Comparative Evaluation of Unsupervised Clustering Methods for Cell-Type Annotation in HuBMAP CODEX Intestinal Tissue

Zachary Deutsch[1]; Yang Miao[1]; John W. Hickey[1]
[1]Department of Biomedical Engineering, Duke University, Durham, NC

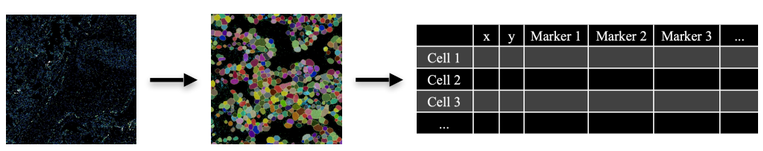
Introduction

- HuBMAP has over 4,300 datasets across 31 organs.
- However, the lack of standardized cell type annotations remains a barrier to downstream analysis.
- Recent work focuses on developing pipelines to preprocess, cluster, visualize and annotate cell types, but developing a useful, end-to-end cell type annotation pipeline requires high annotation accuracy and low human intervention.
- We address this problem by comparing the accuracy of the cell type clustering methods (Leiden, FlowSOM, SpatialSort, and PIXIE) and testing feature engineering methods (use of cell morphology features and use of weighted marker features, respectively).

Dataset

- An annotated HuBMAP healthy intestine with 49 protein markers to retroactively compare accuracy of clustering and feature engineering.
- Clustering method requires different features: Leiden and FlowSOM require protein markers, SpatialSort requires protein markers and spatial coordinates, and PIXIE requires raw expression and cell masks images.

CODEX images input → Segmentation → Z-normalized matrix

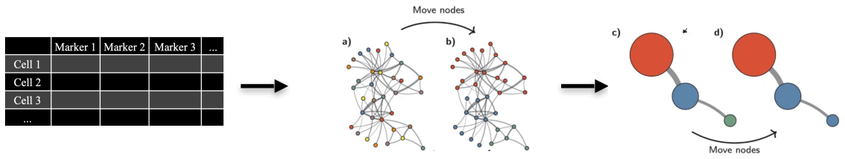


	x	y	Marker 1	Marker 2	Marker 3	...
Cell 1						
Cell 2						
Cell 3						
...						

Clustering Methods

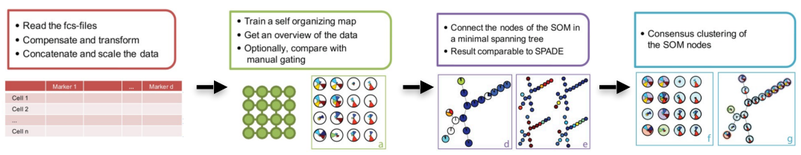
- **Leiden** clustering is a graph-based community-detection algorithm that partitions cells and optimizes a network. [1]

Cell by marker matrix → Move nodes across communities to find a partition and create network → Optimize network by moving nodes



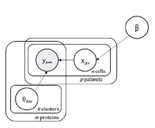
	Marker 1	Marker 2	Marker 3	...
Cell 1				
Cell 2				
Cell 3				
...				

- **FlowSOM** is a self-organizing map–based method that projects single-cell data onto a structured grid and refines the resulting nodes into clusters. [2]



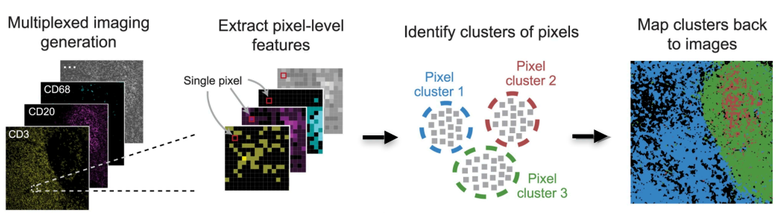
- **SpatialSort** is a Bayesian method that incorporates spatial neighborhood information into clustering. [3]

