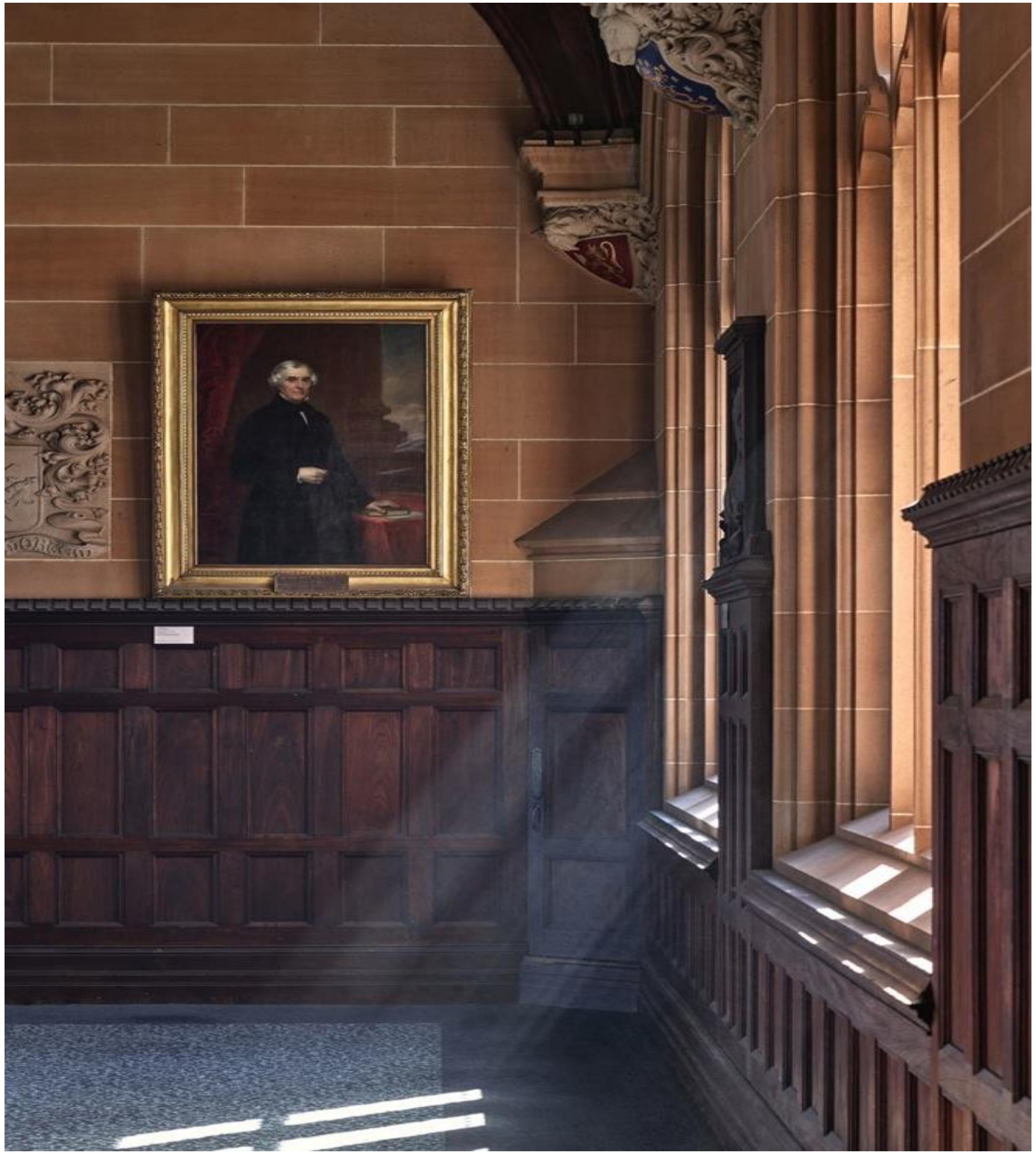
Single-cell analysis workshop

Sydney Precision Bioinformatics Group

THE UNIVERSITY OF SYDNEY



Sydney Precision Bioinformatics Research Group

We share an interest in developing statistical and computational methodologies to tackle the foremost significant challenges posed by modern biology and medicine.

