

Computational Approaches to Dissect Admixed Transcriptome Data

Hao Feng

Assistant Professor

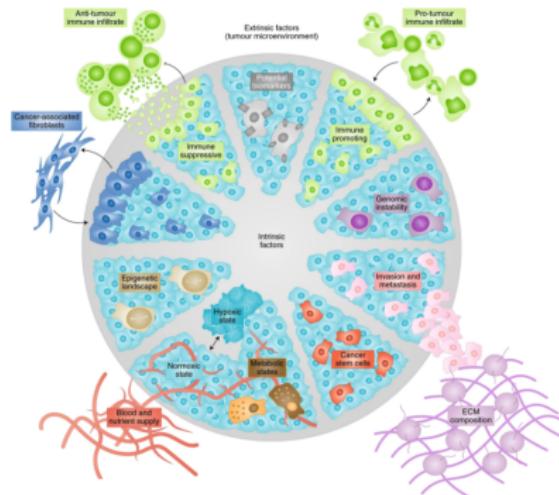
Department of Population and Quantitative Health Sciences
Case Western Reserve University

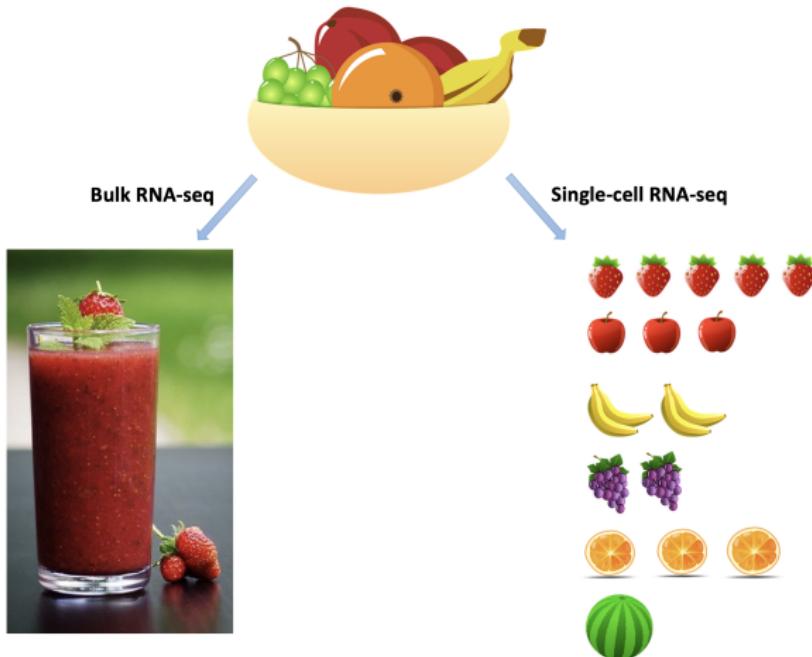
hxf155@case.edu

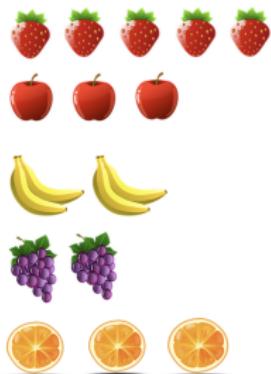
<https://hfenglab.org>

Heterogeneous mixture

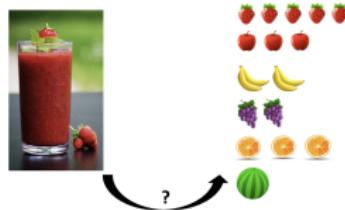
- Human tissues have diverse cell types/states.
- Traditional RNA-seq (“bulk” RNA-seq) can measure **averaged signal** across millions of cells.
- Single-cell RNA-seq (scRNA-seq) give us the first data-driven approach to study the **heterogeneous tissue** at single-cell level.







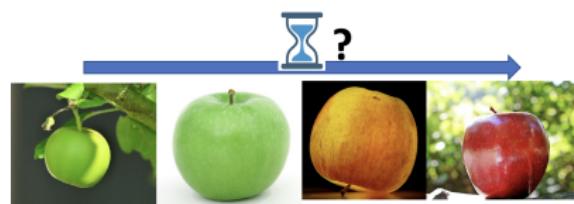
Deconvolution and beyond



Smoothie A
use:

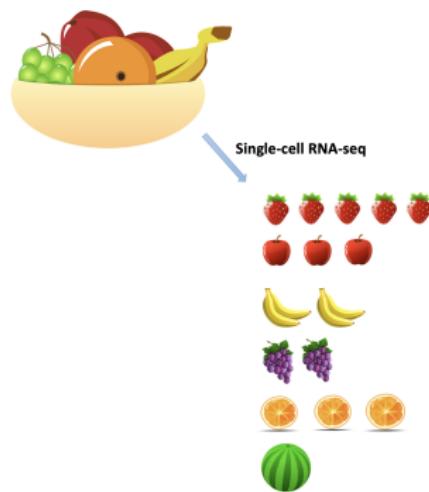


Smoothie B
use:



- ① A hybrid neural network model for single-cell RNA-seq cell type prediction.
- ② A statistical method to recover individual-specific and cell-type-specific transcriptome reference panels.

- ① A hybrid neural network model for single-cell RNA-seq cell type prediction.



scRNA-seq data

	human2_lib1.final_cell_0001	human2_lib1.final_cell_0002	human2_lib1.final_cell_0003	human2_lib1.final_cell_0004	human2_lib1.final_cell_0005	human2_lib1.final_cell_0006
SST	3476	3340	0	2962	3367	3008
INS	24	5	6	6	6	7
GCG	8	11	1995	8	4	10
REG1A	1	0	0	0	0	0
PPY	1	0	0	0	1	2
TTR	10	1	273	13	4	1
IAPP	3	0	2	3	1	0
REG3A	0	2	0	0	0	0
PRSS2	0	0	0	0	0	0
CTR2	0	0	0	0	0	0
REG1B	0	0	0	0	0	0
SPINK1	0	0	3	0	0	0
SERPINA1	89	26	362	20	5	13
SERPINA3	0	0	0	0	0	0
EEF1A1	44	21	66	52	43	30
OLFM4	0	0	0	1	0	0
GNAS	71	15	103	90	67	49
FTL	14	5	24	5	12	3
CTR2	0	0	0	0	0	0
TMSB4X	7	5	5	6	4	5

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**

- *t*-SNE, UMAP

scRNA-seq data analysis questions

- **Data preprocessing**

- Normalization
- Batch effect correction
- Imputation

- **Data analyses**

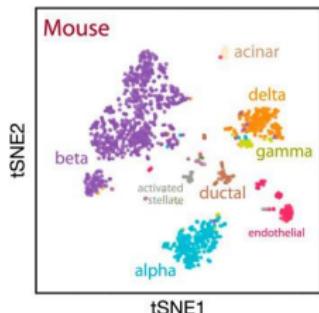
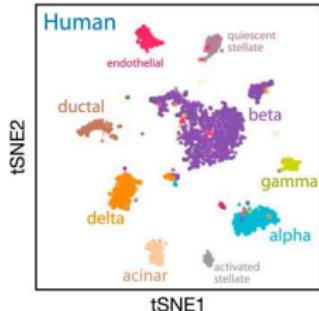
- Cell clustering
- **Cell type identification**
- Differential expression
- Pseudo-time construction
- Rare cell type discovery;
- alternative splicing; allele specific expression
- RNA velocity

- **Visualization**

- *t*-SNE, UMAP

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

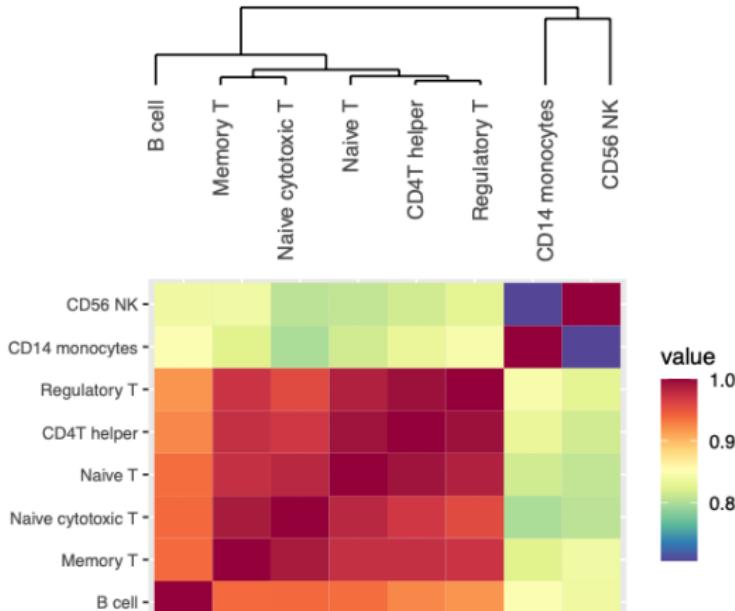
- Two-step approach: Clustering (unsupervised) + labeling.
 - Seurat, SC3, TSCAN, etc...
 - Laborious, time consuming, not best projection, rely on marker gene heavily.
- One-step approach: supervised labeling.
 - scmap, CHETAH, CellAssign, etc...
 - (1) marker-based, (2) correlation-based, and (3) tree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

