

Personalized cell-type-specific -omics Profile Deconvolution and Inference

Hao Feng

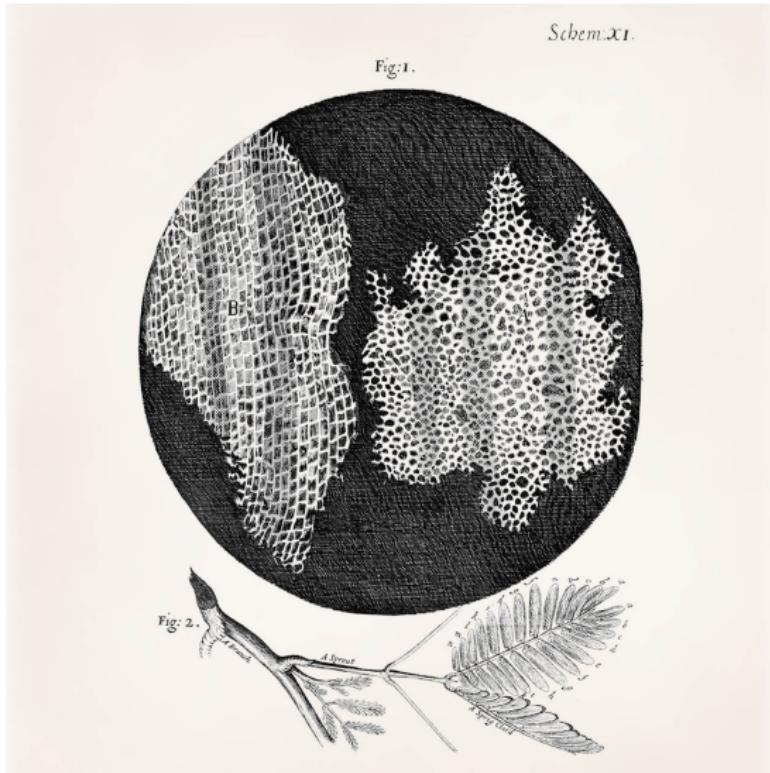
Assistant Professor

Department of Population and Quantitative Health Sciences
Case Western Reserve University

hxf155@case.edu

<https://hfenglab.org>

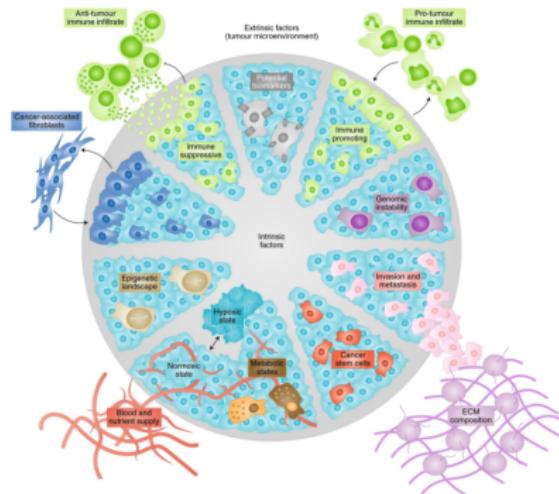
The building block of life



Robert Hooke's drawing of cork cells. Image obtained from Micrographia.

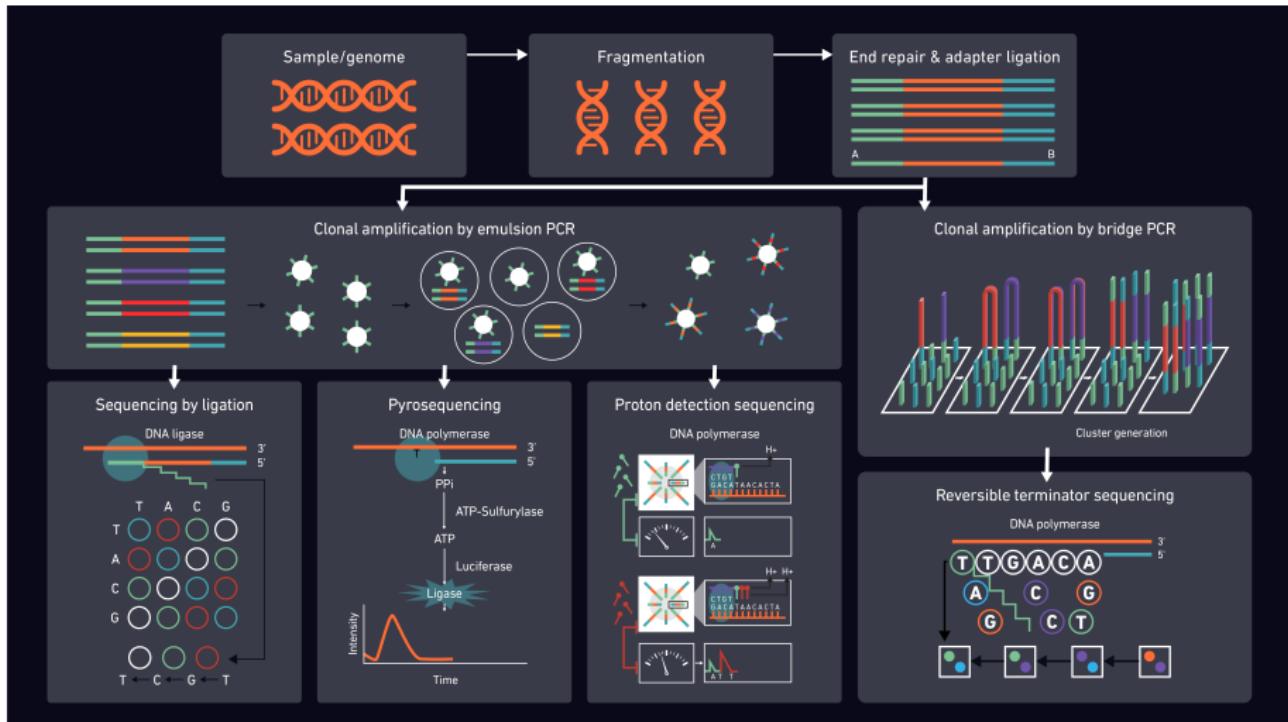
Heterogeneous mixture

- Human tissues have diverse cell types/states.
- Traditional RNA-seq (“bulk” RNA-seq) can measure **averaged signal** across millions of cells.



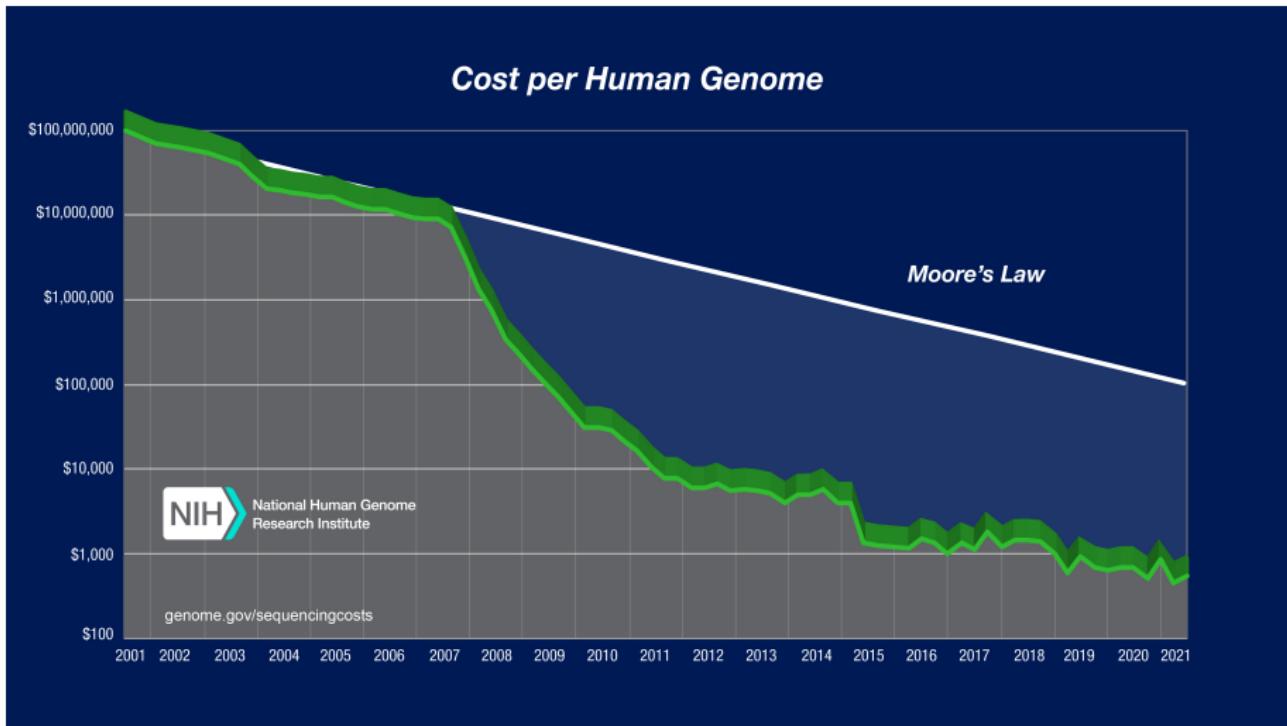
Lawson et al. Nature. <https://www.nature.com/articles/s41556-018-0236-7>

Second Generation Sequencing



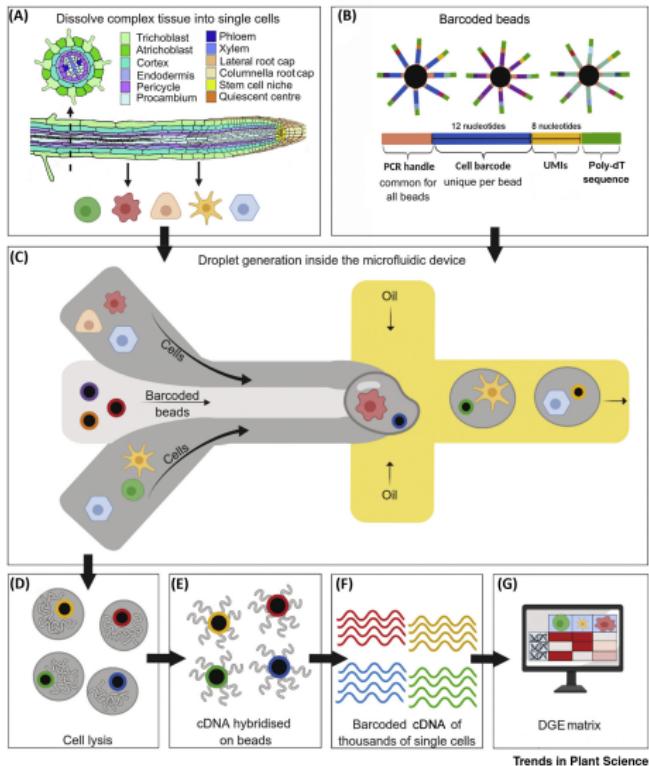
Athina Gkazi, Principle 2G sequencing platforms and chemistries. Technology Networks Genomics Research.

