

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

Hao Feng

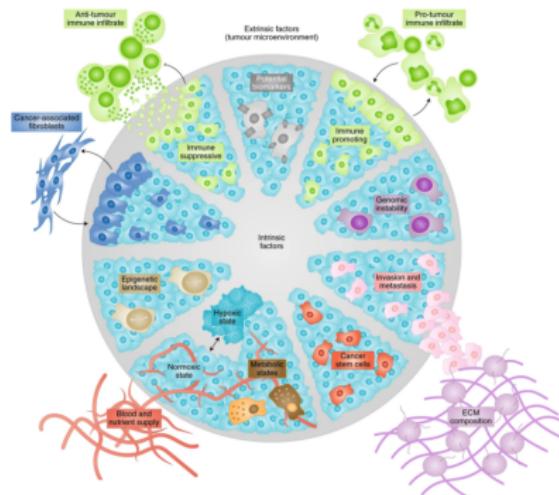
Department of Population and Quantitative Health Sciences
Case Western Reserve University

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.



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

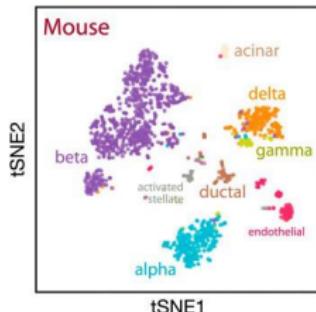
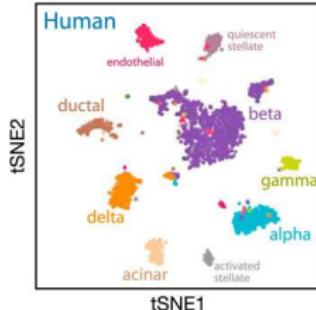
- **Data analyses**

- 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.

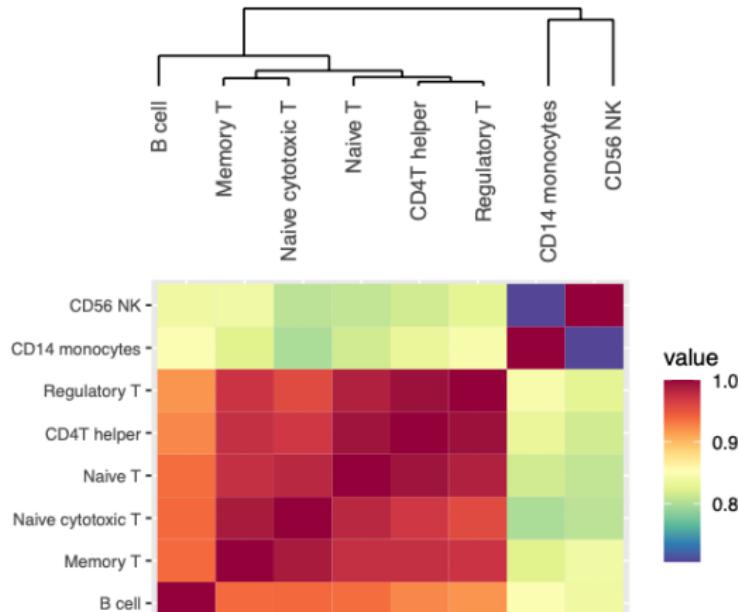


Baron et al. Cell Systems. doi: 10.1016/j.cels.2016.08.011

Cell type identification methods

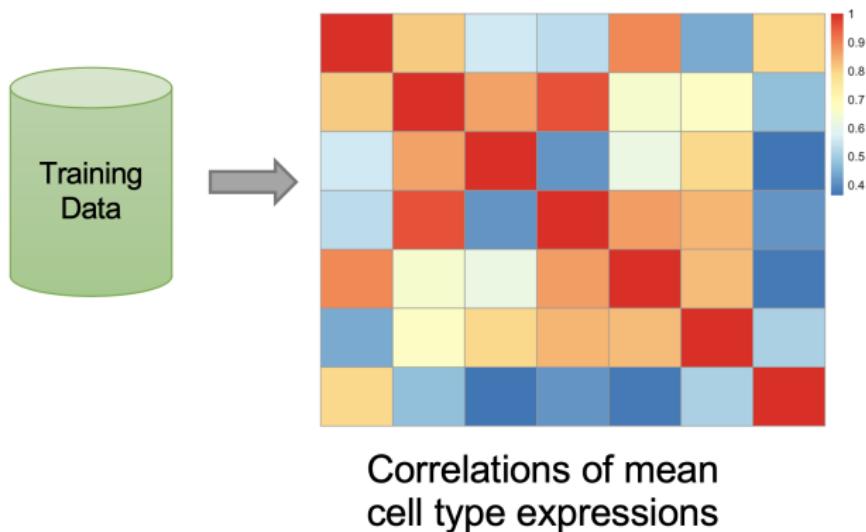
- Two-step approach: Clustering (unsupervised) + labeling.
 - Seurat, SC3, TSCAN, etc...
 - Laborious, time consuming, not best projection, rely on marker gene heavily.
- One-step approach: supervised labeling.
 - scmap, CHETAH, CellAssign, etc...
 - (1) marker-based, (2) correlation-based, and (3) tree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