Our proposed method

NeuCA: a Neural network-based Cell Annotation

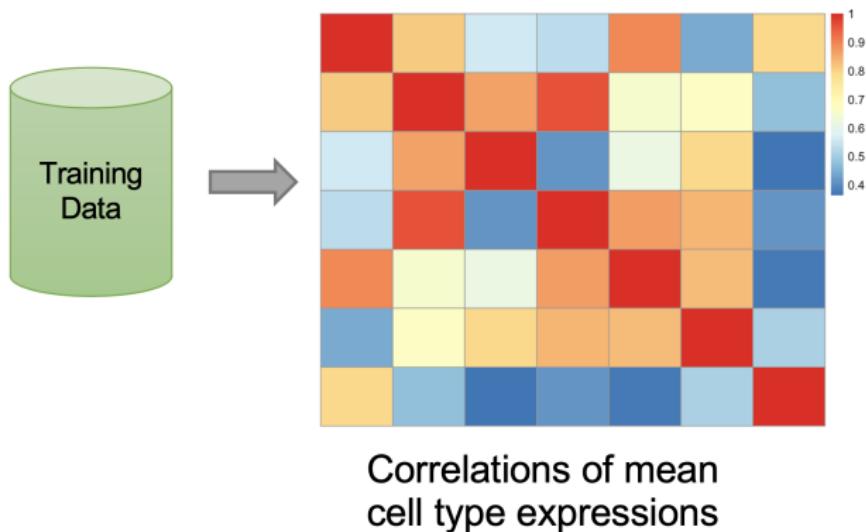
- Exhaustive: Annotate all cells.
- Reliable: Leverage on massive amount of well-studied existing scRNA data.
- Flexible: Adopt different prediction strategies, depending on correlation.

Correlations of cell types



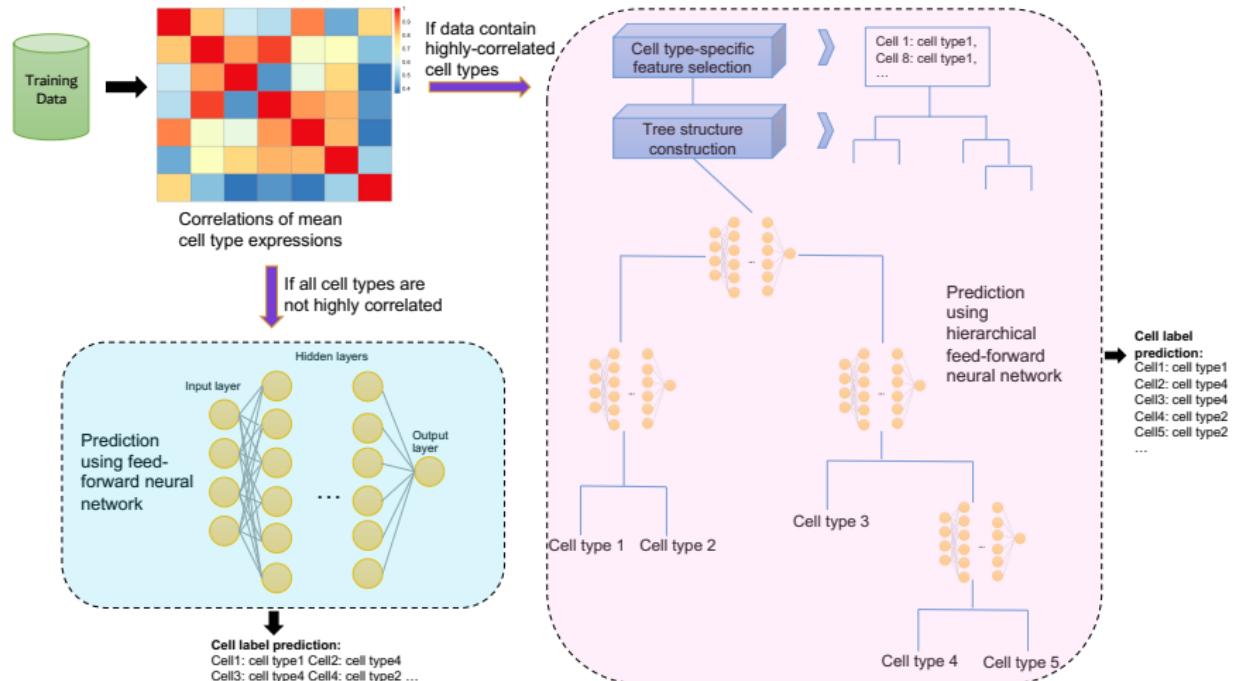
High-correlation cell types pose major challenges.

Tackle the issue of high-correlation



How about adopting a flexible approach?

Method: NeuCA



Wires inside the cogs

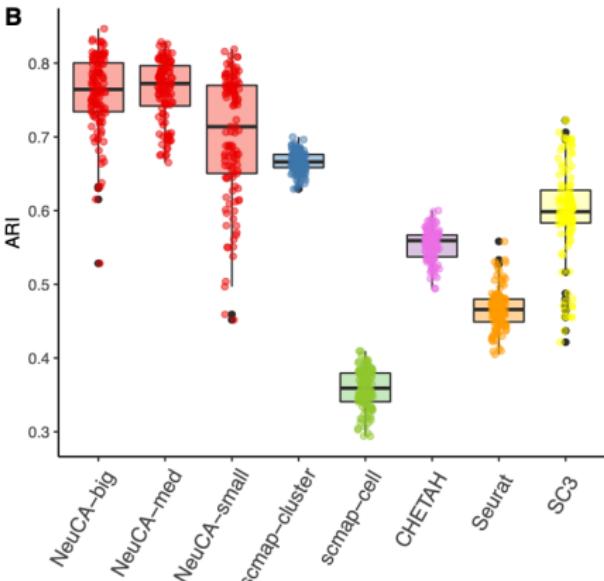
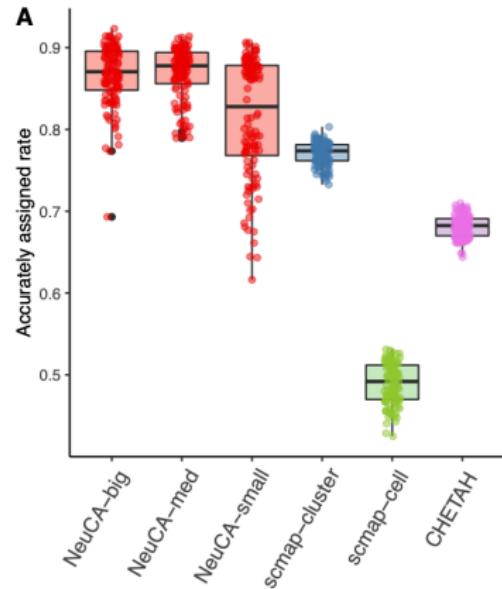
- Route 1: feed-forward neural-network
 - feature selection.
 - 3 model sizes: big, medium, small. (256 to 64 units/nodes)
 - activation function: Rectified Linear Unit (ReLU).
 - output: Softmax.
 - categorical cross-entropy loss.
- Route 2: marker-guided hierarchical neural-network
 - feature selection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