Meet our senior and junior research leaders

Jean Yang

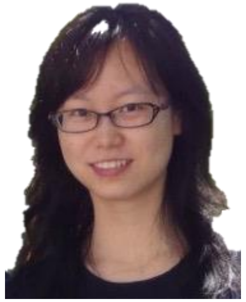
Samuel Muller

John Ormerod

Pengyi Yang

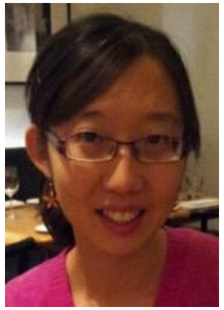
Ellis Patrick

Rachel Wang



Garth Tarr

Kitty Lo



and senior research associates, PhD candidates, Honours and TSP students: 25

Find out more: http://www.maths.usyd.edu.au/bioinformatics/

Get interactive: http://shiny.maths.usyd.edu.au/

Roadmap for the workshop

- Setting up: 1:15 – 1:30 Google cloud set up

- Session 1: 1:30 – 2:00 Single cell analysis overview (scdney)
- Session 2: 2:00 – 2:45 Quality control and data integration
- Session 3: 2:45 – 3:45 Cell type identification via cluster analysis
- Session 4: 3:45 – 4:30 Downstream analysis: identify marker genes & cell type composition

- Extension: cell type identification via supervised classification and single cell trajectory analysis

Workshop presenters in each session: Jean Yang, Kevin Wang, Pengyi Yang, Yingxin Lin

