THE UNIVERSITY OF CHICAGO MEDICINE

The curses of performing differential expression analysis using single cell data

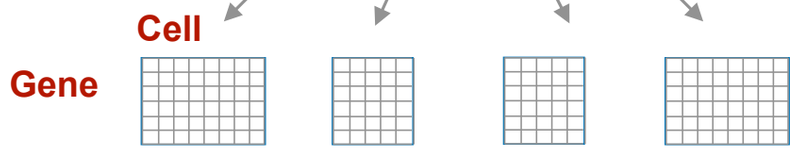
Mengjie Chen

July 4, 2023

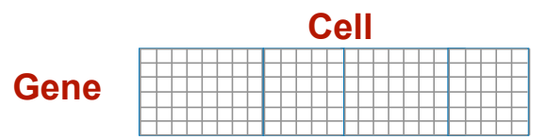
Current practice

scRNA-seq experiments

Raw UMI counts

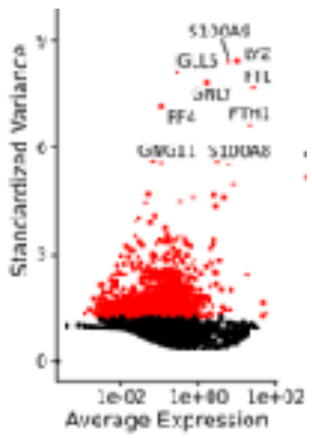


Integration/ Batch effects removal

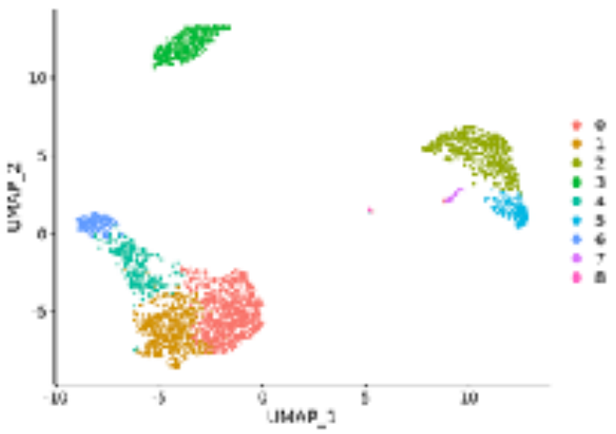


Normalization

Feature Selection based on variability



Dimension Reduction



Clustering/Annotation

Processed expression values

Annotated cell types

Differential expression analysis

Widely used assumptions in scRNAseq DE analysis:

1. Normalization is necessary to adjust differences in library size for each cell.

2. Batch effect removal can mitigate donor effects. DE needs to be performed on integrated data.

3. Success in clustering and annotation will guarantee the success of DE analysis.

Widely used assumptions in scRNAseq DE analysis:

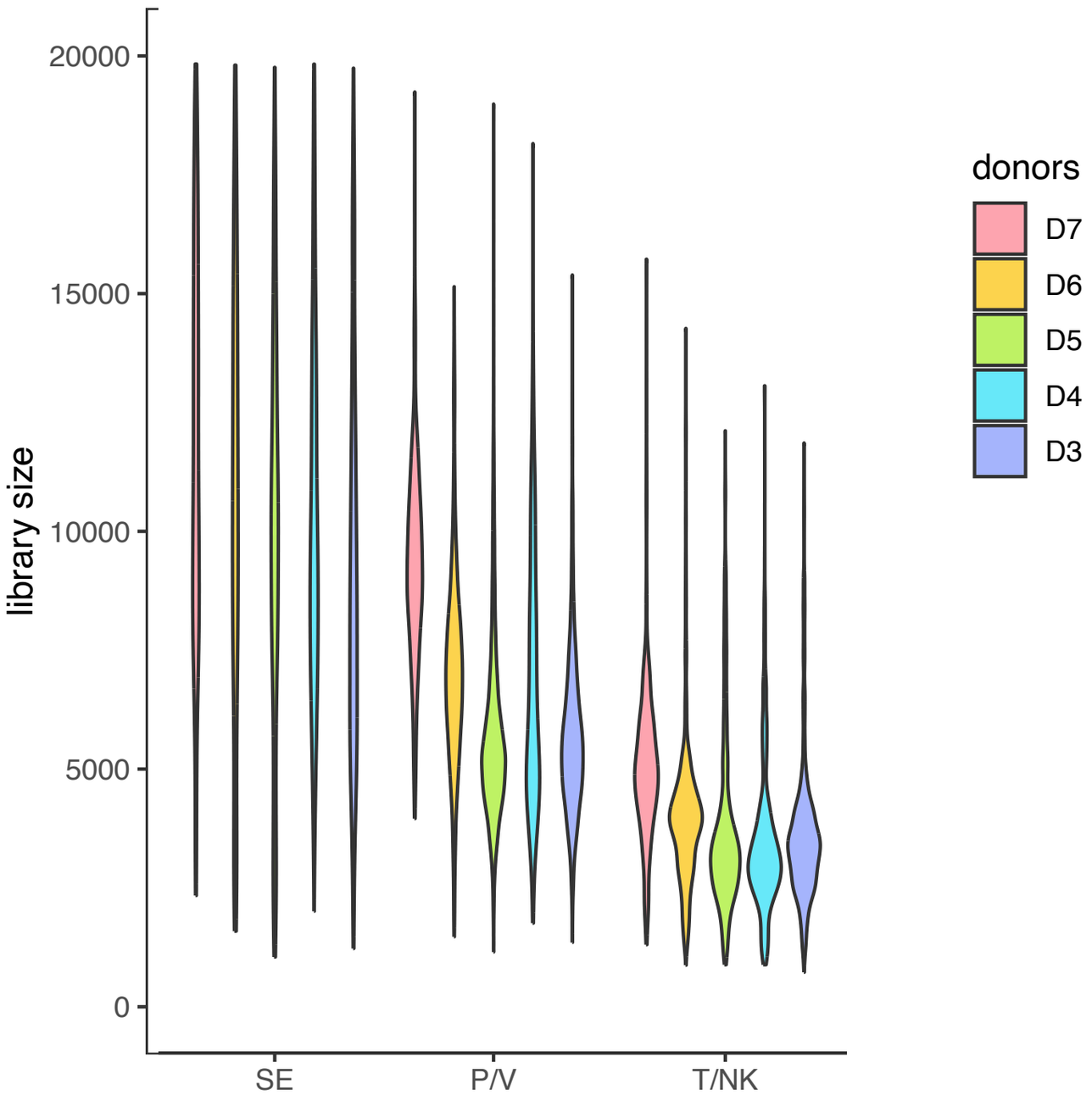
1. Normalization is necessary to adjust differences in library size for each cell.

2. Batch effect removal can mitigate donor effects. DE needs to be performed on integrated data.

3. Success in clustering and annotation will guarantee the success of DE analysis.

Let us check these assumptions in 10X UMI data.

