Benchmarking strategies to identify single-cell phenotypic changes: pyCellPhenoX and miloR

Presented By: Zachary Caterer

Rotated at Fan Zhang Lab at CU Anschutz

2024 / 10 / 18

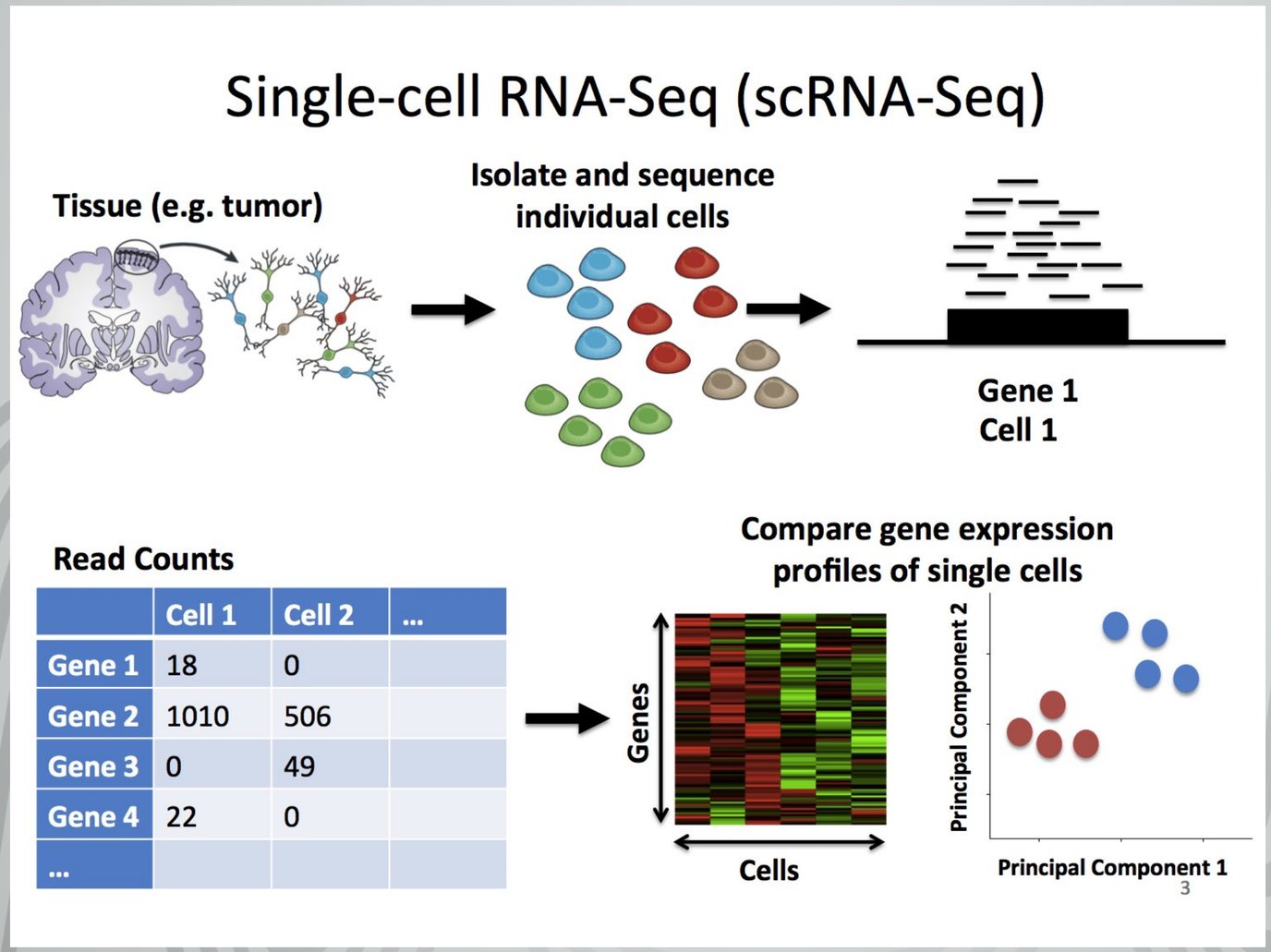
fanzhanglab.org

BioFrontiers Institute