Sequencing costs

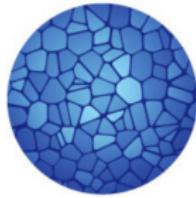


NIH/NHGRI - DNA Sequencing Costs: Data

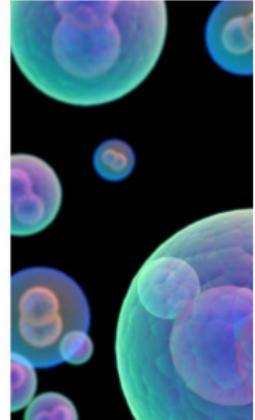
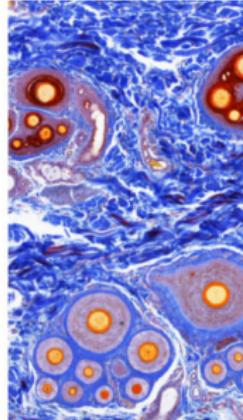
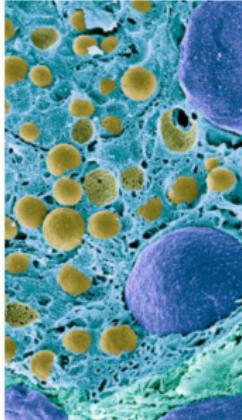
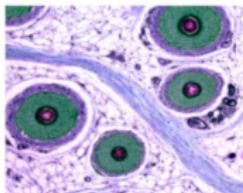
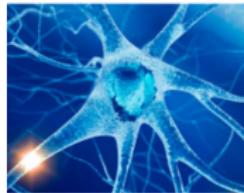
Single-cell sequencing: a high-resolution avenue

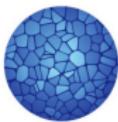


<https://doi.org/10.1016/j.tplants.2019.10.008>

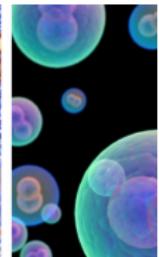
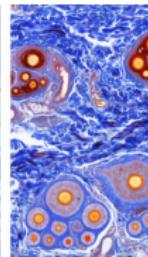
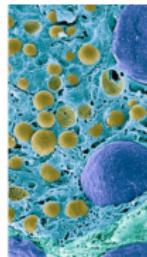
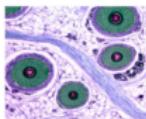


HUMAN CELL ATLAS





HUMAN
CELL
ATLAS



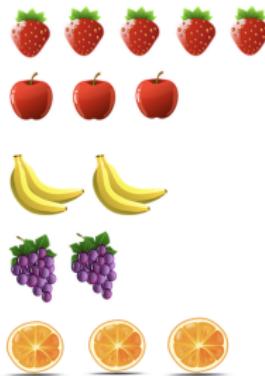
Bulk RNA-seq



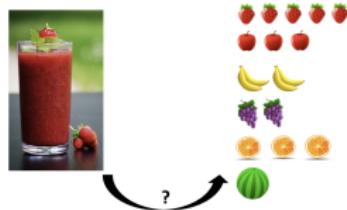
Single-cell RNA-seq



Deconvolution



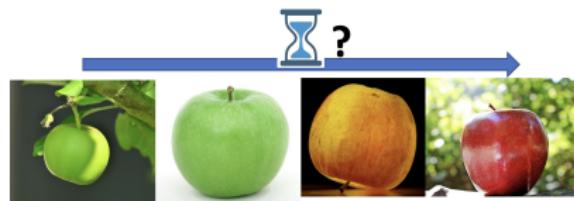
Deconvolution and beyond



Smoothie A
use:



Smoothie B
use:



- ① Personalized and cell-type-specific transcriptome reference panel recovery.
- ② Improved deconvolution using personalized reference panel.
- ③ A hybrid neural network model for scRNA-seq cell type prediction.

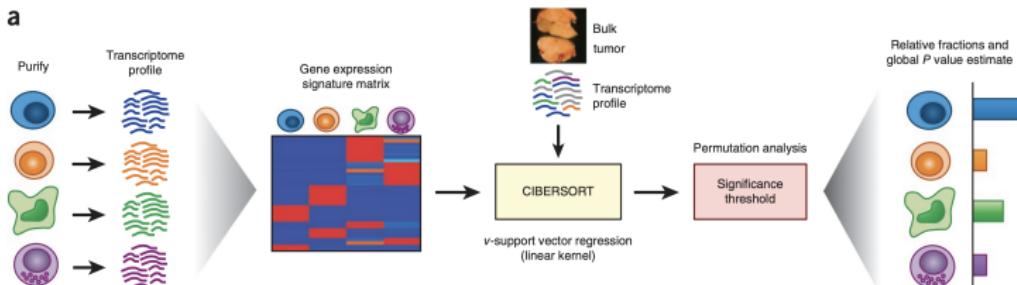
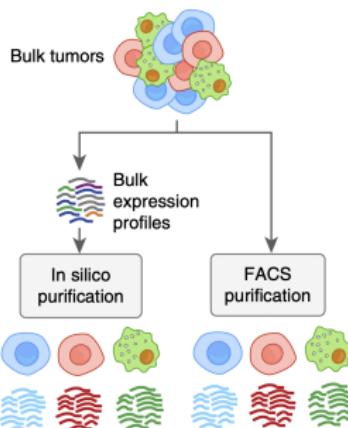
- ① Personalized and cell-type-specific transcriptome reference panel recovery.
- ② Improved deconvolution using personalized reference panel.
- ③ A hybrid neural network model for scRNA-seq cell type prediction.

Bulk RNA-seq data

```
> GeneExpData
```

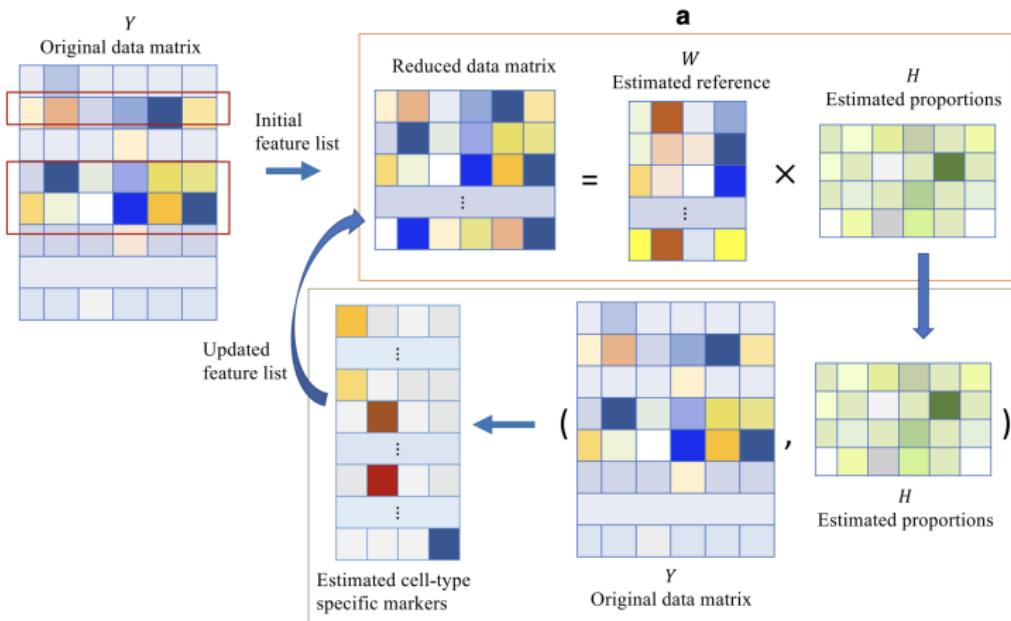
	Case_5_Spl_2	Case_7_Spl_3	Case_3_Spl_2	Case_6_Spl_1	Case_8_Spl_3	Case_3_Spl_3	Case_6_Spl_3	Case_9_Spl_3	Case_4_Spl_1
gene1	467	649	38	42	36	104	103	117	188
gene2	26	26	55	57	57	157	159	177	151
gene3	233	151	289	266	259	67	67	71	68
gene4	94	111	40	42	35	72	72	81	44
gene5	30	39	41	37	41	36	51	36	89
gene6	47	41	88	83	79	38	41	43	29
gene7	138	90	86	78	81	61	55	54	143
gene8	100	78	112	101	101	88	85	86	65
gene9	36	42	36	35	36	77	73	74	165
gene10	29	32	121	136	115	40	40	42	48

