Data Preprocessing

- We standardize the marker names across all datasets to keep the cell type resolution consistent.
- The markers that are present in all datasets are preserved for downstream processing.
- We identify and retain 49 markers that are present in all of the datasets.
- The 64 individual datasets are then merged into one .csv file.
- For unsupervised clustering, all markers are z-normalized for each donor to ensure that one marker is not dominating just because of a higher signal range than another.
- We then remove noise to eliminate cells that stain positive for too many markers in the CODEX multiplex tissue imaging experiments by Z-score thresholding and setting low nuclear intensity cutoffs.

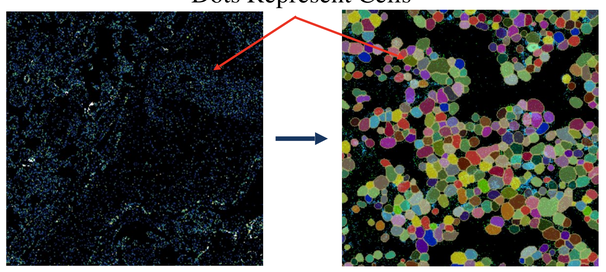


Fig. 4a.
Inputs: Raw CODEX Images

Fig. 4b.
Segmentation
(Mesmer/CellProfiler)

	x	y	Marker 1	Marker 2	Marker 3	...
Cell 1						
Cell 2						
Cell 3						
...						

Fig. 4c. Output Dataframe

	x	y	Marker 1	Marker 2	Marker 3	...
Cell 1						
Cell 2						
Cell 3						
...						

Fig. 5a.
Z-Normalization For Each Marker

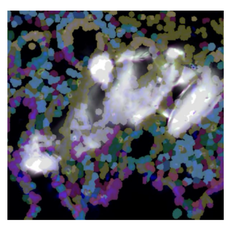


Fig. 5b.
Z-Score Thresholding

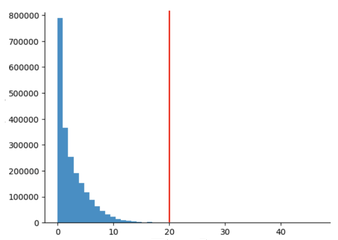


Fig. 5c.
Sum Z-score Frequency With Cutoff

Fig. 5d.
Low Nuclear Intensity Cutoff

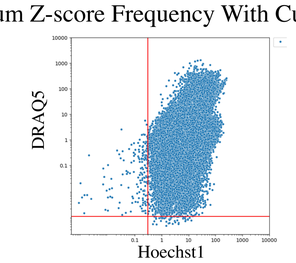


Fig. 5e.
Hoechst1 vs. DRAQ5 With Cutoffs

Methods: Leiden Clustering

- We perform harmony batch correction to adjust for systematic differences that arise between donor, sample, or slide number.
- We employ Leiden Clustering, a type of unsupervised clustering, to annotate the cell types.
- We compute the nearest neighbors distance matrix and a neighborhood graph of observations for use in Leiden Clustering.
- Using the rapids-single cell package, we GPU-accelerate Leiden Clustering for cell types.
- We prefer to overcluster as it helps with separating noisy clusters and makes it easier to have pure clusters and introduce bias into the clusters through subsequent subclustering/reclustering.

	x	y	Marker 1	Marker 2	Marker 3	...
Cell 1						
Cell 2						
Cell 3						
...						

Fig. 6a. Cleaned Dataframe

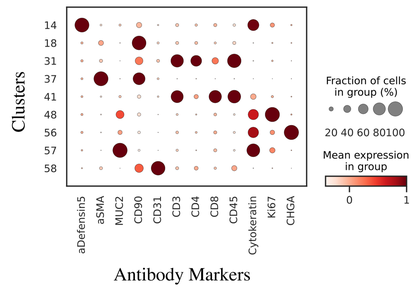


Fig. 6b. Subset of CODEX Cell Type Matrix

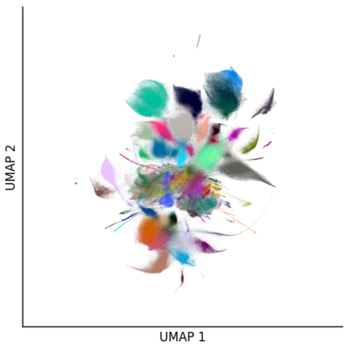


Fig. 6c. UMAP Plot For Visualizing Clusters

Methods: Dotplot Generation And Preliminary Cluster Labeling

- Using the combination of markers expressed and relative expression levels, each cluster is assigned to a cell type, or designated recluster/subcluster.