UNIVERSITY OF COLORADO **BOULDER**

AGENDA

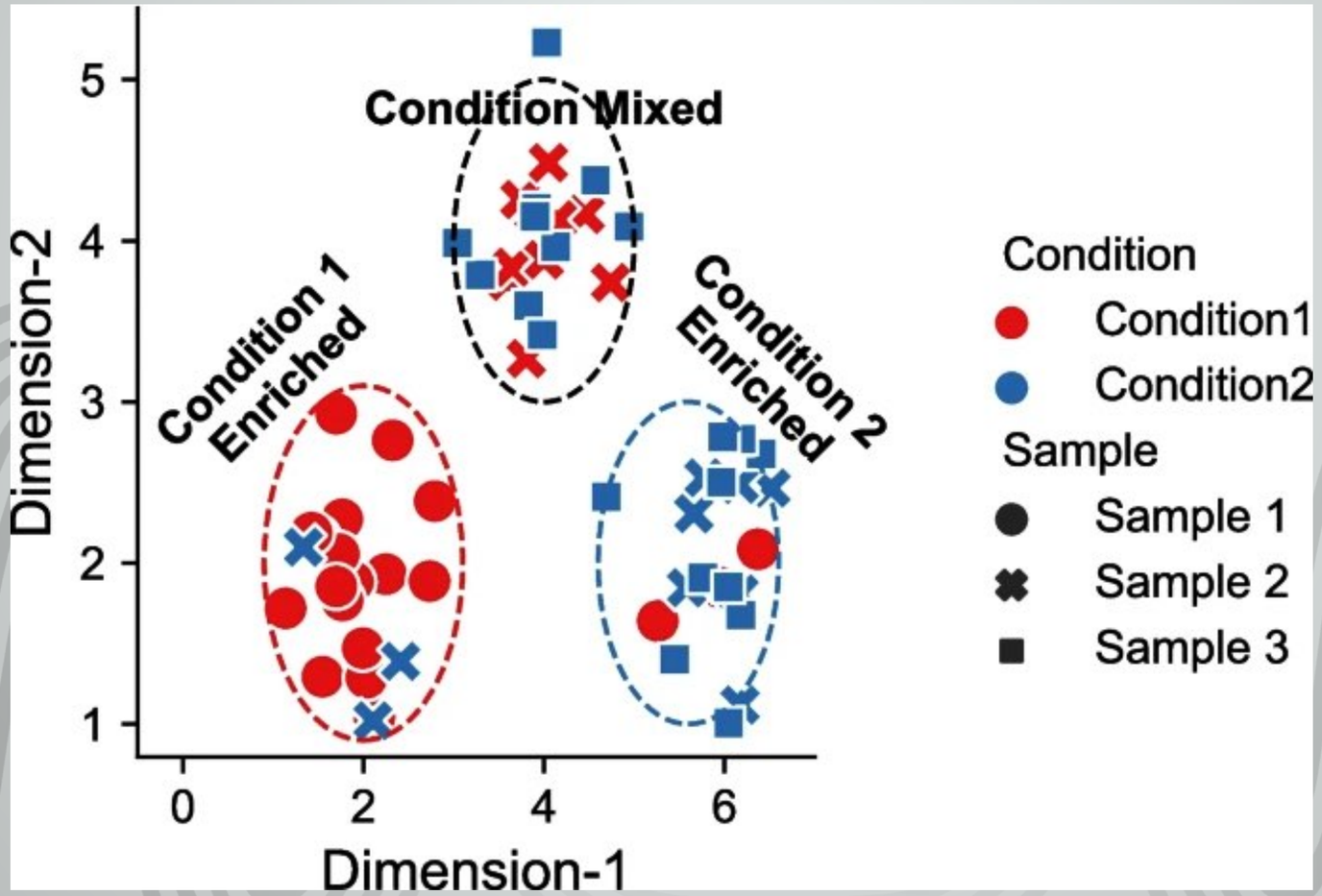
Background

Single-cell RNA-Sequencing



Hicks 2017

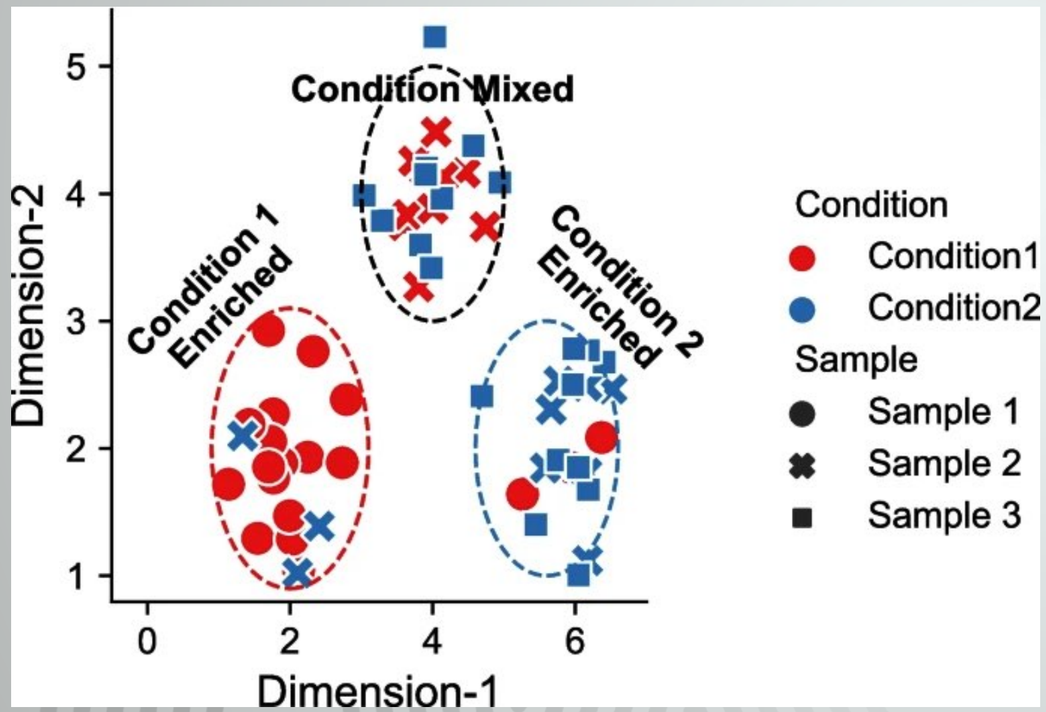
Differential Expression and Abundance



Yi 2024