Cell composition of complex tissues



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

Cell composition of complex tissues

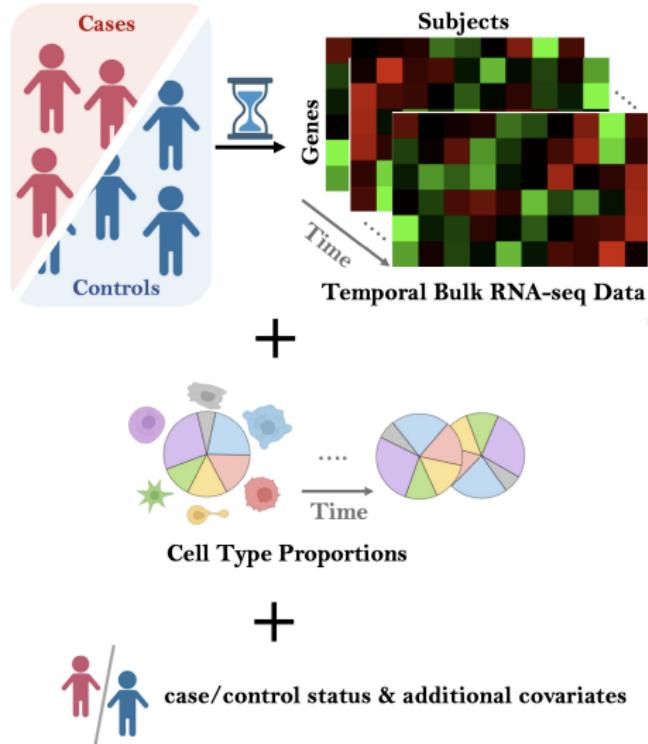


Li et al. Genome Biology 2019

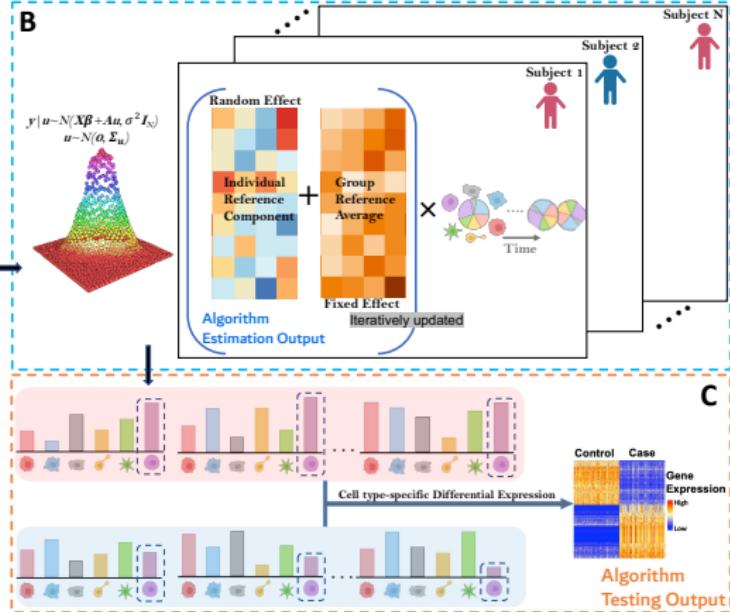
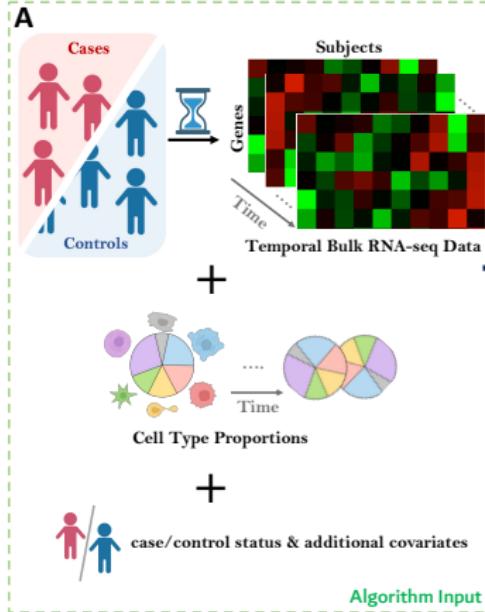
Existing problems

- 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 method



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 .

$$\mathbf{y} = (y_{11}, y_{12}, \dots, y_{1T_1}, \dots, y_{J1}, y_{J2}, \dots, y_{JT_J})_{N \times 1}^T$$

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

Other known inputs:

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

The mixed-effect model

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

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

$$\mathbf{u} = \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{J1} \\ \vdots \\ u_{1K} \\ u_{2K} \\ \vdots \\ u_{JK} \end{pmatrix}_{Q \times 1} \quad \boldsymbol{\beta} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_K \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}_{2K \times 1}$$

Towards EM algorithm setup

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

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

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

So we have:

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

Towards EM algorithm setup

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

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

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

So we have:

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

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 $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

E-step & M-step

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

E-step:

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

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

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

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

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

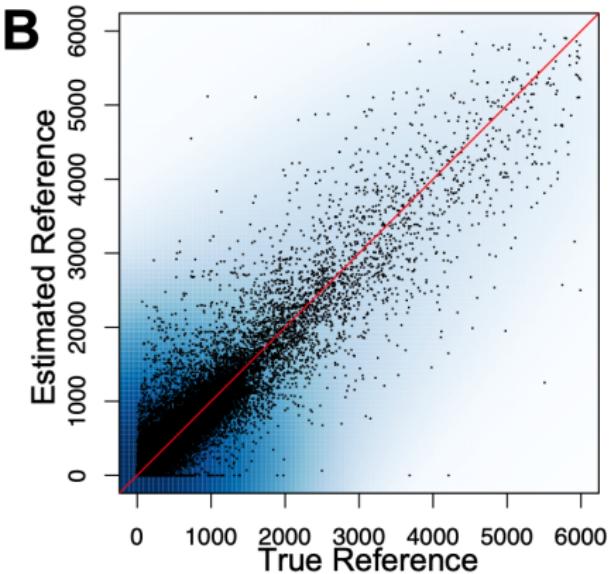
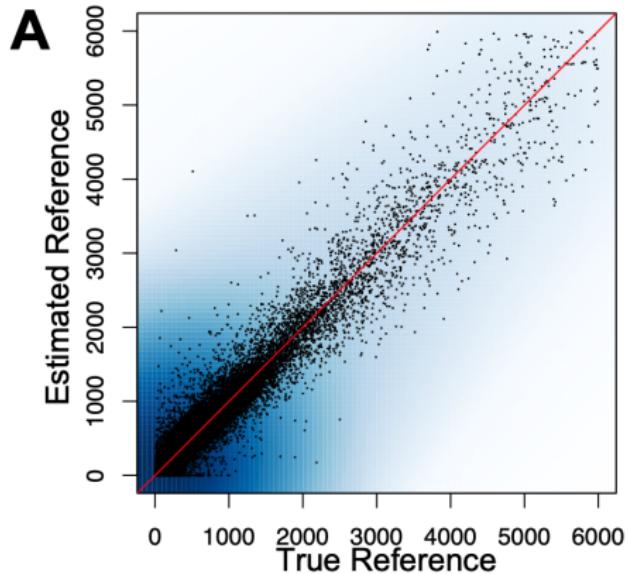
5

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

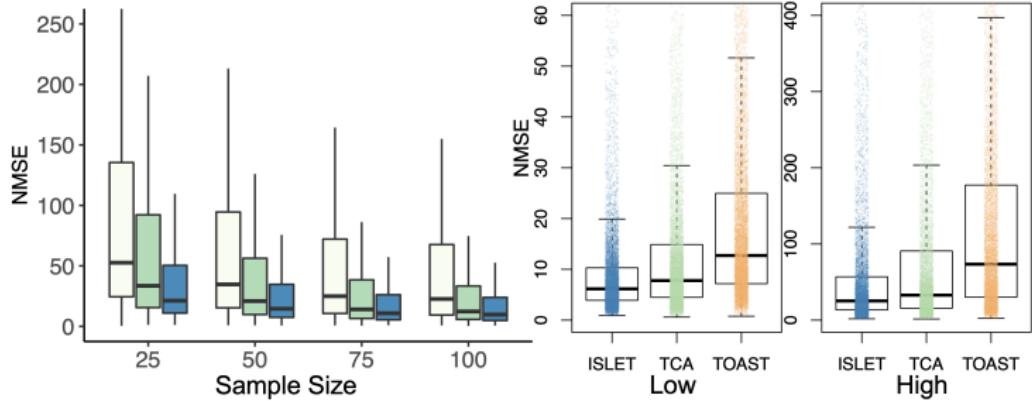
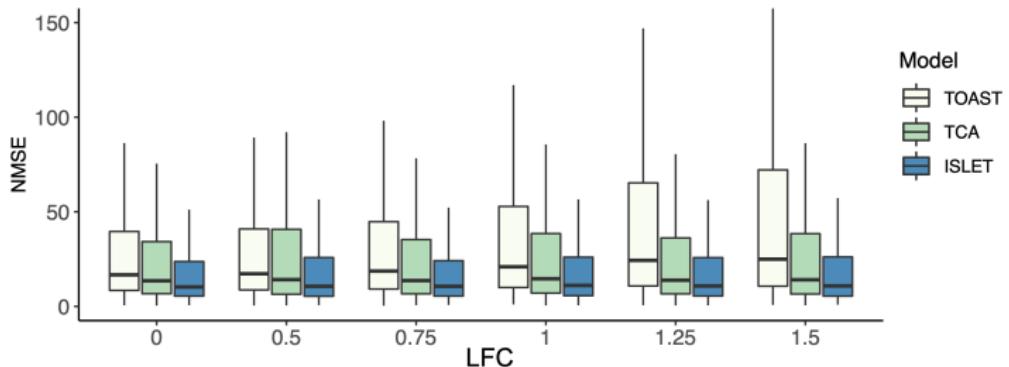
$$\boldsymbol{y}_{it} | \boldsymbol{r}_{it} \sim Poisson(\boldsymbol{r}_{it})$$

Simulation results

Individual reference panel recovery



Individual reference panel recovery



Hypothesis testing

To test cell-type-specific DE genes, we can test each hypothesis:

$$H_0 : \beta_k = 0$$

Using the (observed) marginal likelihood function:

$$l(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \ln f(\mathbf{y}; \mathbf{X}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\sigma})$$

The Likelihood Ratio Test (**LRT**) test statistics can be constructed:

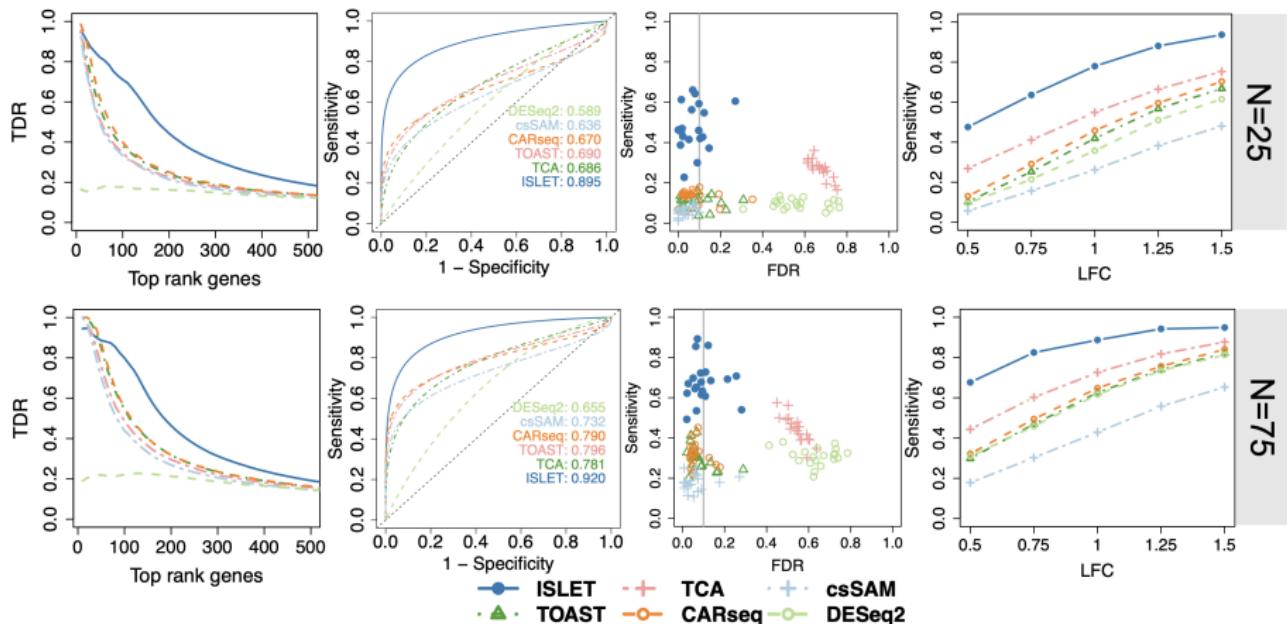
$$\Lambda = 2(l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}) - l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\sigma}}))$$

$\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}$: the EM estimate for the full model.