Simulation

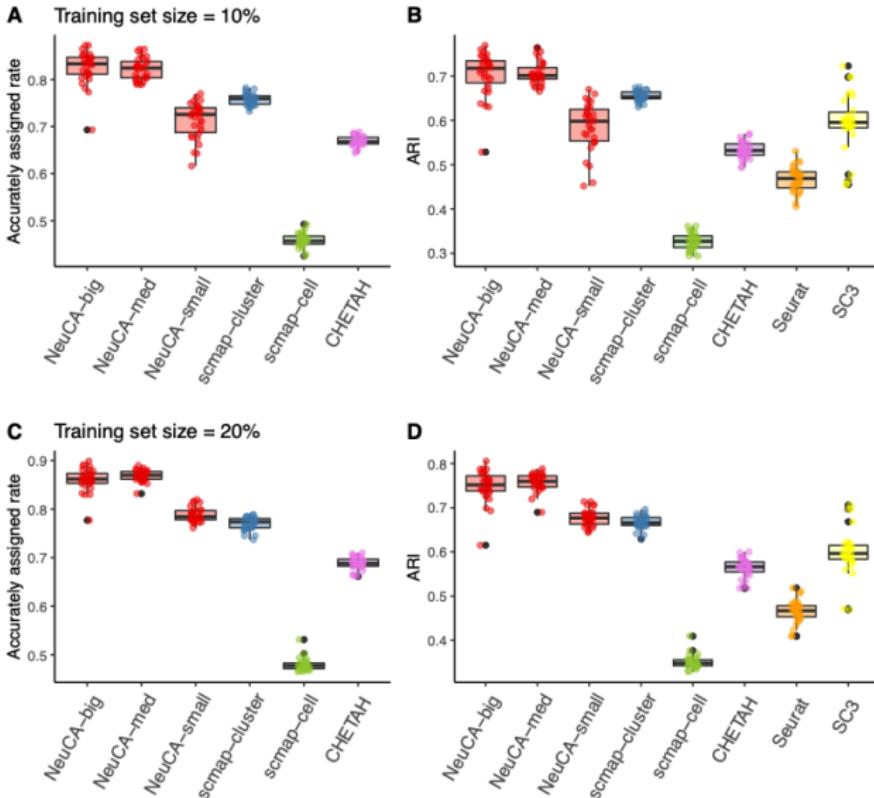
Real-data based simulation

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

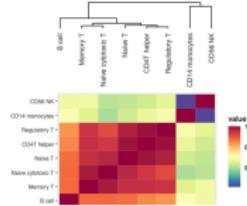
Simulation results: real-data based



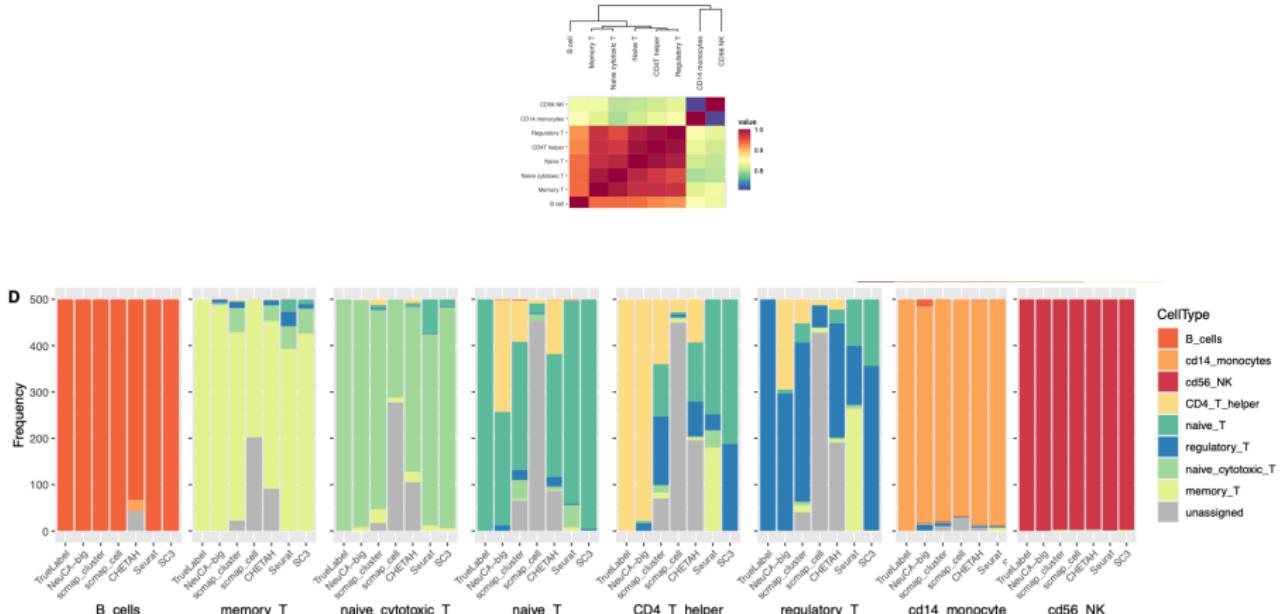
Real-data based simulation



Real-data based simulation

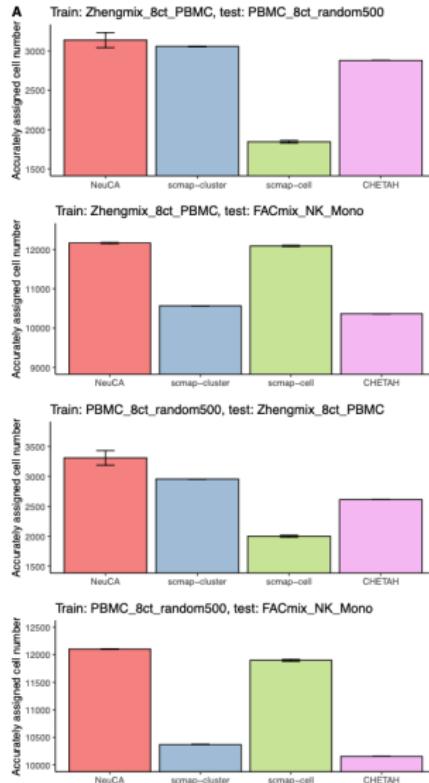


Real-data based simulation

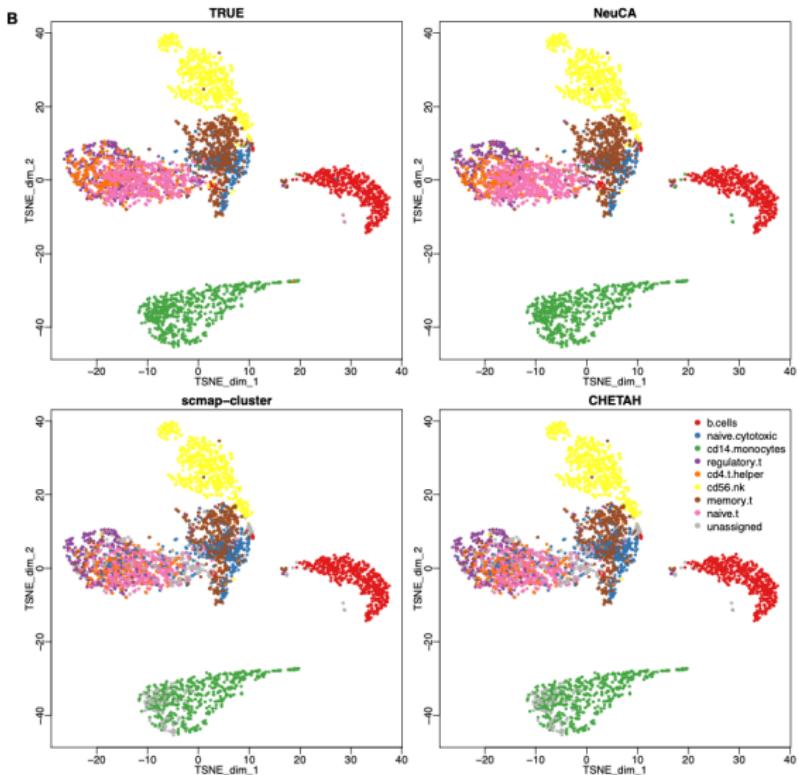


Real data results

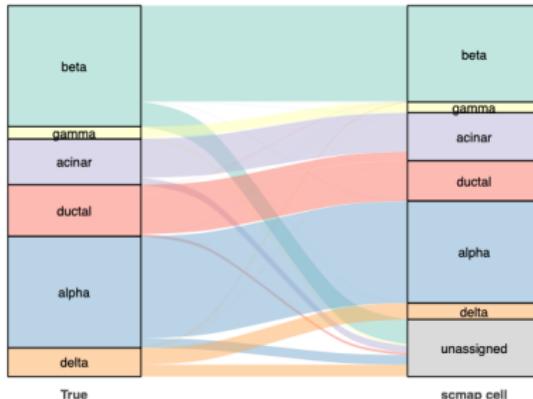
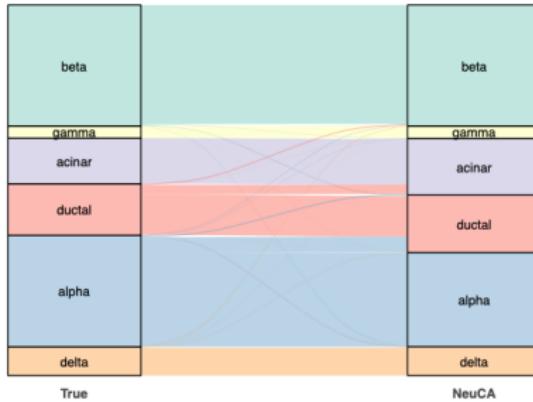
Various PBMC datasets



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 website with the NeuCA package page open. The header includes the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area displays the NeuCA package details, including its version (3.14), platforms (all), rank (2008 / 2083), support (0 / 0), dependencies (61), and build status (unknown). It also shows the DOI (10.18129/B9.bioc.NeuCA) and social media links for Facebook and Twitter. Below this, a brief description of NeuCA as a neural-network based single-cell annotation tool is provided, along with author and maintainer information. To the right, there are two boxes: one for Documentation (listing vignettes, workflows, online books, courses, videos, and community resources) and another for Support (mentioning the posting guide and support locations).

