Reference-based cell type classification using spatial molecular imaging of gene expression

Zhi Yang[2], Patrick Danaher[2], Lidan Wu[2], Evan W. Newell[1], Shanshan He[2], Isabel Lee[2], Aster Wardhani[2], Youngmi Kim[2], Andrew Wu[2], Michael Rhodes[2], Dae Kim[2], Joseph Beechem[2]

1. Fred Hutchinson Cancer Research Center, Vaccine and Infectious Disease Division, Seattle WA 98109
2. Nanostring Technologies, Seattle WA 98109

nanoString

NanoString Technologies
530 Fairview Avenue North, Seattle, WA 98109

Summary

As consortia like the Human Cell Atlas (HCA) authoritatively catalog cell types, single cell gene expression studies are shifting from simply defining cell types to exploring cell behavior. Under this framework, classifying cells via unsupervised clustering becomes unnecessary, and we can instead employ supervised cell classification, which is simpler to implement and less sensitive to subjectively chosen models and cluster names. Here we introduce a toolkit for supervised classification of single cells using reference gene expression profiles, with an emphasis on spatially-resolved single cell data.

Our approach incorporates several advances to the field:

1. It uses a data-generating model for spatial expression data to calculate each cell's likelihood of belonging to each cell type, gaining large performance improvements.
2. This framework enables p-values for cell classifications and flagging of unreliable classifications.
3. It automatically infers a hierarchy from broad to finely-defined cell types, then assigns cells to the most specific cell type possible.
4. It gains accuracy by harnessing fluorescently-labeled protein data, which is collected alongside high sensitivity gene expression on the Spatial Molecular Imager (SMI).

SMI is for research use only and not for use in diagnostic procedure.

Overview of Reference-based Cell Type Classification (RCTC)

Model Specification and Notation

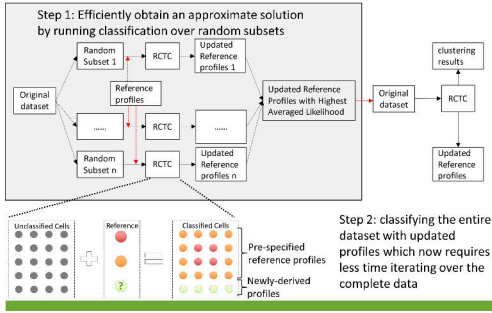
We use a semi-supervised EM algorithm with the below data-generating model:

$$y_{ij} \sim NegBin(s_i X_{k(i),j} + b_{ij}, \theta)$$

A list of notations:

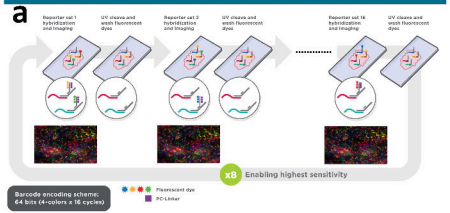
i = cell ID, j = gene ID,
s_i = scaling factor for cell i
$k(i)$ = cell type of cell i
X = cell profile matrix, $X_{k(i),j}$ is the gene expression of gene j in cell type $k(i)$
b_{ij} = expected background
θ = pre − specified dispersion parameter for the Negative binomial distribution

Simplified Diagram of the Classification Algorithm

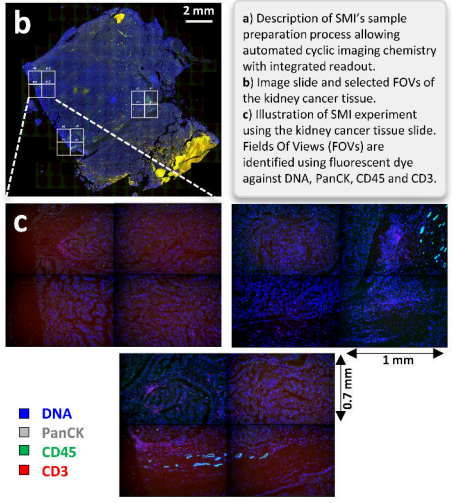


Step 2: classifying the entire dataset with updated profiles which now requires less time iterating over the complete data

Introduction to SMI Workflow and Data Collection



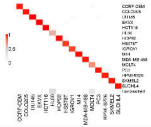
300-plex data on Kidney Cancer Tissue



a) Description of SMI's sample preparation process allowing automated cyclic imaging chemistry with integrated readout.
b) Image slide and selected FOVs of the kidney cancer tissue.
c) Illustration of SMI experiment using the kidney cancer tissue slide. Fields Of Views (FOVs) are identified using fluorescent dye against DNA, PanCK, CD45 and CD3.

