NeuCA: a neural network-based method for exhaustive cell label assignment using single-cell RNA-seq data

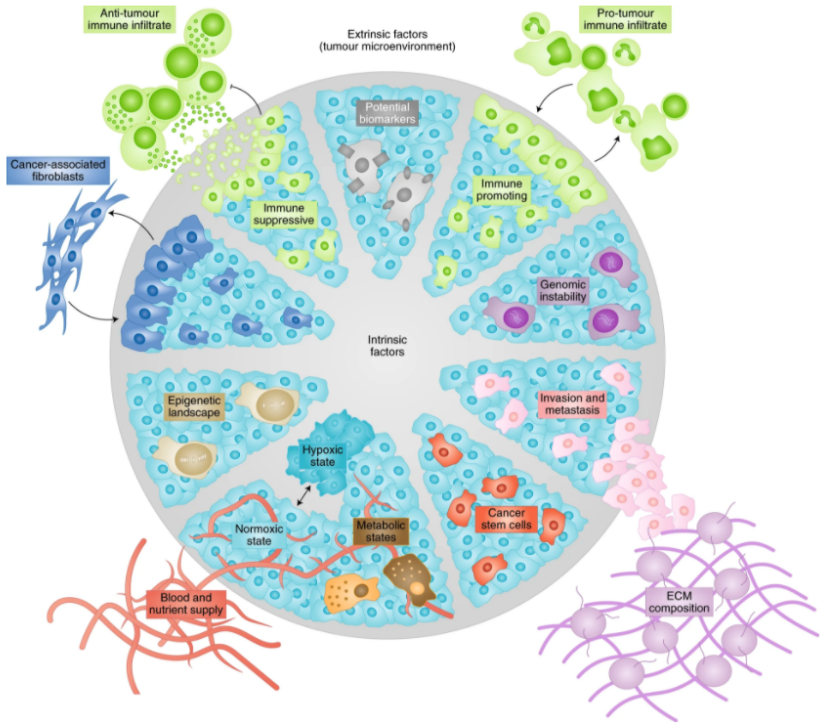
Hao Feng

Department of Population and Quantitative Health Sciences
Case Western Reserve University

hxf155@case.edu
https://hfenglab.org

Single-cell RNA-seq(scRNA-seq)

- Human tissues have diverse cell types/states.
- Traditional RNA-seq ("bulk" RNA-seq) can only measure **averaged signal** across millions of cells.
- Single-cell RNA-seq(scRNA-seq) give us the first data-driven approach to study the **heterogeneous tissue** at single-cell level.



Lawson et al. Nature. https://www.nature.com/articles/s41556-018-0236-7

scRNA-seq data analysis questions

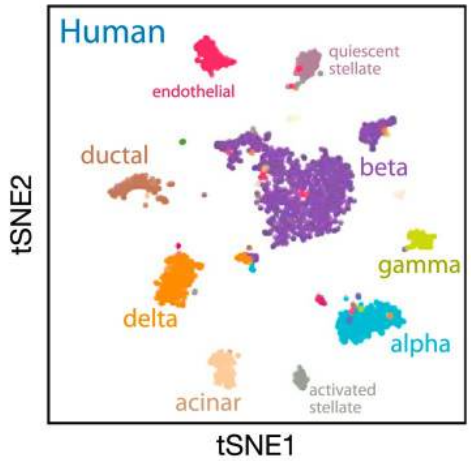
- **Data preprocessing**
  - Normalization
  - Batch effect correction
  - Imputation
- **Data analyses**
  - Cell clustering
  - Cell type identification
  - Differential expression
  - Pseudo-time construction
  - Rare cell type discovery;
  - alternative splicing; allele specific expression
  - RNA velocity
- **Visualization**

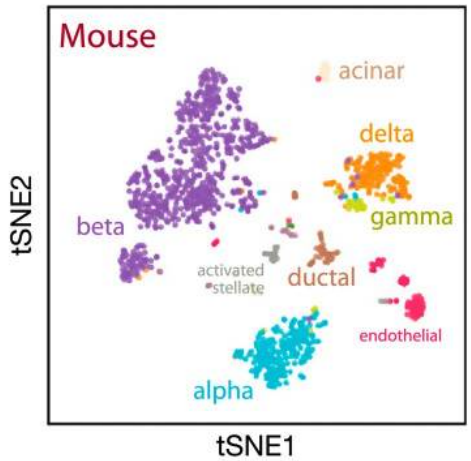
scRNA-seq data analysis questions

- **Data preprocessing**
  - Normalization
  - Batch effect correction
  - Imputation
- **Data analyses**
  - Cell clustering
  - **Cell type identification**
  - Differential expression
  - Pseudo-time construction
  - Rare cell type discovery;
  - alternative splicing; allele specific expression
  - RNA velocity
- **Visualization**

Cell type identification

- Sequencing output of scRNA-seq is anonymous in terms of cell identities.
- Annotating the cells is a key task in scRNA-seq data analysis.