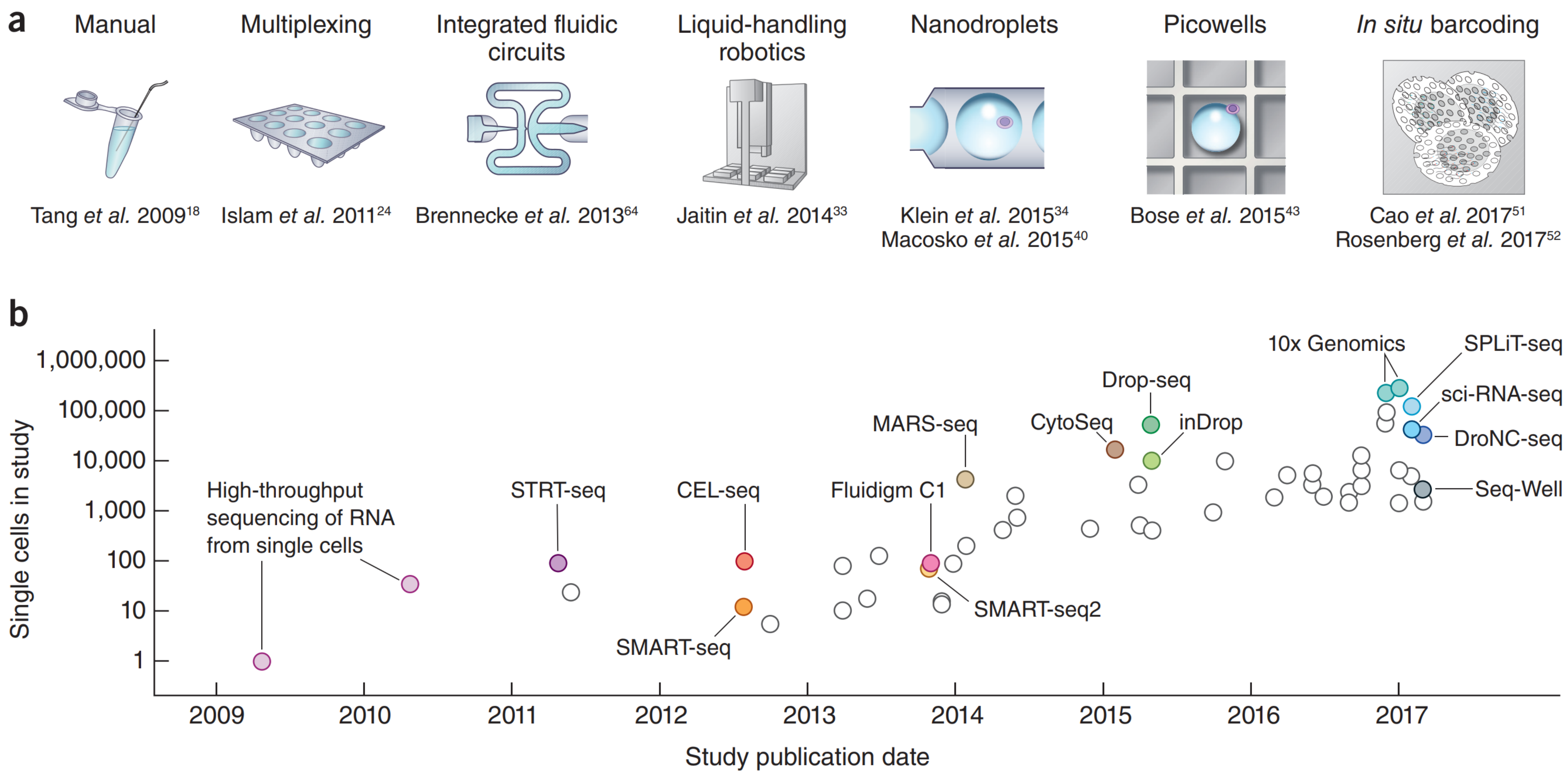
Configuring Google Cloud

- Machine 1: 34.69.169.142

- Machine 2: 34.94.220.230

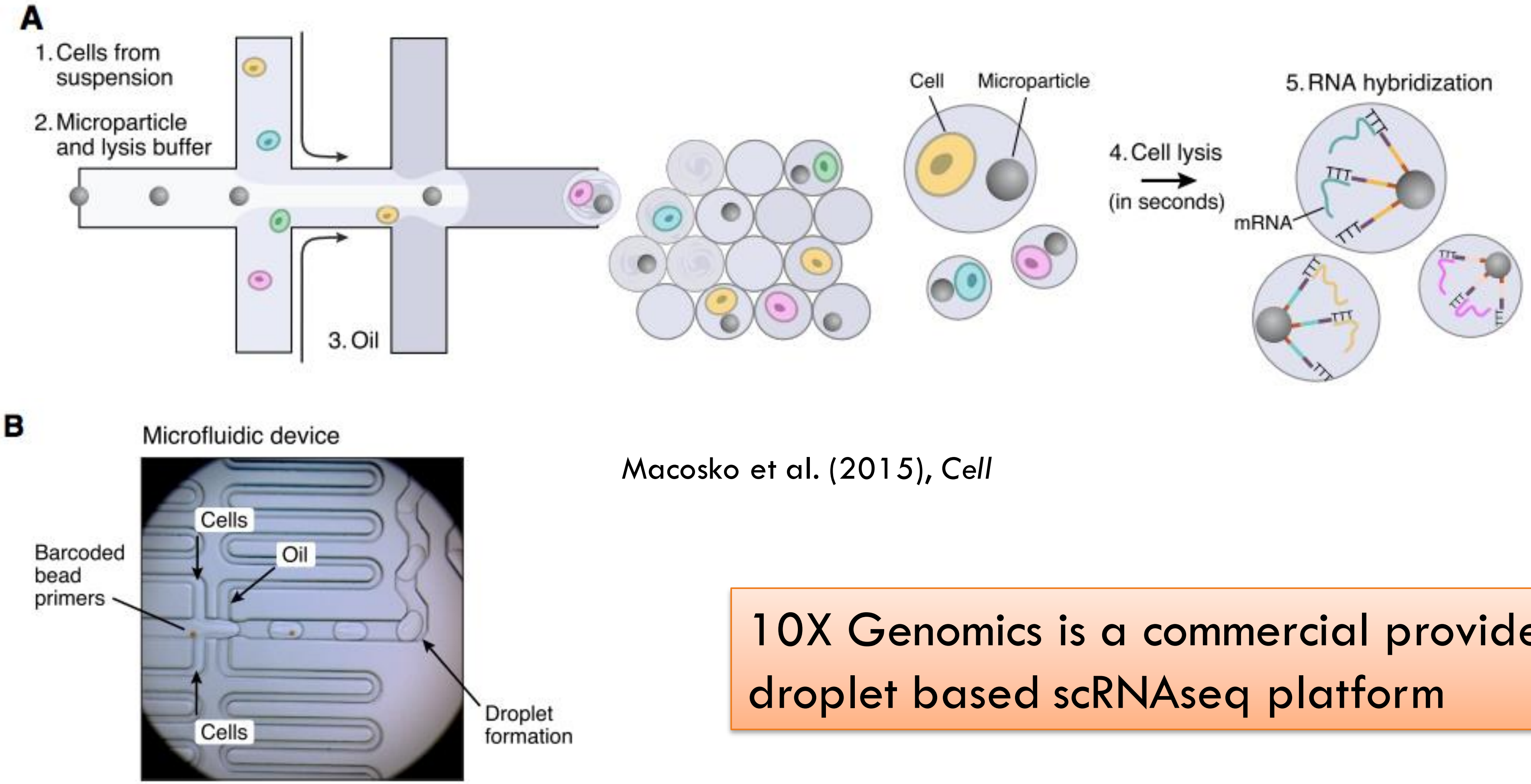
```
source("/home/user_setup.R")
```

Exponential growth in single cell RNA seq technologies



Svensson et al. *Nature Protocols* (2018)