Raw UMI data



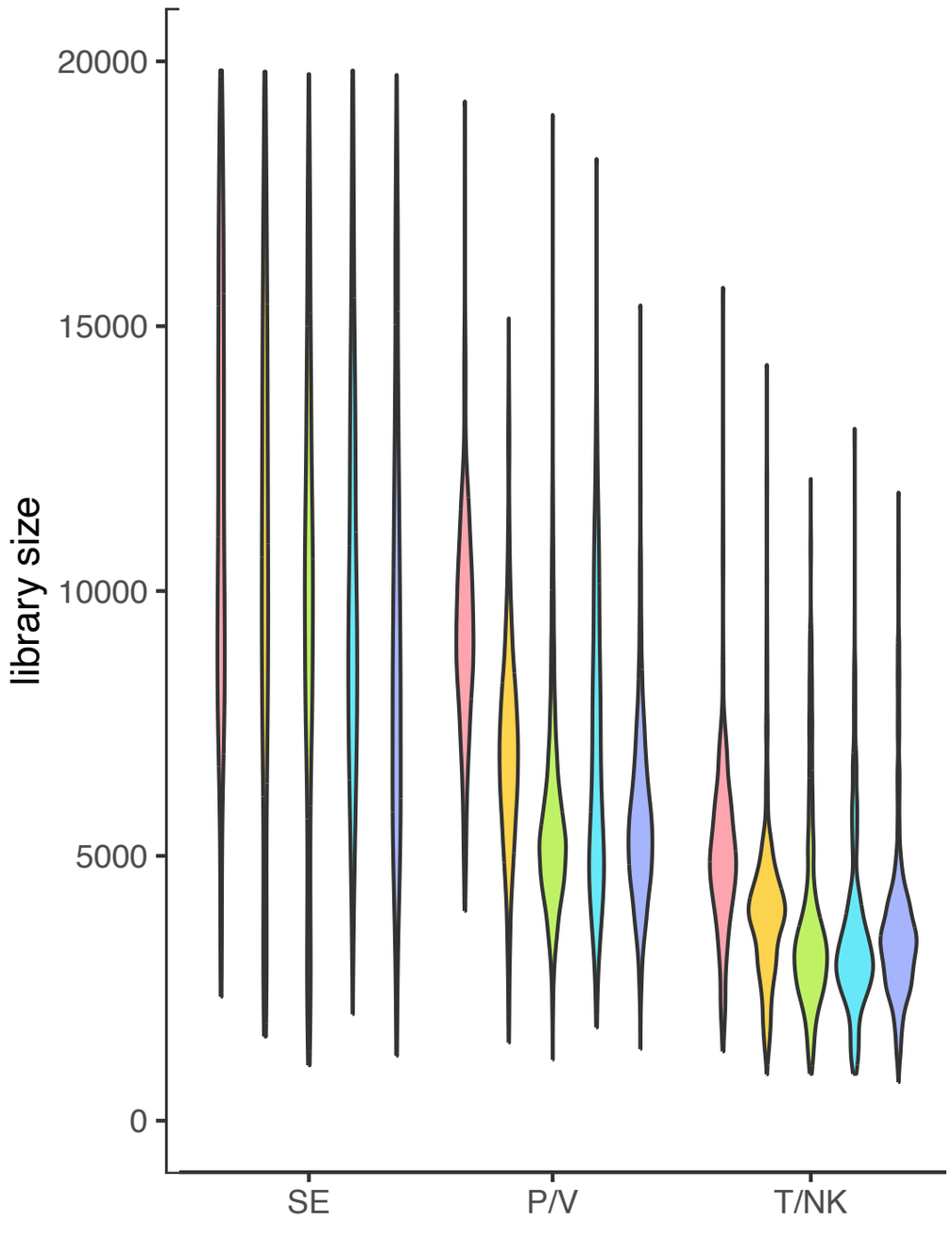
Cell Reports

Resource

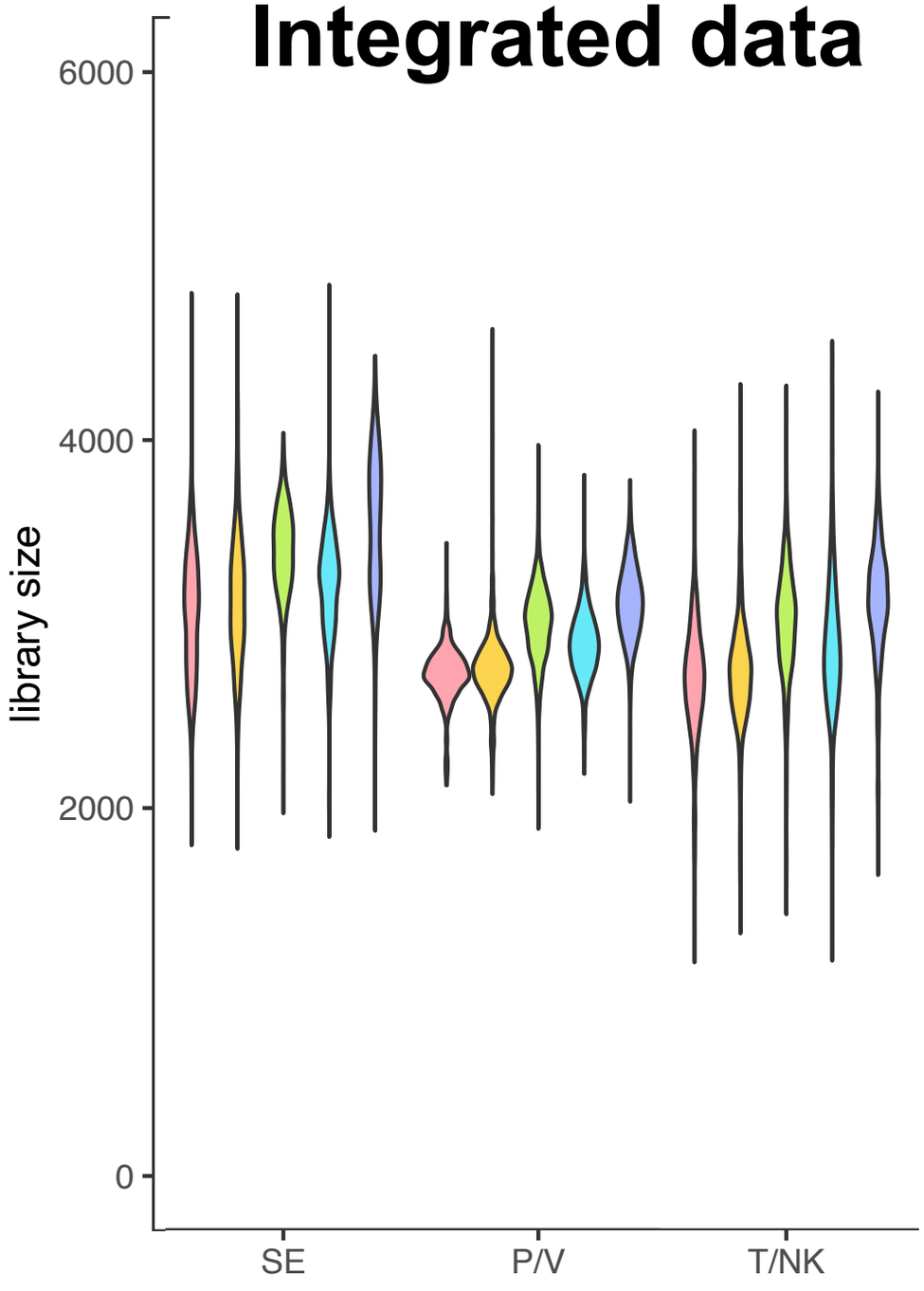
A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing

