DCATS: Differential Composition Analysis of Single-Cell data

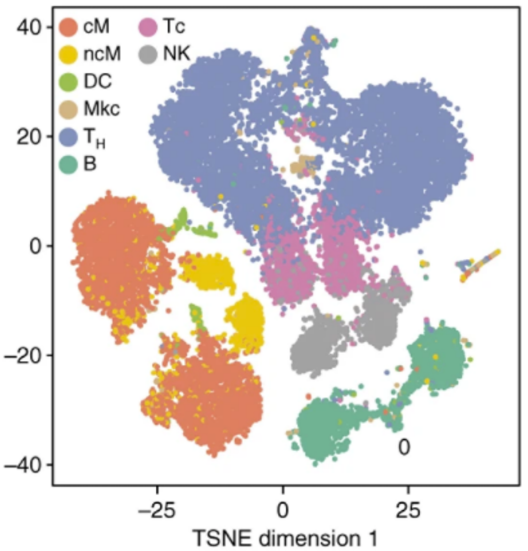
Xinyi Lin

The University of Hong Kong, Ho Lab

2021/11/01

Motivation

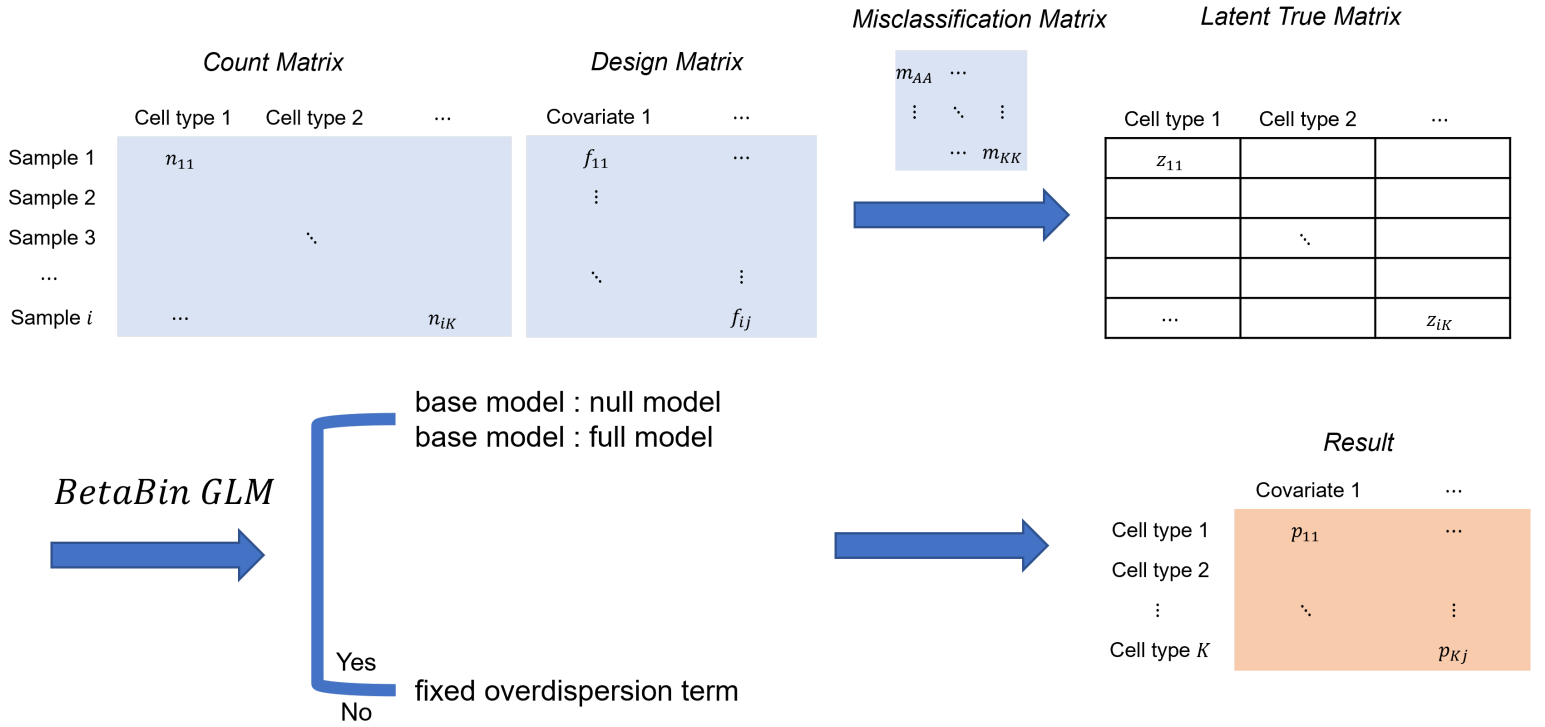
- R package designed for differential composition analysis on single cell data
- Basic assumptions
  - Cell counts follow beta-binomial distribution
  - Misclassification error exists[2]