Search:

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Home » Bioconductor 3.14 » Software Packages » NeuCA

NeuCA

platforms all | rank 2008 / 2083 | support 0 / 0 | in Bioc < 6 months | build unknown | updated before release | dependencies 61

DOI: [10.18129/B9.bioc.NeuCA](https://doi.org/10.18129/B9.bioc.NeuCA) [f](#) [t](#)

NEUral network-based single-Cell Annotation tool

Bioconductor version: Release (3.14)

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: Ziyi Li [aut], Hao Feng [aut, cre]
Maintainer: Hao Feng <hxfl155 at case.edu>

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

Hao Feng Introduction

27 / 68

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>

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), Sijiu
Email: dxd429@case.edu

(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

(Mouse) Enteric

(Mouse) Hippocampus

(Mouse) Spinal Cord

NeuCA web server

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

Built-in Pre-trained Classifier Upload My Own Training Data

Choose the data type
Please select an option below ▼ [What are these data?](#)

Choose your testing file(.RData/.rda)
Browse... No file selected

Choose the model size
small

Generate Predicted Labels

 Download Predicted Labels

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.
- High accuracy.
- R/BioC package and web server with GUI (free!).



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



orts

gdatas

nt

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.

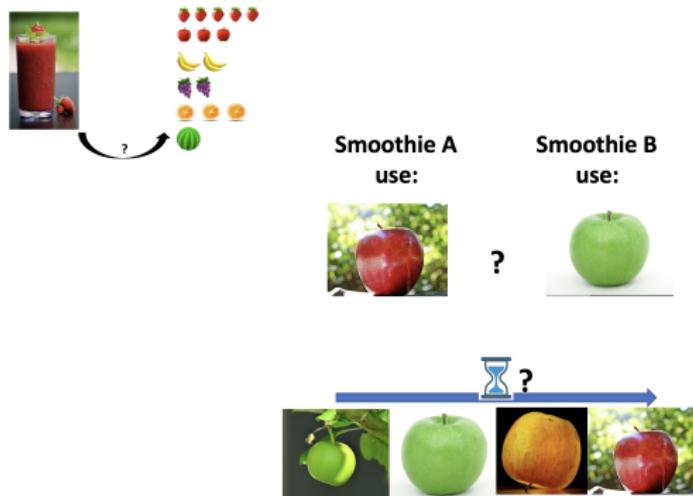
using single cell RNA-seq data

Ziyi Li & Hao Feng

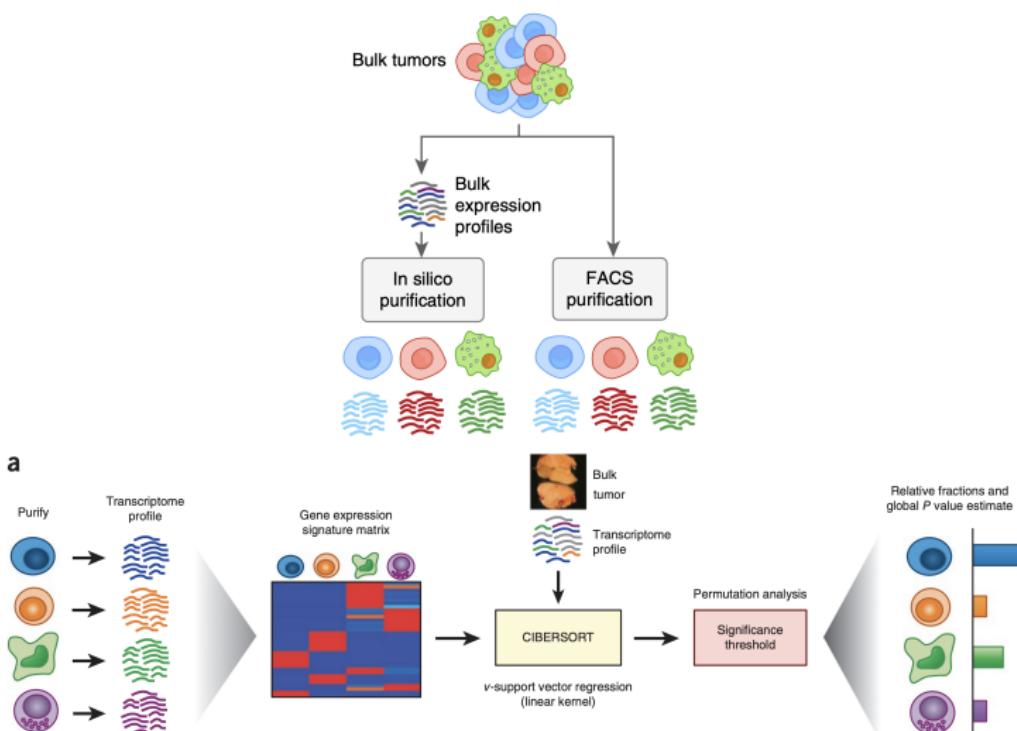
The fast-advancing single cell RNA sequencing (scRNA-seq) technology enables researchers to study the transcriptome of heterogeneous tissues at a single cell level. The initial important step of analyzing scRNA-seq data is usually to accurately annotate cells. The traditional approach of

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

- ② A statistical method to recover individual-specific and cell-type-specific transcriptome reference panels.

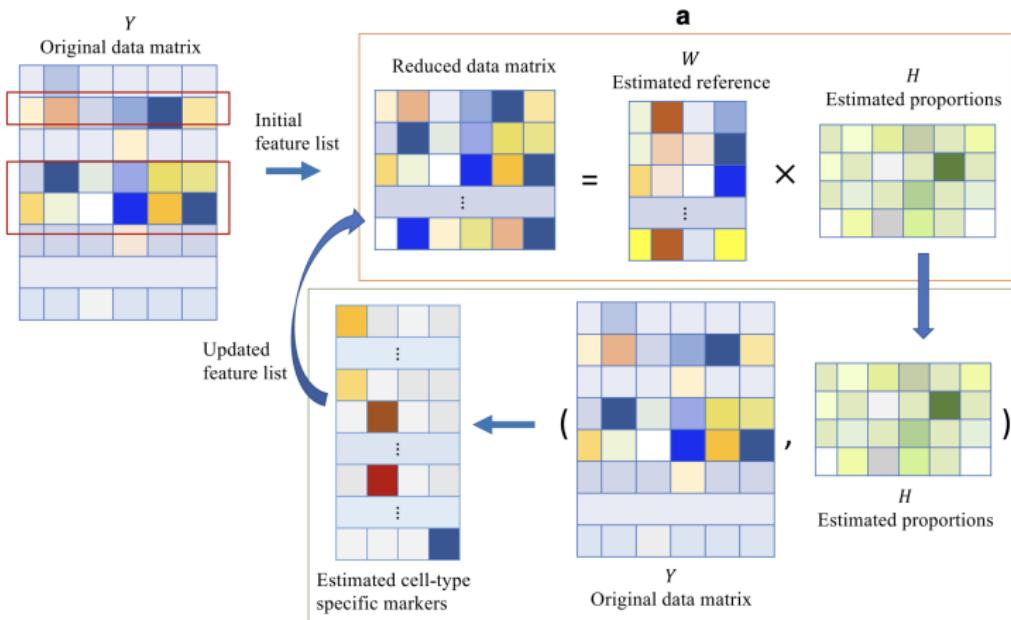


Cell composition of complex tissues



Newman et al. Nat Biotechnol. 2019; Newman et al. Nat Methods. 2015

Cell composition of complex tissues

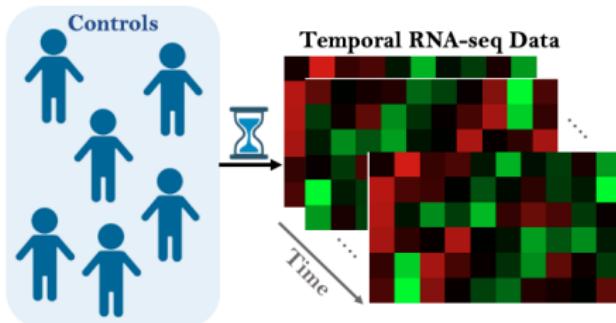
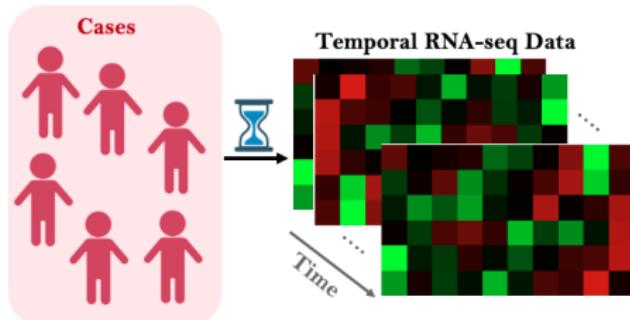


