# Creating a condensed scRNA-seq atlas of the lung

*… and the framework to create other condensed atlases*

# Goals

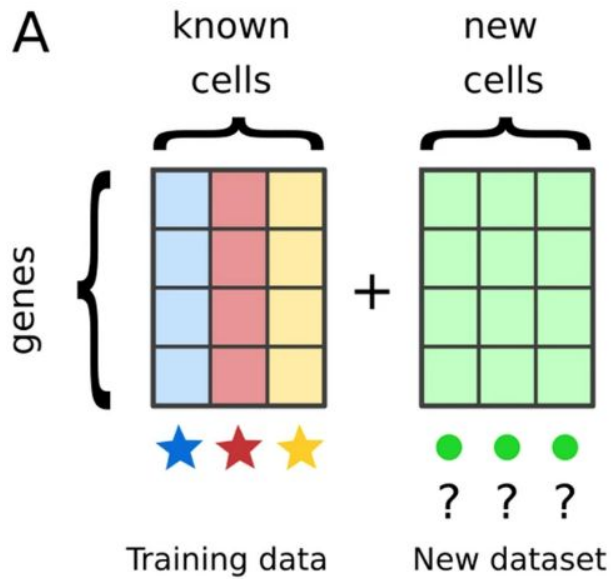
- Condense very large, dense scRNA-seq datasets into smaller, most relevant data
- Compare and contrast different scRNA-seq datasets cohesively
- Create a navigable interface to easily answer questions about scRNA-seq datasets
    - Does cell type A express gene A more than cell type B?
    - Is expression of gene A higher earlier in development or later?
    - In dataset A what does expression of gene X look like?

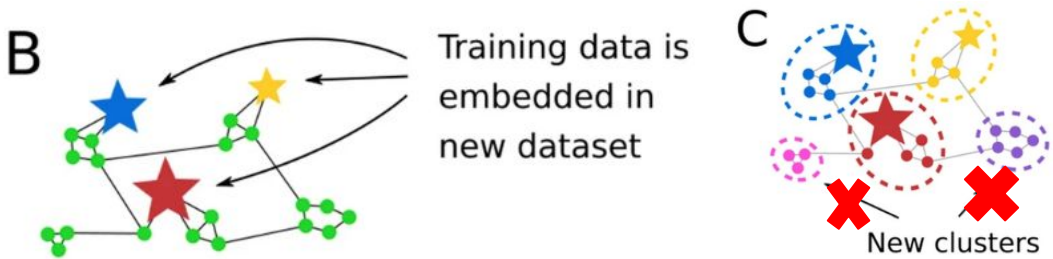
# Outline

1. Align datasets
    a. Assign the same set of cell types to all datasets
    b. Ensure metadata is consistent
2. Extract relevant data from each dataset into a database
    a. Across whole dataset and each metadata
        i. Cell count
        ii. Average expression of each gene
        iii. Percent expressing of each gene
3. Create API to pull and visualize data
    a. Ying’s project

# Align datasets

- Choose a atlas scRNA-seq dataset to align others to
  - This has cell type list to assign to all data
- Run northstar package using atlas above
  - Run only looking for neighbors in the atlas
    - This forces only cell types from atlas to be assigned, no identification of new clusters/cell types
- Make sure metadata naming is consistent across all datasets

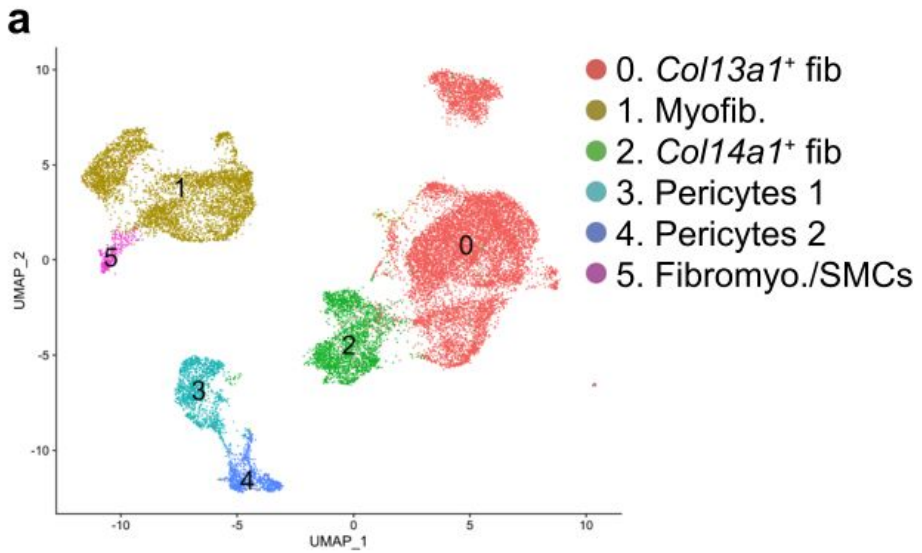



Zanini 2020

# Example reannotation

- Dataset similar to ours in developing lung
  - Has additional timepoints we do not
- Mesenchymal cells annotated with much less diversity than our own
  - E.g. missing vascular smooth muscle cells (VSMC)
- Reannotation using Northstar identifies Pericyte 1 as VSMC



# Extract relevant data

- Standardized pipeline of handling count matrices
- Must have consistent naming conventions for
  - Cell type
  - Timepoint
  - More metadata to be added in the future
- Important plotting info
  - Gene expression averages
  - Gene expression percent of group expressing
  - Abundances
  - More?

# Data Structure ideas/options (work in progress)

Python dictionary

Levels

1. Metadata level 1 (cell type)
    a. By metadata columns
        i. Cell counts
        ii. Average expression
        iii. Percent Expressing
2. Meta data level 2 (cell type _ timepoint)
    a. E.g.
        i. Cell type / timepoint
        ii. Cell type / dataset

Stored as H5 files

| Cell type Avg. Exp. | Gene A | Gene B | Gene C |
|---|---|---|---|
| ASM | | | |
| AT2 | | | |
| | | | |

| Cell type/Timepoint Avg Exp | Gene A | Gene B | Gene C | |
|---|---|---|---|---|
| ASM_P1 | | | | |
| ASM_P7 | | | | |
| … | | | | |

# Web Interface

- Ying’s project
- Creating a user friendly interface to access data and ask relevant biological questions
- Heatmaps / Dotplots / Barplots