Motivation: correlations of cell types



High-correlation cell types pose major challenges.

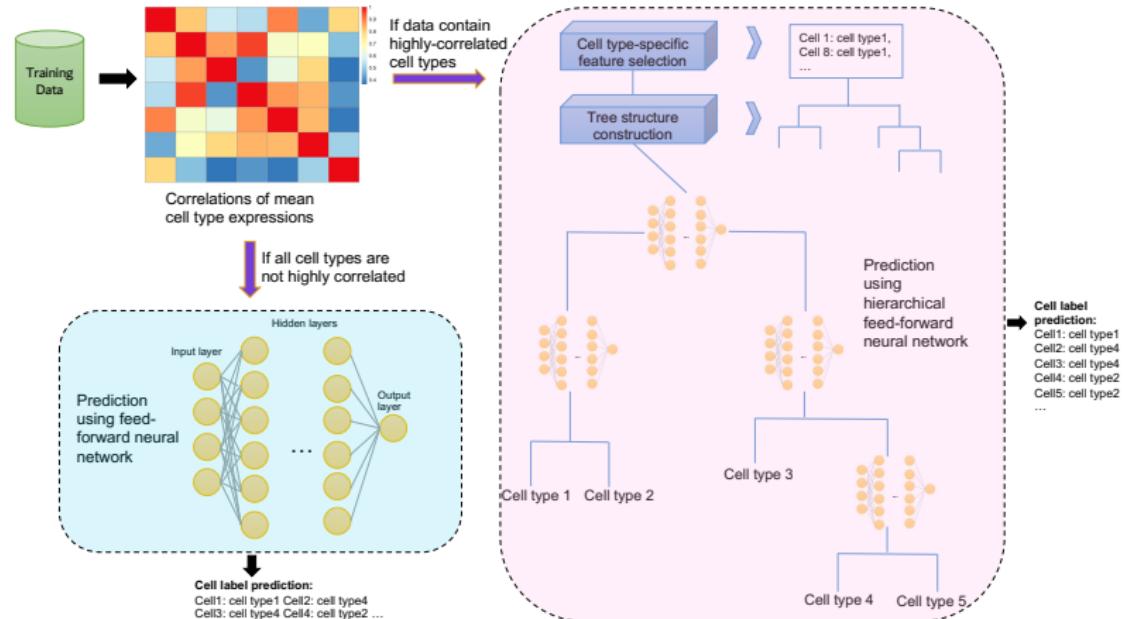
Tackle the issue of high-correlation



How about adopting a flexible approach?