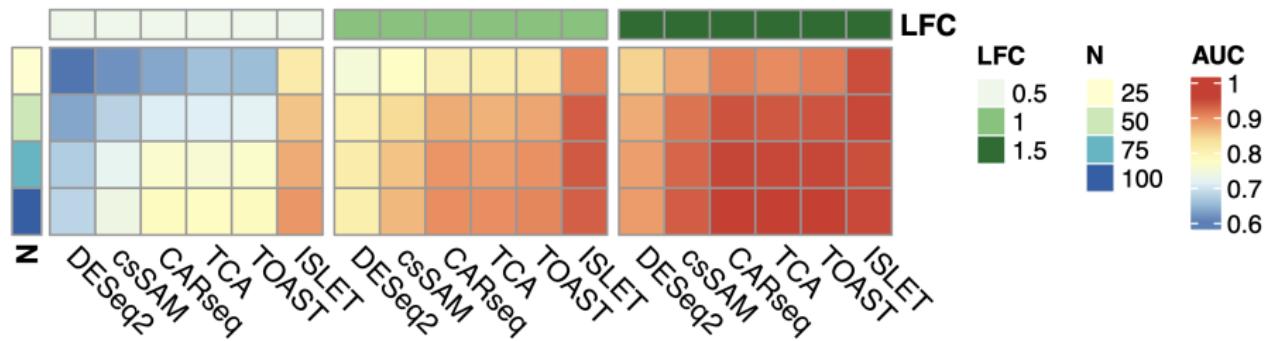
$\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\sigma}}$: the estimate for the reduced model under H_0 .

Ref dist'n: $\lambda \sim \chi_d^2$

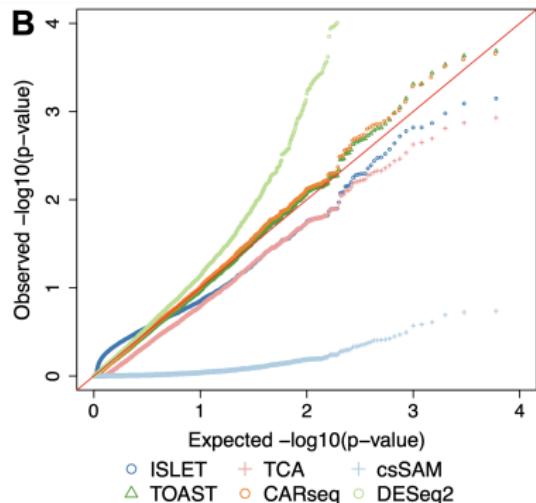
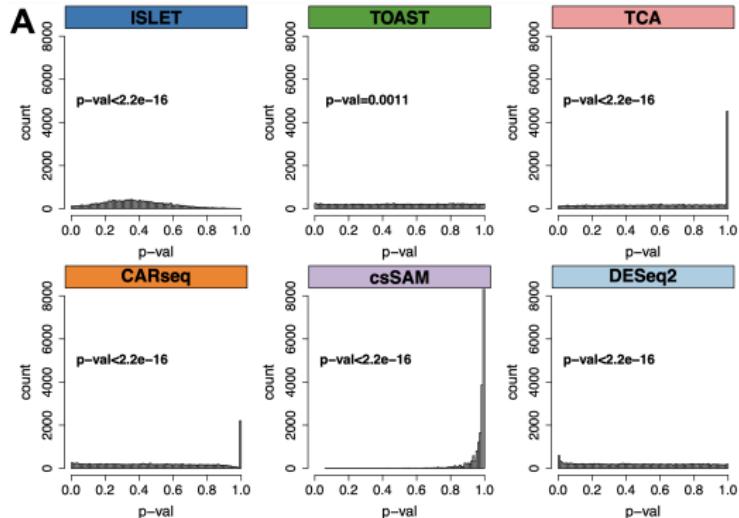
Identifying cell-type-specific DE genes



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 the TEDDY logo with a teddy bear holding a globe. Below it is the tagline "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 navigation 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 three sections: 1) "What is Type-1 Diabetes?" with a photo of a smiling girl and text about insulin and diabetes; 2) "What is the TEDDY Study?" with a photo of a baby crawling and text about the study's goal; 3) a summary section with text about diabetes prevention.

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.

Children with type 1 diabetes must take insulin several times a day to stay alive and healthy. Right now, there is no cure for type 1 diabetes.

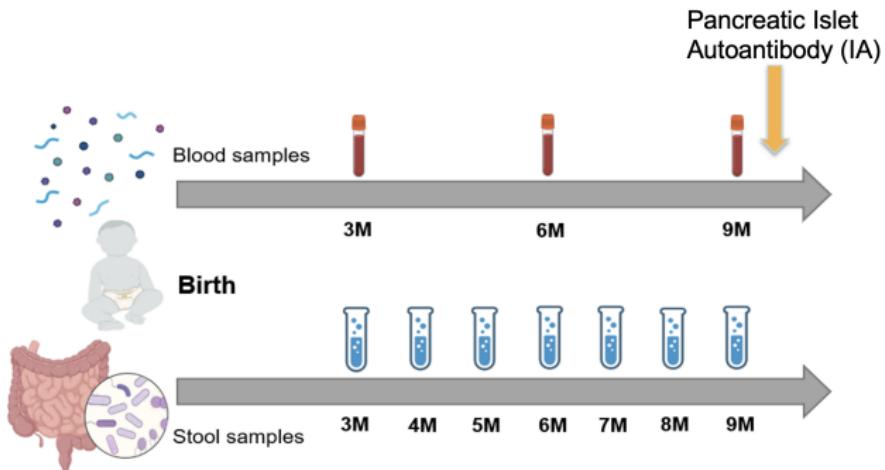
- T1D is a serious disease affecting 1 out of every 300 (1/300) children in the United States.
- T1D occurs when special cells in the pancreas, called beta cells, are destroyed by the body's own immune system. When

What is the TEDDY Study?

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

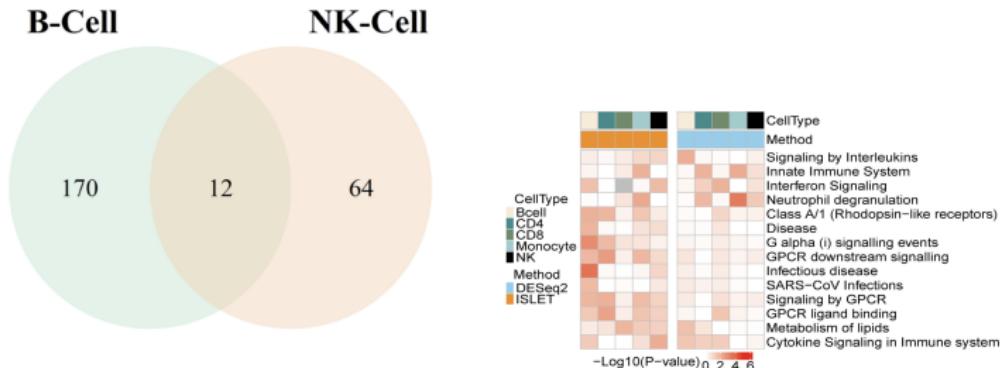
The TEDDY study - The Environmental Determinants of Diabetes in the Young - is looking for the causes of type 1 diabetes mellitus (T1DM). T1DM used to be called childhood diabetes or insulin-dependent diabetes.

TEDDY Data



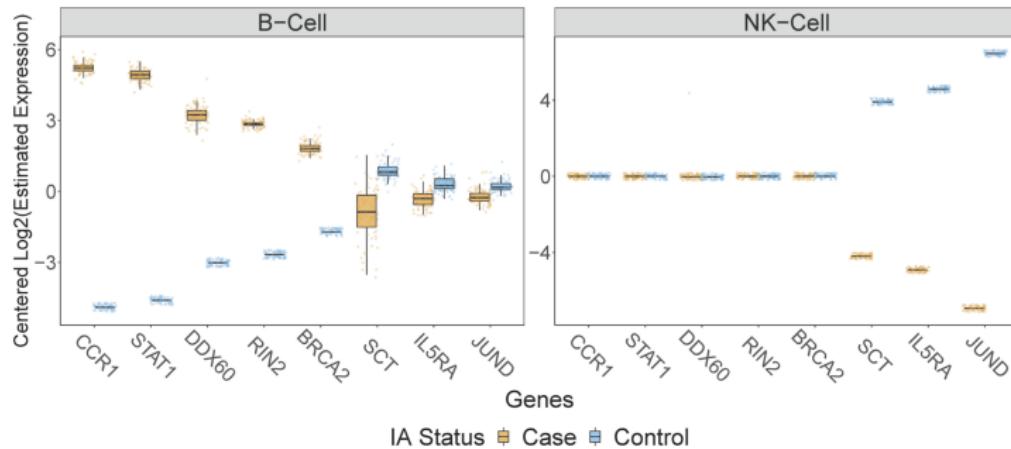
Six clinical centers in four countries: U.S., Finland, Germany, and Sweden.
Prospective cohort: 8,676 high-risk infants were enrolled from birth and followed every 3 months, for blood sample collection and islet autoantibody (IAbs) measurement.
Developing IA (cases) started at 9 months with a plateau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