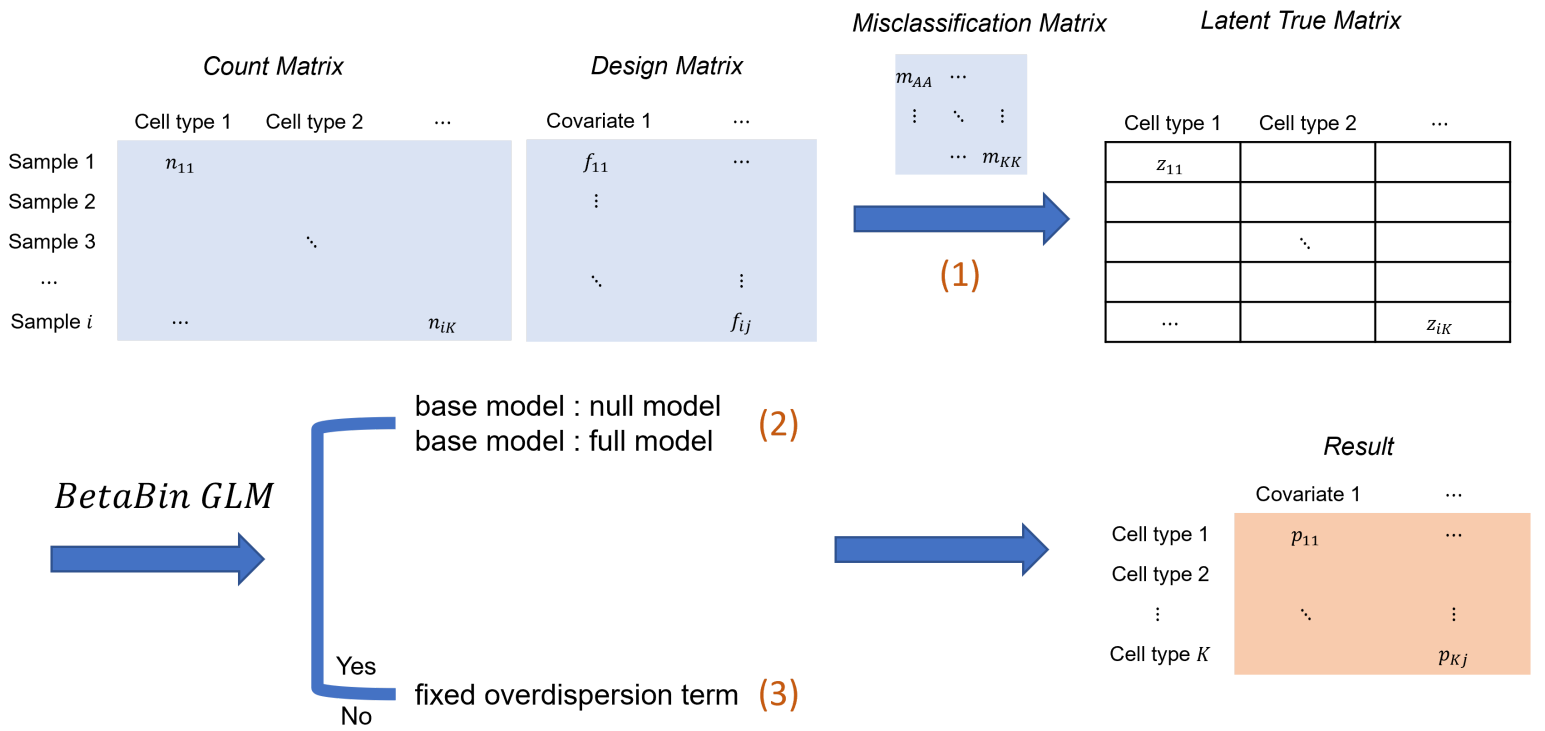
	B	cM	Th	Tc	DC	ncM	Mkc	NK
B	0.921	0.01	0.009	0.002	0.038	0.001	0.024	0.001
cM	0.008	0.829	0.005	0.004	0.042	0.031	0.136	0.003
Th	0.035	0.014	0.946	0.07	0	0.005	0.136	0.011
Tc	0.007	0.004	0.027	0.797	0	0.003	0.021	0.08
DC	0.013	0.027	0.001	0	0.874	0.002	0.006	0
ncM	0.009	0.103	0.003	0.002	0.042	0.956	0.027	0.001
Mkc	0.005	0.007	0.005	0	0.004	0.002	0.614	0.002
NK	0.002	0.006	0.003	0.125	0	0.001	0.036	0.901

cM, CD14+CD16– monocytes; ncM, CD14+CD16+ monocytes; DC, dendritic cells; Mkc, megakaryocytes; Th, CD4+ T cells; B, B cells; Tc, CD8+ T cells; NK, natural killer cells

Workflow


Count Matrix
Design Matrix
Misclassification Matrix
Latent True Matrix
Cell type 1
Cell type 2
Covariate 1
Sample 1
Sample 2
Sample 3
Sample i
n_{11}
n_{iK}
f_{11}
f_{ij}
m_{AA}
m_{KK}
z_{11}
z_{iK}
BetaBin GLM
base model : null model
base model : full model
Yes
No
fixed overdispersion term
Result
Covariate 1
Cell type 1
Cell type 2
Cell type K
p_{11}
p_{Kj}

Workflow


Count Matrix
Design Matrix
Misclassification Matrix
Latent True Matrix
Cell type 1
Cell type 2
Covariate 1
Sample 1
Sample 2
Sample 3
Sample i
n_{11}
n_{iK}
f_{11}
f_{ij}
m_{AA}
m_{KK}
z_{11}
z_{iK}
(1)
BetaBin GLM
base model : null model
base model : full model
(2)
Yes
No
fixed overdispersion term
(3)
Result
Covariate 1
Cell type 1
Cell type 2
Cell type K
p_{11}
p_{Kj}

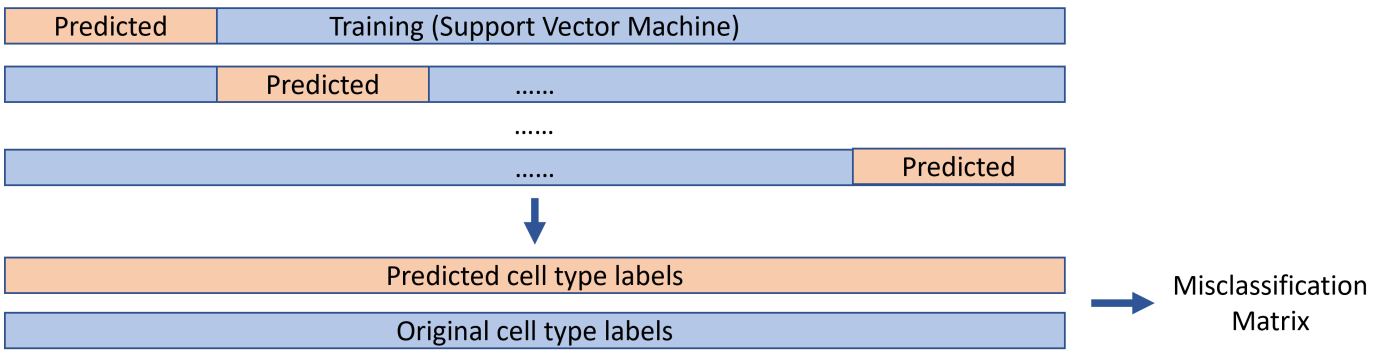
Method: Misclassification Correction

- Uniform type:

$$\begin{bmatrix} a & (1-a)/(K-1) & \dots & (1-a)/(K-1) \\ (1-a)/(K-1) & a & \dots & (1-a)/(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ (1-a)/(K-1) & (1-a)/(K-1) & \dots & a \end{bmatrix}$$