Our method: NeuCA

NeuCA: a Neural network-based Cell Annotation



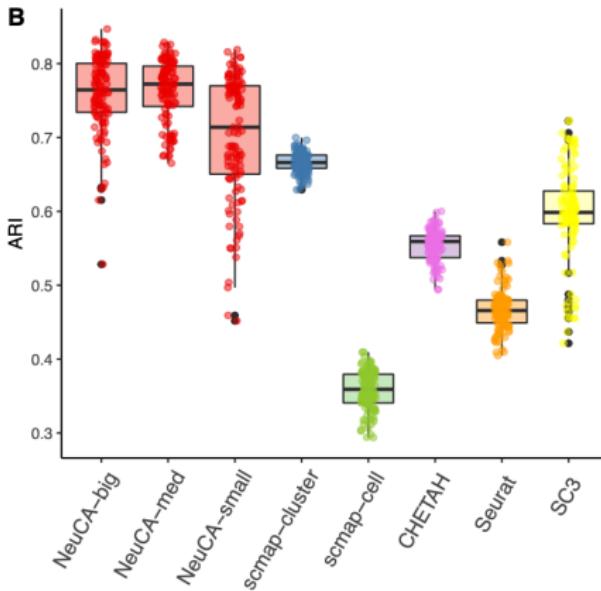
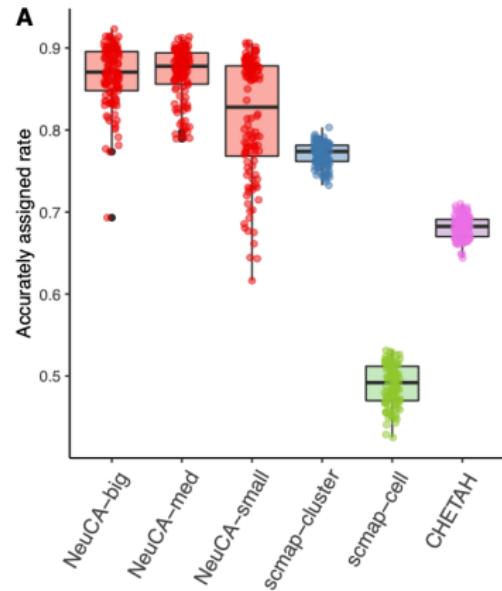
Methodology details

- 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.