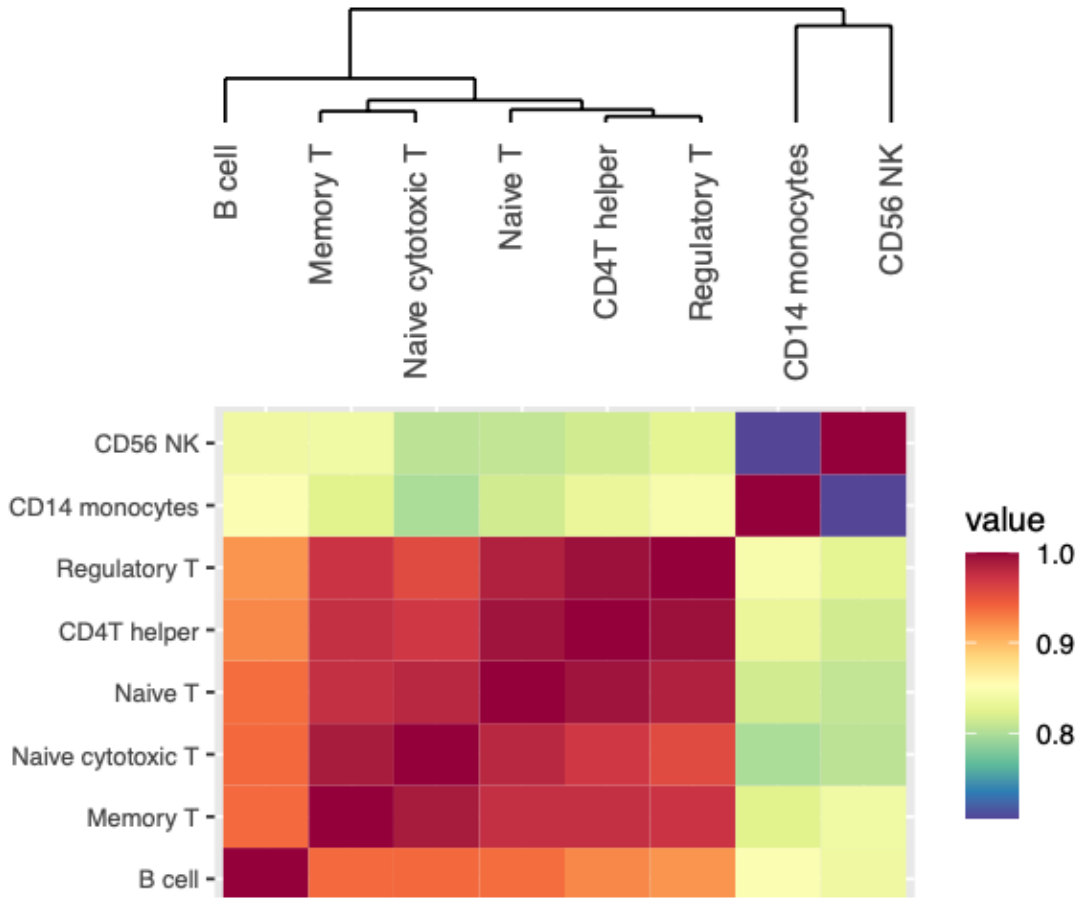


Baron et al. Cell Systems. doi: 10.1016/j.cels.2016.08.011

Cell type identification methods

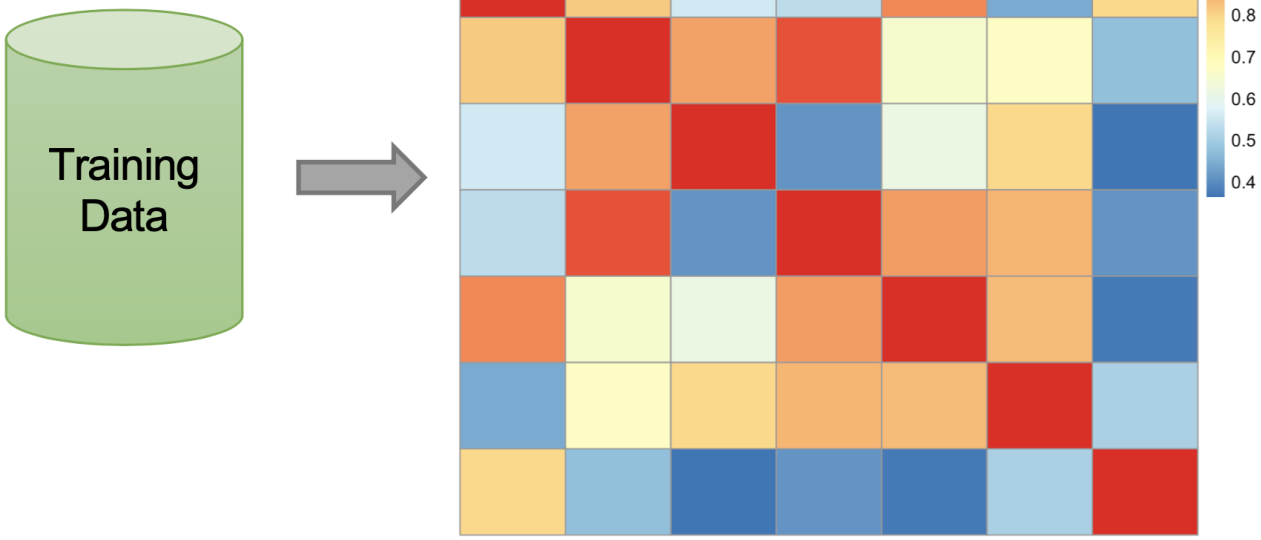
- Two-step approach: Clustering (unsupervised) + labeling.
  - Seurat, SC3, TSCAN, etc...
  - Laborious, time consuming, not best projection, rely on marker gene heavily.
- One-step approach: supervised labeling.
  - scmap, CHETAH, CellAssign, etc...
  - (1) marker-based, (2) correlation-based, and (3) tree structure based.
  - Not suitable for novel cell type discovery.

Motivation: correlations of cell types



High-correlation cell types pose major challenges.