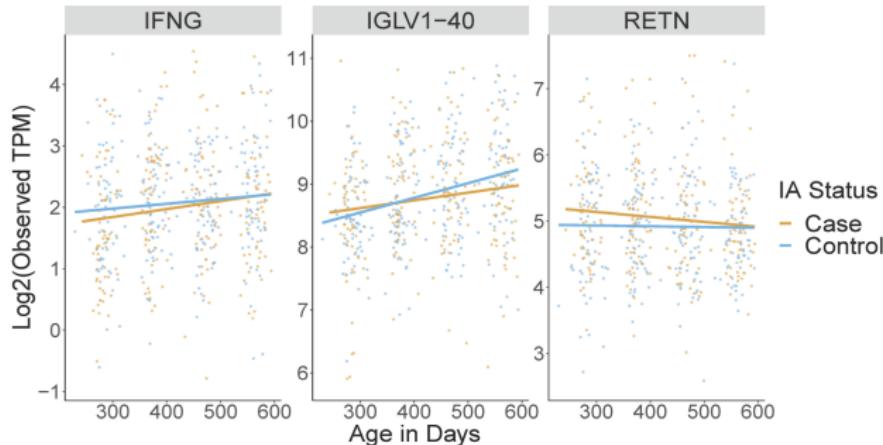


IFN- β -inducible genes in vitro and IFN-indicible transcriptional signatures in human PBMCs were also found prior to the development of autoantibodies (BABY DIET).

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

ISLET: Individual-Specific CeLI TypE Referencing Tool

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



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

Hao Feng* and Qian Li*

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

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

hxlf155@case.edu

9 September 2022

Abstract

This vignette introduces the usage of the Bioconductor package ISLET (Individual-Specific ceLI 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

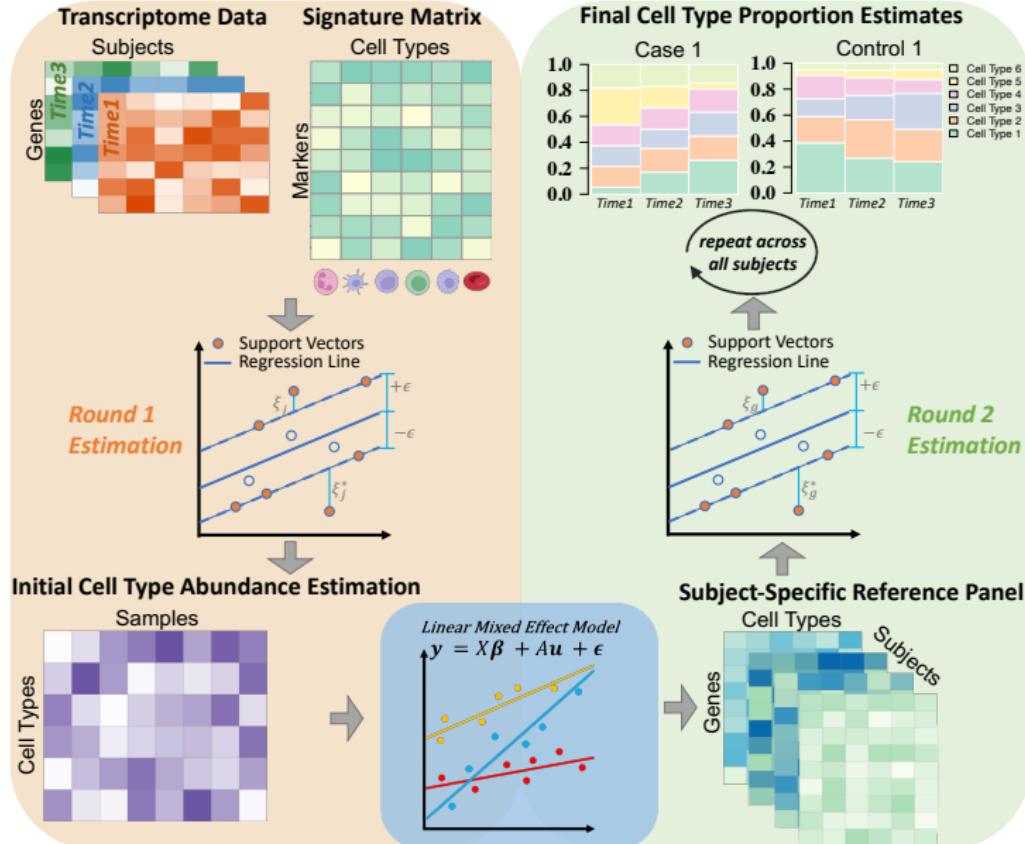
1.2 How to get help

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 "3.16". A "Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS" logo is present. The main content area includes a brief description of the package, its platforms (all), rank (2146 / 2151), support (0 / 0), and dependencies (25). It also shows the DOI (10.18129/B9.bioc.ISLET) and a note that this is the development version of ISLET. Below this, there is a section titled "Individual-Specific ceLI typE referencing Tool" with a detailed description of the package's purpose and methodology.

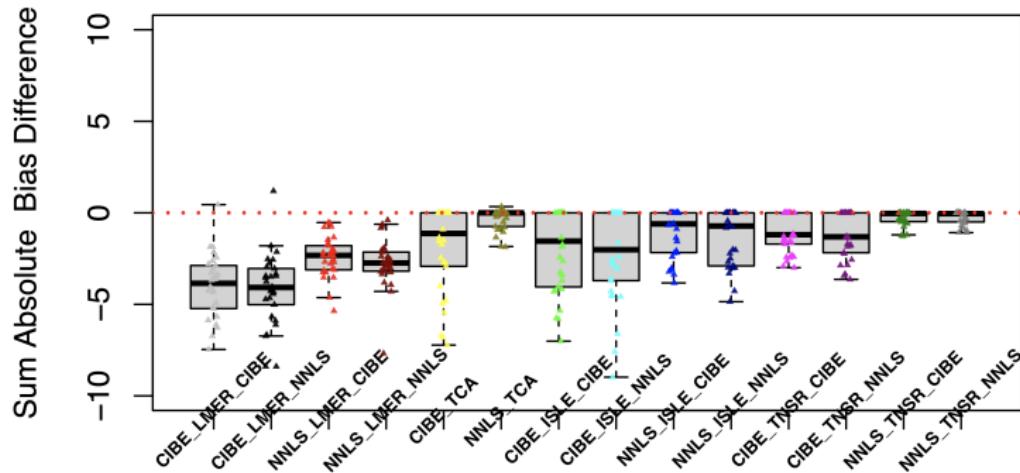
Personalized (better) reference panel \Rightarrow More accurate deconvolution?

- ① Personalized and cell-type-specific transcriptome reference panel recovery.
- ② Improved deconvolution using personalized reference panel.
- ③ A hybrid neural network model for scRNA-seq cell type prediction.

Model for improved deconvolution



Simulation: improved estimation



Initial marker genes:300

	CIBE_LMER_CIBE	CIBE_LMER_NNLS	NNL_S_LMER_CIBE	NNL_S_LMER_NNLS	CIBE_TCA	NNL_S_TCA	CIBE_ISLE_CIBE
err:0-5% N:25 LFC:0.5	-3.51(1.56)	<u>-3.95(1.92)</u>	-1.68(1.3)	-1.8(1.42)	-2.21(1.96)	-0.33(0.85)	1.21(20.15)
err:0-5% N:25 LFC:0.75	<u>-5.2(7)</u>	-4.18(1.95)	-2.41(1.1)	-2.72(1.32)	-1.93(2.39)	-0.34(0.58)	13.82(40.99)
err:10%-20% N:25 LFC:0.5	<u>-2.83(1.34)</u>	-3.27(1.64)	-1.69(1.07)	-1.66(1.16)	-0.1(4.84)	0.65(3.79)	13.74(41.37)
err:10%-20% N:25 LFC:0.75	<u>-3.13(1.76)</u>	-3.3(2.26)	-2.21(0.88)	-2.4(1.23)	0.19(3.09)	1.5(2.9)	9.59(35.68)
err:0-5% N:50 LFC:0.5	-7.97(3.44)	<u>-8.97(3.86)</u>	-2.81(4.52)	-3.7(2.94)	-6.6(4.92)	-0.8(2.2)	5.76(52.49)
err:0-5% N:50 LFC:0.75	<u>-10.77(12.7)</u>	-9.28(4.97)	-4.66(1.82)	-5.3(2.26)	-4.68(5.35)	-0.9(1.66)	9.6(60)
err:10%-20% N:50 LFC:0.5	-6.14(2.27)	<u>-6.48(2.86)</u>	-3.28(1.79)	-3.11(1.95)	1.63(15.12)	2.07(10.22)	2.26(38.29)
err:10%-20% N:50 LFC:0.75	-6.59(3.1)	<u>-7.05(4.28)</u>	-4.51(1.69)	-4.8(2.08)	0.87(5.21)	2.24(7.06)	21.41(77.72)