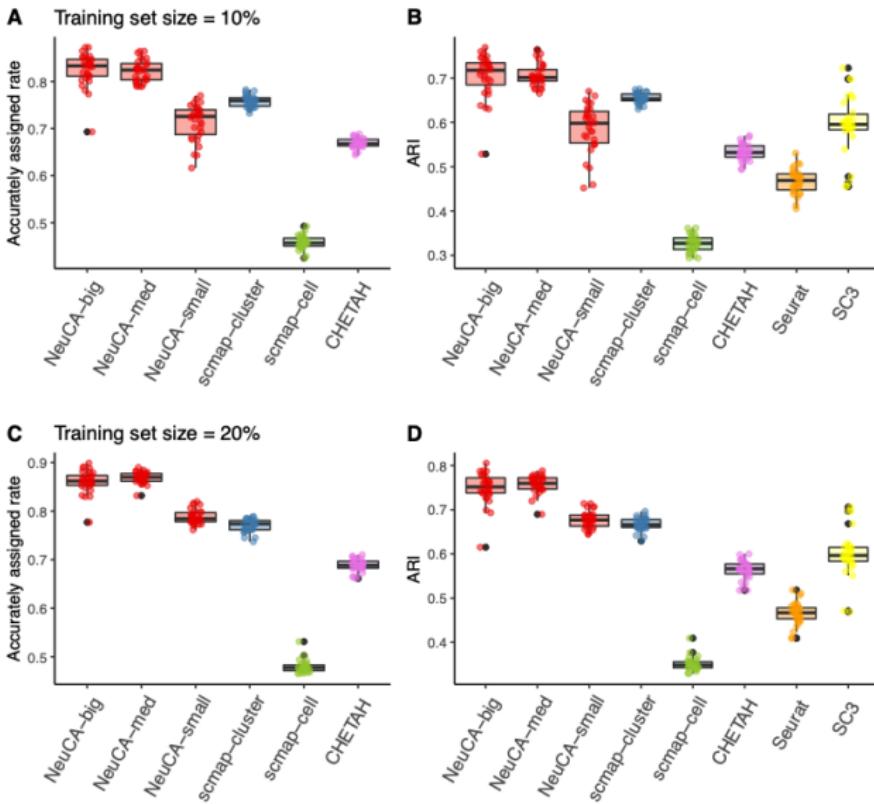
Simulation

Simulation results: real-data based

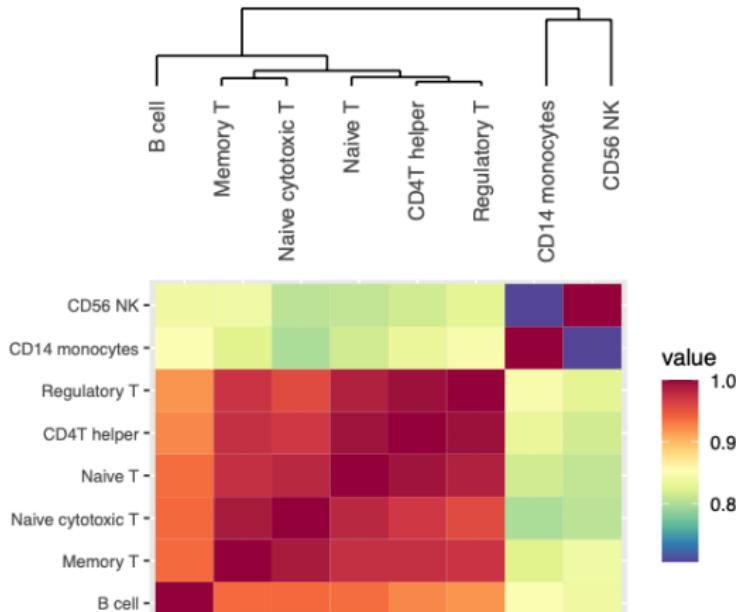
- 10X PBMC scRNA-seq data ([10.1038/ncomms14049](https://doi.org/10.1038/ncomms14049)).
- 80 Monte Carlo simulations are conducted and aggregated.
- Training set proportion ranging from 10% to 80%.



Real-data based simulation

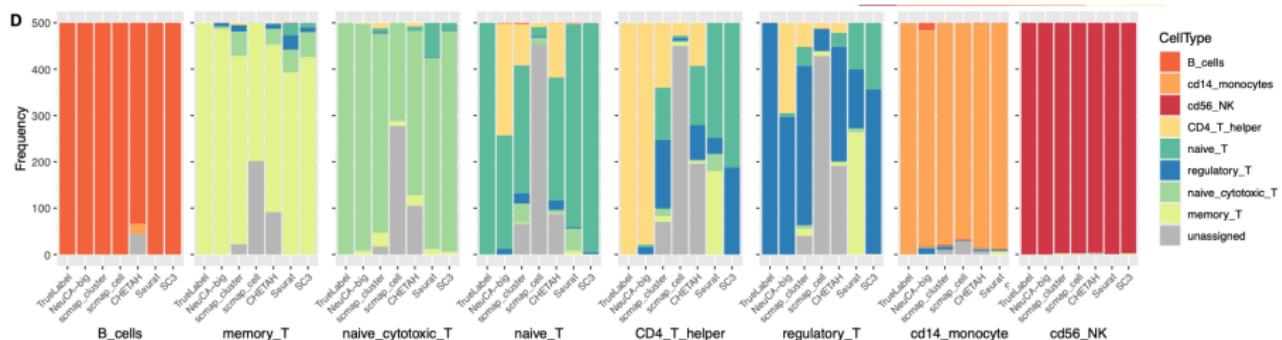


Real-data based simulation



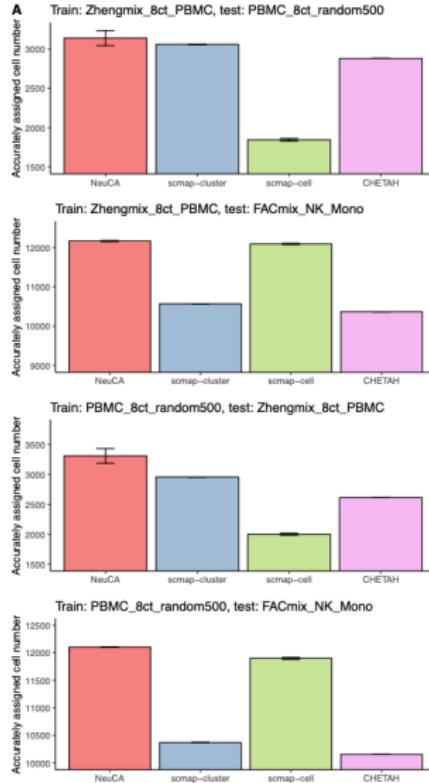
See next page for results on the T cells.