- KNN type: m_{ij} = % of cluster i's neighborhoods $\in$ cluster j

- SVM type:



Method: Beta-binomial GLM

For each cell type:

- Cell count vector

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

- Design matrix

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

- With mean μ of y, a beta-binomial GLM can be fitted as:

$$ln(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 \times \mathbf{x}_1 + \beta_2 \times \mathbf{x}_2 \dots\dots$$

Method: Models Selection

- Type 1: Null model

Model 0: $g(\mu) = \beta_0$

Model 1: $g(\mu) = \beta_0 + \beta_1 * tested\_covariate$

- Type 2: Full model

Model 0: $g(\mu) = \beta_0 + \beta * other\_covariates$

Model 1: $g(\mu) = \beta_0 + \beta * other\_covariates + \beta_i * tested\_covariate$

	age	sex	sample_type	condition
S-HC003	46	M	fresh PBMC	control
S-HC004	34	M	fresh PBMC	control
S-HC005	37	F	fresh PBMC	control
S-HC006	27	M	fresh PBMC	control
S-HC007	27	M	fresh PBMC	control
S-HC008	44	M	fresh PBMC	control

Method: Determine Over-dispersion

- Numbers of cells follow beta-binomial distribution:

$$P(Y = y|n, p) = \binom{n}{y} = p^y(1-p)^{n-y}$$

$$f(p|a, b) = \frac{1}{B(a, b)} p^{a-1}(1-p)^{b-1}$$

$$E(Y|n, \pi, \phi) = n\pi, Var(Y|n, \pi, \phi) = n\pi(1-\pi)[1 + (n-1) \times \phi]$$

- Without fixed over-dispersion term :
  - ϕ is estimated in each beta-binomial GLM for each cell type

- With fixed over-dispersion term :
  - ϕ is estimated across all cell types before testing
  - The estimated ϕ is given in each beta-binomial GLM for each cell type

How to Use DCATS

- **Count Matrix**

```
data("Haber2017")
rbind(Haber2017$count_ctrl, Haber2017$count_Hpoly3)
```

```
##     Endocrine Enterocyte Enterocyte.Progenitor Goblet Stem  TA TA.Ear
## B1         36         59                   136     36  239 125      1
## B2          5         46                    23     20   50  11      
## B3         45         98                   188    124  250 155      3
## B4         26        221                   198     36  131 130      1
## B5         52         75                   347     66  323 263      3
## B6         65        126                   115     33   65  39      1
```

How to Use DCATS

- **Design Matrix**

```
sim_design = data.frame(condition = c(rep("control", 4), rep("Hpoly3",
print(sim_design)
```

```
##   condition
## 1   control
## 2   control
## 3   control
## 4   control
## 5    Hpoly3
## 6    Hpoly3
```

How to Use DCATS

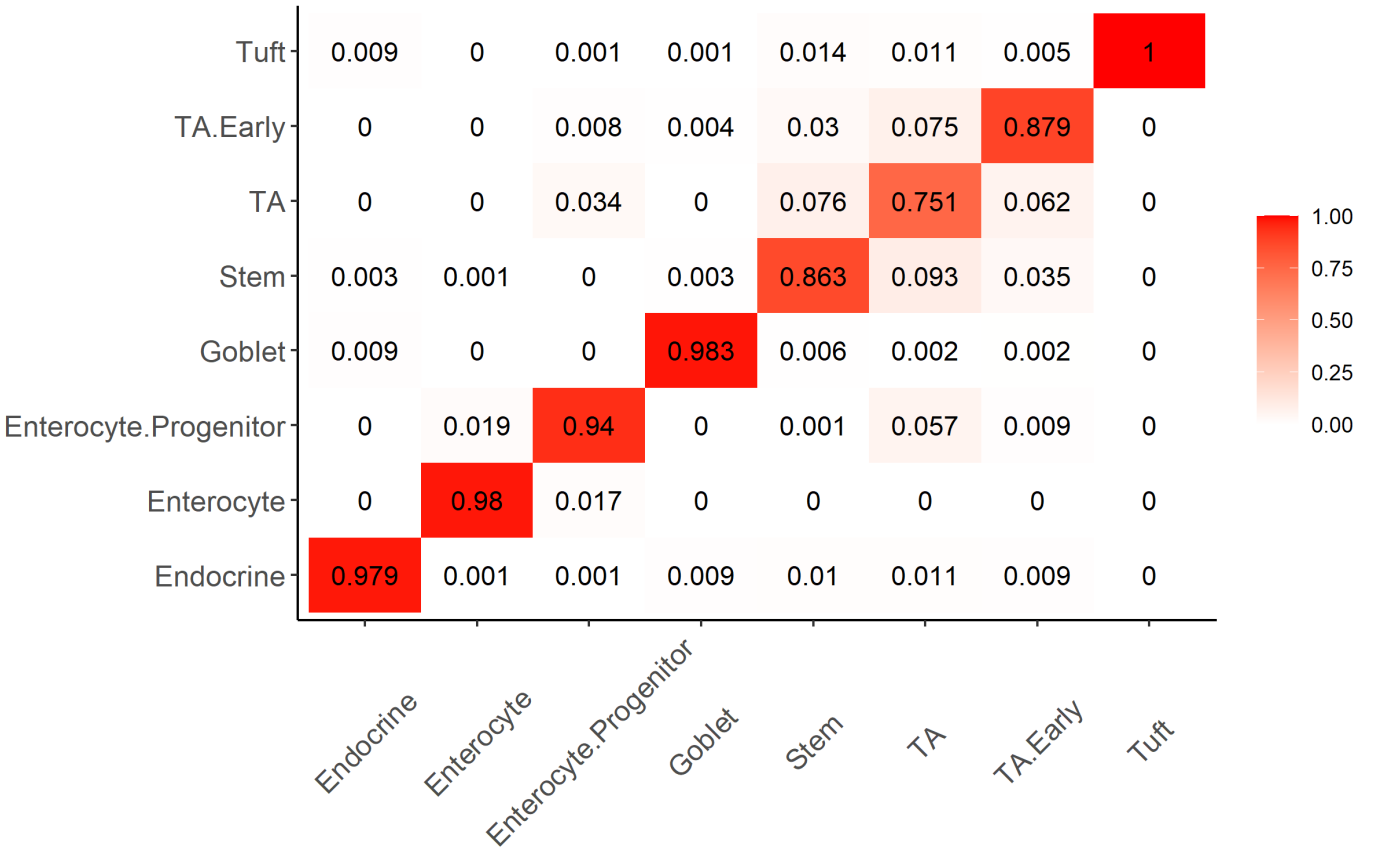
- **Design Matrix**

```
data("Ren2021")
print(head(Ren2021$designM, 10))
```

```
##          age sex sample_type   state
## S-HC003   46   M  fresh PBMC control
## S-HC004   34   M  fresh PBMC control
## S-HC005   37   F  fresh PBMC control
## S-HC006   27   M  fresh PBMC control
## S-HC007   27   M  fresh PBMC control
## S-HC008   44   M  fresh PBMC control
## S-HC009   29   M  fresh PBMC control
## S-HC010   58   M  fresh PBMC control
## S-HC011   35   M  fresh PBMC control
## S-HC012   33   M  fresh PBMC control
```