Differential Abundance Clustering

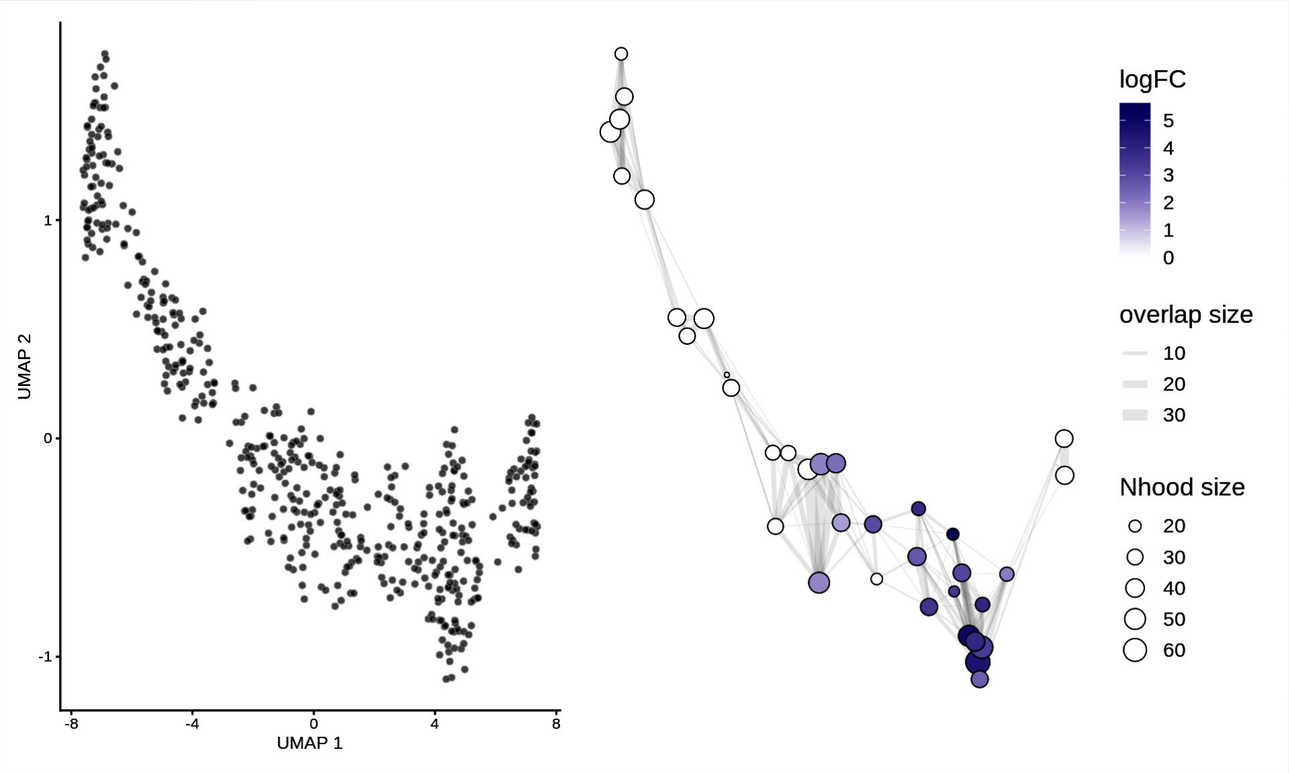
Cluster Based Differential Abundance



Yi 2024

1. Group similar cells into clusters
2. Compare the abundance of clusters
3. Organizing cells into groups that share common features