Real-data based simulation

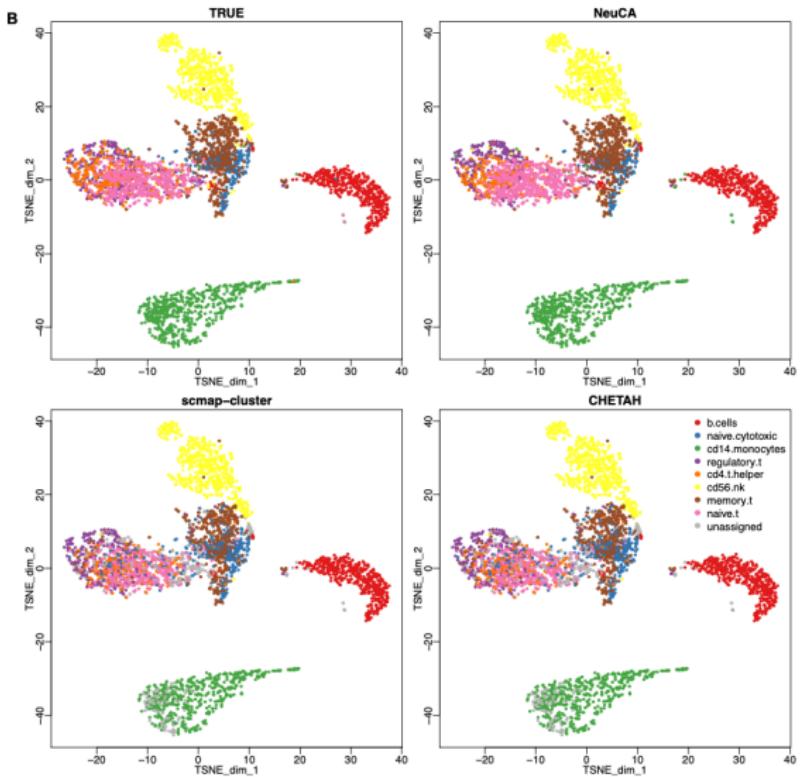


Real data results

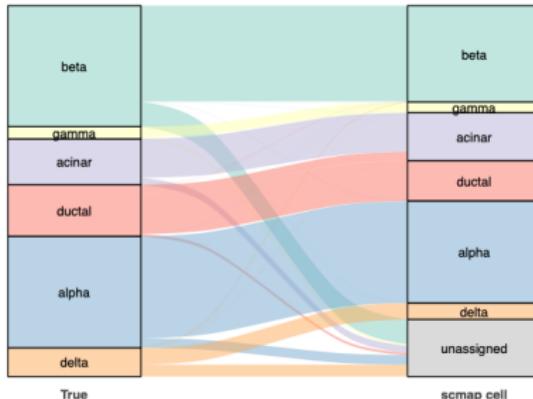
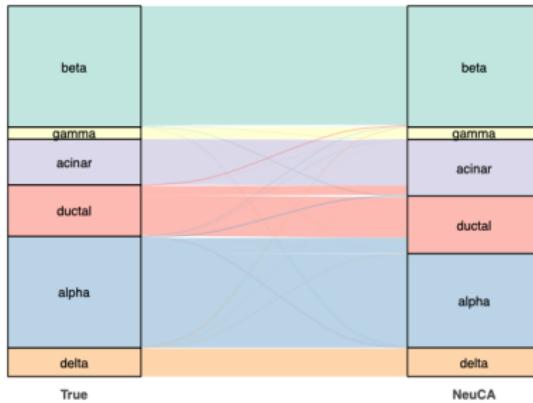
Various PBMC datasets