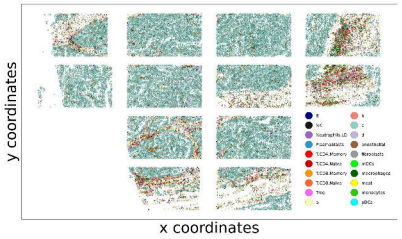
High-accuracy Cell Type Classification in High-plex Cell Line Data

RCTC yielded high accuracy in classifying the cell types among 1000-plex SMI CPA data using the averaged expression profiles. All the accuracy rates are greater than 98% except for HL60 (61%), MOLT4 (73%), IGROV1 (92%), and HS578T (94%). The misclassification rates are less than 4% by considering the unclassified category.

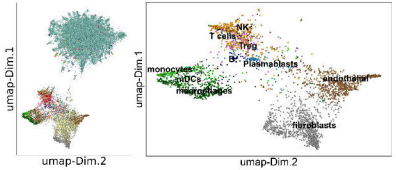


300 plex Panel Spatially Classifies Individual Cells in Kidney Cancer

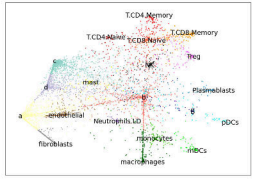
High sensitivity cell characterization of kidney cancer



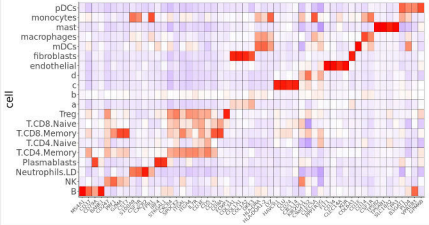
UMAP projection separates cell types



Cells are positioned according to their probabilities of being each cell type

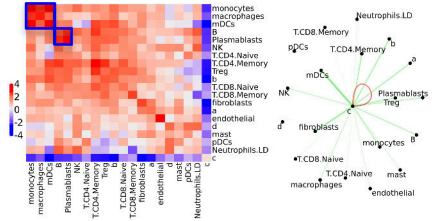


Mean expression of top marker genes by cell

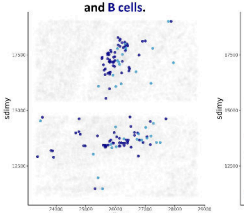


Cell-cell Proximity Interaction

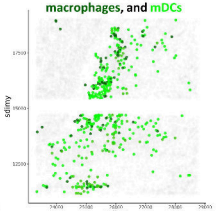
The heatmap based on cell-cell proximity scores displays the frequency of physical contact between cell populations. The enriched/depleted interaction between cells agrees with the expected cell behaviors.

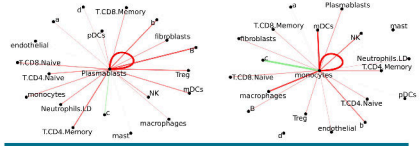


Enrichment between plasmablasts and B cells.



Enrichment among monocytes, macrophages, and mDCs





Conclusions

1. The semi-supervised EM-based cell classification algorithm provides **an easy-to-use, reliable, and flexible clustering method** allowing users to incorporate reference profiles and to construct unknown cell profiles in addition to the references.

2. SMI measures **high plex** and **high sensitivity** spatial gene expression of kidney cancer on a molecular level in which we identified a total of 20 cell types including 16 known cells and 4 unknown cells. These derived cells types are consistent with the immunofluorescent image and cell behaviors.

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.
www.nanostring.com | info@nanostring.com
© 2021 NanoString Technologies, Inc.