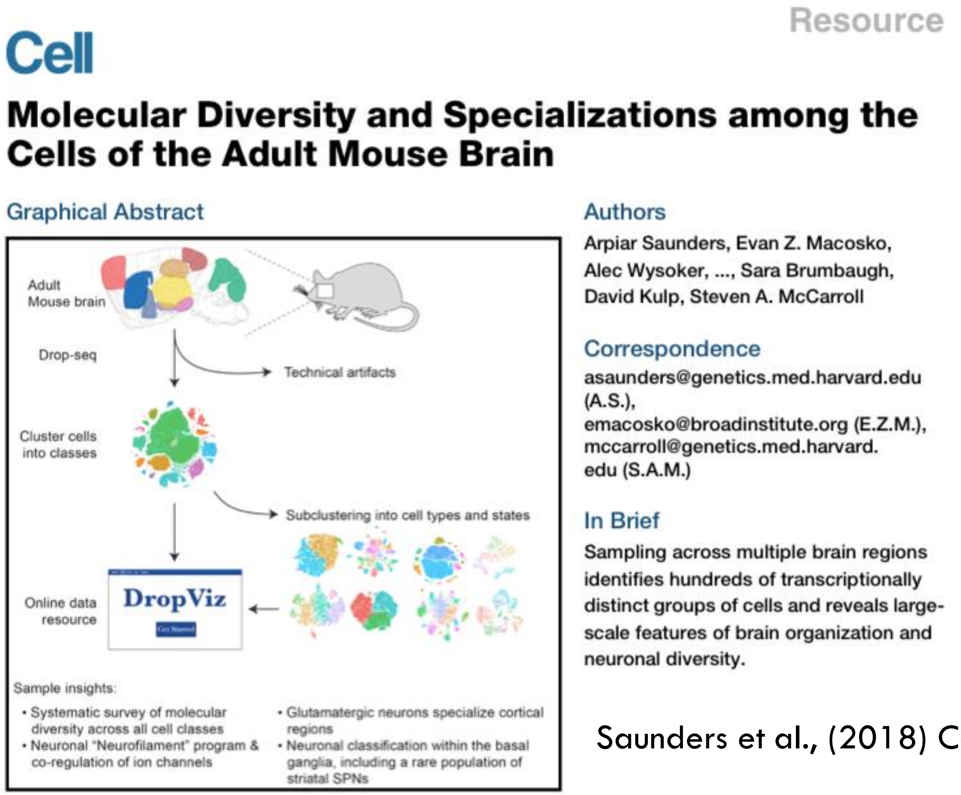
Droplet based technologies are now dominating



Macosko et al. (2015), *Cell*

10X Genomics is a commercial provider of droplet based scRNAseq platform

scRNAseq experiments approaching 1 million cells



Resource

Cell

Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain

Graphical Abstract

Sample insights:

- Systematic survey of molecular diversity across all cell classes
- Neuronal "Neurofilament" program & co-regulation of ion channels
- Glutamatergic neurons specialize cortical regions
- Neuronal classification within the basal ganglia, including a rare population of striatal SPNs

Authors

Arpiar Saunders, Evan Z. Macosko, Alec Wysoker, ..., Sara Brumbaugh, David Kulp, Steven A. McCarroll