8 Patients
18 Samples
60,574 Cells

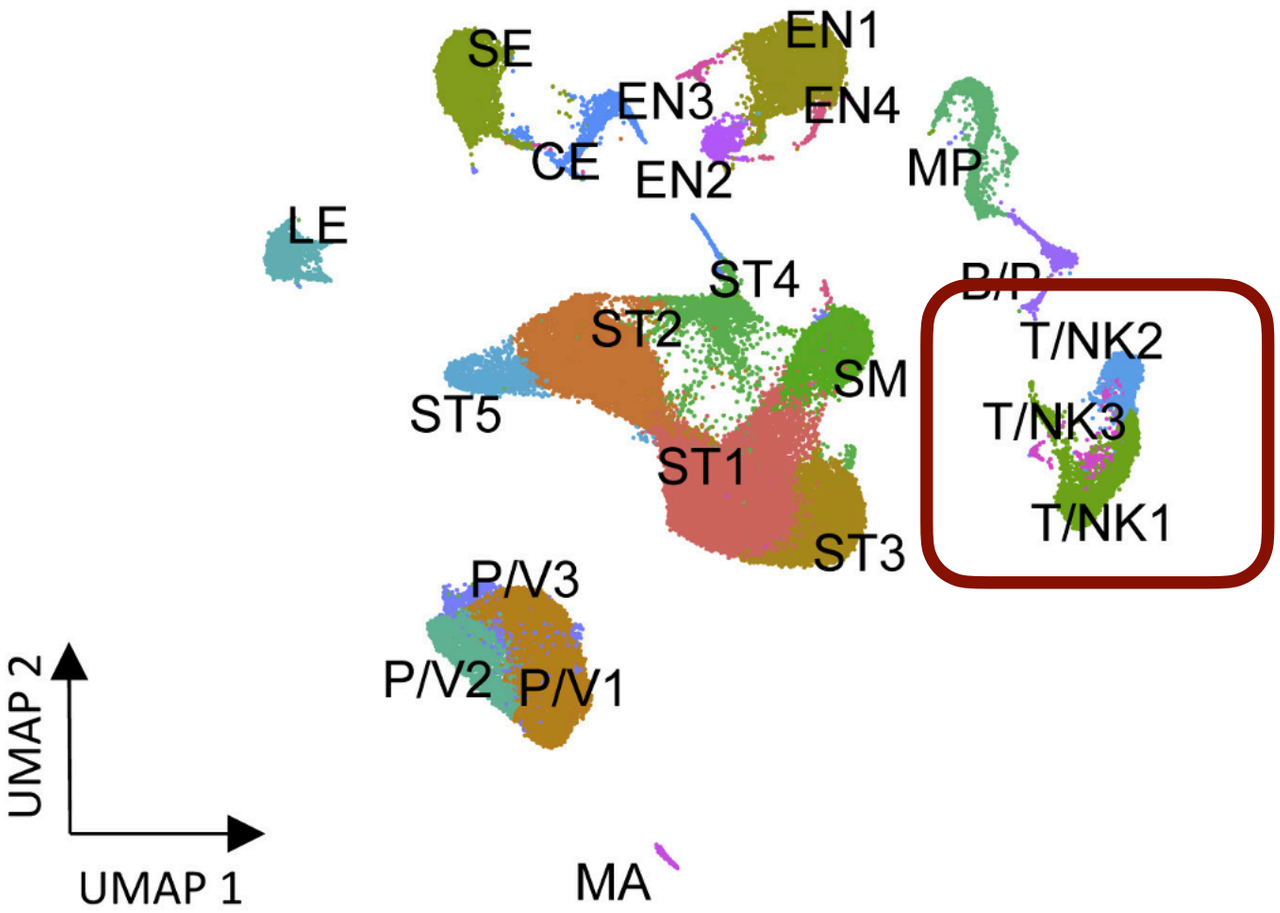
Raw UMI data



Seurat Normalized Integrated data



Fallopian tube



Cell Reports

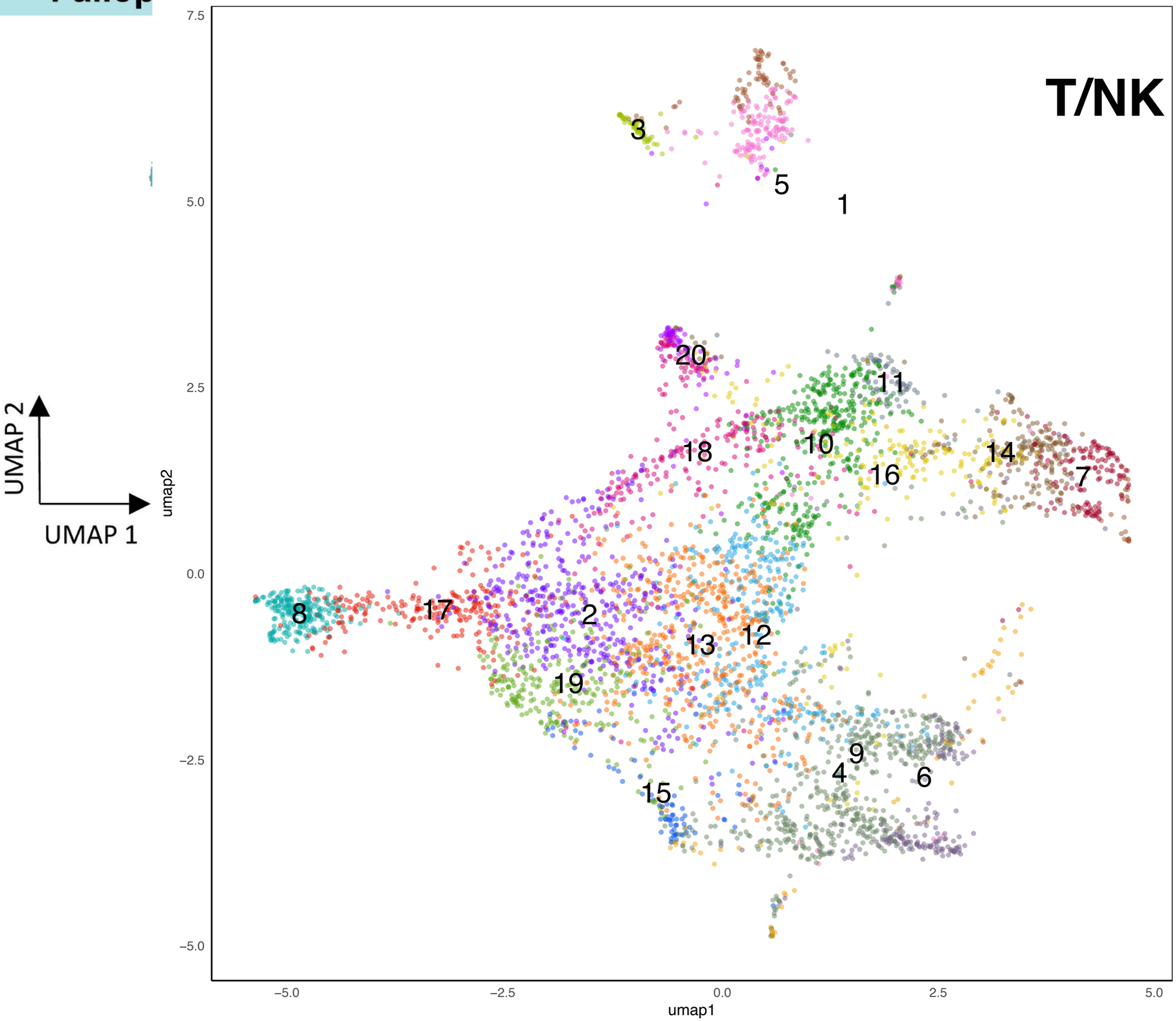
Resource

A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing

8 Patients
18 Samples
60,574 Cells

Fallopian tube

T/NK cells



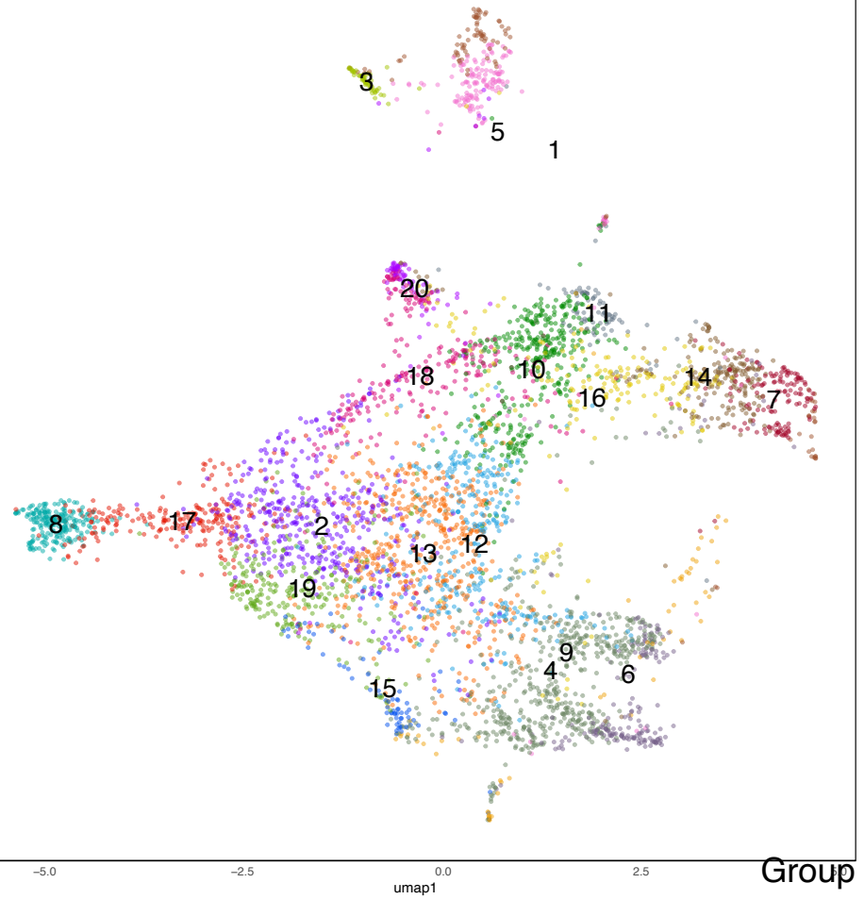
Resource

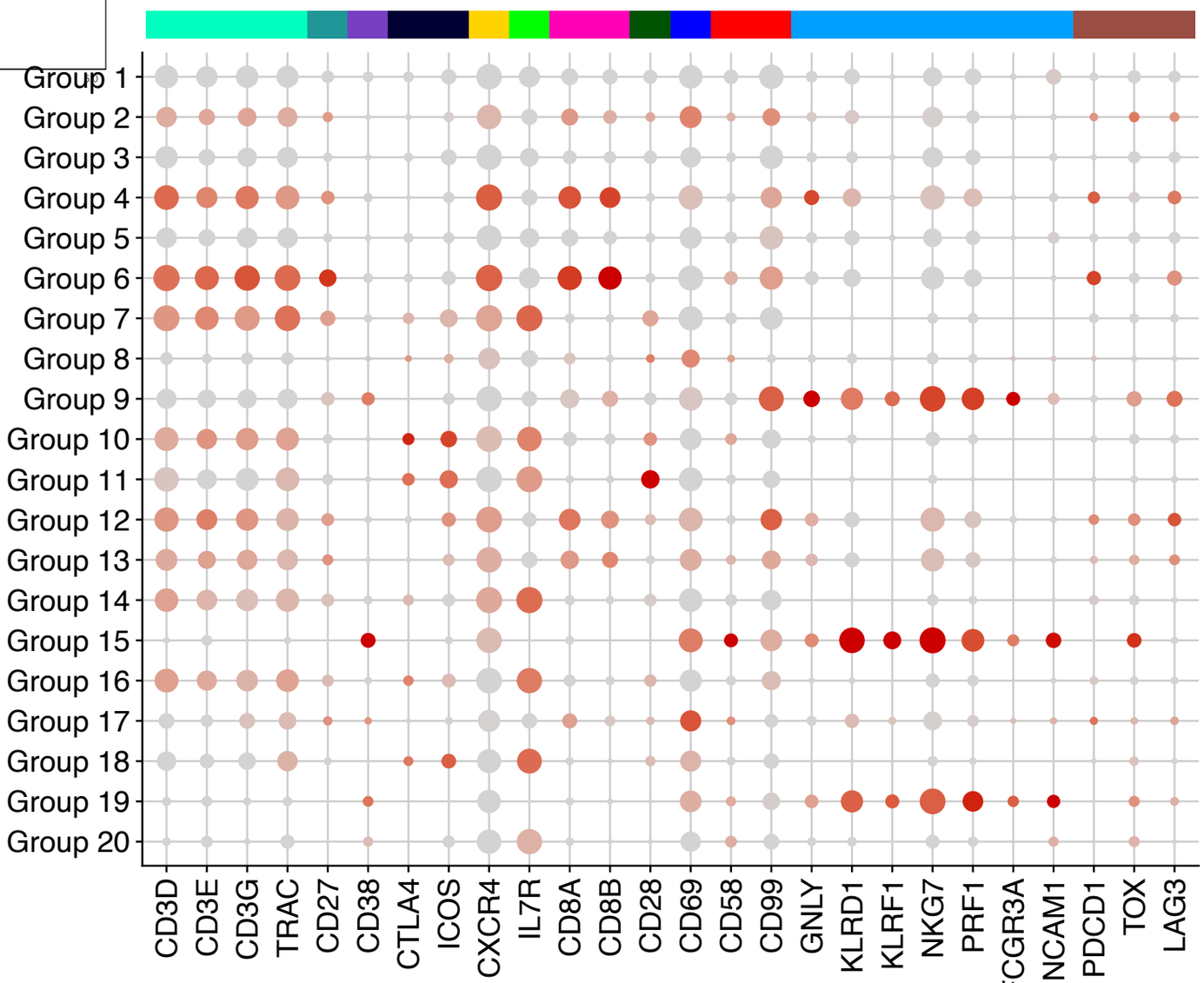
8 Patients

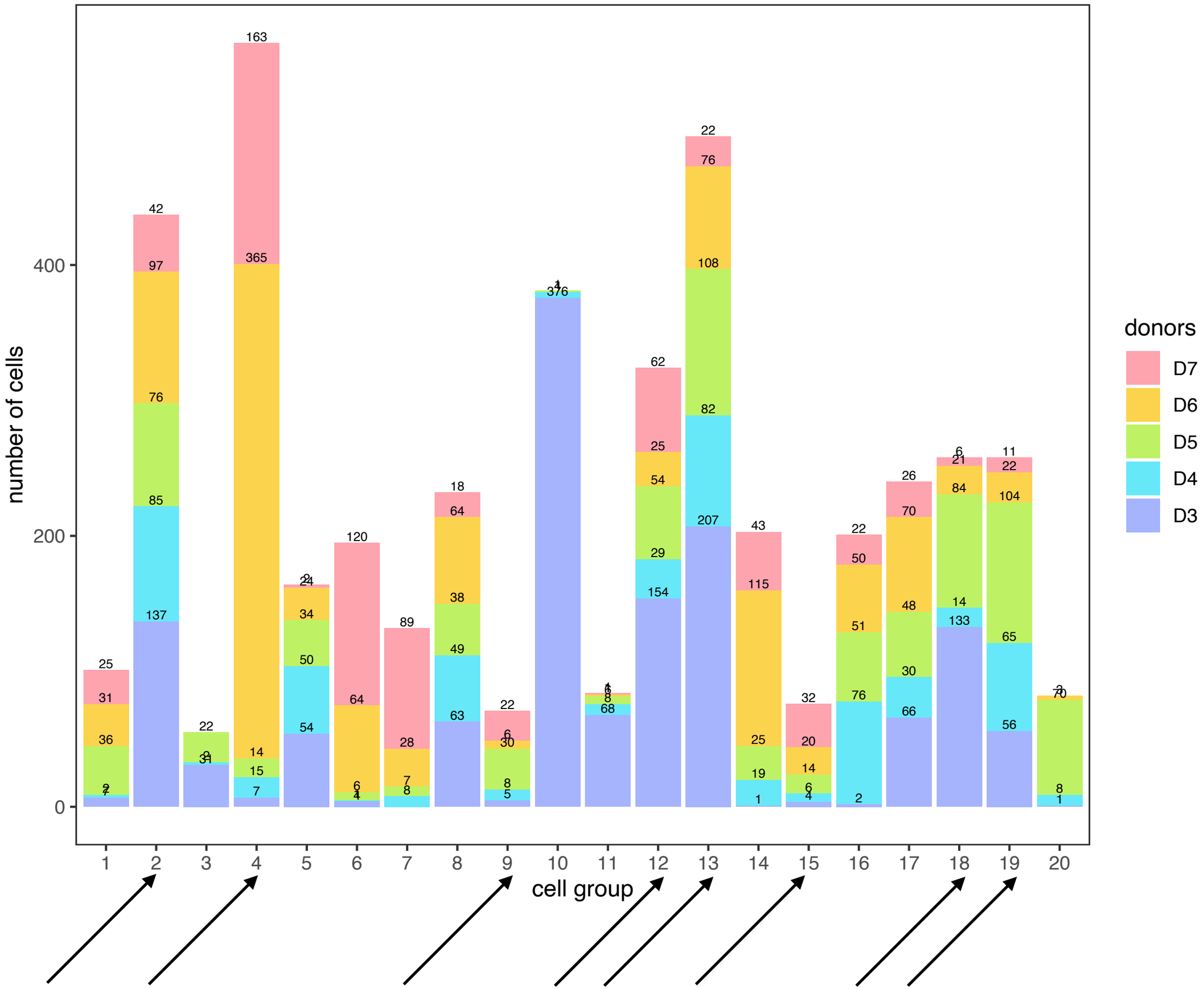
8 Samples

0,574 Cells

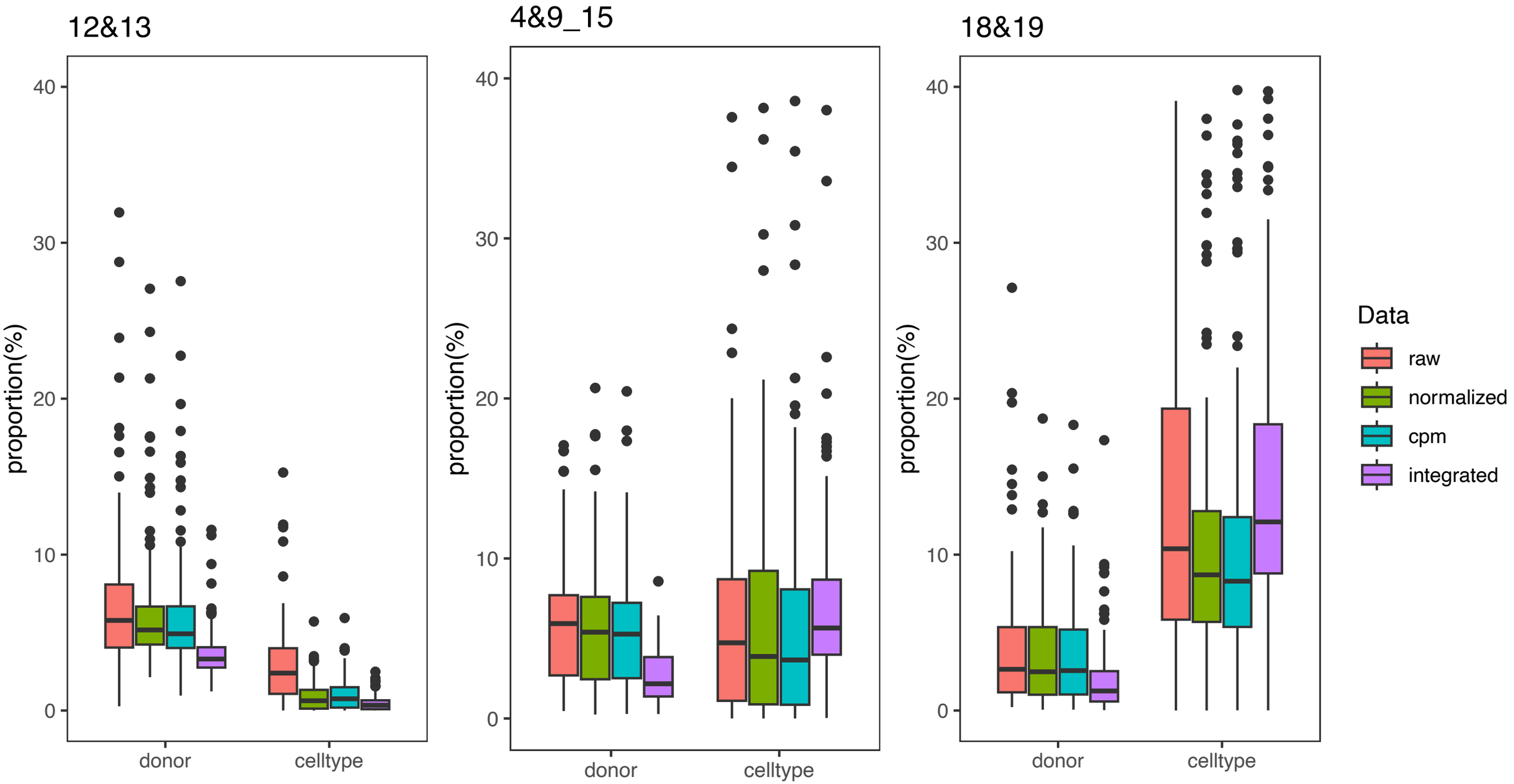