Tackle the issue of high-correlation

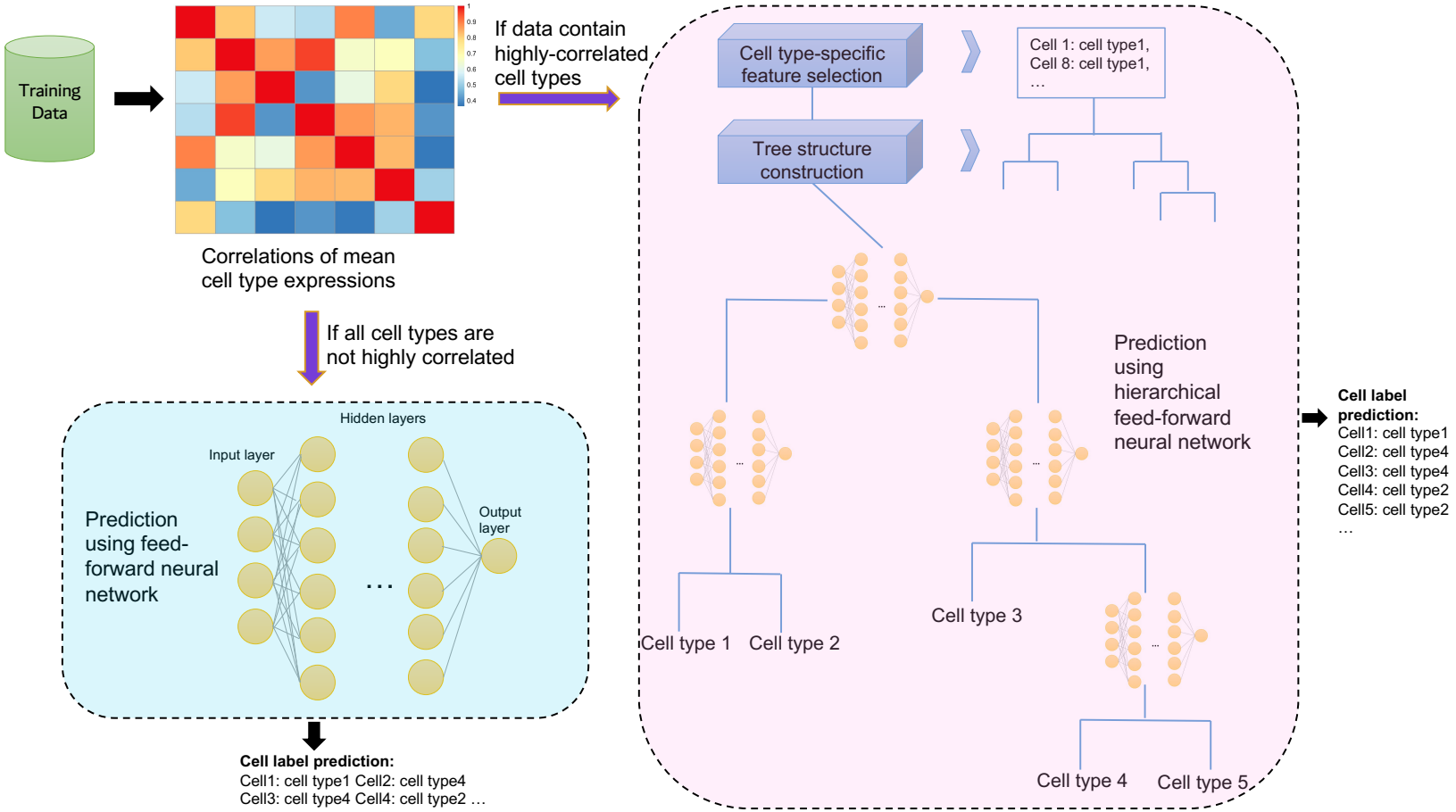


Correlations of mean cell type expressions

How about adopting a flexible approach?

Our method: NeuCA

NeuCA: a **Neu**ral network-based **C**ell **A**nnotation



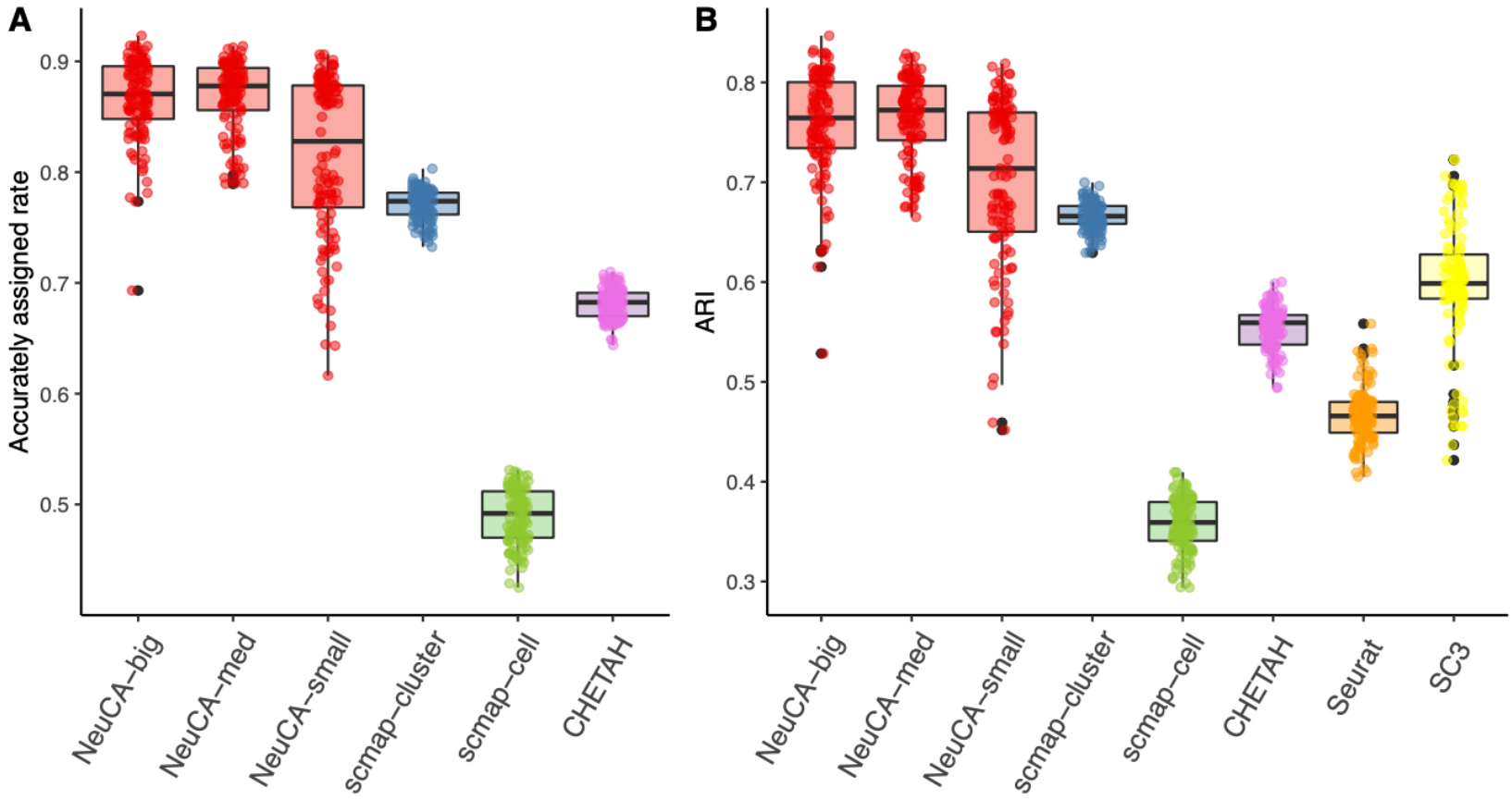
Methodology details

- Route 1: feed-forward neural-network
  - feature selection.
  - 3 model sizes: big, medium, small. (256 to 64 units/nodes)
  - activation function: Rectified Linear Unit (ReLU).
  - output: Softmax.
  - categorical cross-entropy loss.
- Route 2: marker-guided hierarchical neural-network
  - feature selection (gene-specific sensitivities).
  - cell type hierarchical tree.
  - hierarchical neural-network tree.
  - similar model sizes as Route 1.