Correspondence

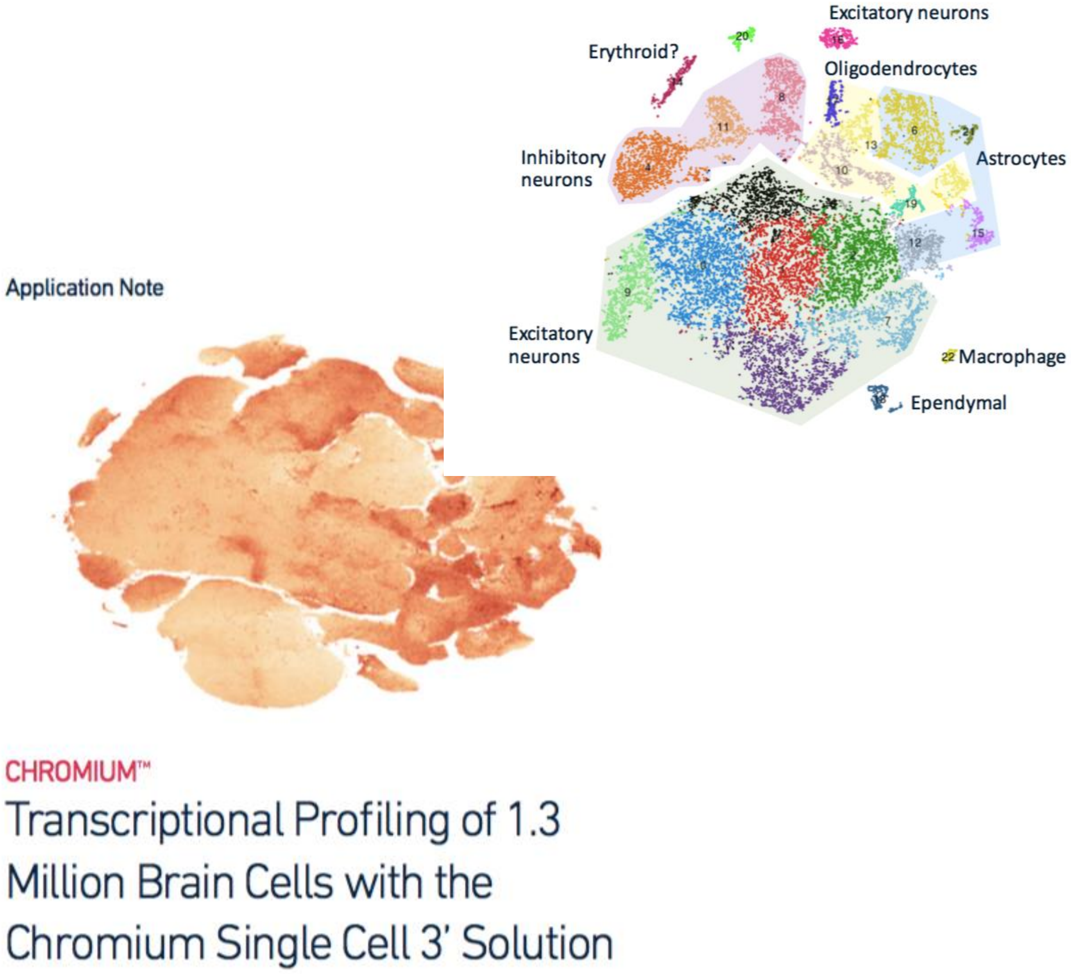
asaunders@genetics.med.harvard.edu (A.S.), emacosko@broadinstitute.org (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief

Sampling across multiple brain regions identifies hundreds of transcriptionally distinct groups of cells and reveals large-scale features of brain organization and neuronal diversity.