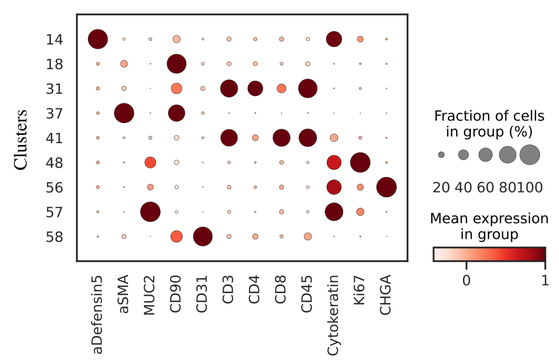


Fig. 7a. Subset of CODEX Cell Type Matrix

Fig. 7b. Subset of CODEX Cell Type Matrix

Methods: Spatial Visualization With Vitessce

- We use Vitessce, a spatial dataset visualization tool, to visualize the clusters on the original image and evaluate the clustering results by overlaying cell marker expression and cluster assignments.
- This is important because sometimes a cluster will not show high staining for any markers, appear as an artifact, or is impure.
- After identifying and reclustering clusters that contain mixed cell types or artificial cells, we revise our initial cell type annotations.

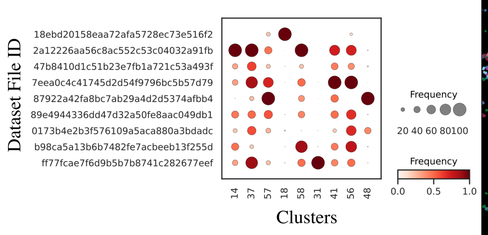


Fig. 8. Cell Counts Per Subset of Clusters Per Subset of Datasets

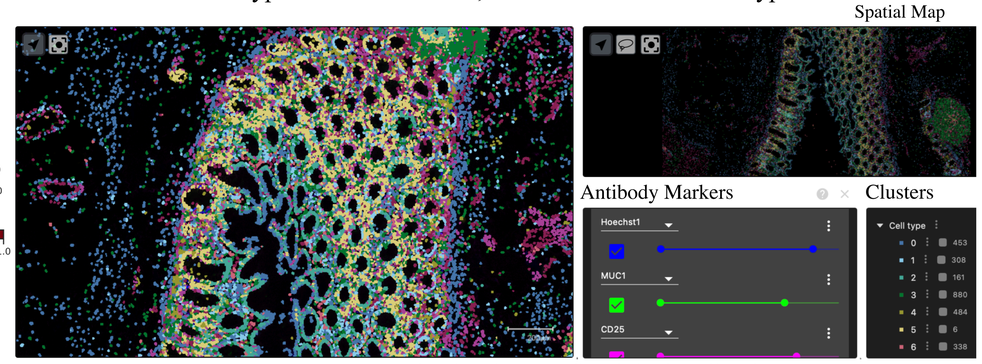


Fig. 9. Vitessce Spatial Visualization

Conclusion and Future Work

- We have created a computational pipeline for cell type annotation of small and large intestine HuBMAP spatial-omics data.
- In the future, we aim to annotate other organs on the HuBMAP Data portal, such as the lymph nodes, spleen, and thymus.
- These will provide valuable annotations that will enable scientists to both study how the body is organized and act as a healthy reference to diseased datasets they collect.

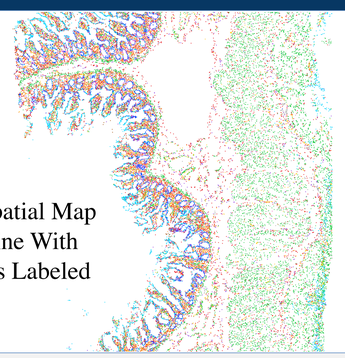
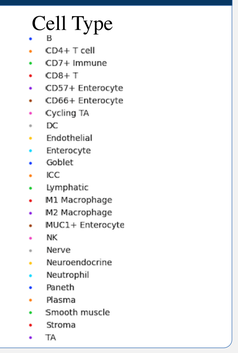


Fig. 10. Spatial Map of Intestine With Cell Types Labeled



References

1. Black, S., Phillips, D., Hickey, J.W. et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nat Protoc 16, 3802–3835 (2021).
2. Hickey, J.W., Becker, W.R., Nevins, S.A. et al. Organization of the human intestine at single-cell resolution. Nature 619, 572–584 (2023).

Acknowledgments

1. This research was supported by NIH 3OT2OD033759-01S4.
2. The results here are in whole or part based upon data generated by the NIH Human BioMolecular Atlas Program.