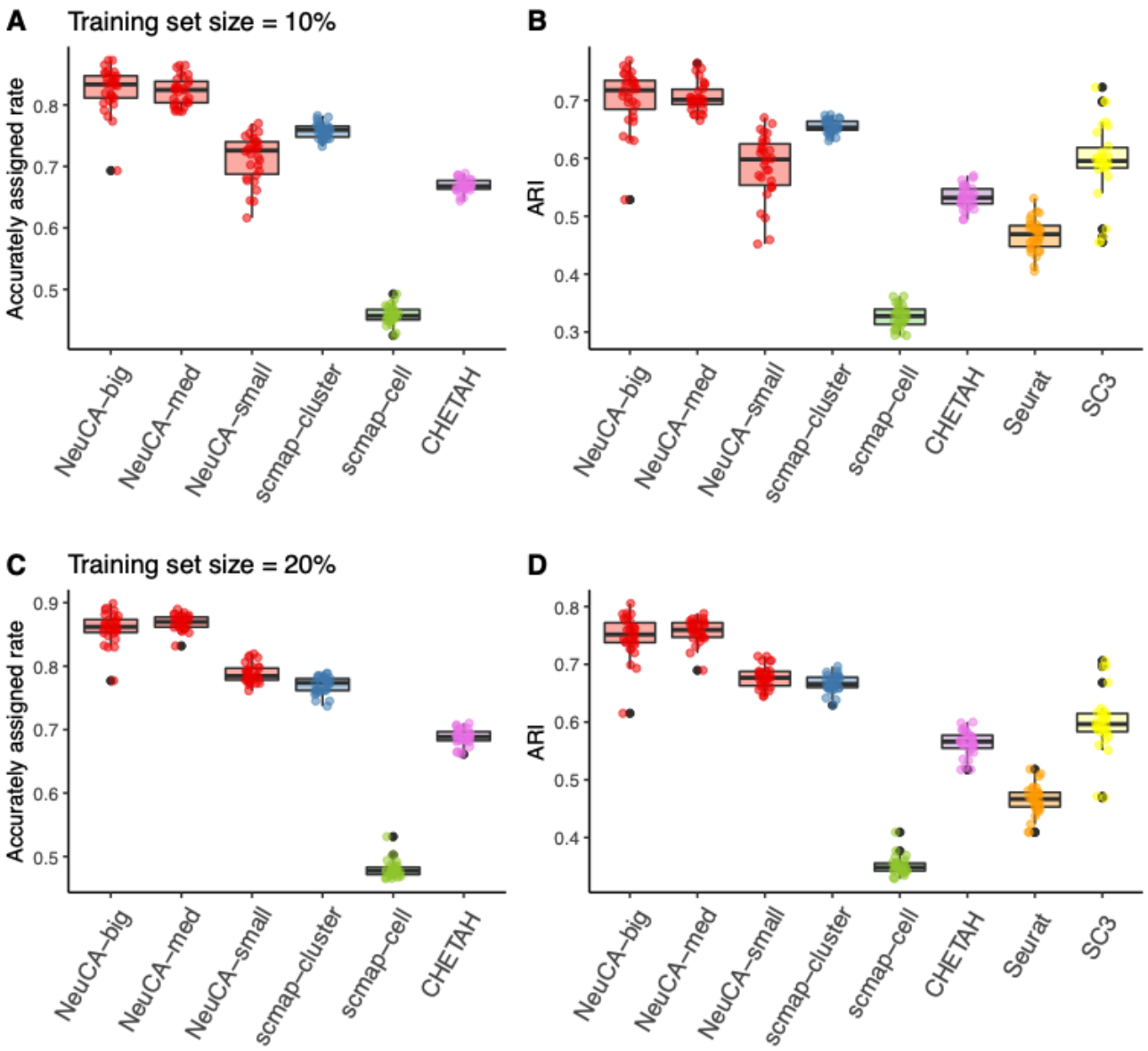
Simulation

Simulation results: real-data based

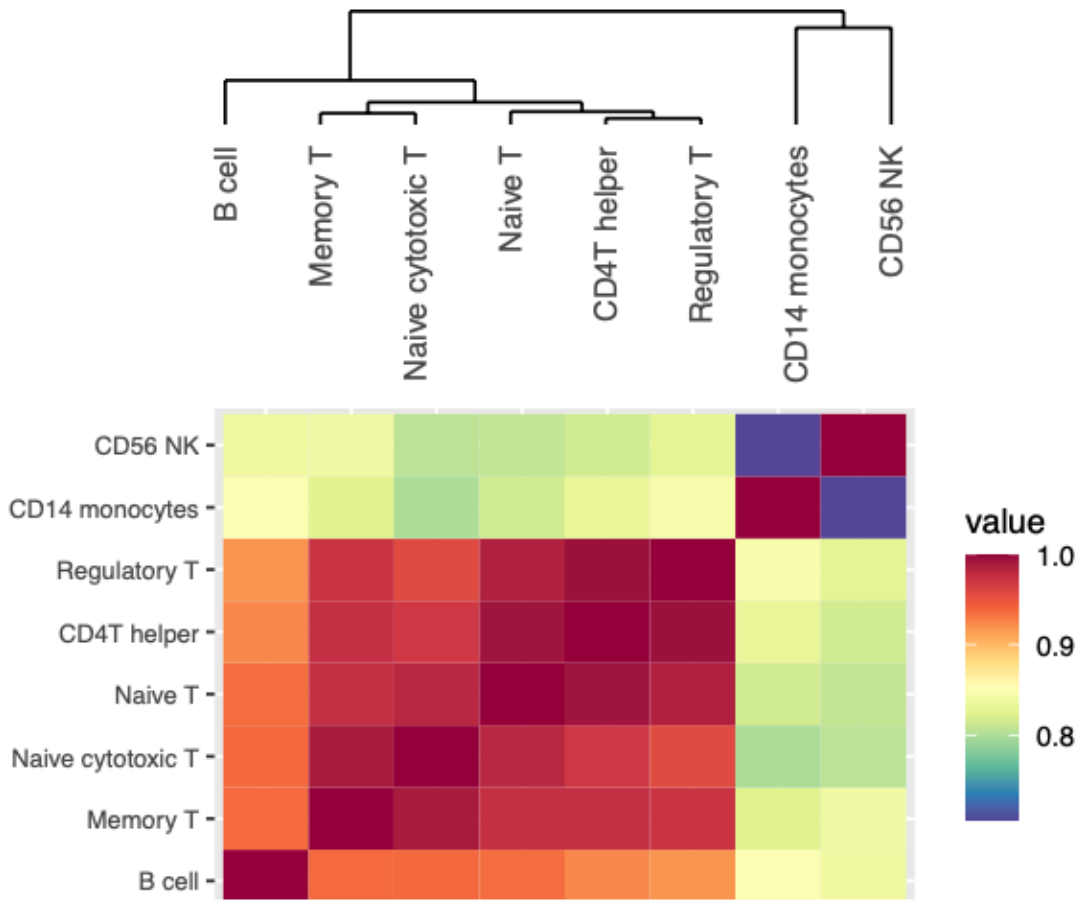
- 10X PBMC scRNA-seq data (10.1038/ncomms14049).
- 80 Monte Carlo simulations are conducted and aggregated.
- Training set proportion ranging from 10% to 80%.