How to Use DCATS

- **Misclassification Matrix (a K $\times$ K matrix)**



How to Use DCATS

- **Misclassification Matrix**

```
data("Kang2017")
data("simulation")

# Three ways to calculate similarity matrices
## Uniform type
simil_mat = create_simMat(K = 3, confuse_rate = 0.2)

## KNN type
knn_mat = knn_simMat(KNN_matrix = simulation$knnGraphs, clusters = simu

## SVM type
svm_mat = svm_simMat(dataframe = Kang2017$svmDF)
```

How to Use DCATS

- **Main Function**

```
print(sim_count)
```

```
##      [,1] [,2] [,3]
## [1,]   36   35   29
## [2,]  271  279  250
## [3,]  518  379  403
## [4,]  152  220  228
## [5,]   84   87   79
## [6,]  259  203  238
## [7,]  345  376  379
```

```
print(sim_design)
```

```
##   condition gender
## 1        g1 Female
## 2        g1 Female
## 3        g1 Female
## 4        g1 Female
## 5        g2   Male
## 6        g2 Female
## 7        g2 Female
```

```
## null model, flexible phi
res = dcats_GLM(sim_count, sim_design, similarity_mat = simil_mat)
## full model, flexible phi
res = dcats_GLM(sim_count, sim_design, simil_mat, base_model='FULL')
## null model, fixed phi
phi = getPhi(sim_count, sim_design)
res = dcats_GLM(sim_count, sim_design, simil_mat, fix_phi = phi)
```

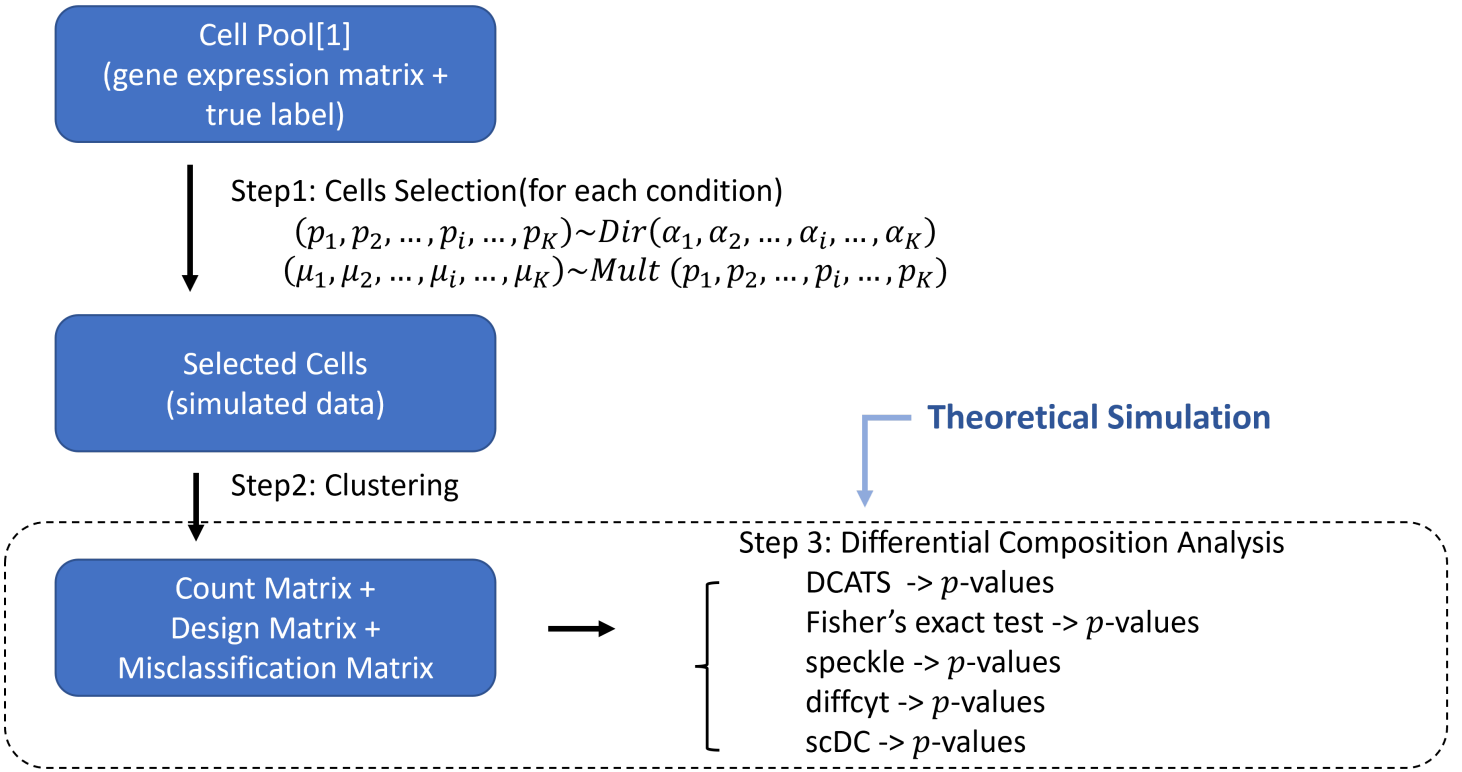
How to Use DCATS

- **Main Function**

```
res = dcats_GLM(sim_count, sim_design, similarity_mat = simil_mat)
print(res$LRT_pvals)
```

```
##              condition    gender
## cell_type_1 0.8939193 0.9903591
## cell_type_2 0.6906093 0.6115926
## cell_type_3 0.5738177 0.6273852
```