Real data analysis: PDBP



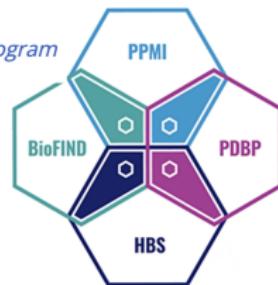
PDBP



National Institute of
Neurological Disorders
and Stroke

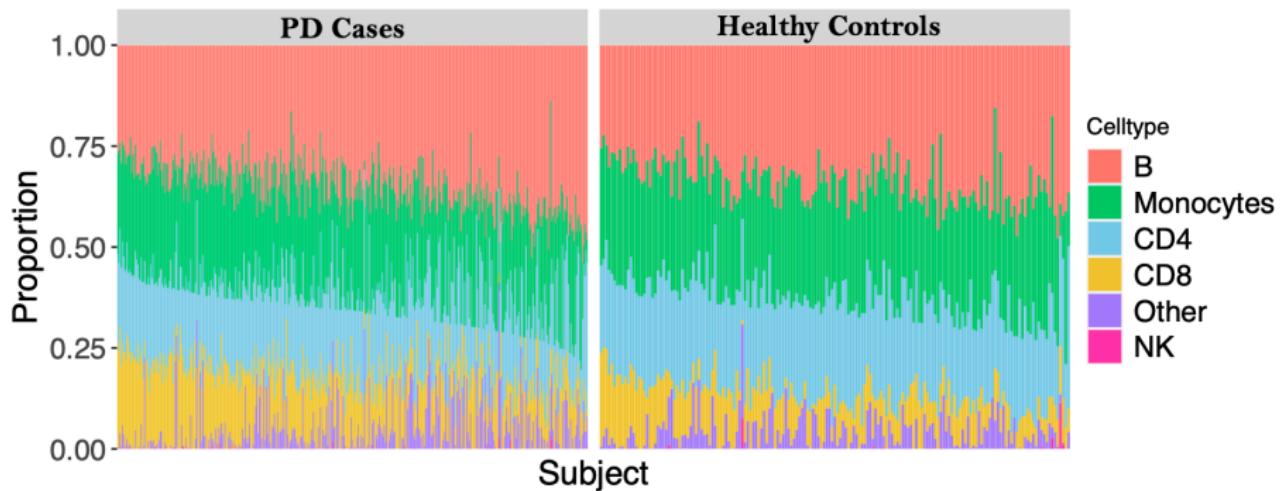


an Accelerating Medicines Partnership® (AMP®) program



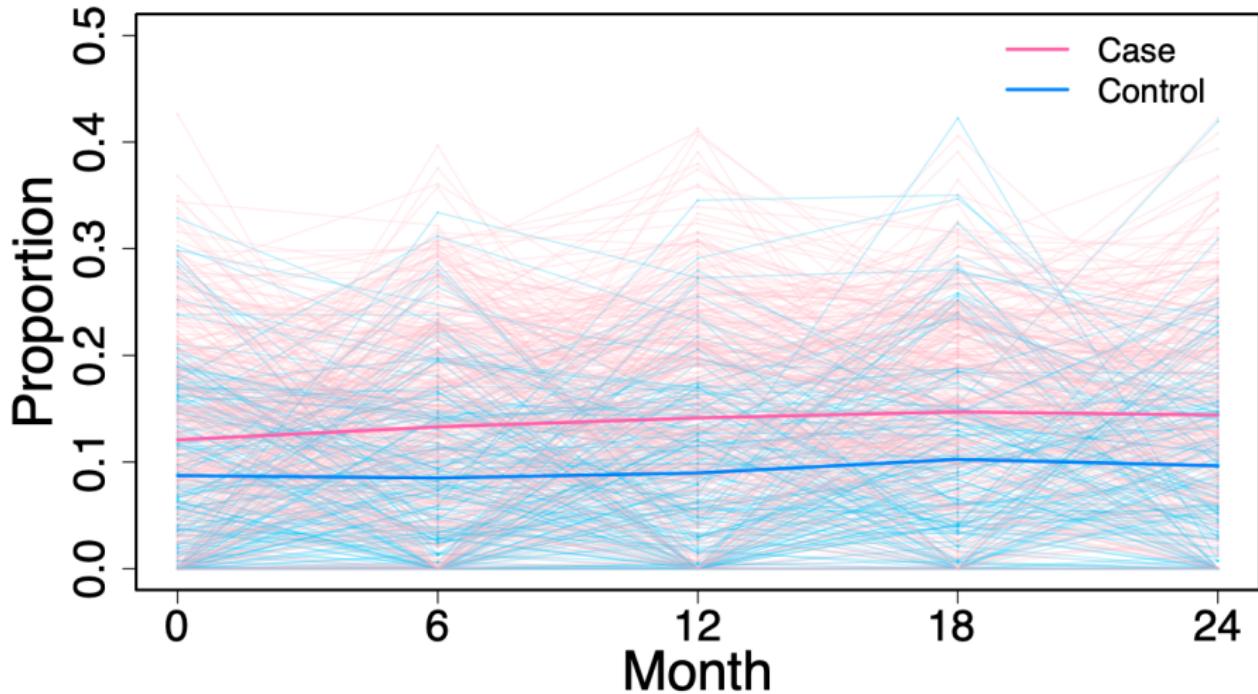
- 399 PD subjects and 173 controls
- 2,599 longitudinal samples over 2 years
- Whole blood transcriptome

Real data analysis: PDBP



Real data analysis: PDBP

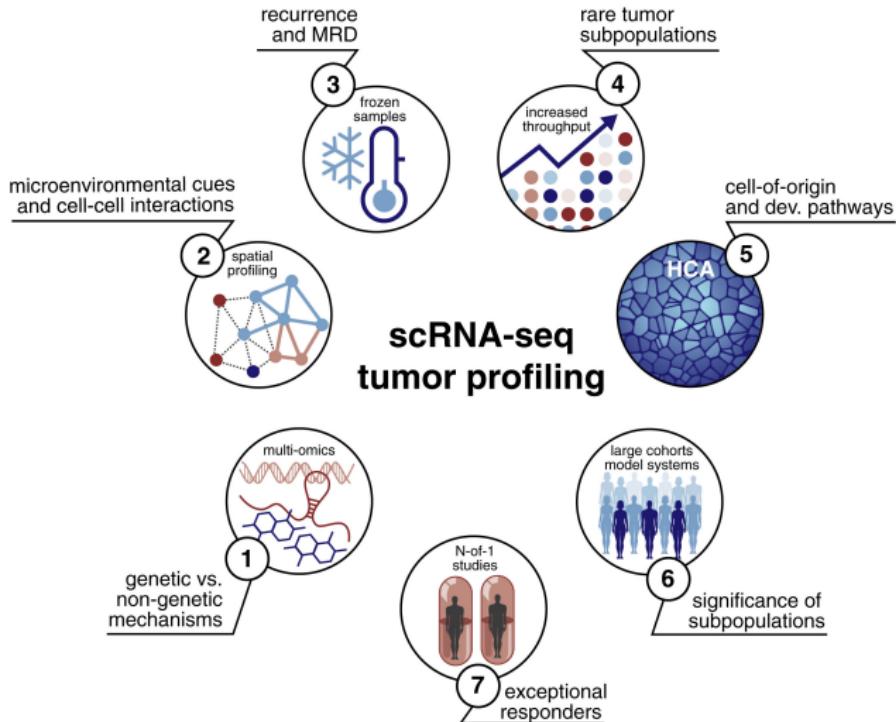
CD8



interim summary for personalized deconvolution

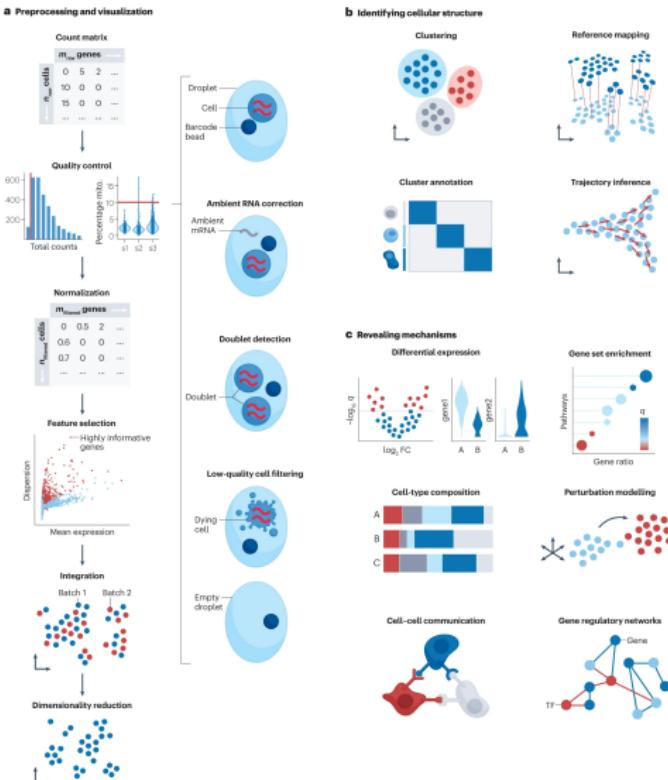
- Statistical frameworks to recover individual-specific and cell-type-specific reference panel.
- Optimize the usage of longitudinally measured samples within the same subject.
- Accurate, robust and powerful performance.
- A R/Bioconductor package ISLET.
- Our model successfully identified gene signatures in B, NK and CD4+ T cells, prior to the onset of pancreatic β -cell autoantibody.
- Our model deconvoluted cell type proportions aligned with low-throughput experiments in PD.

scRNA-seq questions in cancer biology



Suvà, M. L., Tirosh, I. (2019). Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Molecular cell*, 75(1), 7-12.

scRNA-seq data analysis



• Data preprocessing

- Normalization
- Batch effect correction
- Imputation

• Data analyses

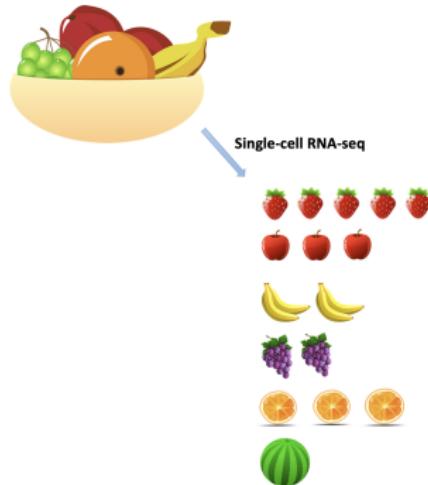
- Cell clustering
- **Cell type identification**
- Differential expression
- Pseudo-time construction
- Rare cell type discovery;
- alternative splicing; allele specific expression
- RNA velocity

• Visualization

- t-SNE, UMAP

Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet (2023).

- ① Personalized and cell-type-specific transcriptome reference panel recovery.
- ② Improved deconvolution using personalized reference panel.
- ③ A hybrid neural network model for scRNA-seq cell type prediction.