T/NK cells


3
5
1
20
11
18
10
14
16
7
8
17
2
13
12
19
9
4
6
15
−5.0
−2.5
0.0
2.5
umap1


Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7
Group 8
Group 9
Group 10
Group 11
Group 12
Group 13
Group 14
Group 15
Group 16
Group 17
Group 18
Group 19
Group 20
CD3D
CD3E
CD3G
TRAC
CD27
CD38
CTLA4
ICOS
CXCR4
IL7R
CD8A
CD8B
CD28
CD69
CD58
CD99
GNLY
KLRD1
KLRF1
NKG7
PRF1
FCGR3A
NCAM1
PDCD1
TOX
LAG3
Activation/Mature naïve
Adhesion
CD4+ T cells
CD4+ T regulatory
CD8+ T cells
Costimulation/Survival
Migration/CD4+ T regulatory
NK
T Cell infiltration markers
T cells
Transitional/Activation
Transitional/Costimulation
Percent Expressed
25
50
75
100
Average Expression
2.5
2.0
1.5
1.0
0.5
0.0


number of cells
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
cell group
donors
D7
D6
D5
D4
D3

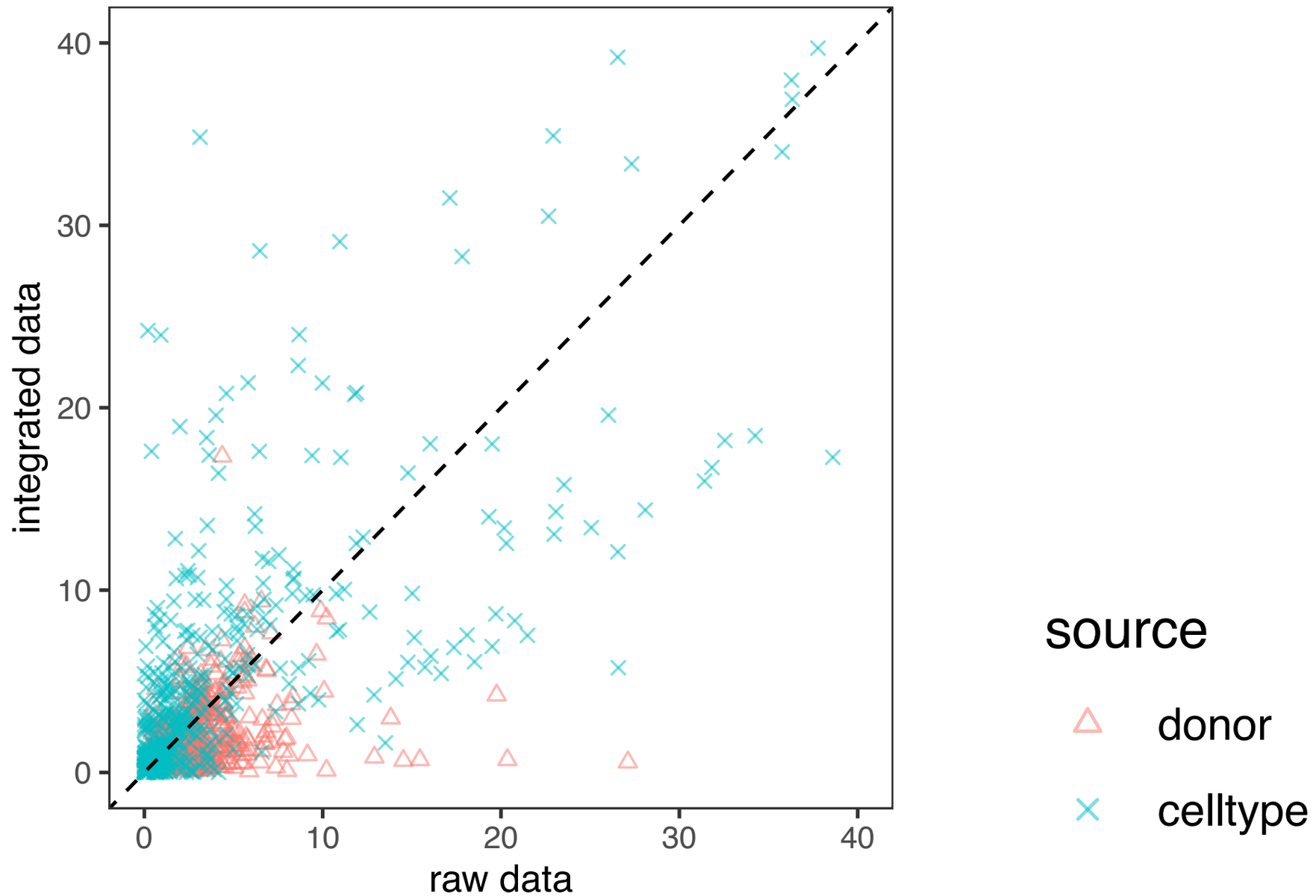
Can integration (batch effects removal) mitigate donor effects?

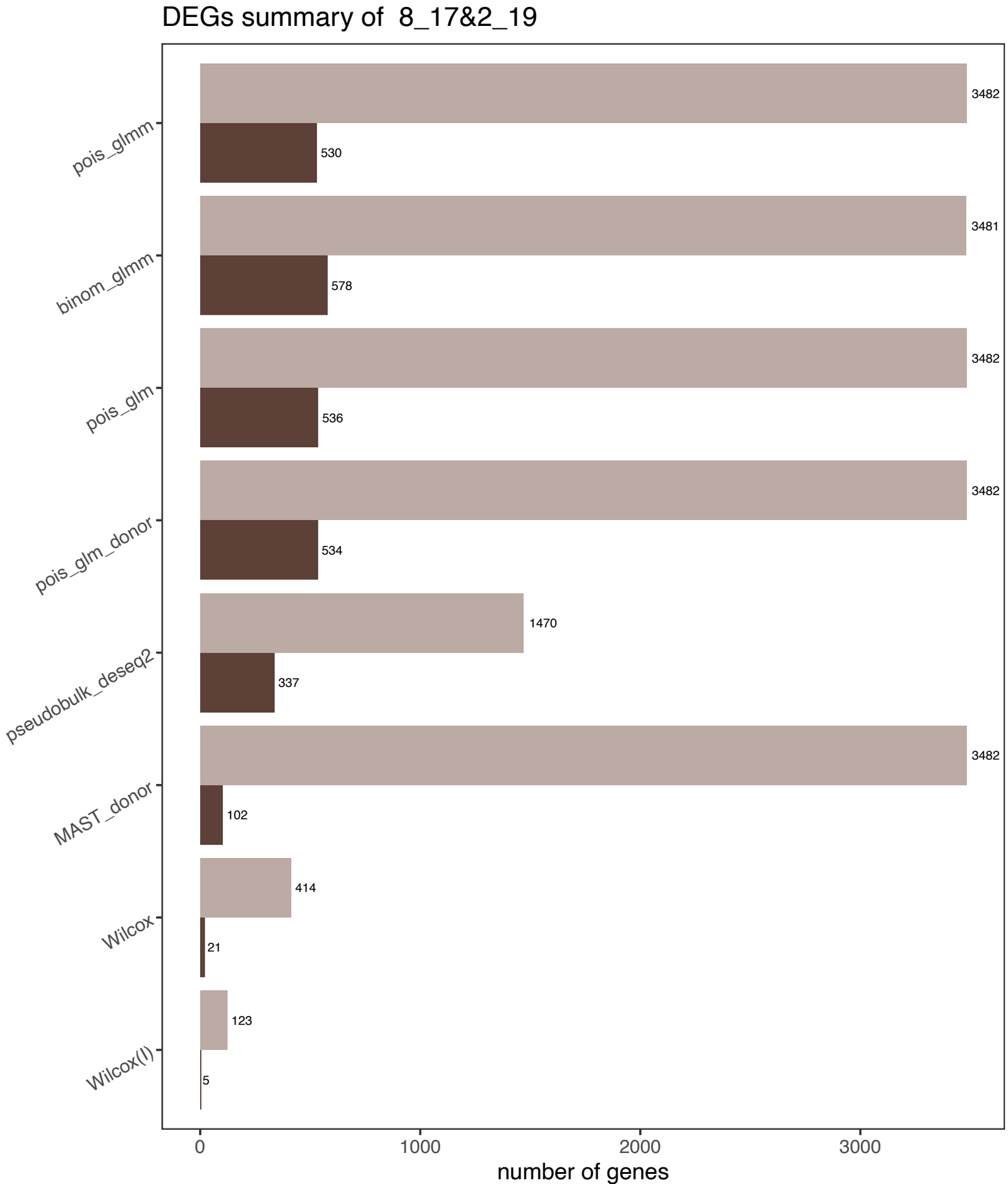


Proportions of variation estimated by a linear model

Can integration (batch effects removal) mitigate donor effects?