Cluster Free Differential Abundance

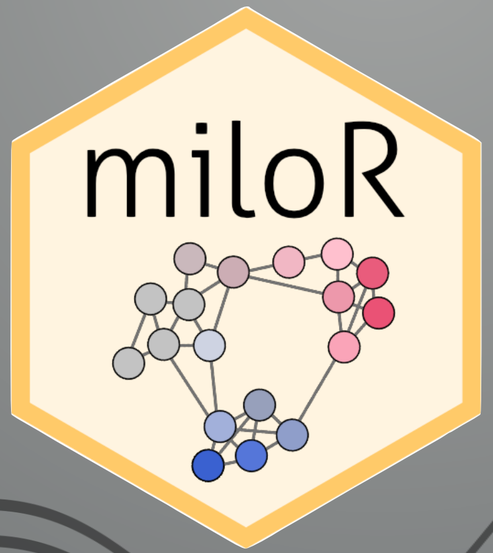


1. Evaluate abundance of each individual cell
2. Inferred which cells are associated with different conditions
3. More robust analysis of cellular heterogeneity

Dann 2022

miloR

Dann, *Nat Biotechnol*, 2022.
doi.org/10.1038/s41587-021-01033-z

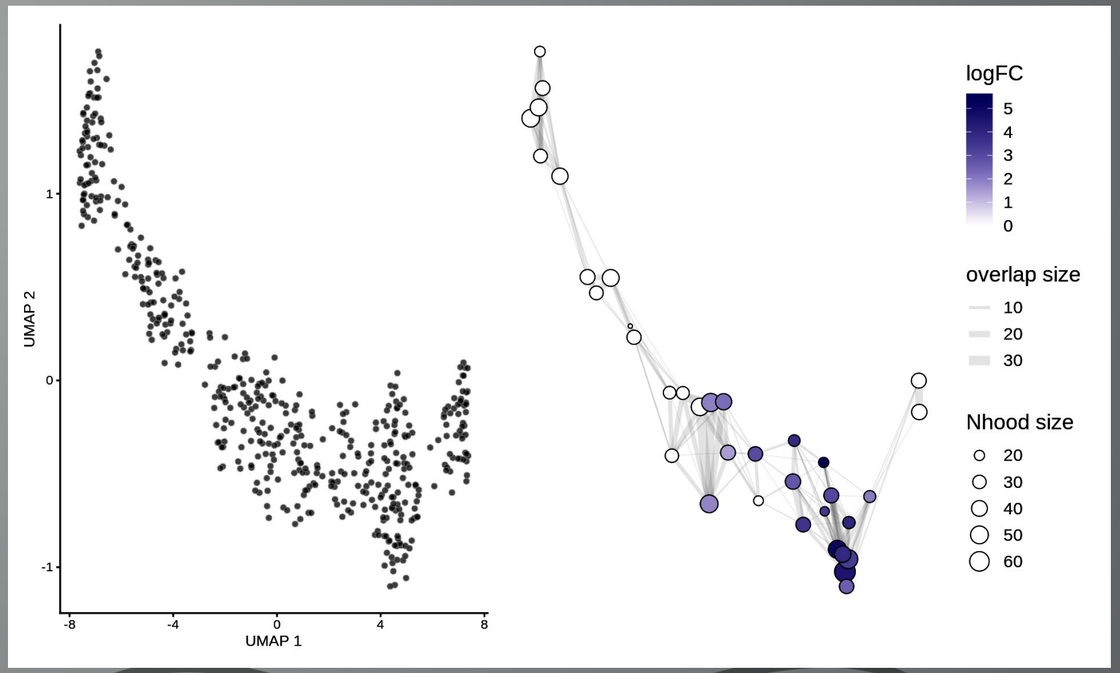


BioFrontiers Institute
UNIVERSITY OF COLORADO **BOULDER**

Cell Neighborhoods using kNN

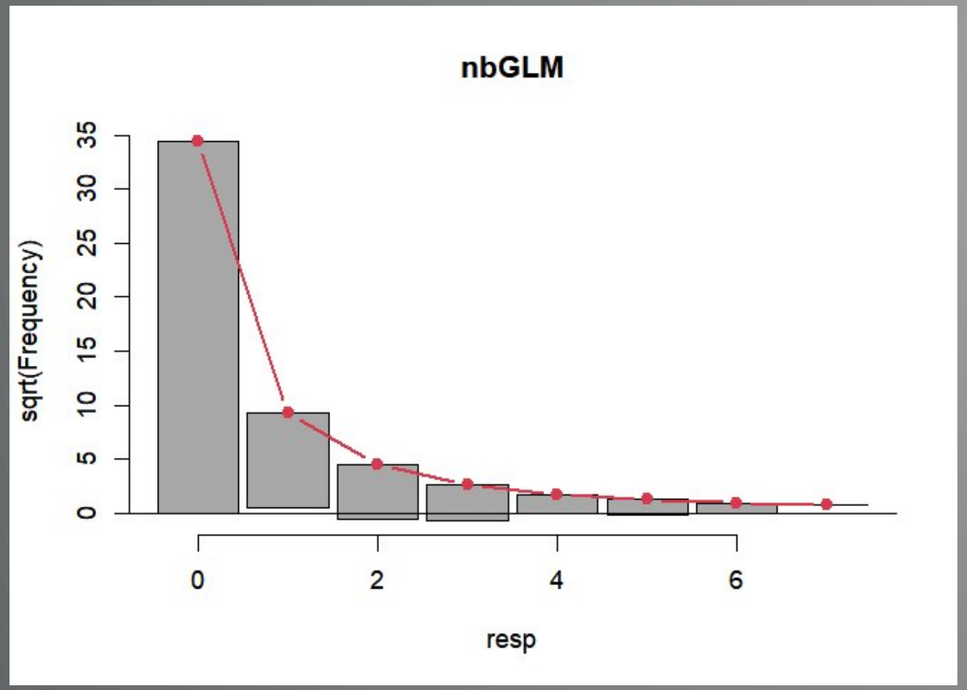
NEIGHBORS OVER CLUSTERS

- miloR operates on the concept of **cell neighborhoods** rather than clusters