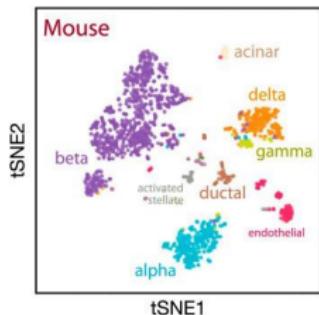
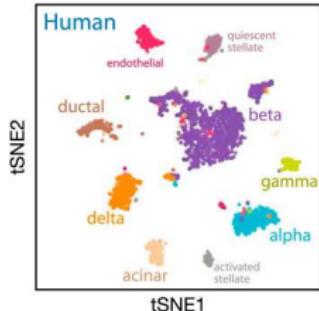


scRNA-seq data

```
> Baron_counts[1:20,1:9]
   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            3088
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               0              2              0              0               0
PRSS2           0               0              0              0              0               0
CTR82           0               0              0              0              0               0
REG1B           0               0              0              0              0               0
SPINK1          0               0              0              3              0               0
SERPINA1        89              26            362             20              5            13
SERPINA3        0               0              0              0              0               0
EEF1A1          44              21              66             52             43            30
OLF4M           0               0              0              1              0               0
GNAS            71              15            103             90             67            49
FTL             14               5              24              5              12              3
CTR81           0               0              0              0              0               0
TMSB4X          7                5              5              6              4               5
```

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

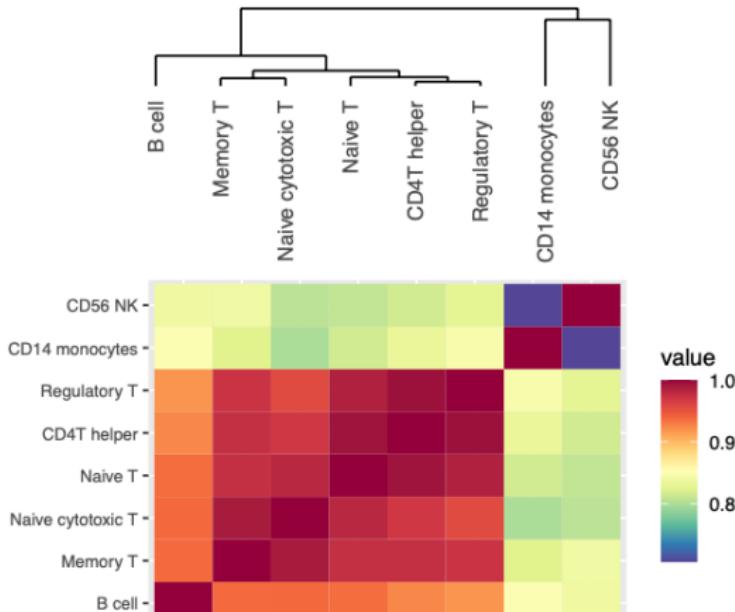
Cell type identification methods

- 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 aims

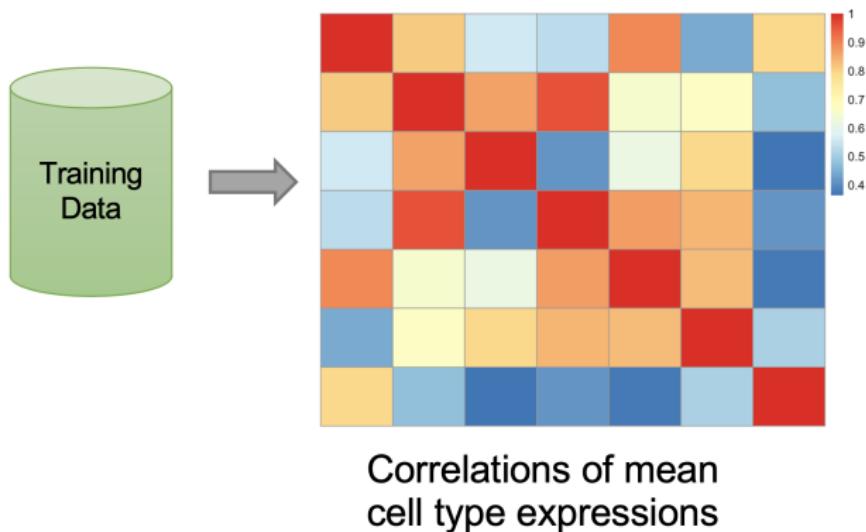
- **Exhaustive**: Annotate all cells.
- **Reliable**: Leverage on massive amount of well-studied existing scRNA data.
- **Flexible**: Adopt different prediction strategies, depending on correlation.

Correlations of cell types



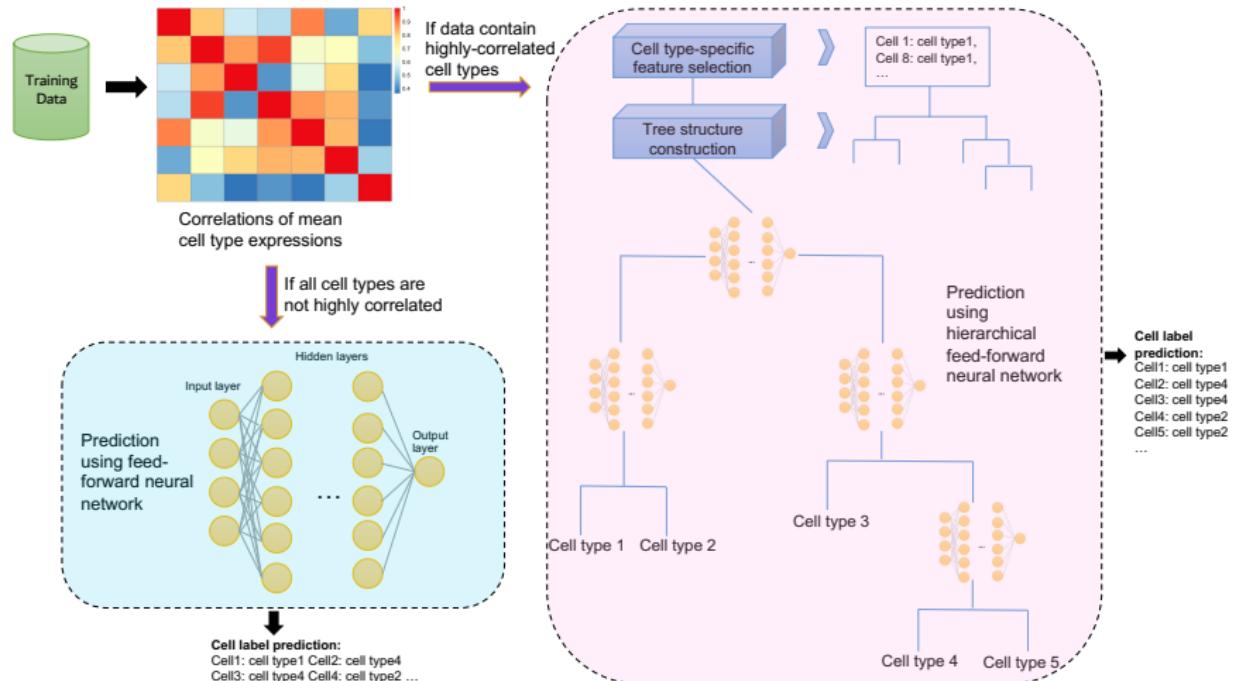
High-correlation cell types pose major challenges.

Tackle the issue of high-correlation



How about adopting a flexible approach?

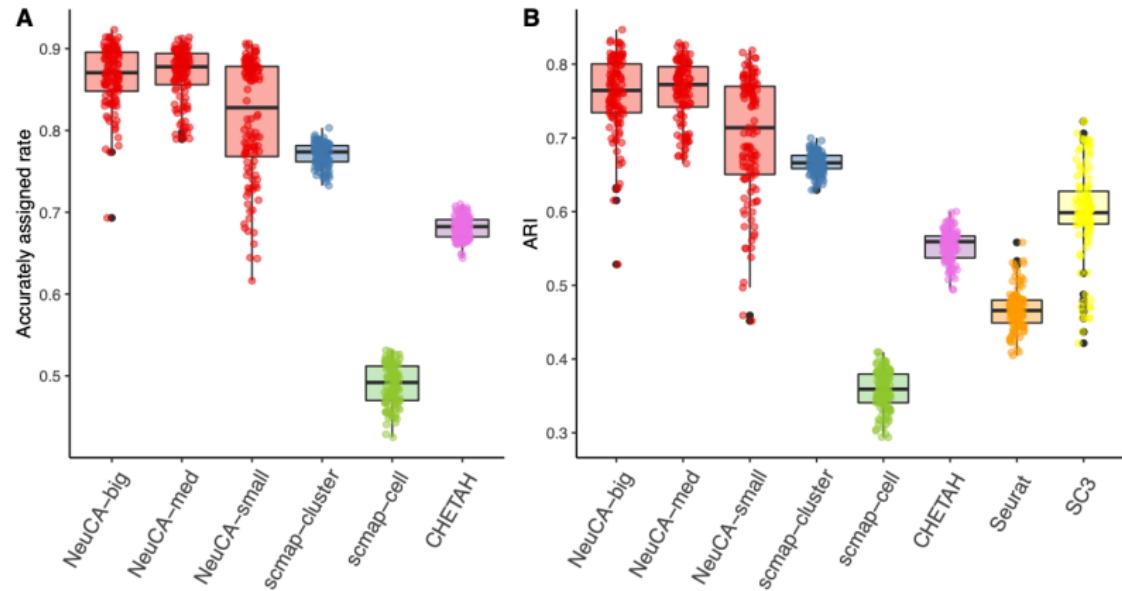
Our method: a neural network-based cell annotation



Wires inside the cogs

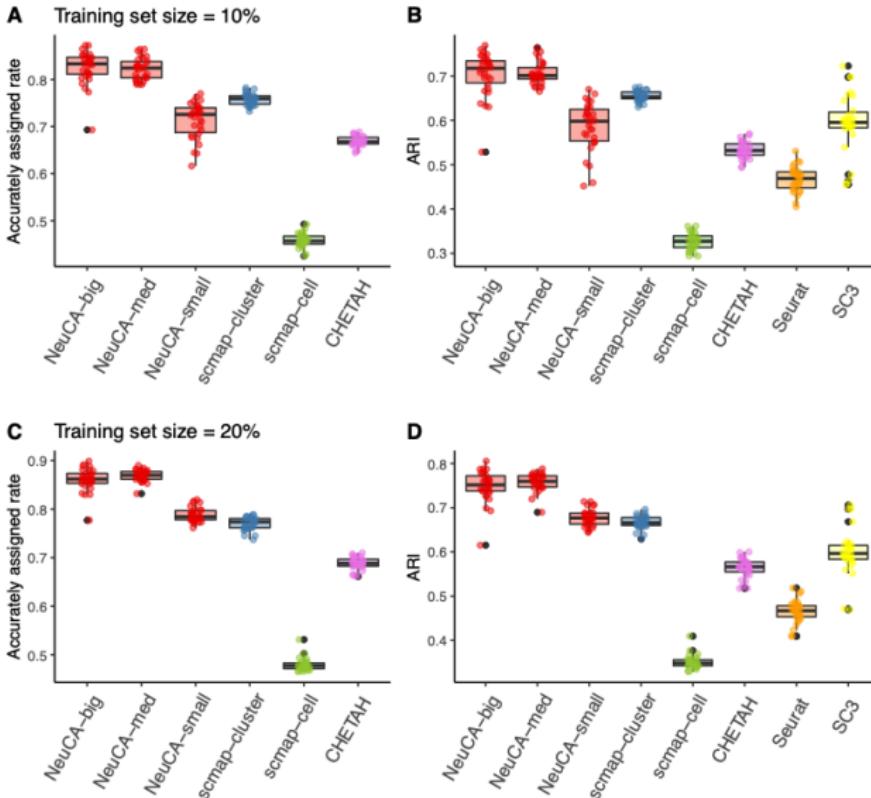
- Route 1: feed-forward neural-network
 - feature selection.
 - 3 model sizes: big, medium, small. (256 to 64 units/nodes)
 - activation function: Rectified Linear Unit (ReLU).
 - output: Softmax.
 - categorical cross-entropy loss.
- Route 2: marker-guided hierarchical neural-network
 - feature selection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