Method: Simulation Design

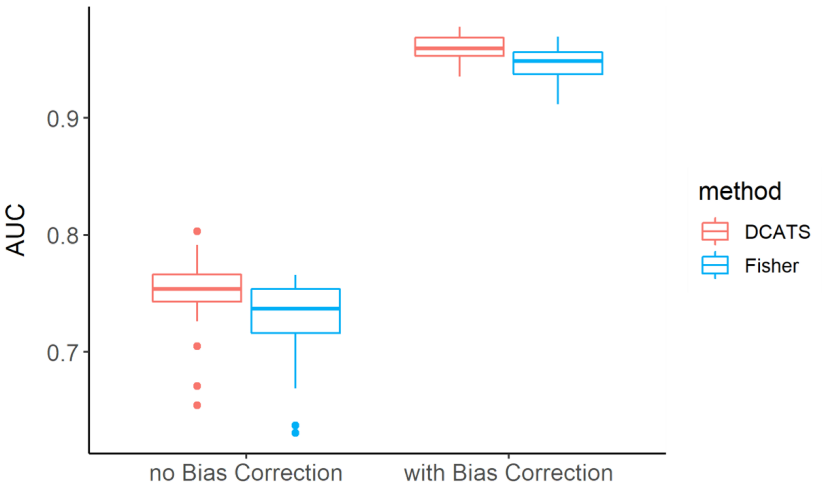


Results: Simulation

- **Theoretical Simulation**

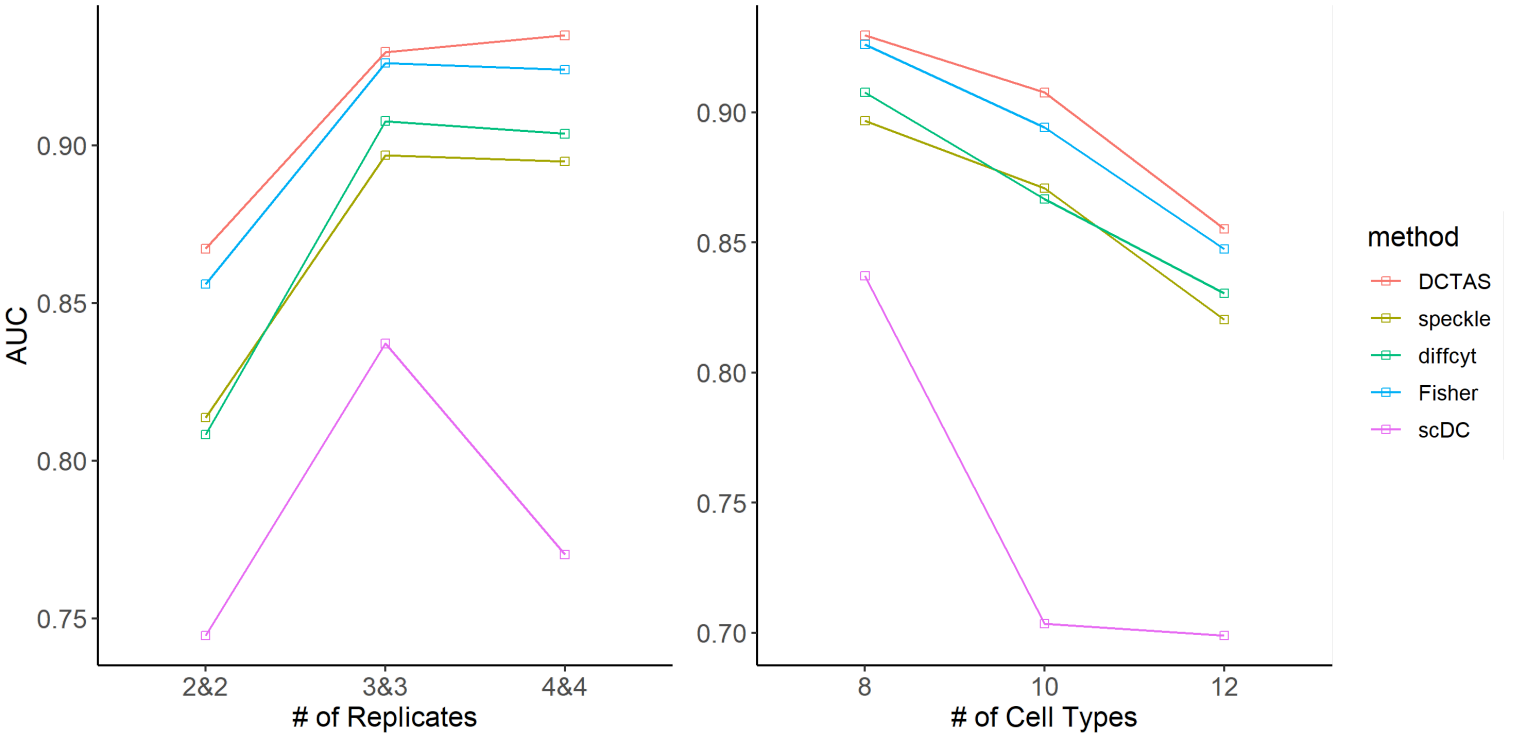
similarity/misclassification matrix

	Cell Type 1	Cell Type 2	Cell Type 3
Cell Type 1	1.0	0.0	0.0
Cell Type 2	0.0	0.7	0.3
Cell Type 3	0.0	0.3	0.7