Real-data based simulation

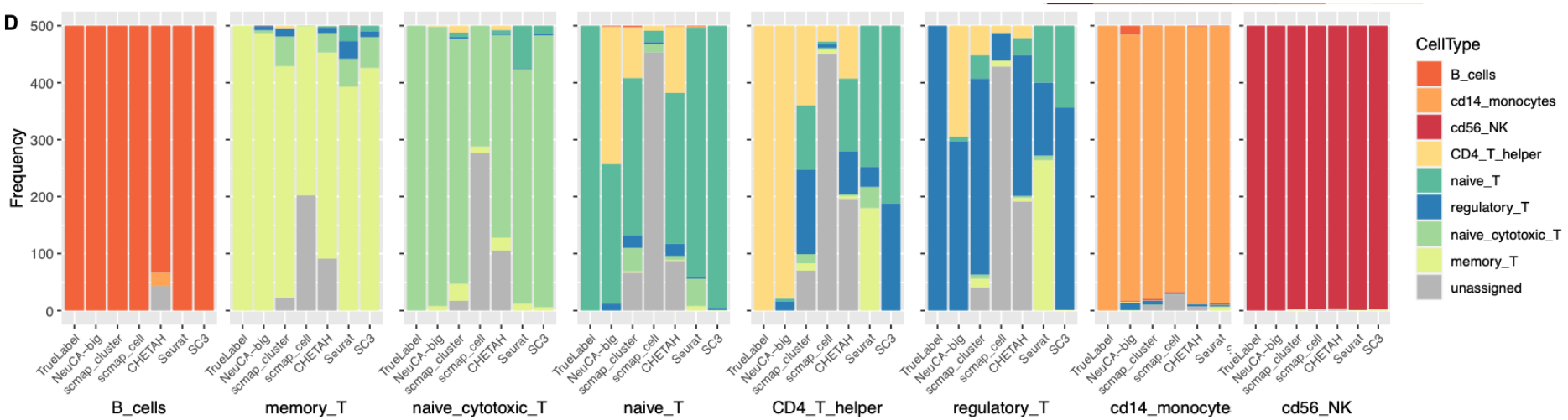


Real-data based simulation



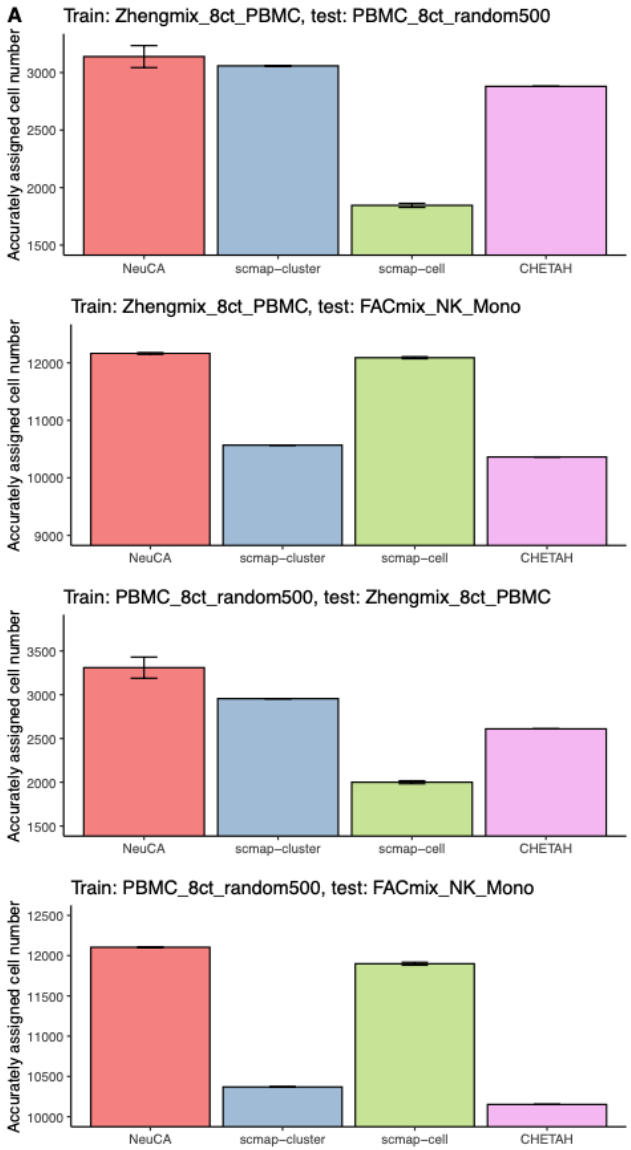
See next page for results on the T cells.