Proportions of variation estimated by a linear model





Input: 1. Raw counts

2. Zero proportions

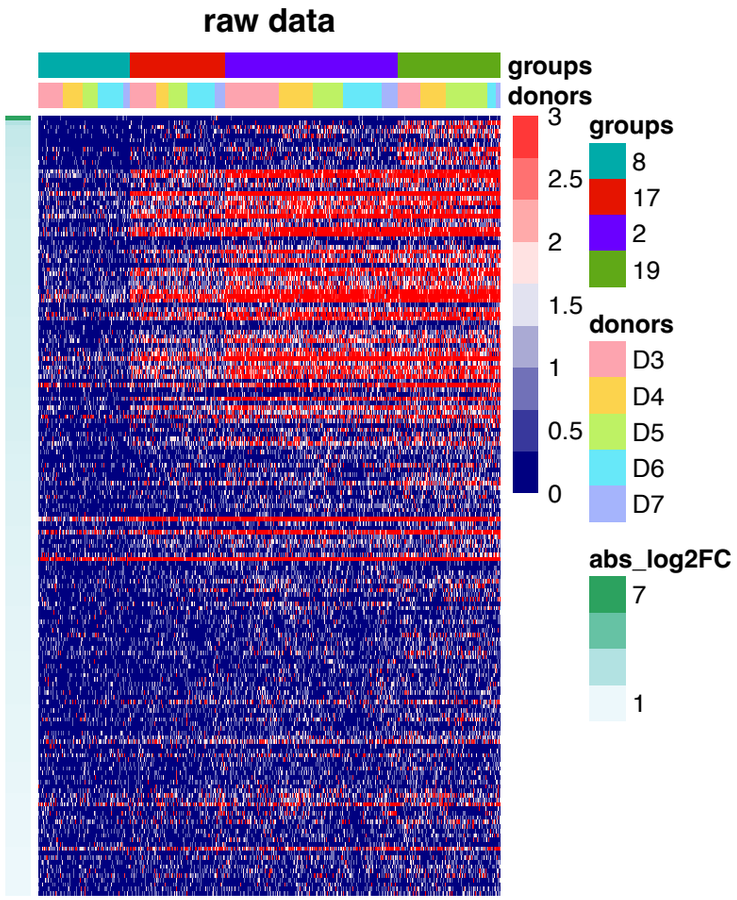
Model: 1. GLMM

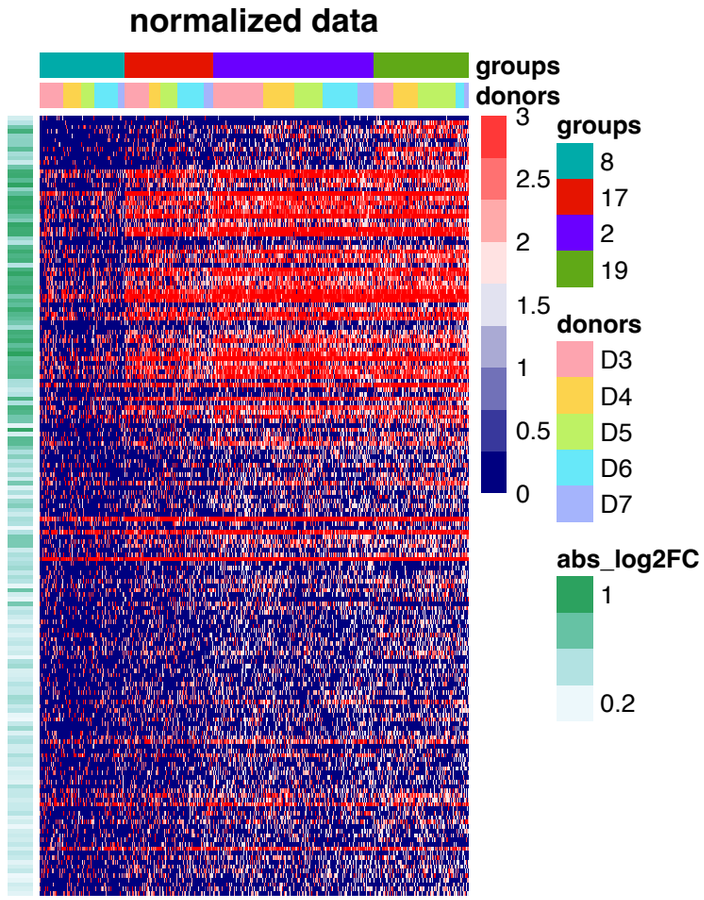
- fixed donor effects

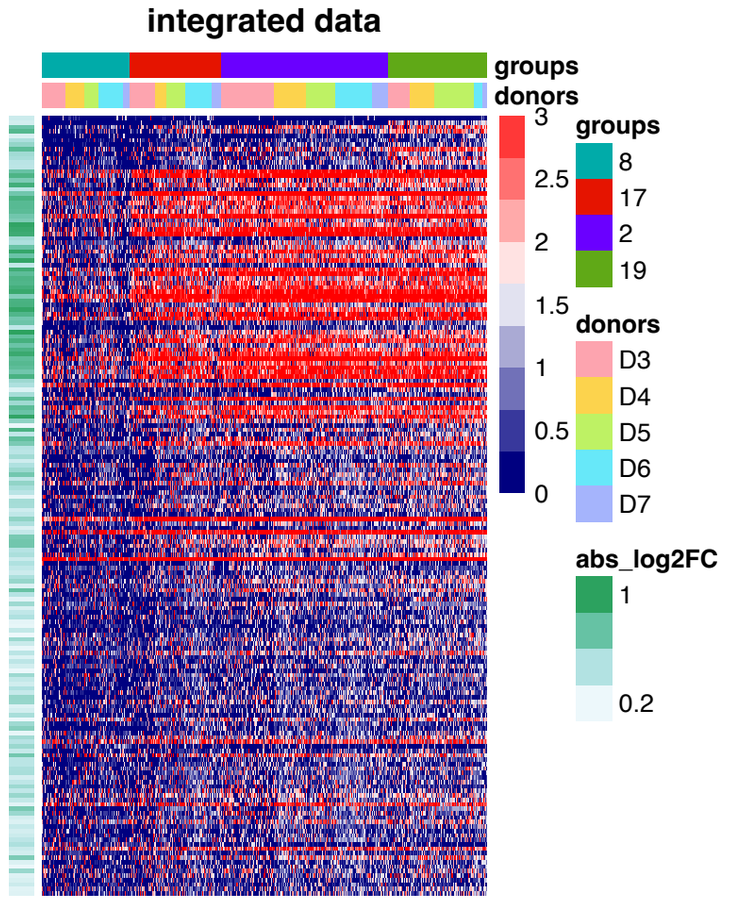
2. GLMM

- random donor effects