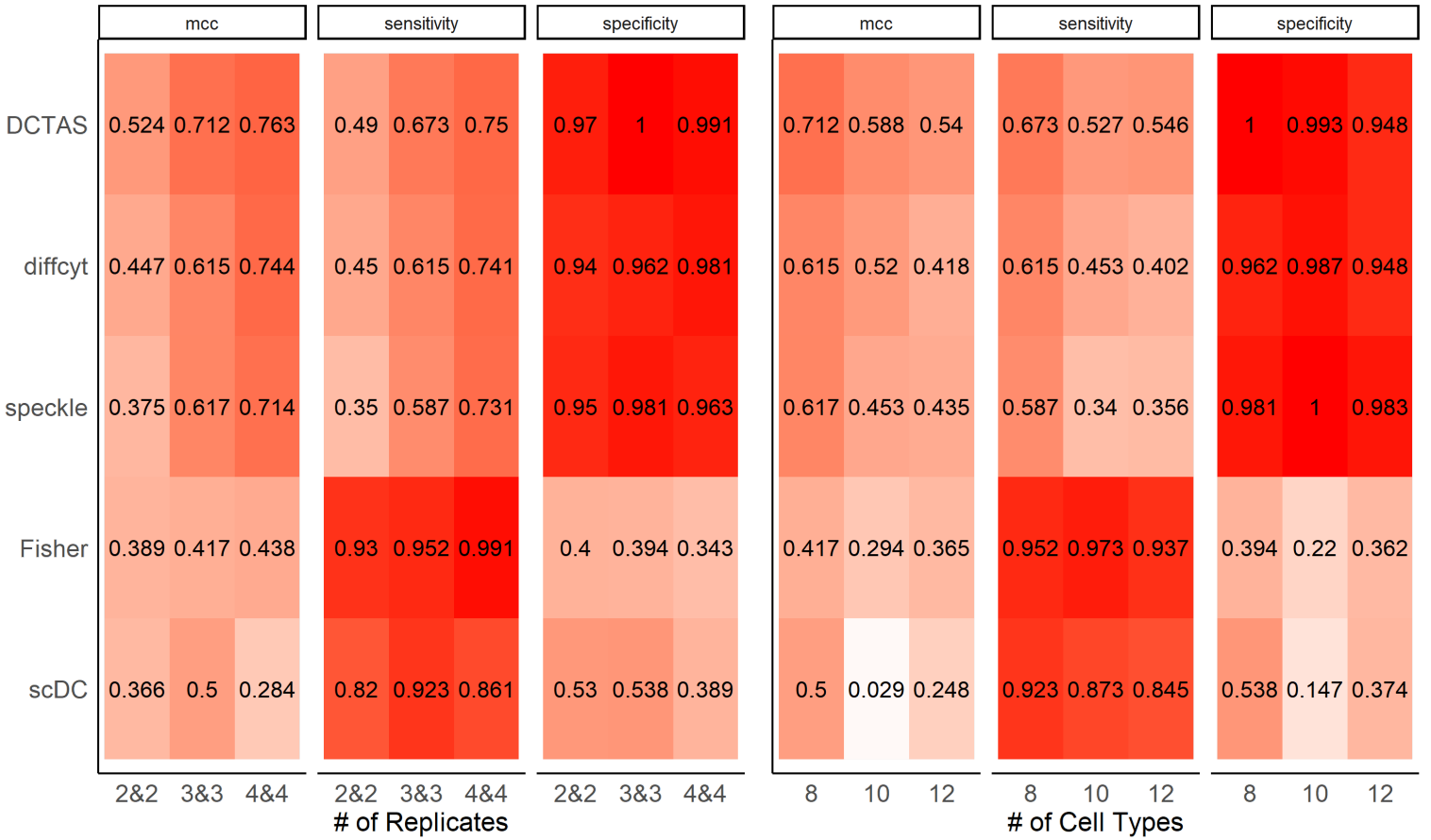
Results: Simulation

- **Simulation with Expression Info**



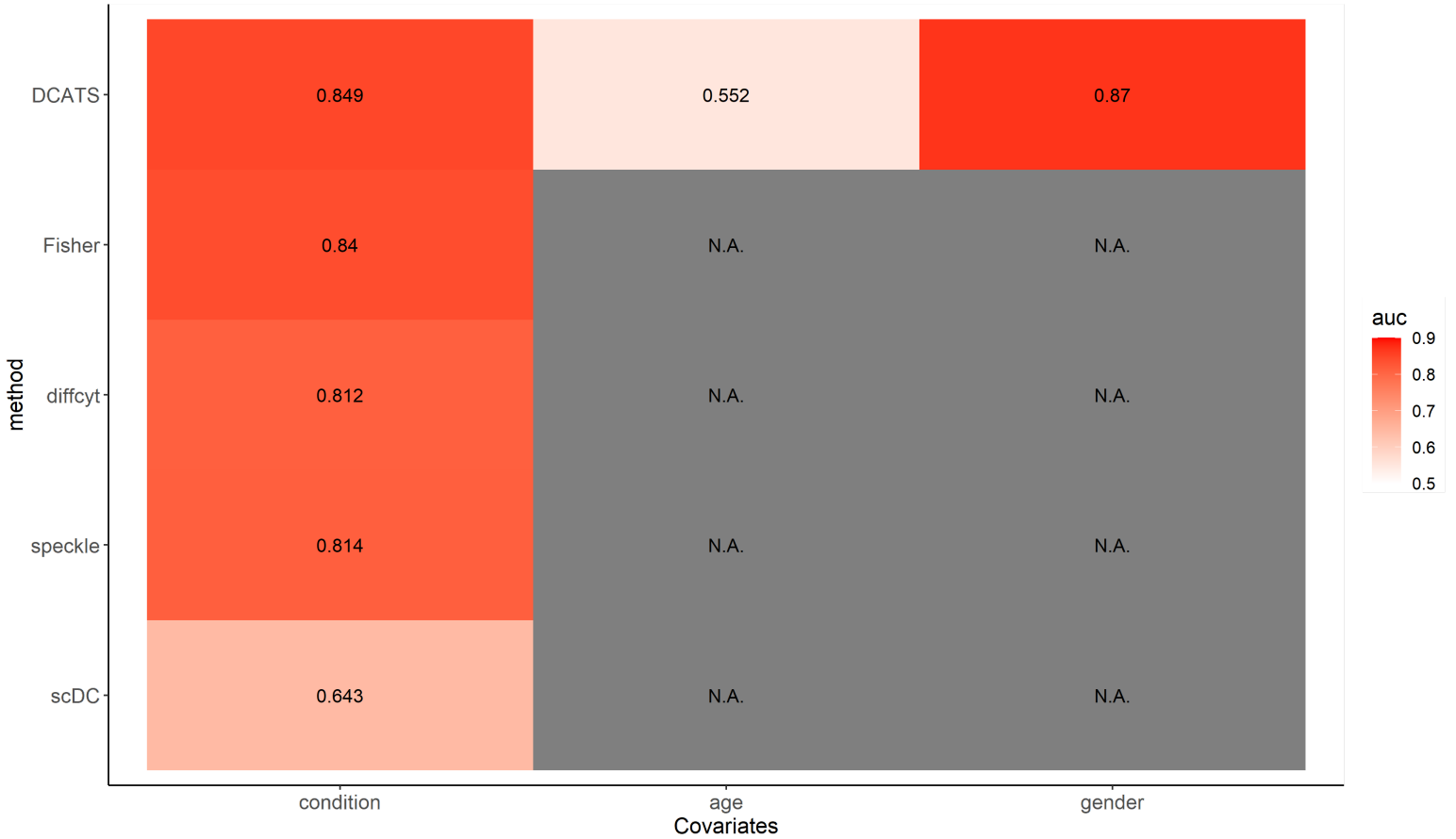
Results: Simulation

- **Simulation with Expression Info**