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

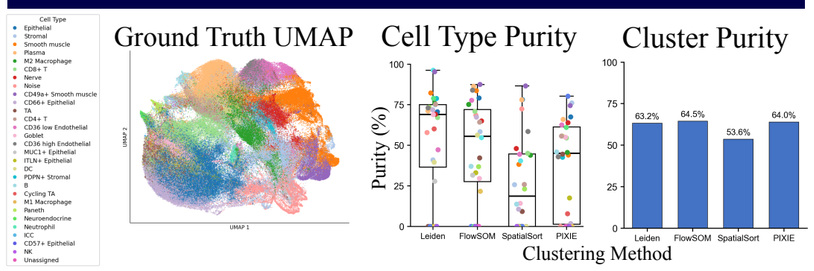


Variable	Represents	Distribution
y_{pnm}	Marker m expression of cell n of patient p	Gaussian
x_{pn}	Cell type label of cell n of patient p	Modified Potts
θ_{km}	Combined mean and variance of marker m expression of cluster k	Normal-Gamma
β	Cell-cell interaction parameter	Beta

- **PIXIE** is pixel-to-cell method that uses pixel-level phenotype features with cell-level aggregation. [4]



Clustering Results



Feature Engineering Methods

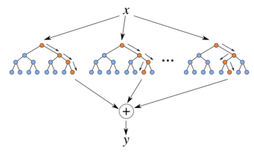
- Clustering with protein markers and **cell morphology features**: area, perimeter, major axis length, minor axis length, eccentricity, solidity, circularity, aspect ratio, equivalent diameter.

	Marker 1	Marker 2	...	Marker 49	Area	Perimeter	...
Cell 1							
Cell 2							
Cell 3							
...							

→ Clustering

- Random Forest **Mean Decrease in Impurity** (MDI) feature weighting.

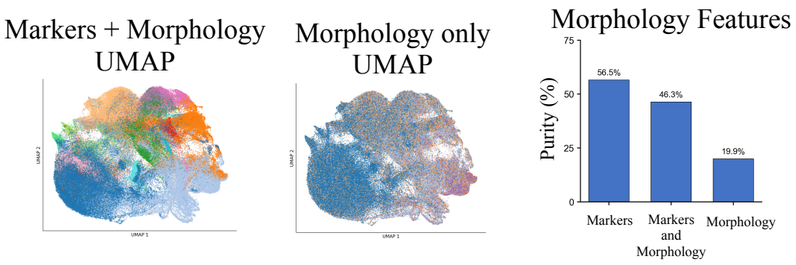
	Marker 1	Marker 2	Marker 3	...
Cell Type 1				
Cell Type 2				
Cell Type 3				
...				



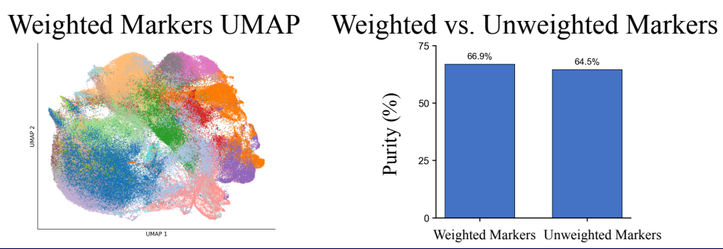
Marker 1	Weight 1
Marker 2	Weight 2
...	...

Feature Engineering Results

- Morphology features with markers and morphology features alone, when clustered using FlowSOM, had lower purity compared to the markers alone.



- Random Forest MDI feature weighting with FlowSOM outperformed unweighted markers



Conclusion

- In conclusion, FlowSOM clustering yielded the highest purity at 64.5% with unweighted markers and 66.90% with weighted markers on an annotated healthy human intestine dataset
- Future work should focus on analyzing how improved protein marker prediction and cell segmentation affects clustering

References

[1] T.raag V. A. Traag, L. Waltman and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Scientific Reports, vol. 9, Article no. 5233, 2019.
[2] S. Van Gassen et al., "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data," Cytometry A, vol. 87, no. 7, pp. 636–645, 2015.
[3] E. Lee et al., "SpatialSort: A Bayesian model for clustering and cell population annotation of spatial proteomics data," Bioinformatics, vol. 39, suppl. 1, pp. i131–i139, 2023.
[4] C. C. Liu et al., "Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering," Nature Communications, vol. 14, p. 4618, 2023.
[5] J. L. Rumberger et al., "Automated classification of cellular expression in multiplexed imaging data with Nimbus," Nature Methods, vol. 22, pp. 2161–2170, 2025.
[6] J. W. Hickey et al., "Organization of the human intestine at single-cell resolution," Nature, vol. 619, pp. 572–584, 2023.

HICKEY LAB

Duke Health Data Science Poster Showcase
Duke University, December 12, 2025

DukeHealth