Li et al. Genome Biology 2019

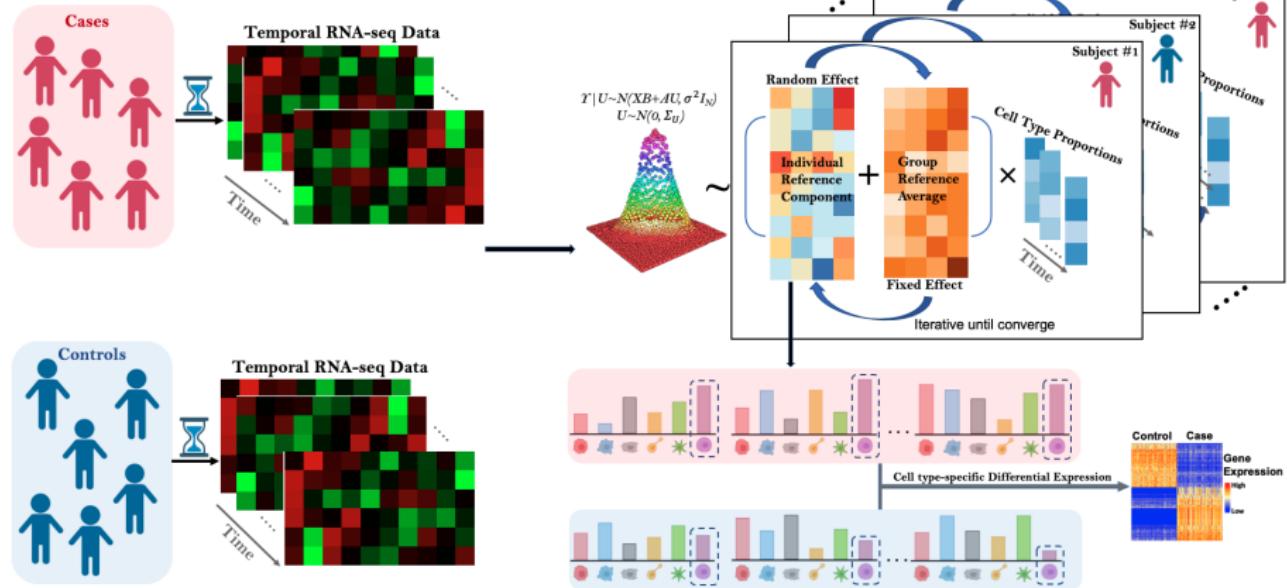
Existing problems

- Lack of individual-level reference profile.
- Repeatedly measured samples over the same individual are not optimally used.
- Bias in subsequent differential expression testing.

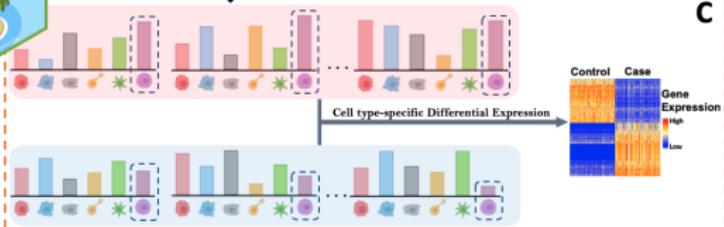
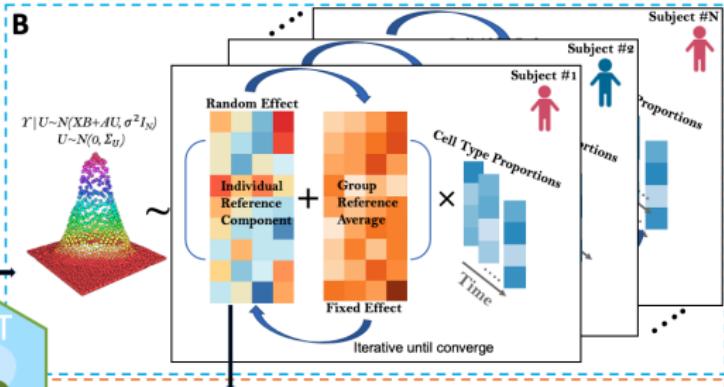
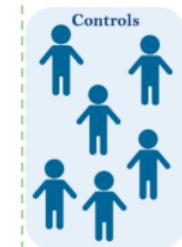
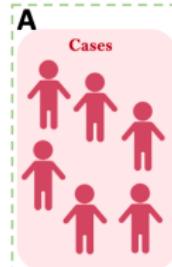
Data structure



Proposed solution



ISLET





ISLET: Individual-Specific CeLI TypE Referencing Tool

Data: Gene expression data from bulk RNA-seq

For one specific gene:

- Subject is index by j , where $j \in 1, 2, \dots, J$.
- For each subject j , there are T_j longitudinal observations.
- y_{jt} : the observed gene expression for subject j at time t .

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T_1} \\ \vdots \\ y_{J1} \\ y_{J2} \\ \vdots \\ y_{JT_J} \end{pmatrix}_{N \times 1}$$

Here, $N = \sum_{j=1}^J T_j$.

Other known input

- Number of cell types: K .
- Cell type proportions: θ_{jT_jk} , naturally $\sum_{k=1}^K \theta_{jT_jk} = 1$.
- Binary scalar z_j to indicate the subject's disease status: (e.x. cancer vs. normal).

The mixed-effect model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I})$$

$$\mathbf{X} = \left(\begin{array}{ccccccc} \theta_{111} & \theta_{112} & \dots & \theta_{11K} & z_1\theta_{111} & z_1\theta_{112} & \dots & z_1\theta_{11K} \\ \theta_{121} & \theta_{122} & \dots & \theta_{12K} & z_1\theta_{121} & z_1\theta_{122} & \dots & z_1\theta_{12K} \\ \dots & \dots \\ \theta_{1T_11} & \theta_{1T_12} & \dots & \theta_{1T_1K} & z_1\theta_{1T_11} & z_1\theta_{1T_12} & \dots & z_1\theta_{1T_1K} \\ \dots & \dots \\ \theta_{J11} & \theta_{J12} & \dots & \theta_{J1K} & z_J\theta_{J11} & z_J\theta_{J12} & \dots & z_J\theta_{J1K} \\ \theta_{J21} & \theta_{J22} & \dots & \theta_{J2K} & z_J\theta_{J21} & z_J\theta_{J22} & \dots & z_J\theta_{J2K} \\ \dots & \dots \\ \theta_{JT_J1} & \theta_{JT_J2} & \dots & \theta_{JT_JK} & z_J\theta_{JT_J1} & z_J\theta_{JT_J2} & \dots & z_J\theta_{JT_JK} \end{array} \right)_{N \times 2K}$$

Design matrix for the random effect

$$\mathbf{a}_{jk} := \begin{pmatrix} \theta_{j_1 k} \\ \theta_{j_2 k} \\ \vdots \\ \theta_{j T_j k} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{11} & 0 & 0 & 0 & \dots & \mathbf{a}_{1K} & 0 & 0 & 0 \\ 0 & \mathbf{a}_{21} & 0 & 0 & \dots & 0 & \mathbf{a}_{2K} & 0 & 0 \\ 0 & 0 & \ddots & 0 & \dots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{a}_{J1} & \dots & 0 & 0 & 0 & \mathbf{a}_{JK} \end{pmatrix}_{N \times Q}$$

$$Q = JK$$

Fixed and random effects

$$\boldsymbol{\beta} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_K \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}_{2K \times 1}$$

$$\boldsymbol{u} = \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{J1} \\ u_{12} \\ u_{22} \\ \vdots \\ u_{J2} \\ \vdots \\ u_{1K} \\ u_{2K} \\ \vdots \\ u_{JK} \end{pmatrix}_{Q \times 1}$$

Towards EM algorithm setup

$$\boldsymbol{w} = (\boldsymbol{y}, \boldsymbol{u}) := (\boldsymbol{w}_{obs}, \boldsymbol{w}_{mis}) \quad \boldsymbol{w}_{obs} := \boldsymbol{y} \quad \boldsymbol{w}_{mis} := \boldsymbol{u}$$

$$\boldsymbol{w}_{obs} | \boldsymbol{w}_{mis} = \boldsymbol{y} | \boldsymbol{u} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{u}, \sigma_0^2 \boldsymbol{I})$$

$$\boldsymbol{w}_{mis} = \boldsymbol{u} \sim N(\mathbf{0}, \Sigma_U)$$

$$\Sigma_U = \begin{pmatrix} \sigma_1^2 \boldsymbol{I}_J & 0 & \dots & 0 \\ 0 & \sigma_2^2 \boldsymbol{I}_J & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \sigma_K^2 \boldsymbol{I}_J \end{pmatrix}$$

Towards EM algorithm setup

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad (1)$$

$$COV(\mathbf{y}) = \mathbf{A}\Sigma_U\mathbf{A}^T + \sigma_0^2\mathbf{I} \equiv \mathbf{V} \quad (2)$$

$$COV(\mathbf{y}, \mathbf{u}) = \mathbf{A}\Sigma_U \quad (3)$$

Ready for EM algorithm

So we have:

$$\begin{pmatrix} \mathbf{w}_{obs} \\ \mathbf{w}_{mis} \end{pmatrix} = N \left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{A}\Sigma_U \\ \Sigma_U^T \mathbf{A}^T & \Sigma_U \end{pmatrix} \right]$$

Ready for EM algorithm

So we have:

$$\begin{pmatrix} \mathbf{w}_{obs} \\ \mathbf{w}_{mis} \end{pmatrix} = N \left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{A}\Sigma_U \\ \Sigma_U^T \mathbf{A}^T & \Sigma_U \end{pmatrix} \right]$$

Theorem

If $X = (X_1, X_2)$, and $X \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$, then

$[X_1|X_2] \sim N(\mu_{1|2}, \Sigma_{1|2})$, where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Thus, we have:

$$[\mathbf{w}_{mis} | \mathbf{w}_{obs}] = [\mathbf{u} | \mathbf{y}] \sim N(\boldsymbol{\mu}_p, \Sigma_p)$$