- each neighborhood is created by defining the first-order neighbors of cell based on similarity using **kNN graphs**
- kNN (k-Nearest Neighbors): ML model that identifies the k most similar data points to a target point based on a defined metric (eg Euclidean distance)



Dann 2022

Negative Binomial General Linear Models



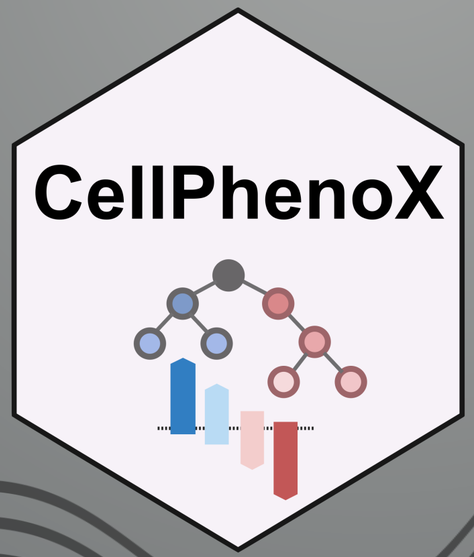
Ford 2016

DA TESTING WITH NB-GLMS

- Once the neighborhoods are established, milo uses **Negative Binomial Generalized Linear Models** to test for differential abundance
- Statistical pipelines of **Cydar** (Lun, 2017) and **edgeR** (Chen, 2008) are employed to compute NB-GLM and calculate DA