Simulation results: real-data based

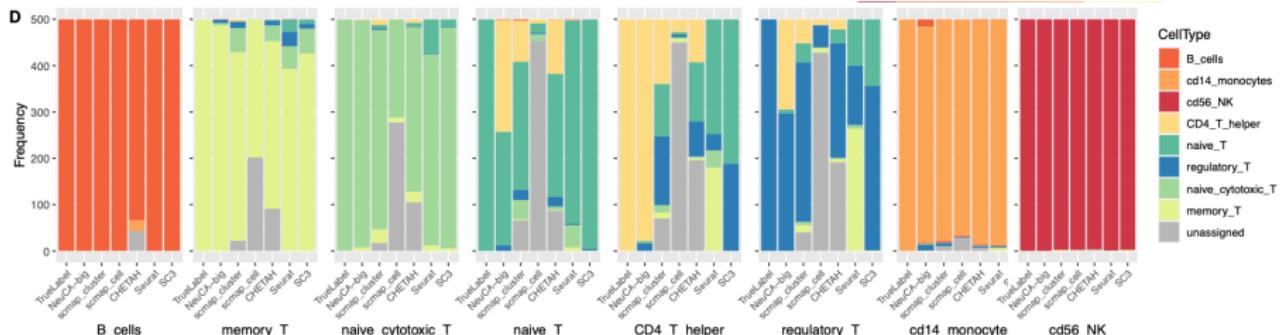
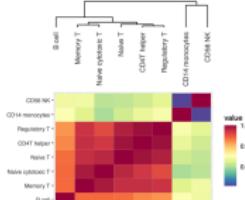


- 10X PBMC scRNA-seq data.
- 80 Monte Carlo simulations are conducted and aggregated.
- Training set proportion ranging from 10% to 80%.

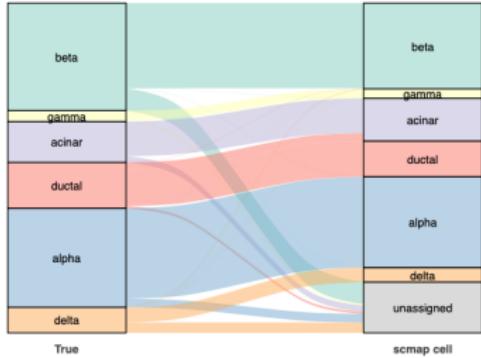
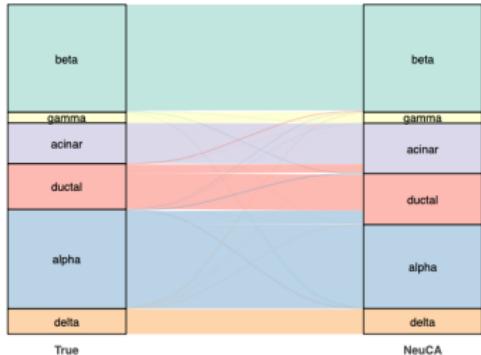
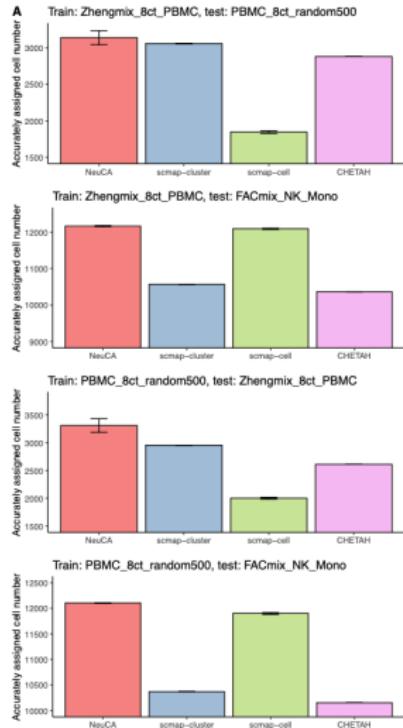
Simulation results



Real data results

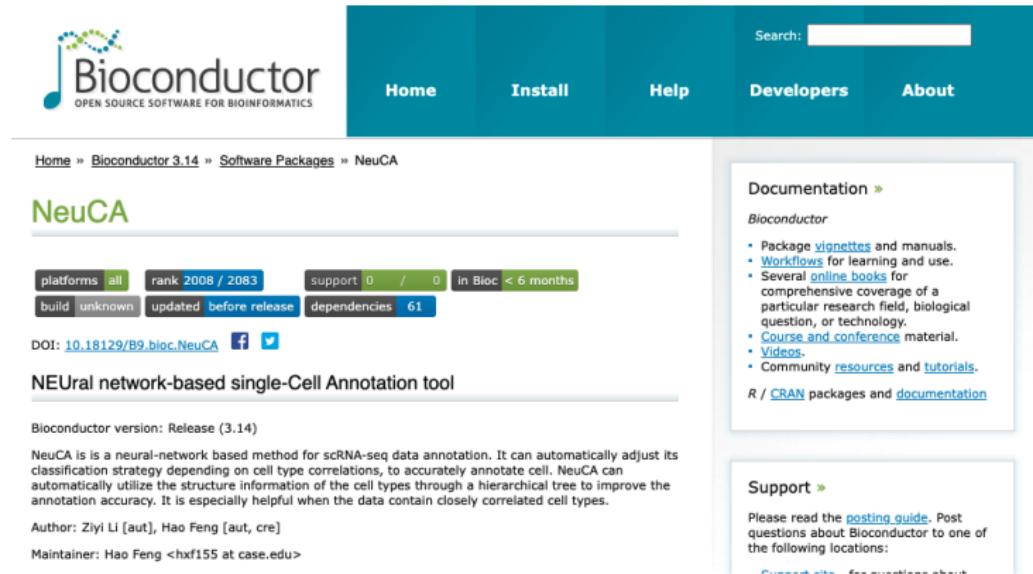


Real data results



Software: NeuCA

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 page title is "NeuCA". The main content area includes a "Documentation" sidebar with links to vignettes, workflows, online books, course material, videos, and community resources. It also mentions R/Cran packages and documentation. The main content area displays package statistics: platforms (all), rank (2008 / 2083), support (0 / 0), build status (unknown), update status (before release), dependencies (61), and a DOI link (10.18129/B9.bioc.NeuCA). There are social media sharing buttons for Facebook and Twitter. Below this, a section titled "NEural network-based single-Cell Annotation tool" provides details about the package version (Release 3.14) and its purpose. It also lists the author (Ziyi Li [aut], Hao Feng [aut, cre]) and maintainer (Hao Feng <hxfl55 at case.edu>).

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: 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 Home Tutorial Run NeuCA FAQ About

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

Choose the data type
Please select an option below

Choose your testing file(.RData/.rda)

Browse... No file selected

What are these data?

Choose the model size
small

Generate Predicted Labels

Navigation icons: back, forward, search, etc.

Publication

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



orts

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

NeuCA web server: doi.org/10.1093/bioinformatics/btac108
NeuCA methodology: doi.org/10.1038/s41598-021-04473-4

... but so what?

scRNA-tools tsunami

<https://www.scrna-tools.org/>

Zappia and Theis *Genome Biology* (2021) 22:301
<https://doi.org/10.1186/s13059-021-02519-4>

Genome Biology

REVIEW

Open Access

Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape



Luke Zappia^{1,2} and Fabian J. Theis^{1,2,3*} 

scRNA-tools tsunami

<https://www.scrna-tools.org/>

Zappia and Theis *Genome Biology* (2021) 22:301
<https://doi.org/10.1186/s13059-021-02519-4>

Genome Biology

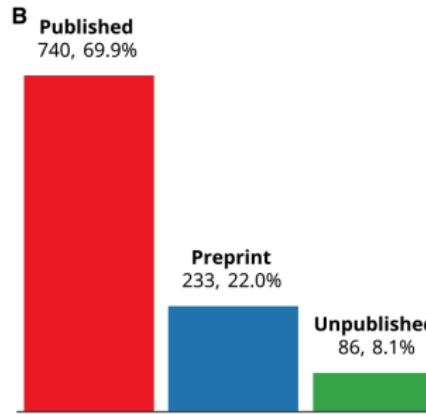
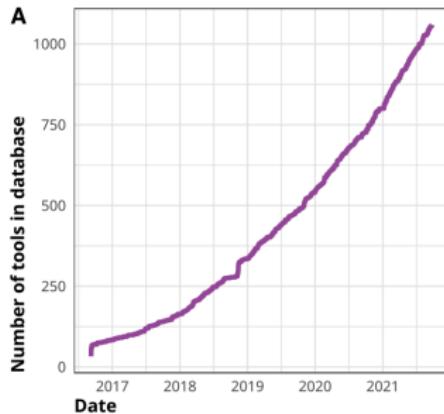
REVIEW

Open Access

Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape



Luke Zappia^{1,2} and Fabian J. Theis^{1,2,3*}



The grand trajectory

"the evolution of the field and a change of focus from ordering cells on continuous trajectories to integrating multiple samples and making use of reference datasets"

"open science practices reward developers with increased recognition and help accelerate the field"

Acknowledgement



- Leslie Meng
- Daoyu Duan
- Wen Tang



- Qian Li



Making Cancer History®

- Ziyi Li



CASE
COMPREHENSIVE
CANCER CENTER

➤ IRG-16-186-21



@HHarryFeng



<https://hfenglab.org/>