E-step

Define: $\mathbf{s} = \mathbf{A}\mathbf{u} + \mathbf{X}\boldsymbol{\beta} - \mathbf{y}$

Then we have:

$$E[\mathbf{u} | \mathbf{w}_{obs} = \mathbf{y}] = \Sigma_U^T \mathbf{A}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$E[\mathbf{s}^T \mathbf{s} | \mathbf{w}_{obs} = \mathbf{y}] = \text{tr}(\mathbf{A} \Sigma_p \mathbf{A}^T) + (\mathbf{A}\boldsymbol{\mu}_p + \mathbf{X}\boldsymbol{\beta} - \mathbf{y})^T (\mathbf{A}\boldsymbol{\mu}_p + \mathbf{X}\boldsymbol{\beta} - \mathbf{y})$$

$$E[\mathbf{u}_k^T \mathbf{u}_k | \mathbf{w}_{obs} = \mathbf{y}] = \text{tr}(\Sigma_{p_k}) + \boldsymbol{\mu}_{p_k}^T \boldsymbol{\mu}_{p_k}$$

M-step

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{A} E_{\eta(t)}(\mathbf{u}^{(t)}))$$

$$\hat{\sigma}_0^{2(t+1)} = \frac{E_{\eta(t)} [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{A}\mathbf{u})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{A}\mathbf{u})]}{N}$$

$$\hat{\sigma}_k^{2(t+1)} = \frac{E_{\eta(t)} (\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k)}{J}$$

Simulation setup

- $N = 50, 100, 150, 200$
- $LFC = 0(\text{null}), 0.5, 0.75, 1.0, 1.25, 1.5.$
- 10% or 0% (null) csDEG.
- 6 cell types
- Reference panel generated from real bulk cell line.
- Proportions from Dirichlet with parameters from scRNA-seq data.
- Gamma-Poisson for observed counts.

Simulation procedure

1

$$\boldsymbol{\mu}_{g,K \times 1} \sim MVN(\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Sigma}}_m)$$

$$\boldsymbol{\phi}_{g,K \times 1} \sim MVN(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_d)$$

2

$$\boldsymbol{M}_{G \times K} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G]^T; \boldsymbol{\Phi}_{G \times K} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_G]^T$$

3

$$\boldsymbol{X}_{G \times K}^i \sim Gamma\{shape = \frac{1}{\exp(\boldsymbol{\Phi})}, scale = \exp(\boldsymbol{M}) \cdot \exp(\boldsymbol{\Phi})\}$$

4

$$\boldsymbol{\theta}_{it} \sim Dir(\boldsymbol{\alpha})$$

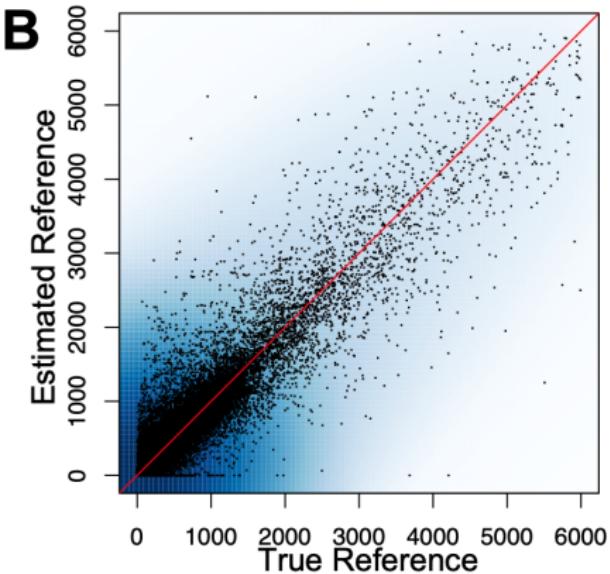
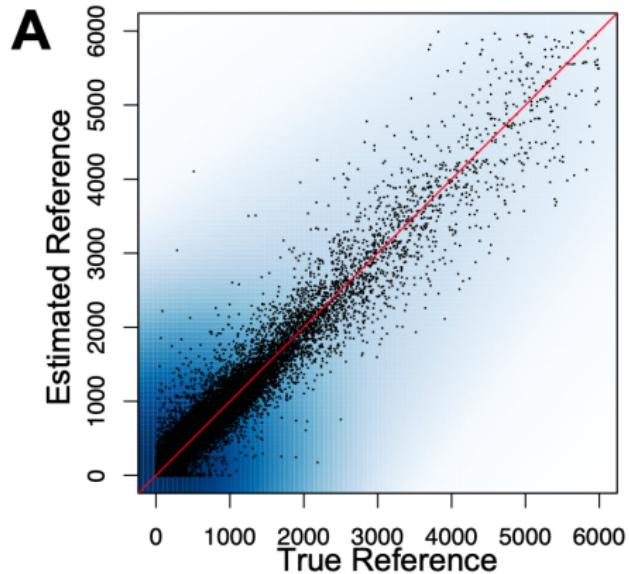
5

$$\boldsymbol{r}_{it} = \boldsymbol{X}^i \boldsymbol{\theta}_{it}$$

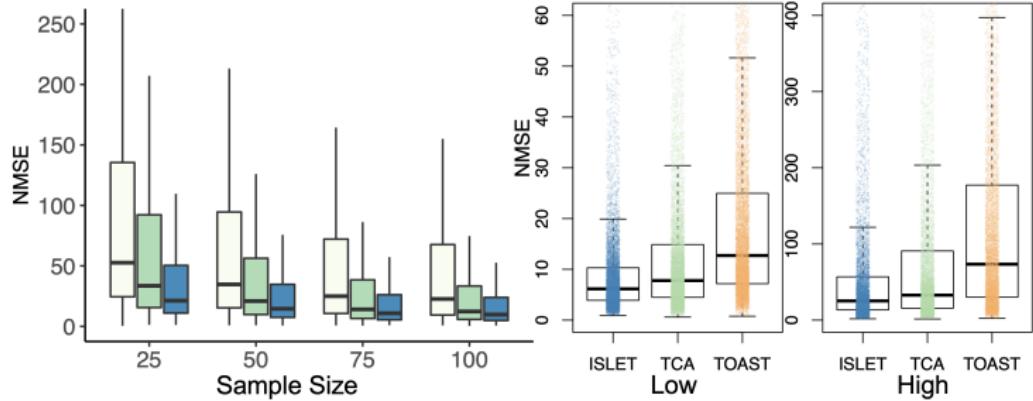
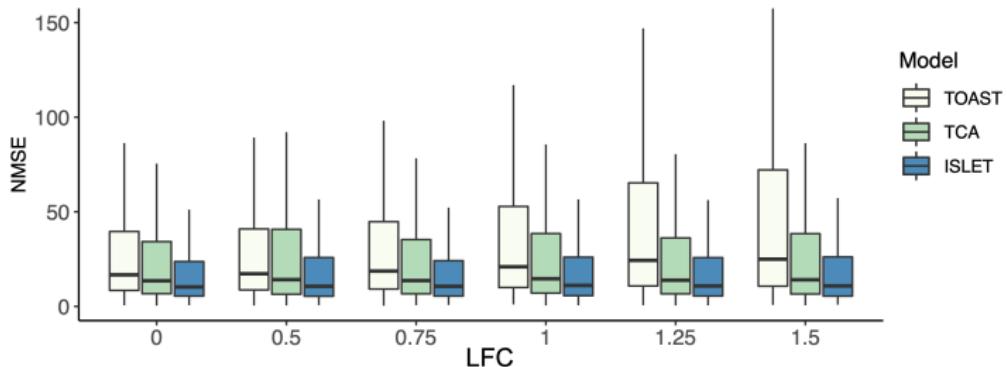
$$\boldsymbol{y}_{it} | \boldsymbol{r}_{it} \sim Poisson(\boldsymbol{r}_{it})$$