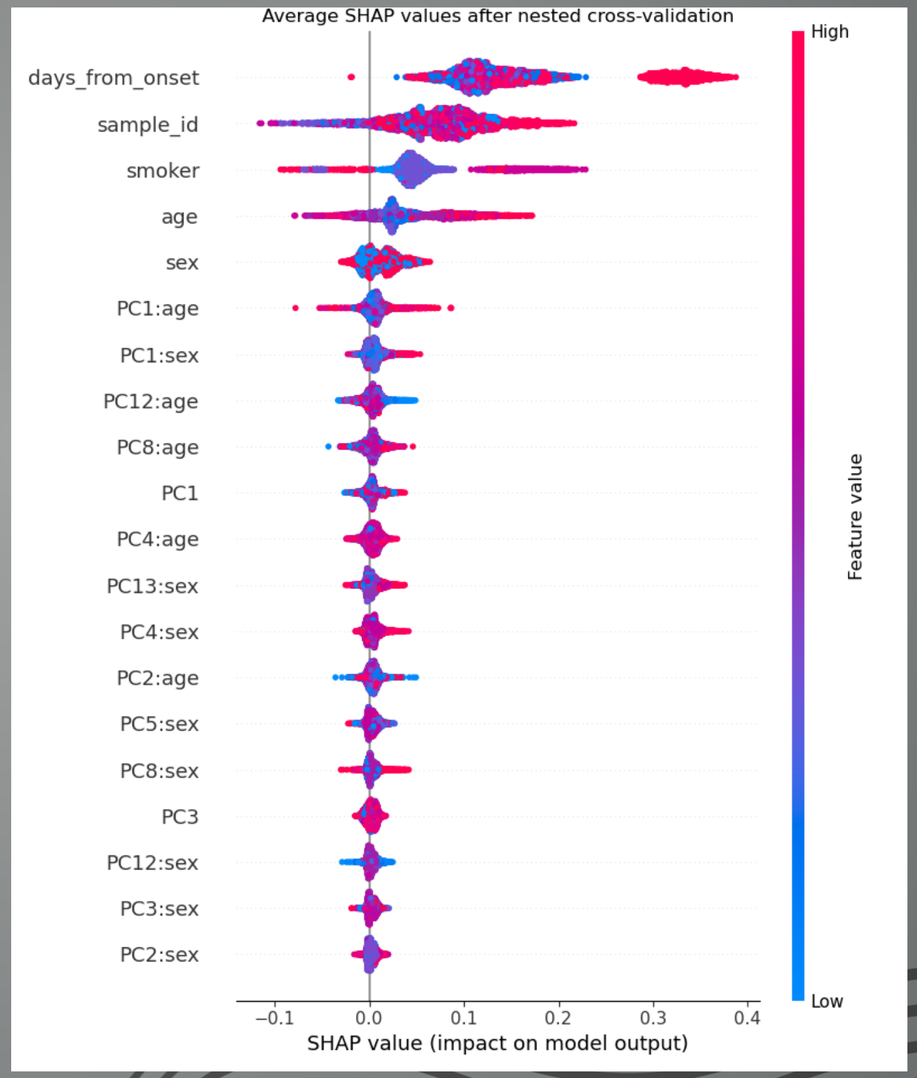
pyCellPhenoX

In development by Jade Young and Zhang Lab at the University of Colorado Anschutz