Real data results: a t -SNE visualization

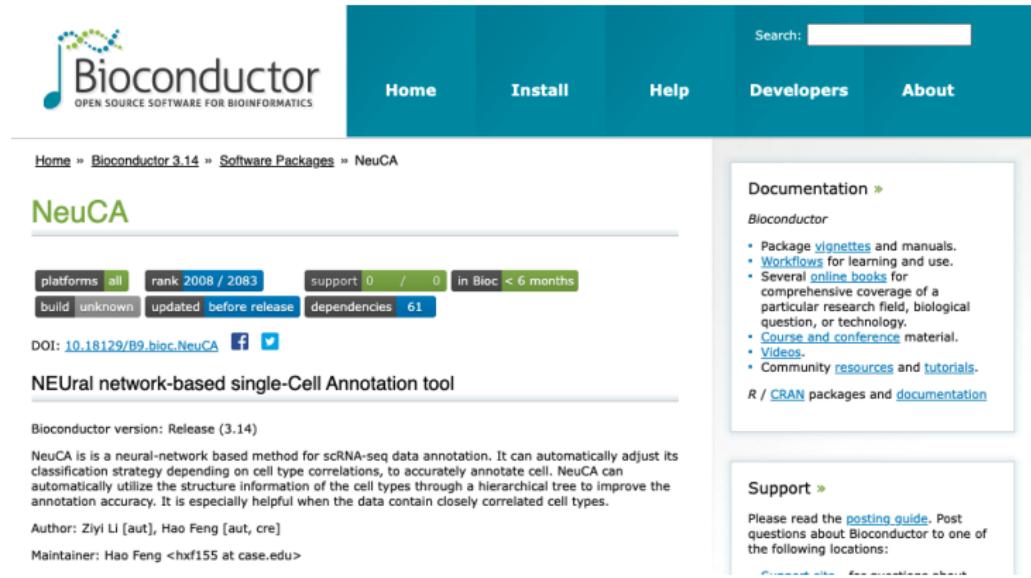


Real data results: pancreas data



Software

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



The screenshot shows the Bioconductor NeuCA package page. At the top, there's a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation bar, the Bioconductor logo is displayed with the text "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". The main content area has a breadcrumb navigation: Home > Bioconductor 3.14 > Software Packages > NeuCA. The title "NeuCA" is prominently displayed. Below the title, there are several status indicators: platforms all, rank 2008 / 2083, support 0 / 0, in BioC < 6 months; build unknown, updated before release, dependencies 61. A DOI link (DOI: 10.18129/B9.bioc.NeuCA) and social media links for Facebook and Twitter are also shown. The package description "NEUral network-based single-Cell Annotation tool" follows. Below this, it says "Bioconductor version: Release (3.14)". A detailed description of NeuCA's functionality is provided: "NeuCA is a neural-network based method for scRNA-seq data annotation. It can automatically adjust its classification strategy depending on cell type correlations, to accurately annotate cell. NeuCA can automatically utilize the structure information of the cell types through a hierarchical tree to improve the annotation accuracy. It is especially helpful when the data contain closely correlated cell types." Author and maintainer information is listed: "Author: Ziyi Li [aut], Hao Feng [aut, cre]" and "Maintainer: Hao Feng <hxfl55 at case.edu>". To the right, there are two boxes: "Documentation" which links to vignettes, workflows, online books, courses, videos, and community resources; and "Support" which links to posting guides and support locations.

Usage

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

NeuCA web server

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

NeuCA web server Home Tutorial Run NeuCA FAQ About



NeuCA: Neural-network based Cell Annotation tool

Introduction

NeuCA is a cell annotation tool in scRNA-seq data. It is a supervised cell label assignment method that uses existing scRNA-seq data with known labels to train a neural network-based classifier, and then predict cell labels in single-cell RNA-seq data of interest. NeuCA web server is based on the [Bioconductor package NeuCA](#). Here, NeuCA web server provides GUI for users who want to use NeuCA to predict cell types, without configuring and deploying deep learning environment/API in local computers.

How to use

Follow instructions provided at the [Tutorial](#) tab. This process can be broken down into two major steps:

Step 1. Data Preparation: Prepare the data for upload as an R object. Training data (labeled, cell type known) and testing data (unlabeled, cell type known) will need to be converted to a [SingleCellExperiment](#) object in R. See [Tutorial](#) for details.

Links
[NeuCA as a Bioconductor package](#)
[Github Page](#)
[Our group's website](#)

Contact
Author: Daoyu Duan(Maintainer), Sijia He
Email: ddd429@case.edu

(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

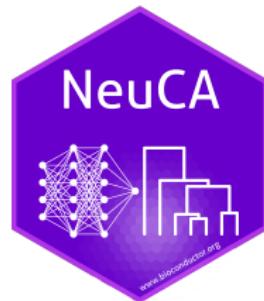
NeuCA web server

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

The screenshot shows the NeuCA web server interface. At the top, there is a navigation bar with the title "NeuCA web server" and links for Home, Tutorial, Run NeuCA, FAQ, and About. Below the navigation bar, there are two radio button options: "Built-in Pre-trained Classifier" (selected) and "Upload My Own Training Data". A section titled "Choose the data type" contains a dropdown menu with the placeholder "Please select an option below" and a link "What are these data?". Another section titled "Choose your testing file(.RData/.rda)" includes a "Browse..." button and a message "No file selected". A third section titled "Choose the model size" has a dropdown menu with the option "small". Below these sections are two buttons: "Generate Predicted Labels" (with a checked checkbox icon) and "Download Predicted Labels" (with a download icon). At the bottom of the page, there is a footer with navigation icons and a copyright notice: "Hao Feng (ENAR Spring 2022) Introduction 22 / 26".

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!).



scientific reports



OPEN

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

Ziyi Li¹ & 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,*}

¹Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, ²College of Arts and Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, ³Department of Surgery, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA and ⁴Department 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 Li¹ & Hao Feng^{2,3,4}

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

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/>