Simulation results

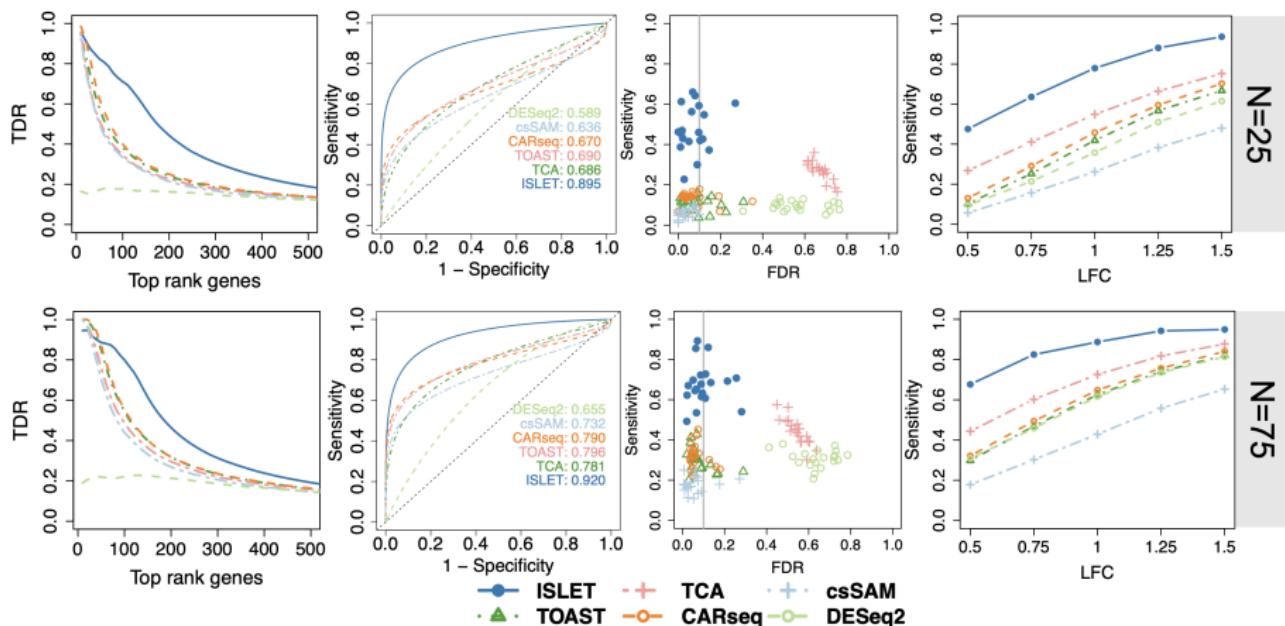
Individual reference panel recovery



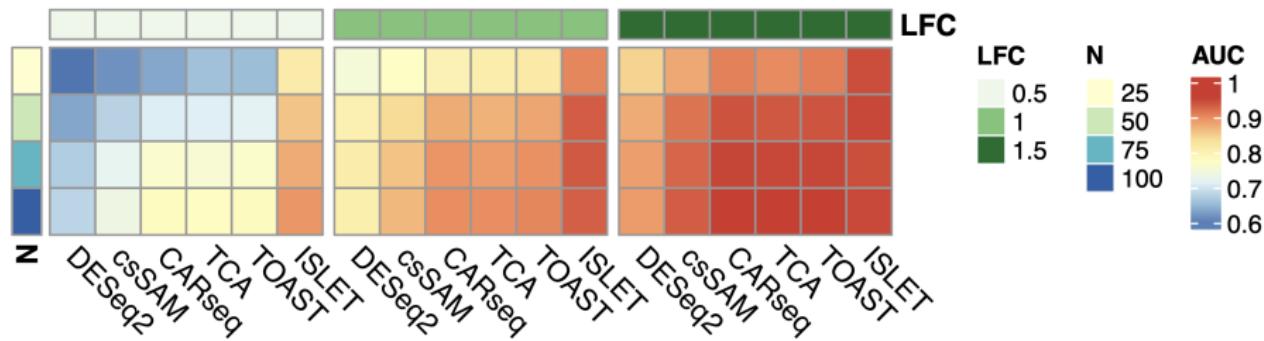
Individual reference panel recovery



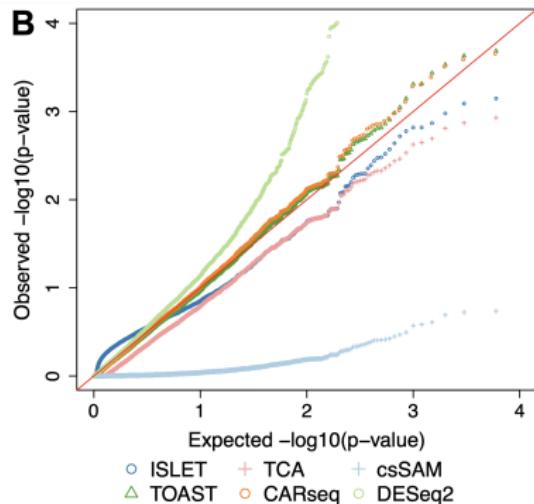
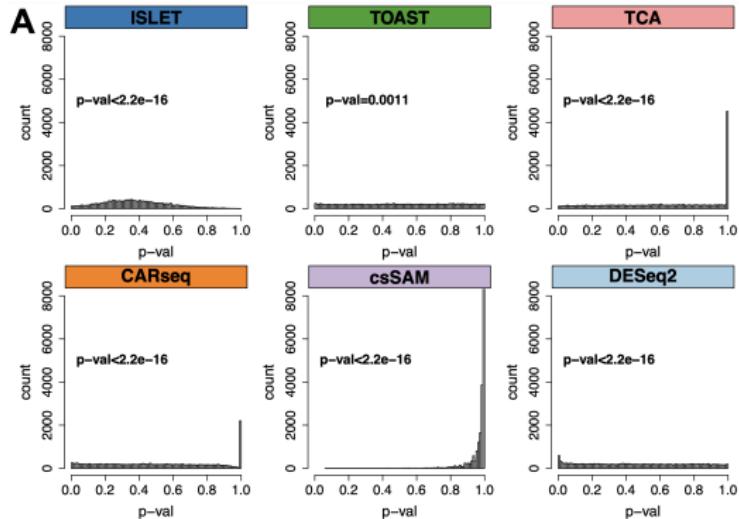
Identifying cell-type-specific DE genes



Identifying cell-type-specific DE genes



Type I error under the null



Real data: TEDDY

The Environmental Determinants of Diabetes in the Young

The screenshot shows the TEDDY Study website. At the top right is a logo featuring a teddy bear holding a small globe. Below the logo is the text "The Environmental Determinants of Diabetes in the Young". To the left is a photo of a baby. A message at the top center says: "Thank you for your interest in the TEDDY Study! We have reached our screening goal and are no longer accepting any new TEDDY subjects". On the left sidebar, there's a vertical menu with links: "Information for Participants and Families", "What is Type-1 Diabetes?", "What is the TEDDY Study?", "Clinical Centers", "News and Publications", "Information for Researchers", "TEDDY Participant Portal", and "TEDDY Staff Members Website". The main content area has two columns. The left column, titled "What is Type-1 Diabetes?", contains text about the disease and a bulleted list of facts. The right column, titled "What is the TEDDY Study?", contains text about the study and a photo of a baby. A sidebar on the right side of the main content area also features a photo of a baby and text about the study's goal.

Information for Participants and Families

What is Type-1 Diabetes?

What is the TEDDY Study?

Clinical Centers

News and Publications

Information for Researchers

TEDDY Participant Portal

TEDDY Staff Members Website

Thank you for your interest in the TEDDY Study! We have reached our screening goal and are no longer accepting any new TEDDY subjects

Finding diabetes early can prevent serious illness and complications

Most of the new cases of type 1 diabetes occur in children who have no family history of the disease.

What is Type-1 Diabetes?

Type 1 diabetes is one of the most common and serious long-term diseases in children. It is a disease where the body's immune system attacks the cells that make insulin. Insulin helps sugar (glucose) get into your cells so it can be used as energy.

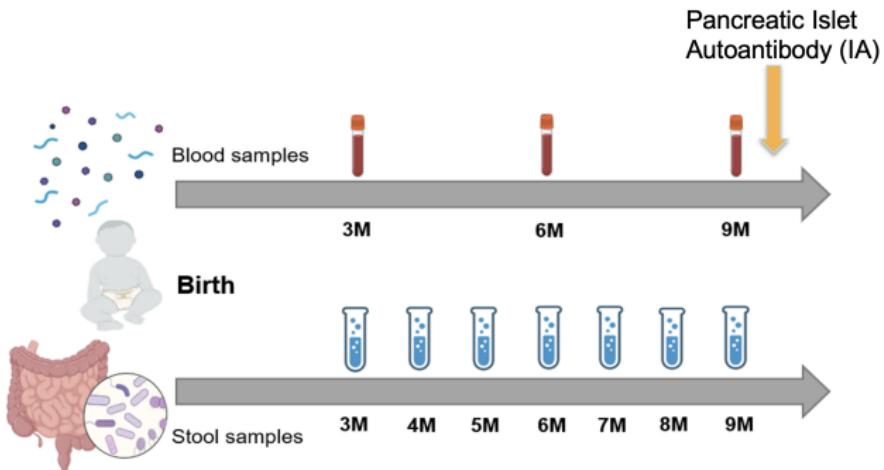
- T1D is a serious disease affecting 1 out of every 300 (1/300) children in the United States.
- T1D occurs when special cells in the pancreas, called beta cells, are destroyed by the body's own immune system. When

What is the TEDDY Study?

Every child in TEDDY helps us come closer to preventing this disease.