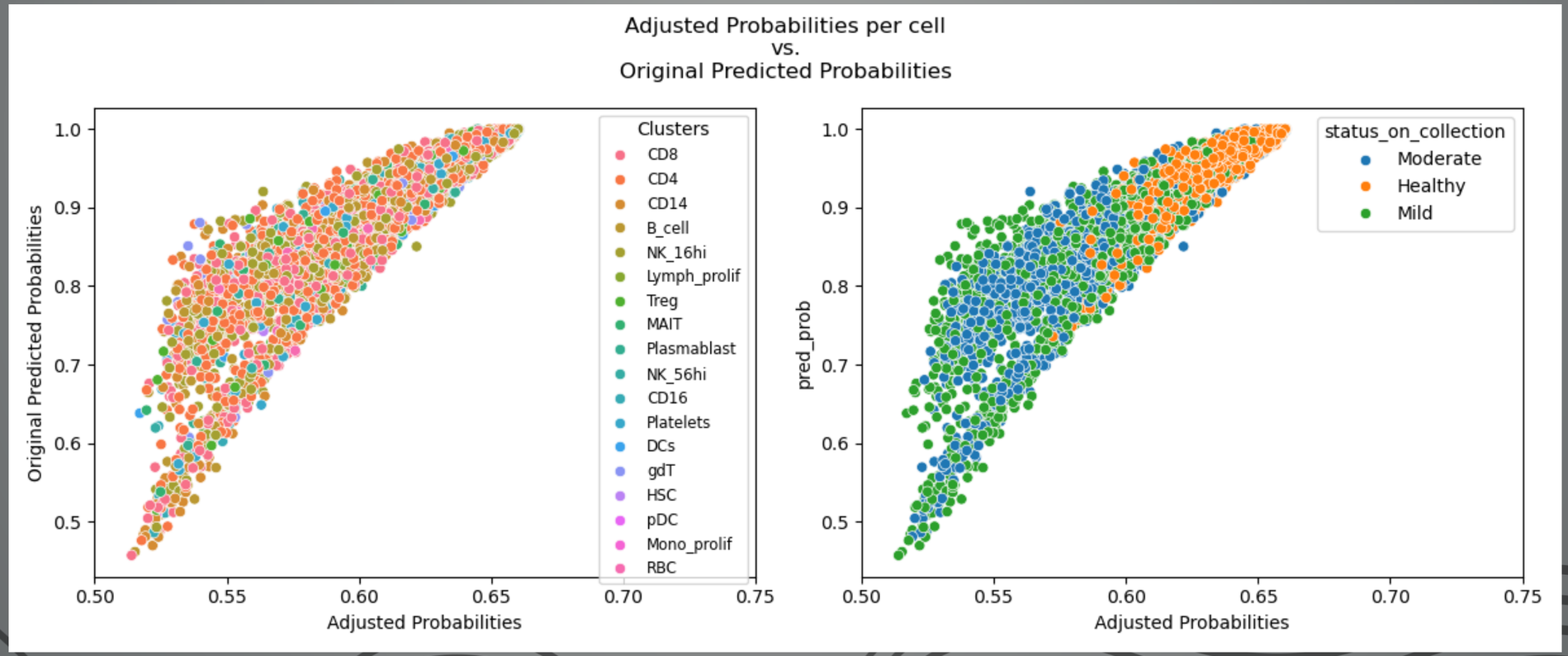


BioFrontiers Institute
UNIVERSITY OF COLORADO **BOULDER**

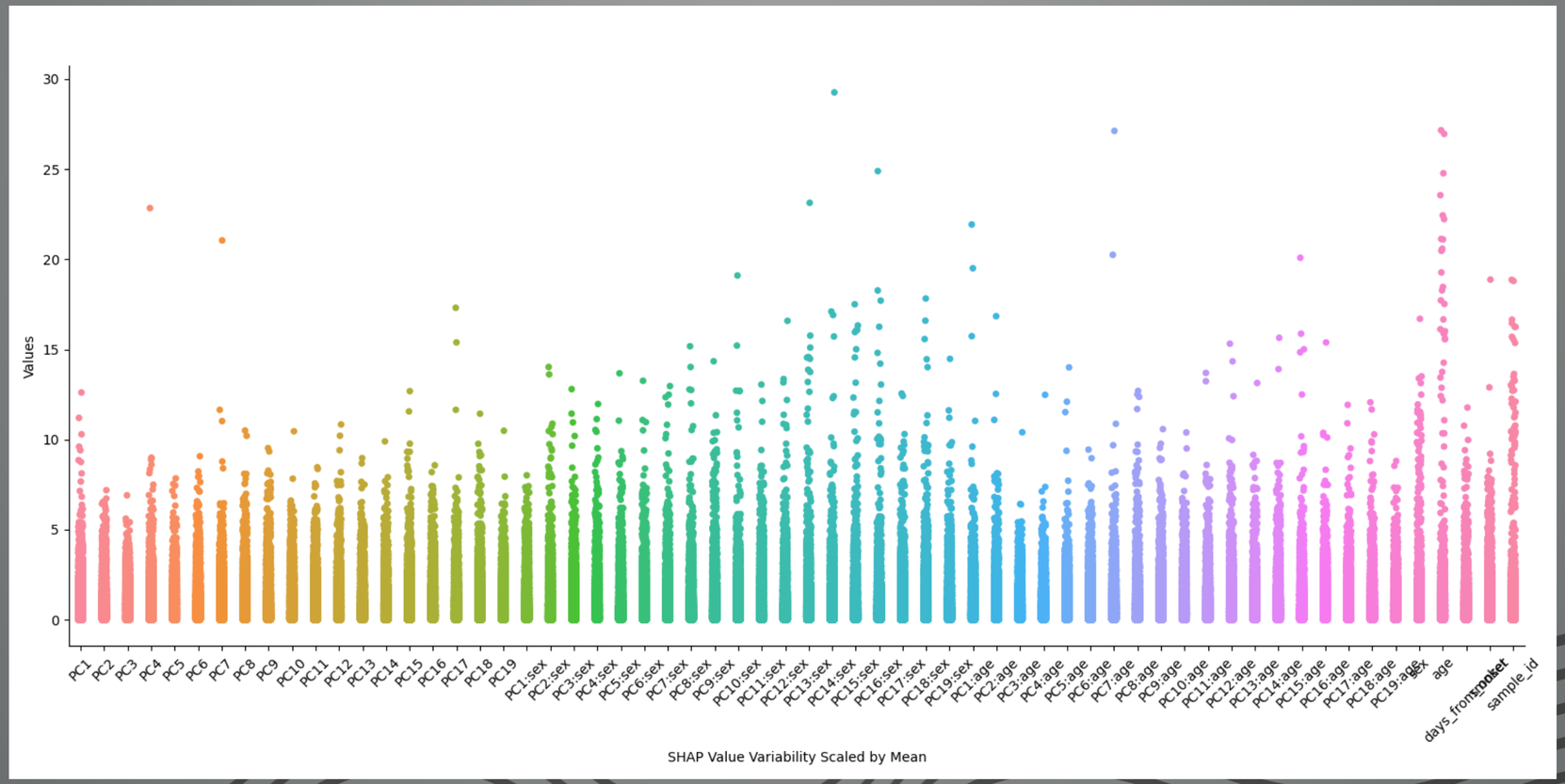
Shapely Additive exPlanations (SHAP) values