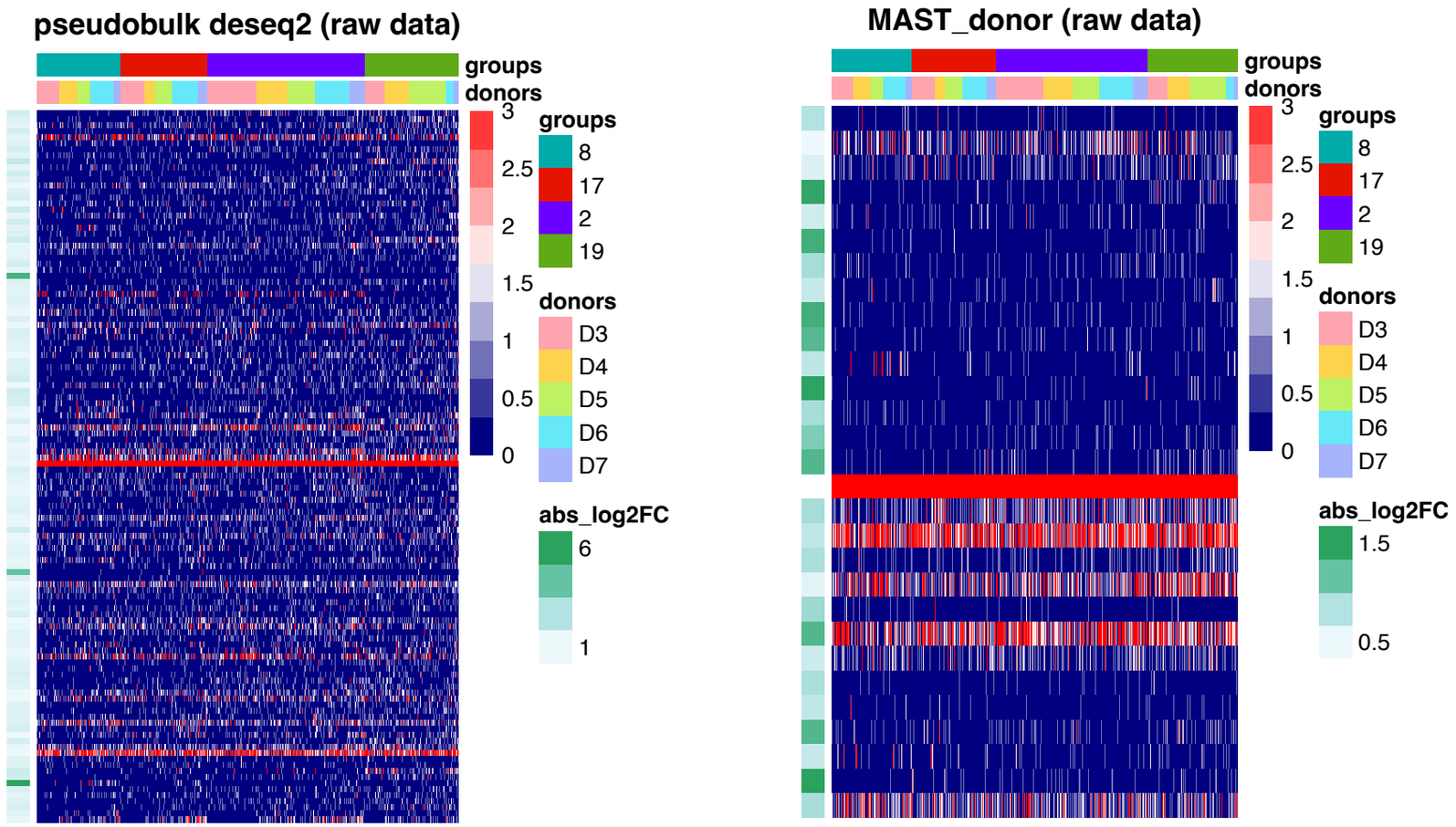
Integration and normalization can create arbitrary patterns.



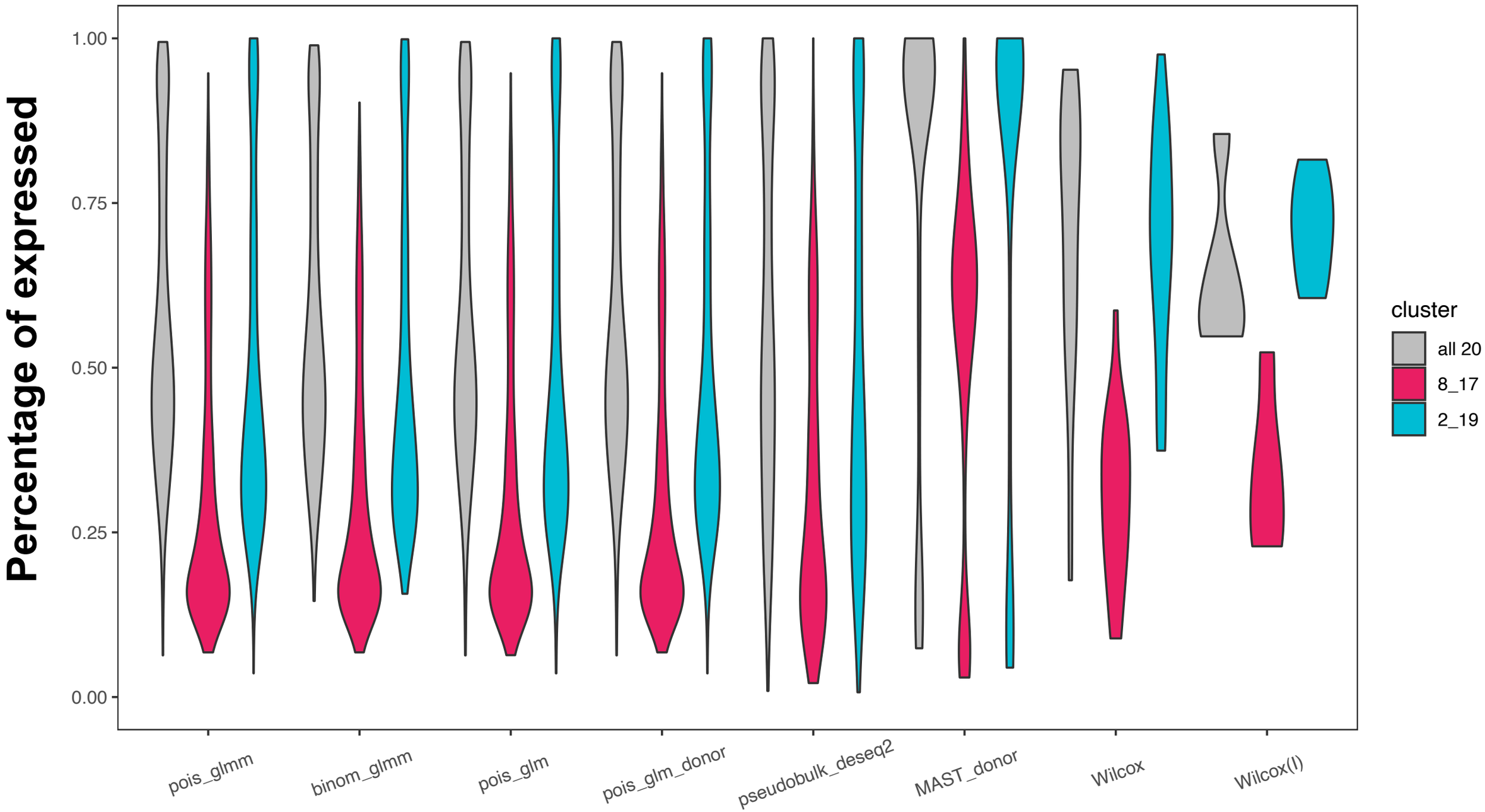




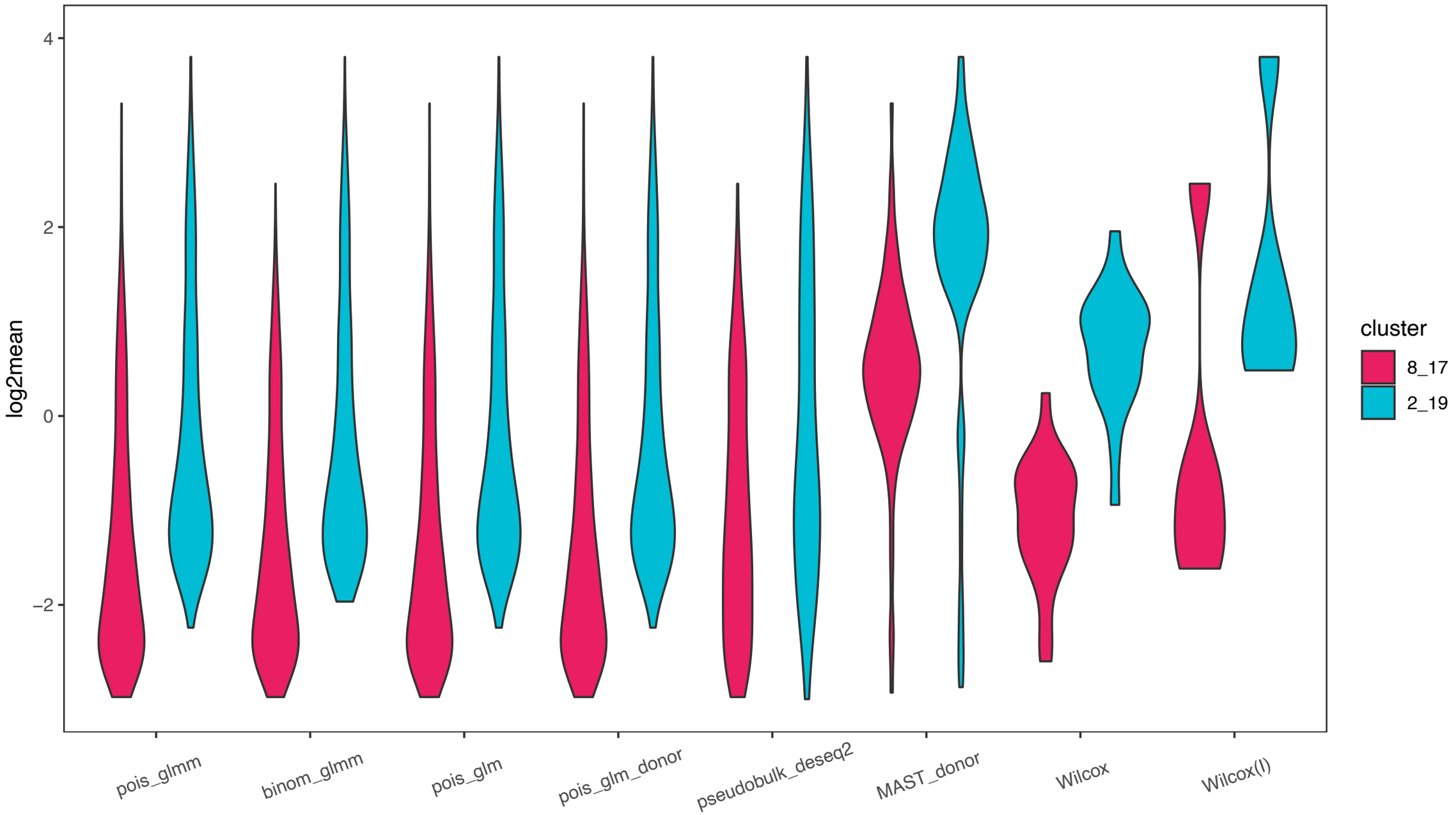
Integration and normalization can create arbitrary patterns, leading to detection of false signals.