Saunders et al., (2018) Cell

690,000 individual cells from 9 regions of adult mouse brain

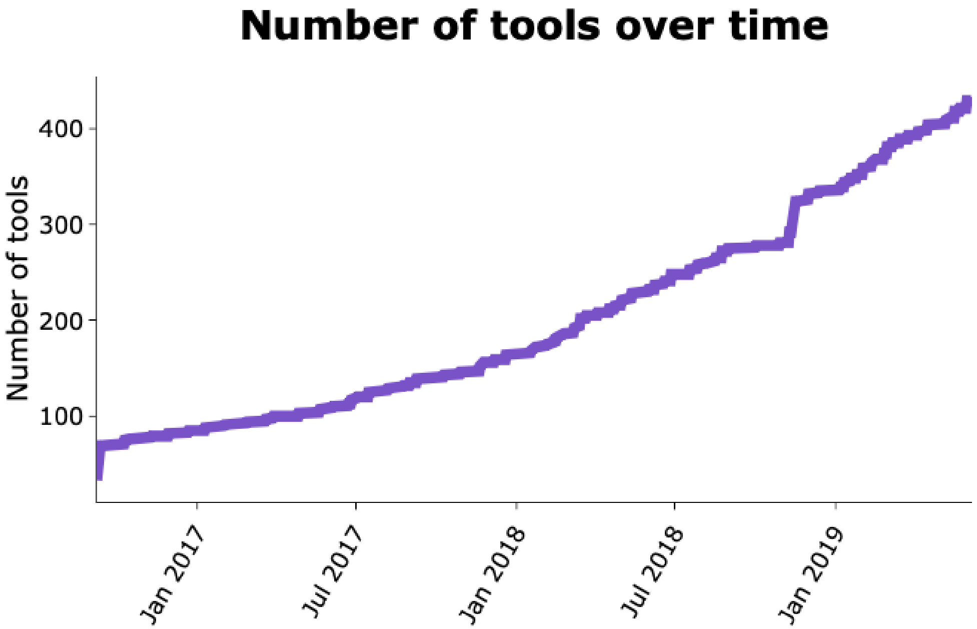


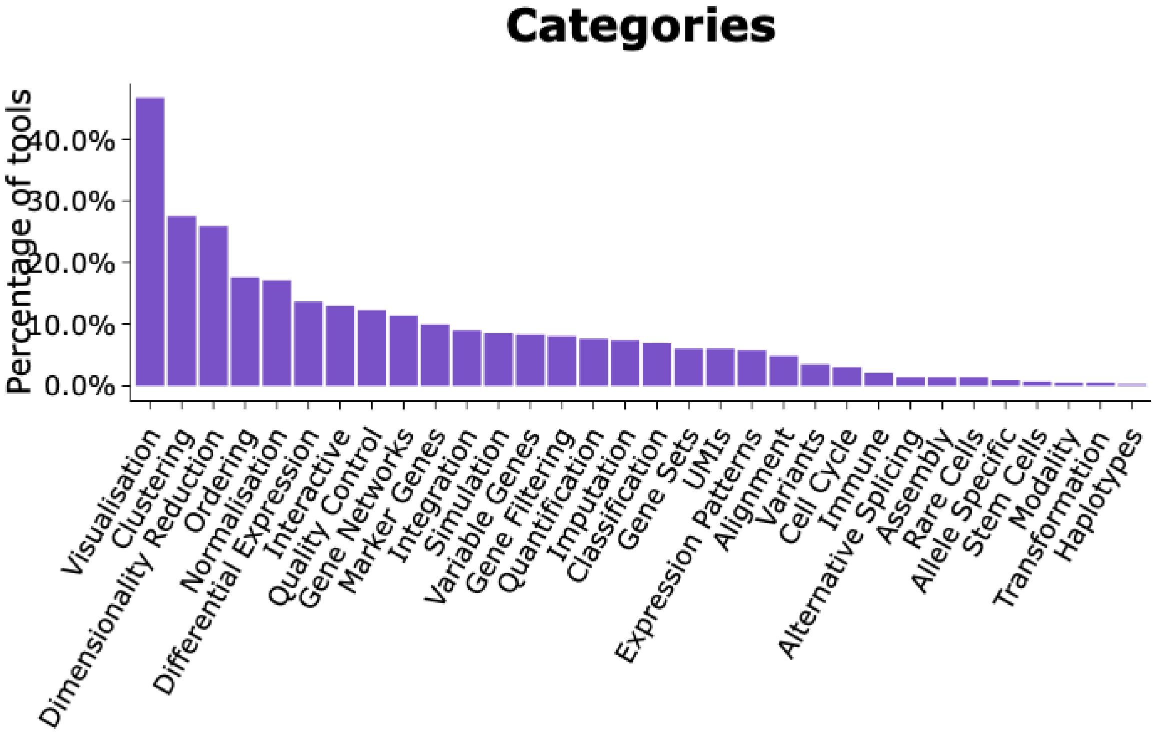
Application Note

CHROMIUM™

Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium Single Cell 3' Solution

Number of scRNAseq tools also increasing rapidly



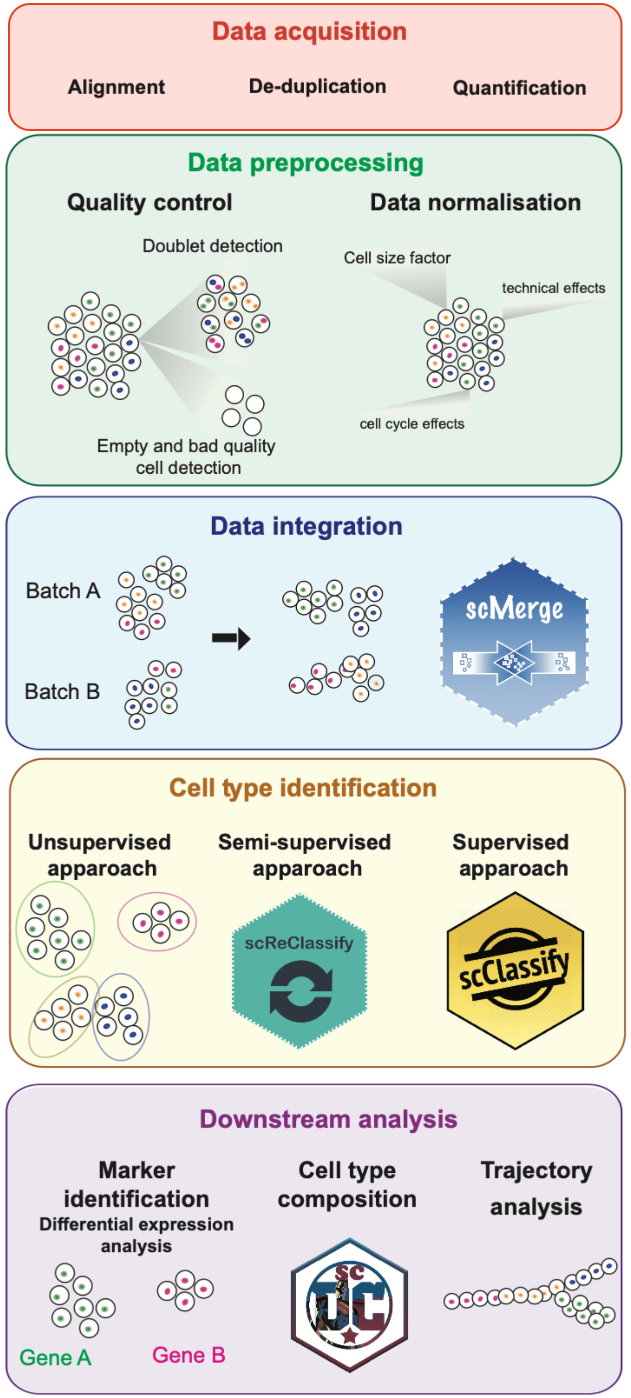


Downloaded from www.scrna-tools.org

Single-cell RNA-seq analysis

THE UNIVERSITY OF SYDNEY

Components of a typical scRNA-seq analysis process



Data acquisition

Alignment

De-duplication

Quantification

Data preprocessing

Quality control

Doublet detection

Empty and bad quality cell detection

Data normalisation

Cell size factor

technical effects

cell cycle effects

Data integration

Cell type identification

Unsupervised apparaoch

Semi-supervised apparaoch

Supervised apparaoch

Downstream analysis

Marker identification
Differential expression analysis

Cell type composition