Real-data based simulation

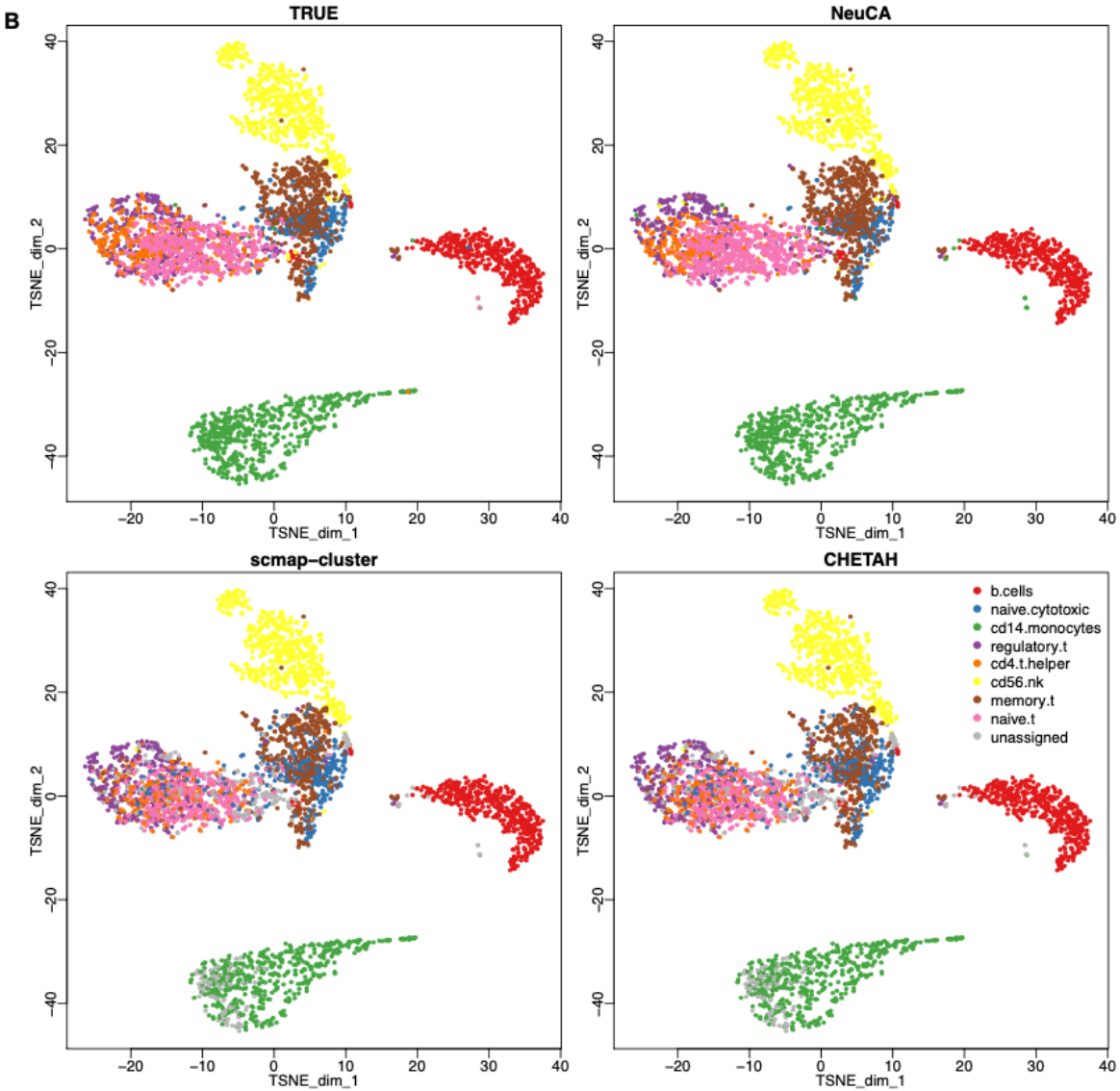


Real data results

Various PBMC datasets



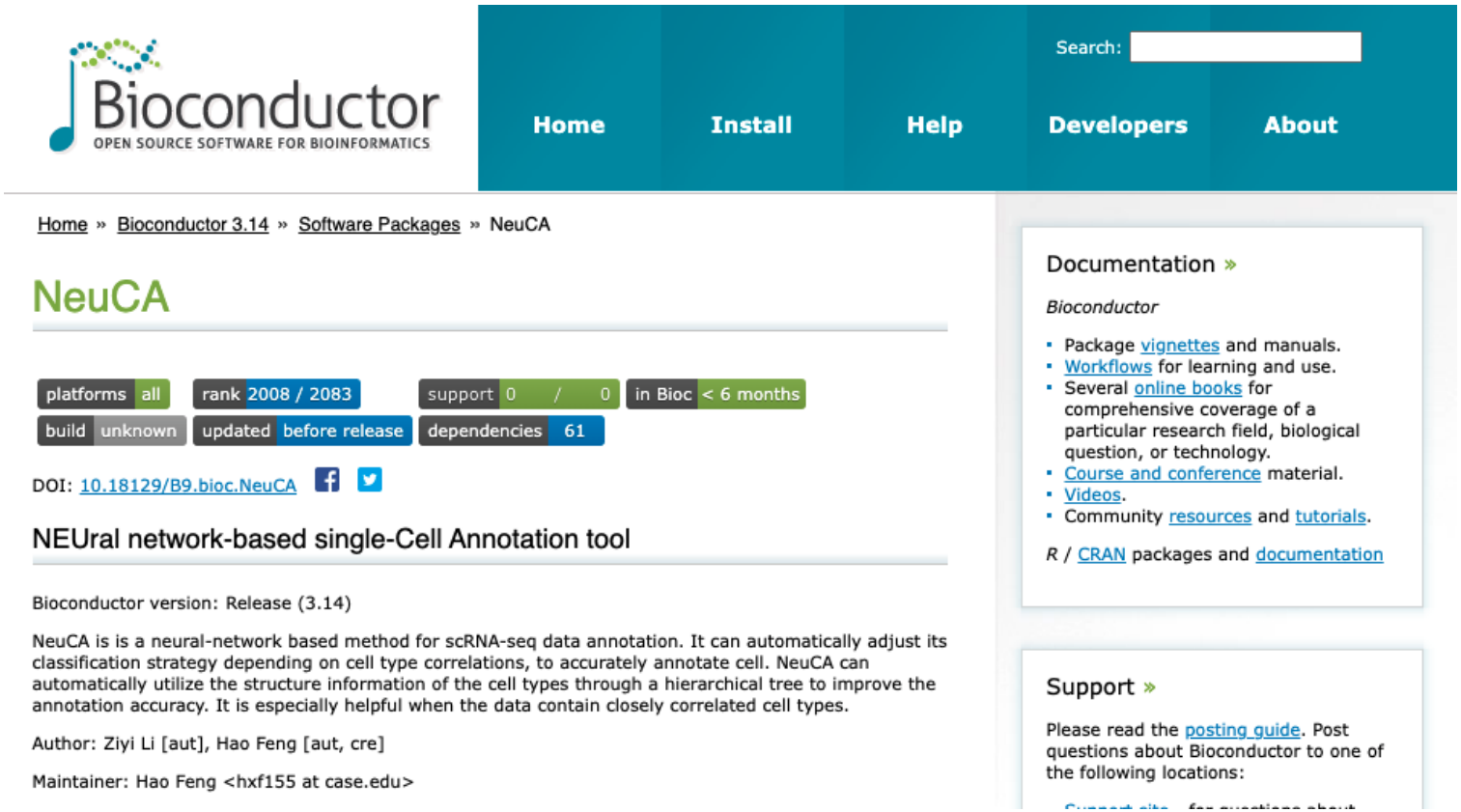
Real data results: a t-SNE visualization