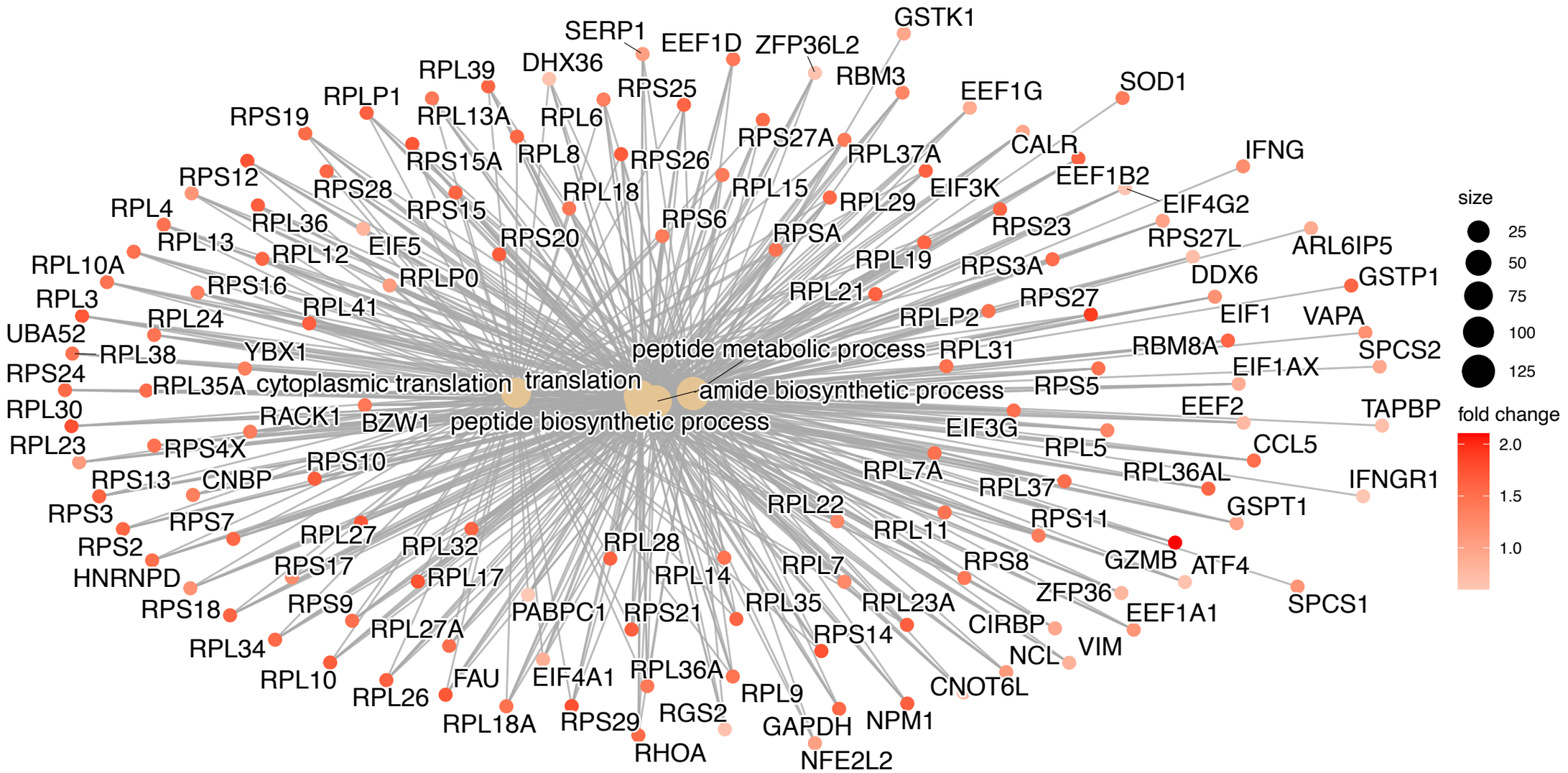


Current methods are biased towards frequently expressed genes.



Current methods are biased towards highly expressed genes.





Comparison of (8, 17) vs (2, 19), i.e, mature T cell vs (CD4, NK) cells

Mature T cells exhibit lower levels of ribosomal RNA activity compared to their immature counterparts. This is mainly due to the state of activation and the metabolic requirements of the cells.

Widely used assumptions in scRNAseq DE analysis:

Cells exhibit varying RNA contents that reflect their biological functions.

1. ~~Normalization is necessary to adjust differences in library size for each cell.~~

2. ~~Batch effect removal can mitigate donor effects. DE needs to~~ be performed on integrated data.

Batch effect removal is not enough to remove donor effects loaded on each gene.

3. ~~Success in clustering and annotation will guarantee the success of DE analysis.~~

   Clustering and DE are two separate problems.

https://c-hw.github.io/DEanalysis/index.html

Current practice

scRNA-seq experiments

Raw UMI counts

Cell

Gene

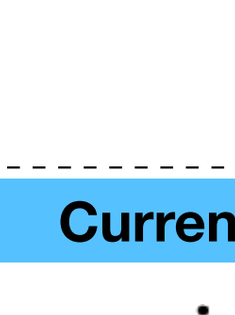
Integration/ Batch effects removal

Cell

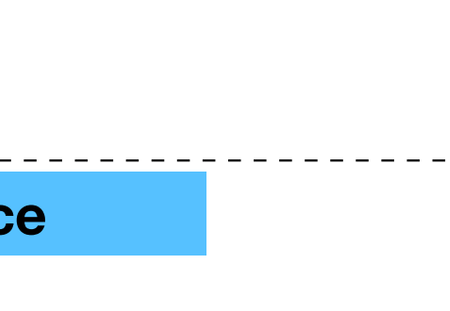
Gene

Normalization

Feature Selection based on variability



Dimension Reduction



Clustering/Annotation

Processed expression values

Annotated cell types

Differential expression analysis

New paradigm

scRNA-seq experiments

Raw UMI counts

Cell

Gene

Sample information

Zero proportions

Sample ID

(Random effects)

Confounding factors (Age, Race etc.)

(Fixed effects)

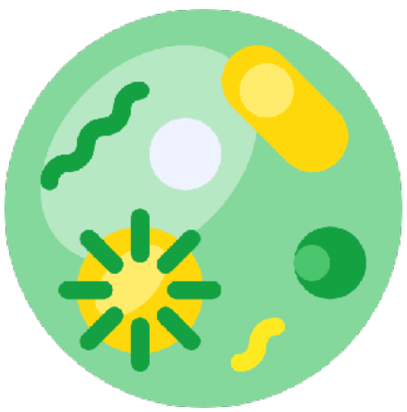
Annotated cell types

(Fixed effects)

Differential expression analysis using a generalized linear mixed model

Acknowledgement

Single cell genomics



Chih-Hsuan Wu
Tae Kim
Yifan Zhou

Ziyou Ren
Lisha Chen
Yan Li

Epitranscriptomics

Scott Zhang
Shun Liu
Xiaoyang Dou
Yong Peng

Weixin Tang
Chuan He
Peng Jin (Emory)

Xiang Zhou (U Michigan)
Oni Basu
Ernst Lengyel
Jian Shu (Harvard)
Rohit Kulkarni
Yang Li
Xin He

National Human Genome Research Institute

AT THE FOREFRONT UChicago Medicine Comprehensive Cancer Center

CHAN ZUCKERBERG INITIATIVE

NIGMS

Alfred P. Sloan FOUNDATION

NSF