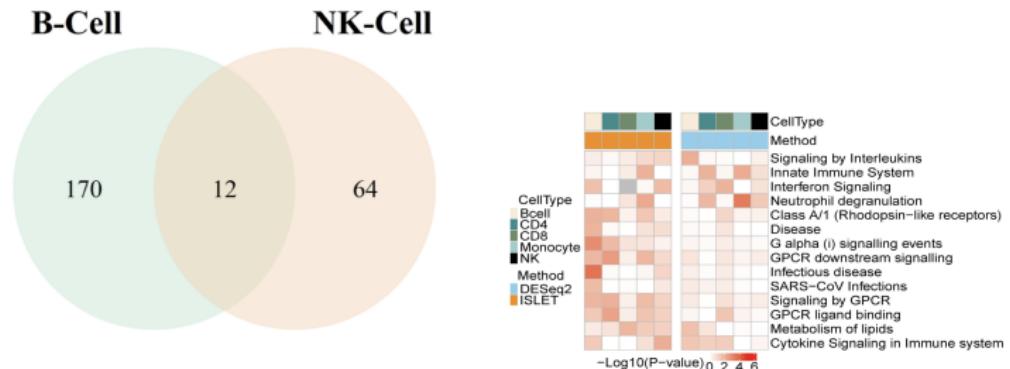
The TEDDY study - The Environmental Determinants of Diabetes in the Young - is looking for the causes of type 1 diabetes mellitus (T1DM). T1DM used to be called childhood diabetes or insulin-dependent diabetes.

TEDDY Data



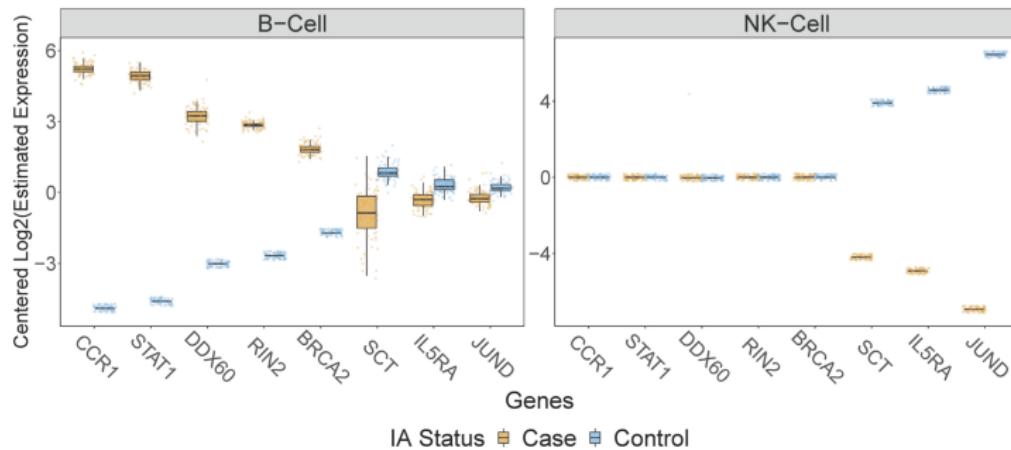
Six clinical centers in four countries: U.S., Finland, Germany, and Sweden. Prospective cohort: 8,676 high-risk infants were enrolled from birth and followed every 3 months, for blood sample collection and islet autoantibody (IAbs) measurement. Developing IA (cases) started at 9 months with a plateau between 1-2 years of age.

Real data: age-independent effect



IA-signatures strongly enriched in B cells and NK cells transcripts, kinases, and TFs.

Real data: age-independent effect

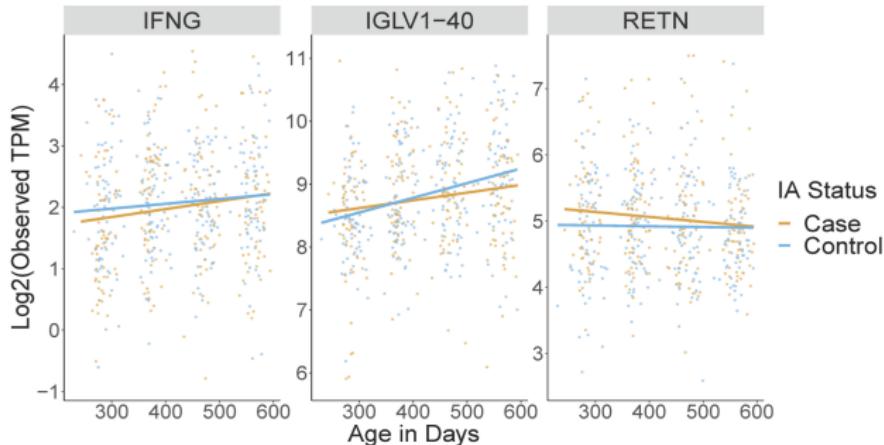


IFN- β -inducible genes in vitro and IFN-indicible transcriptional signatures in human PBMCs were also found prior to the development of autoantibodies (BABY DIET).

CCR1 (chemokine receptor): increased risk of IA in TEDDY among participants who experienced virus infection.

SCT, IL5RA, JUND: regulation of immune response.

Real data: change rate



RETN: Diabetes study found SNPs associated with plasma resistin and glucose levels.

IGLV1-40: neutralizing activity related to protective antibody responses in infants.

More significant and less ambiguous than TEDDY microarray.

ISLET available on Bioconductor

<https://bioconductor.org/packages/ISLET/>



Individual-specific and cell-type-specific deconvolution using ISLET

Hao Feng* and Qian LF

¹Department of Population and Quantitative Health Sciences, Case Western Reserve University

²Department of Biostatistics, St. Jude Children's Research Hospital

*hxf155@case.edu

9 September 2022

Abstract

This vignette introduces the usage of the Bioconductor package ISLET (Individual-Specific ceLl typE referencing Tool). Complementary to classic deconvolution algorithms, ISLET can take cell type proportions as input, and infer the individual-specific and cell-type-specific reference panels. ISLET also offers functions to detect cell-type specific differential expression (cDE) genes. Additionally, it can test for cDE genes change rate difference between two groups, given an additional covariate of time points or age. ISLET is based on rigorous statistical framework of Expectation–Maximization(EM) algorithm, and has parallel computing embedded to provide superior computational performance.

Package

ISLET 0.99.8

Contents

1 Install and help

1.1 Install ISLET

1.2 How to get help

The screenshot shows the Bioconductor package page for ISLET. At the top, there is a navigation bar with links for Home, Install, and Help. Below the navigation bar, the package name "ISLET" is displayed along with its version "0.99.8". The page includes sections for "platforms", "rank", "support", and "dependencies". It also shows the DOI: [10.1186/BioProject/ISLET](https://doi.org/10.1186/BioProject/ISLET). A note indicates that this is the development version of ISLET, and users should install the [devel version](#) of Bioconductor. The page also contains a detailed description of the package, author information (Hao Feng, Qian Li), and citation details.

ISLET summary

- First method to recover individual-specific and cell-type-specific reference panel.
- ISLET: an EM algorithm based deconvolution and testing framework.
- Optimize the usage of repeatedly measured samples within the same subject.
- Accurate, robust and powerful performance.
- A R/Bioconductor package.
- ISLET successfully identified gene signatures in B, NK and CD4+ T cells, prior to the onset of pancreatic β -cell autoantibody.

Acknowledgement



- Daoyu Duan
- Leslie Meng



- Qian Li



- Ziyi Li



CASE
COMPREHENSIVE
CANCER CENTER

❖ IRG-16-186-21

@HHarryFeng
 <https://hfenglab.org/>

ENAR 2023 Spring Meeting

March 19-22

JW Marriott Nashville | Nashville, TN



Monday, March 20 | 10:30 am – 12:15 pm

T2 | Cell-type-aware Differential Analysis for Bulk Transcriptome Data

Instructor:

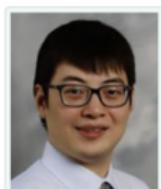
Hao Feng, Department of Population and Quantitative Health Sciences, Case Western Reserve University

Course Description:

The real-world clinical tissue samples are composed of diverse cell types. Recent statistical method advances in signal decomposition have enabled transcriptome studies at cell type resolution. In this tutorial, we will cover common software packages for cell type aware analysis in bulk transcriptome data. First, we will briefly discuss signal deconvolution, to introduce the basics in estimating pure cell type reference profiles and estimating cell type proportions. Next, we will mainly focus on widely adopted tools to conduct cell-type-specific Differential Expression Genes (csDEG) analysis. Novel methods and research progress in this domain will also be covered. This tutorial will have both the lecture component and the hands-on coding practice. We will provide R code implementation for methods introduced in this tutorial.

Statistical/Programming Knowledge Required:

Basic R programming.



Hao "Harry" Feng, is an Assistant Professor in the Department of Population and Quantitative Health Sciences at Case Western Reserve University. His research focuses on the development and the application of statistical bioinformatics methods to better understand high-throughput -omics data, especially in epigenomics. He proposed several novel statistical models in epigenomics data modeling, single-cell data analysis, cell-free DNA methylation and signal deconvolution. He developed a number of open-source software tools that are available on R-CRAN and Bioconductor, with > 24,000 downloads annually. He received his Ph.D. in Biostatistics and Bioinformatics from Emory University

ENAR 2023 Spring Meeting

March 19–22

JW Marriott Nashville | Nashville, TN



Decomposing Admixed Genomics Data: Cell-type-aware Analysis Methodology Advances

Chair & Organizer: Hao Feng, Case Western Reserve University

Speakers:

Aaron Newman, Stanford University

Stephanie Hicks, Johns Hopkins Bloomberg School of Public Health

Wenyi Wang, The University of Texas MD Anderson Cancer Center

Rafael Irizarry, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health.