Real data results: pancreas data

Software

On Bioconductor: https://bioconductor.org/packages/NeuCA

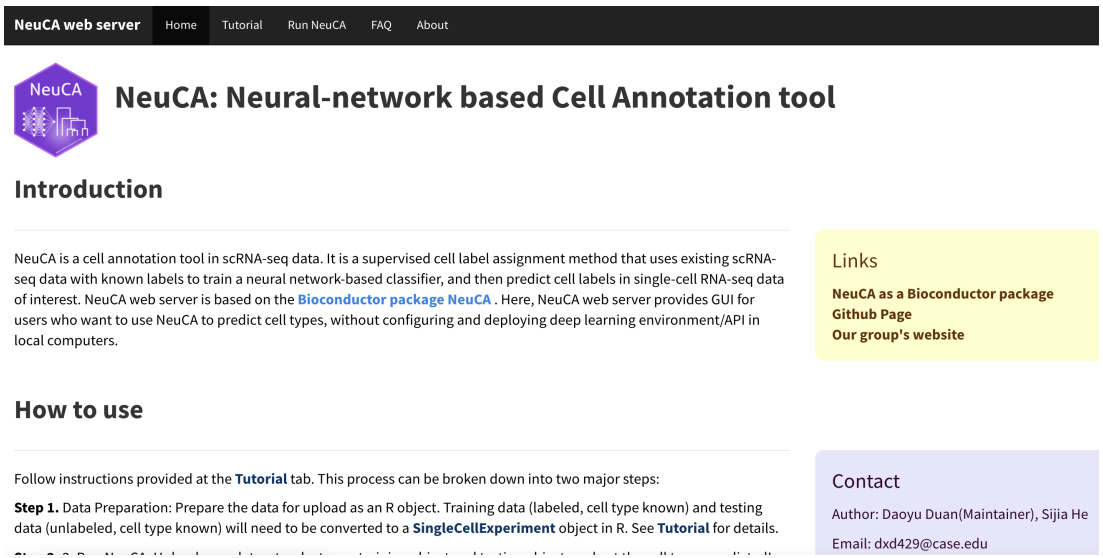


Usage

```
NeuCA(train, test, model.size = "big", verbose = FALSE)
```

NeuCA web server

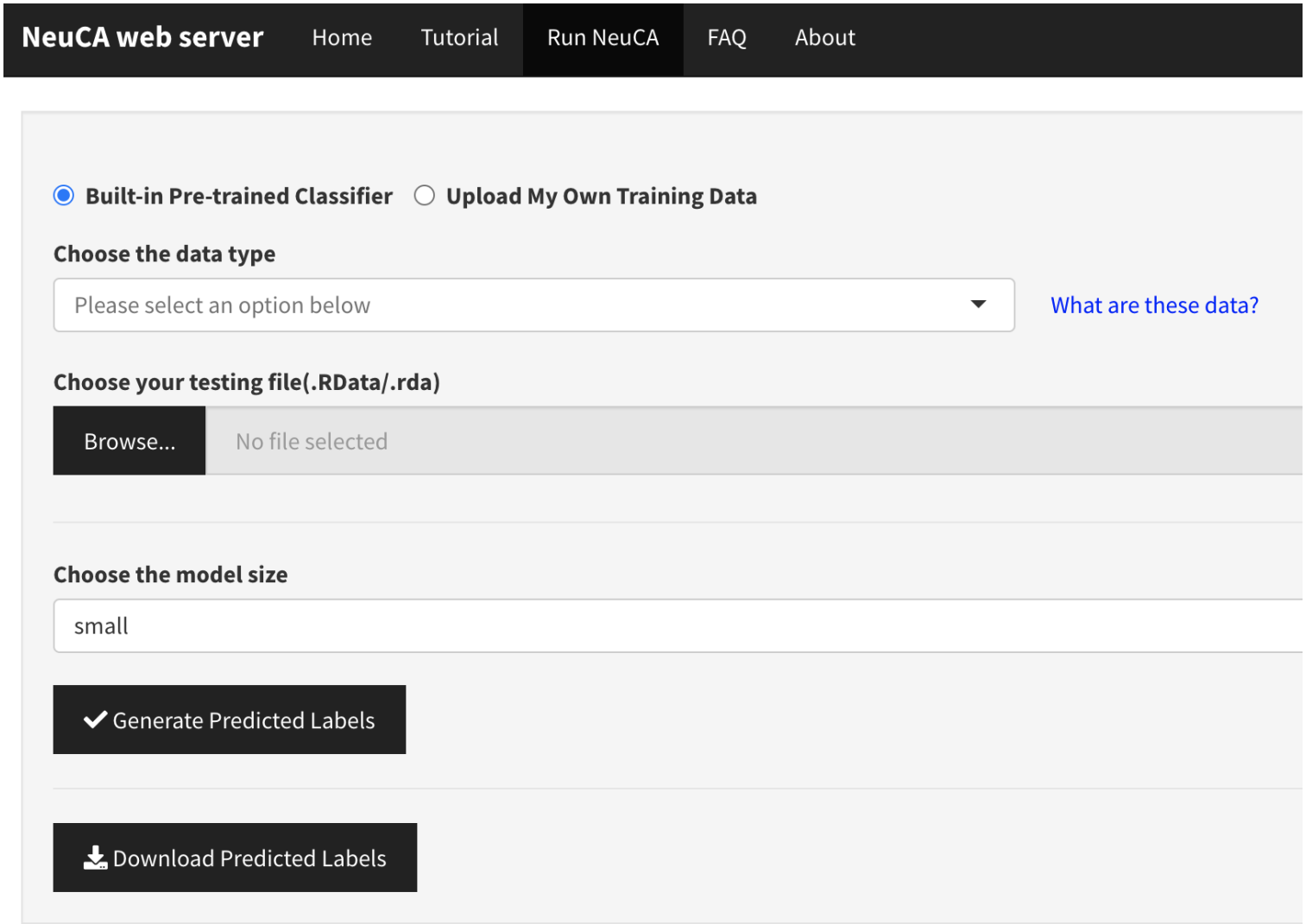
R Shiny App: https://statbioinfo.shinyapps.io/NeuCA