SHAP combined with ML models predict clinical outcomes



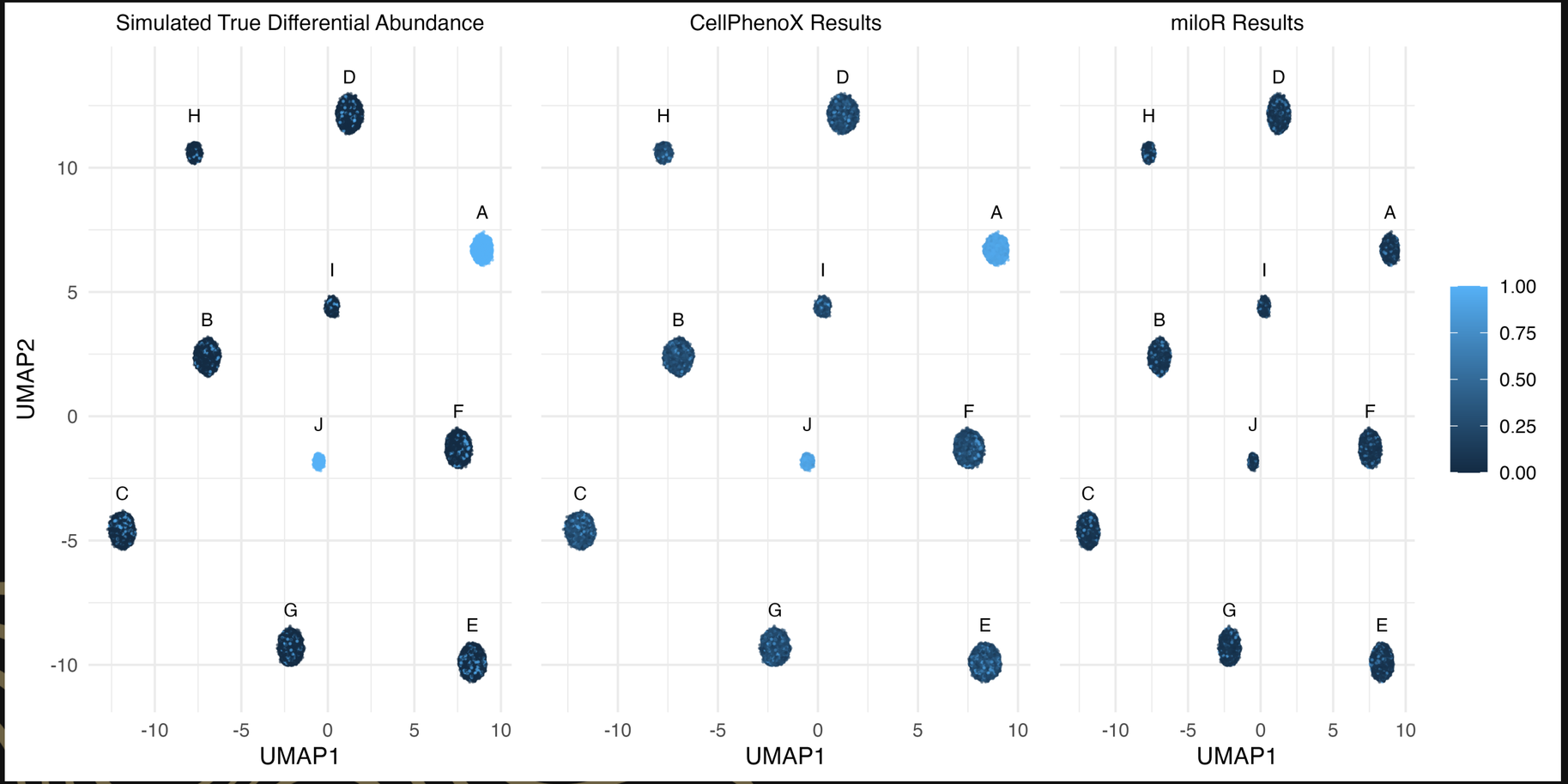
SHAP enables CellPhenoX to understand what drives DA



BioFrontiers Institute
UNIVERSITY OF COLORADO **BOULDER**

Results

Comparing pyCellPhenoX to miloR



Software development of pyCellPhenoX

Importing Dependencies

```
import pyCellPhenoX

import pandas as pd
```

[1] Python

```
/Users/zhanglab/Documents/Python Projects/pyCellPhenoX/pycpx/lib/python3.12/site-packages/tqdm/auto.py:21: TqdmWarning: IProgress not four
  from .autonotebook import tqdm as notebook_tqdm
```

Step 1: import data

```
# paths to expression data and meta data files
expression_file = "./../input/uc_fibroblast_exp.csv"
meta_file = "./../input/uc_fibroblast_meta.csv"
output_path = "./../output/"
# read in data
expression_mat = pd.read_csv(expression_file, index_col=0)
meta = pd.read_csv(meta_file, index_col=0)
```

[3]

Conclusions

miloR vs CellPhenoX

- Employed **miloR** for differential abundance analysis
- Ran miloR on **Alpine supercomputer** with minimal dependency issues
- Compared results to **SHAP** on same dataset

Software development

- Refactored **pyCellPhenoX** by cleaning up code and removing unused elements.
- Used **Vulture** for cleanup, **Black** for linting, and increased **testing coverage**.
- Documented with **Sphinx**, deployed to **PyPI, Anaconda** and **Github**.

Acknowledgements

University of Colorado
Anschutz Medical Campus

Department of Medicine Rheumatology
Department of Biomedical Informatics
Center of Health AI

EMAIL fan.3.zhang@cuanschutz.edu

GITHUB github.com/fanzhanglab

WEBSITE fanzhanglab.org

Questions