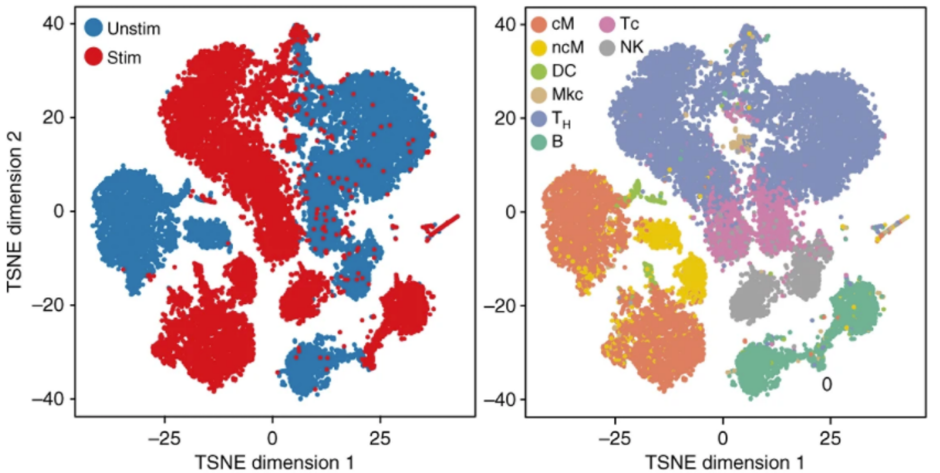


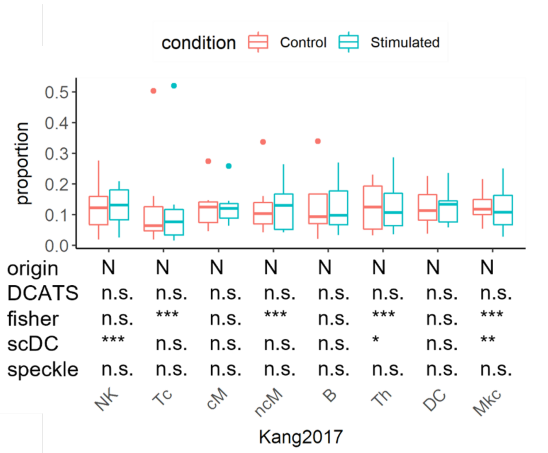
Results: Simulation

- **Simulation with Covariates**



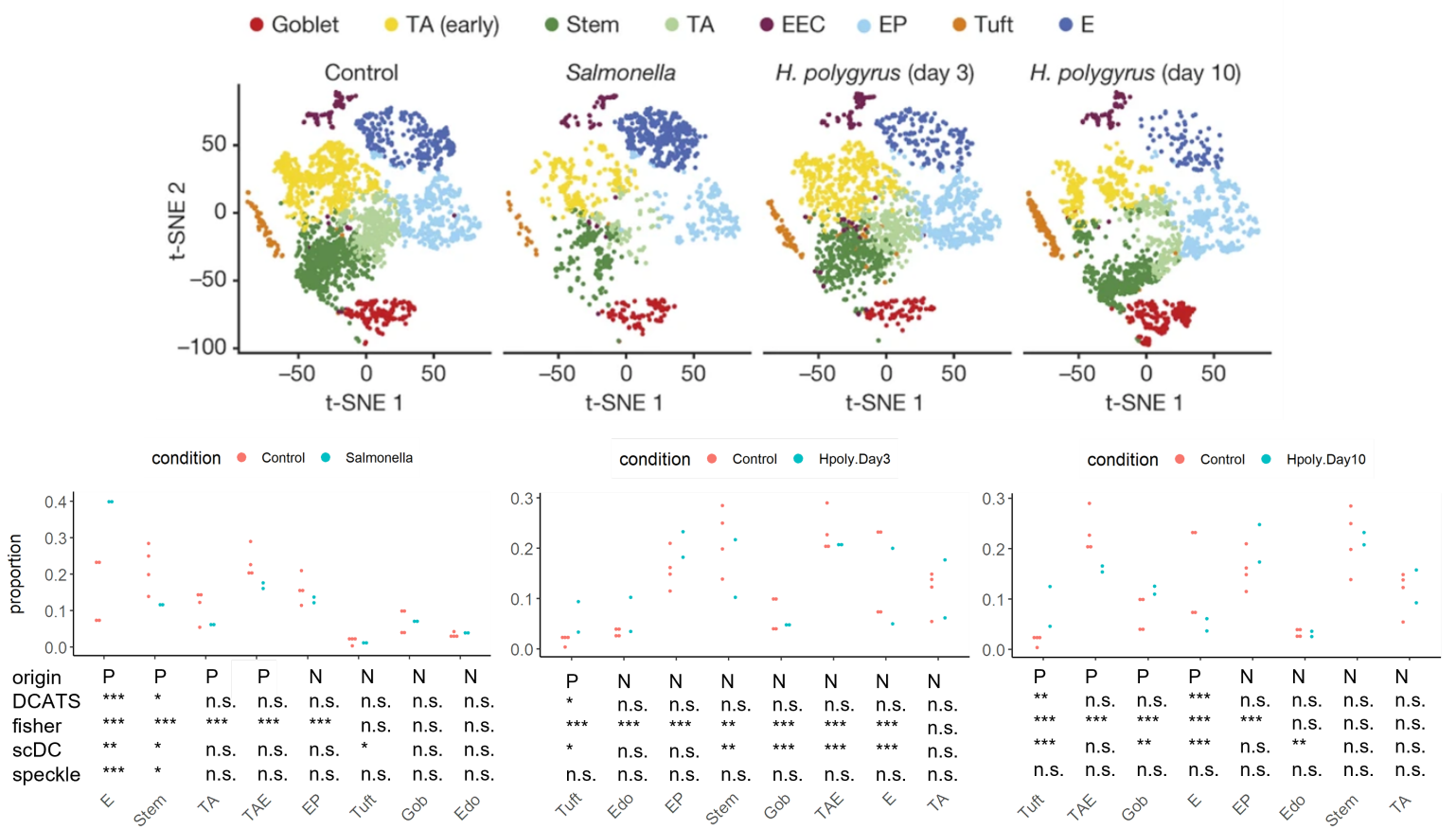
Results: Real World Data [2]