Trajectory analysis

Component 1: Data acquisition

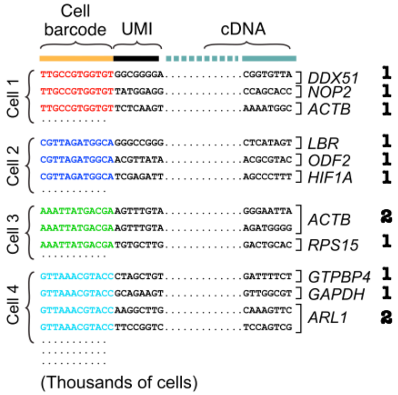
Data acquisition
Alignment
De-duplication
Quantification

Input

- BCL or fastq file from the sequencer

Output

- Gene/cell counts matrix


Cell barcode
UMI
cDNA
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTTA
TTGCCGTGGTGTTATGGAGG............CCAGCACC
TTGCCGTGGTGTTCTCAAGT............AAAATGGC
DDX51 1
NOP2 1
ACTB 1
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAGT
CGTTAGATGGCAACGTTATA............ACGCGTAC
CGTTAGATGGCATCGAGATT............AGCCCTTT
LBR 1
ODF2 1
HIF1A 1
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATTA
AAATTATGACGAAGTTTGTA............AGATGGGG
AAATTATGACGATGTGCTTG............GACTGCAC
ACTB 2
RPS15 1
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTCT
GTTAAACGTACCGCAGAAGT............GTTGGCGT
GTTAAACGTACCAAGGCTTG............CAAAGTTC
GTTAAACGTACCTTCCGGTC............TCCAGTCG
GTPBP4 1
GAPDH 1
ARL1 2
(Thousands of cells)