(Human) PBMC
(Human) Pancreas
(Human) Autism
(Human) Molar
(Human) Choroid Plexus
(Human) Healthy Lung
(Human) Aging Skin
(Human) Fetal Maternal Decidual
(Human) Muscle
(Human) Bronchoalveolar from COVID-19 Patients
(Human) Adult Retina
(Human) Fetal Gut
(Human) Human_Ovary
(Human) Human_Glioblastoma
(Mouse) Lung
(Mouse) Enteric
(Mouse) Hippocampus
(Mouse) Medulla
(Mouse) Spinal Cord
Drosophila Ovary

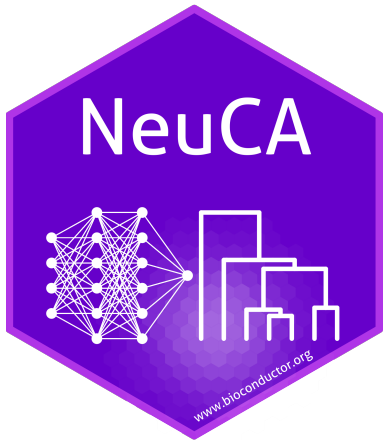
NeuCA web server

R Shiny App: https://statbioinfo.shinyapps.io/NeuCA



Summary

- One-step supervised learning method for cell label assignment.
- A neural-network based classifier.
- Flexible: adopt different approaches depending on correlation level.
- Perform well even with low amount of training set.
- BioC package and web server with GUI (free!).



Publication

www.nature.com/scientificreports

scientific reports

Check for updates

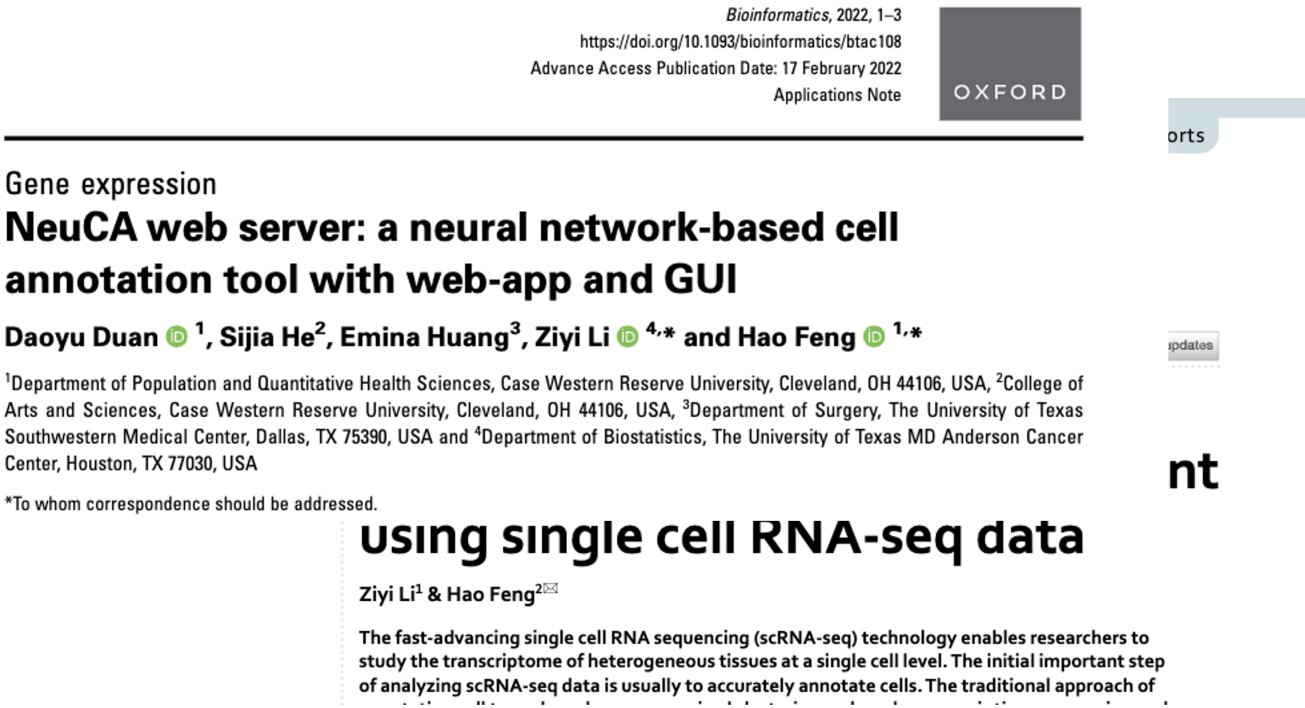
OPEN

A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data

Ziyi Li[1] & Hao Feng[2]✉

The fast-advancing single cell RNA sequencing (scRNA-seq) technology enables researchers to study the transcriptome of heterogeneous tissues at a single cell level. The initial important step of analyzing scRNA-seq data is usually to accurately annotate cells. The traditional approach of

Publication



Bioinformatics, 2022, 1–3
https://doi.org/10.1093/bioinformatics/btac108
Advance Access Publication Date: 17 February 2022
Applications Note

OXFORD

Gene expression

NeuCA web server: a neural network-based cell annotation tool with web-app and GUI

Daoyu Duan [1], Sijia He[2], Emina Huang[3], Ziyi Li [4,*] and Hao Feng [1,*]

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, [2]College of Arts and Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, [3]Department of Surgery, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA and [4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

*To whom correspondence should be addressed.

using single cell RNA-seq data

Ziyi Li[1] & Hao Feng[2]✉

The fast-advancing single cell RNA sequencing (scRNA-seq) technology enables researchers to study the transcriptome of heterogeneous tissues at a single cell level. The initial important step of analyzing scRNA-seq data is usually to accurately annotate cells. The traditional approach of

doi.org/10.1093/bioinformatics/btac108
doi.org/10.1038/s41598-021-04473-4

Authors and Acknowledgement

Ziyi Li, Assistant Professor at MD Anderson Cancer Center

Daoyu Duan, PhD student at Case Western Reserve University

hfeng@case.edu

@HHarryFeng

https://hfenglab.org/