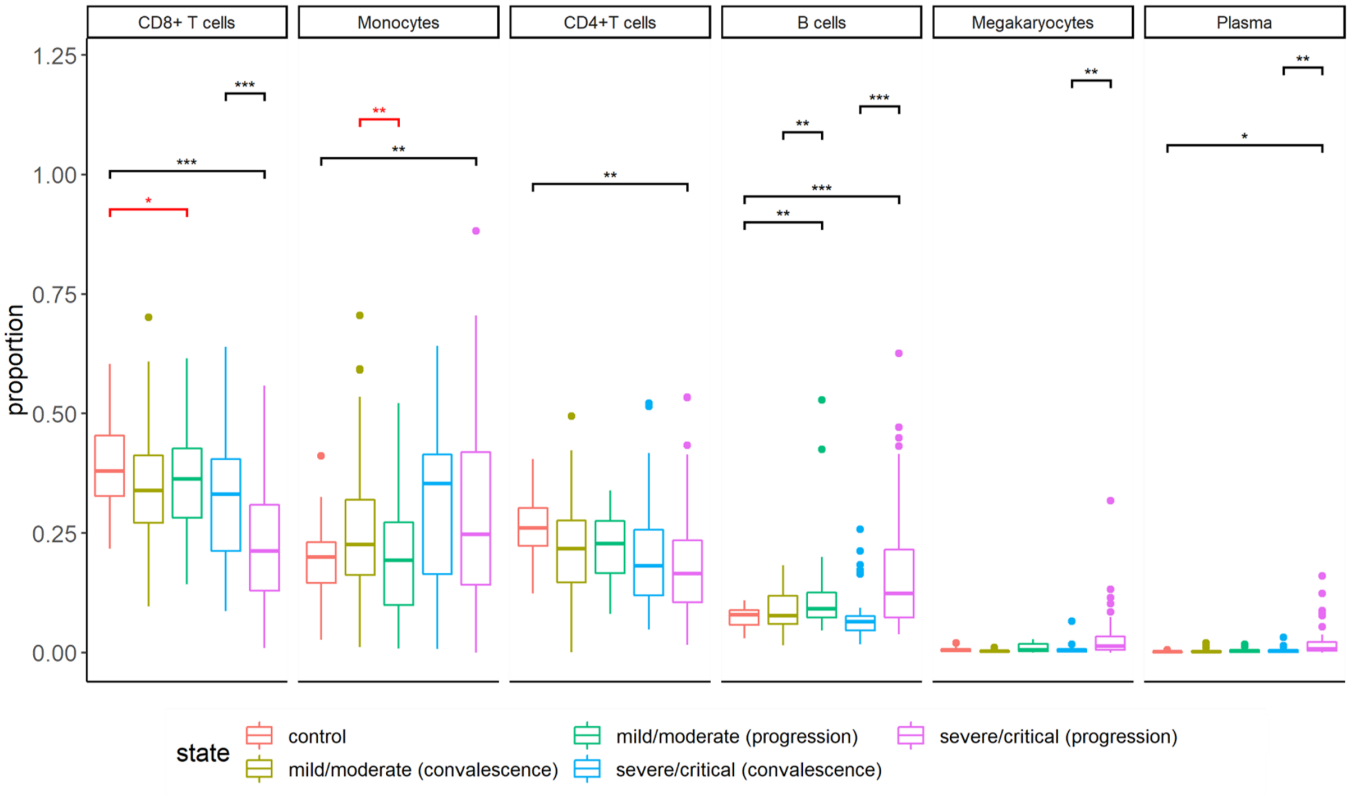
cM, CD14+CD16– monocytes; ncM, CD14+CD16+ monocytes; DC, dendritic cells; Mkc, megakaryocytes; Th, CD4+ T cells; B, B cells; Tc, CD8+ T cells; NK, natural killer cells

Results: Real World Data [3]



E, Enterocyte; TA, transit amplifying; TAE, TA.Early; EP, Enterocyte.Progenitor; Gob, Goblet

Results: Real World Data [4]



Acknowledgments

Joshua Ho Lab

Weizhong Zheng

Arron Kwok

Junyi Chen

All lab members

Yuanhua Huang Lab

All lab members

References

[1] L. Zappia, B. Phipson, and A. Oshlack. "Splatter: simulation of single-cell RNA sequencing data". In: *Genome biology* 18.1 (2017), pp. 1-15.

[2] H. M. Kang, M. Subramaniam, S. Targ, et al. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". In: *Nature biotechnology* 36.1 (2018), p. 89.

[3] A. L. Haber, M. Biton, N. Rogel, et al. "A single-cell survey of the small intestinal epithelium". In: *Nature* 551.7680 (2017), pp. 333-339.

[4] X. Ren, W. Wen, X. Fan, et al. "COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas". In: *Cell* 184.7 (2021), pp. 1895-1913.