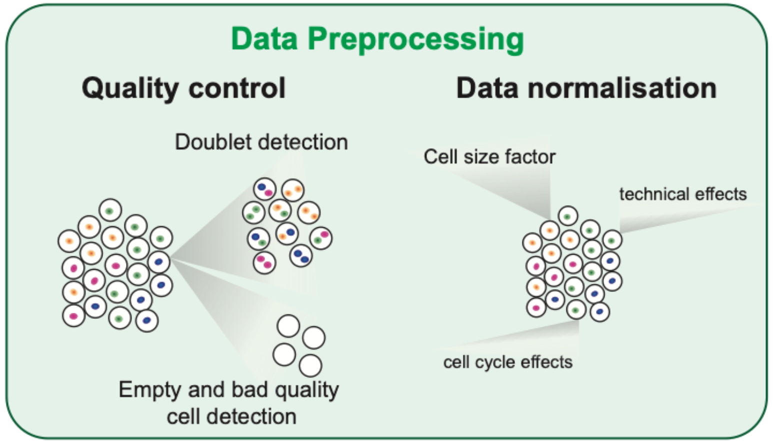
Software

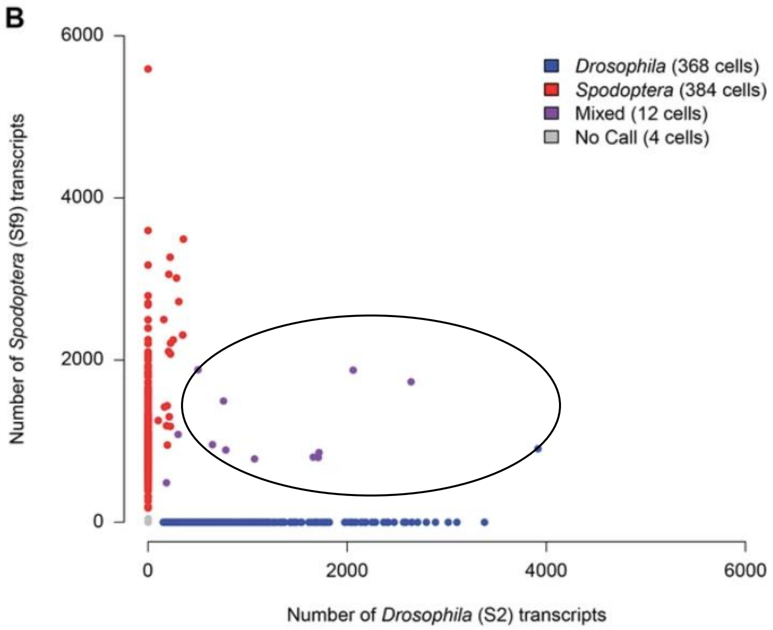
- CellRanger for 10X Genomics data
- Macosko’s custom scripts for DropSeq data
- STAR for alignment plus custom scripts (or there is STAR-solo)

Considerations

- Single or mix of species? Does it include ERCC spike-ins? May need to build a custom reference
- Barcode and/or UMI sequencing errors – CellRanger takes care of this automatically
- Align to exon or exon and intron?

Component 2: Data preprocessing – Quality control





Croset (2018), eLife

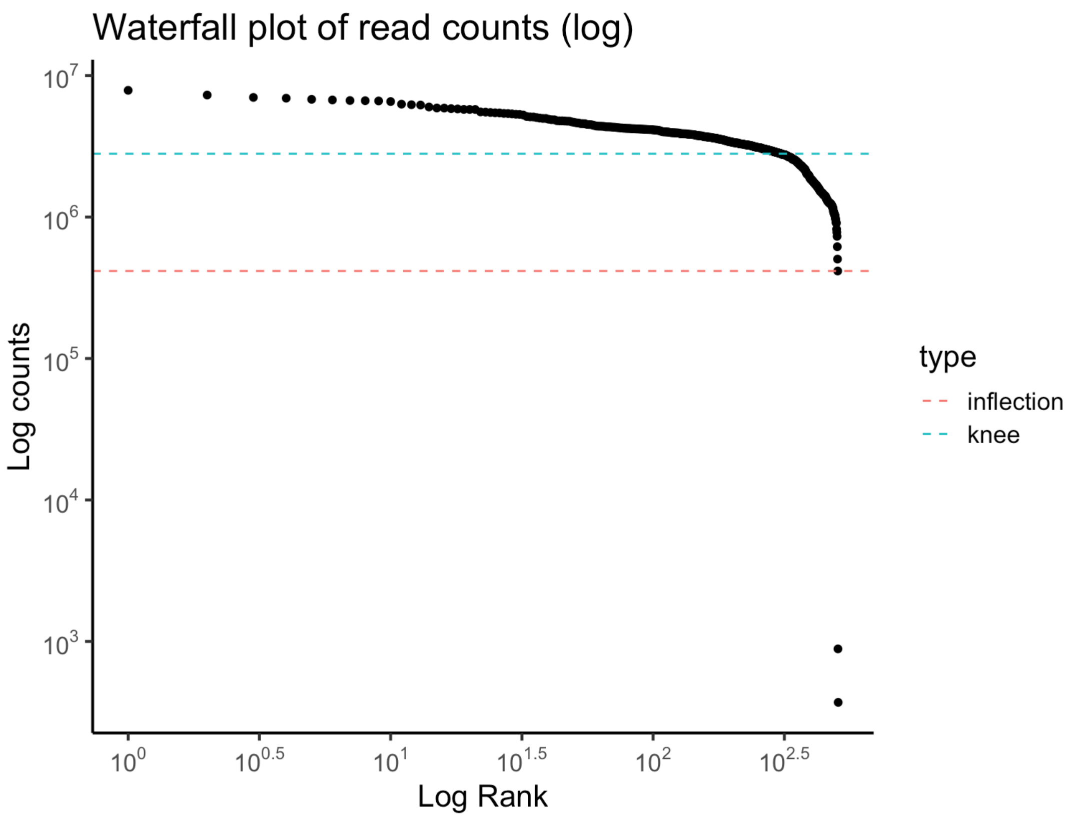
Software

- Seurat (all-purpose single cell R package)
- Scater
- DropletUtils (R package with a number of handy utility functions)
- Your own custom scripts

Considerations

- Filter out droplets with doublets – may be difficult to find. Can estimate expected rate by doing species mixture experiment

Component 2: Data preprocessing – Quality control



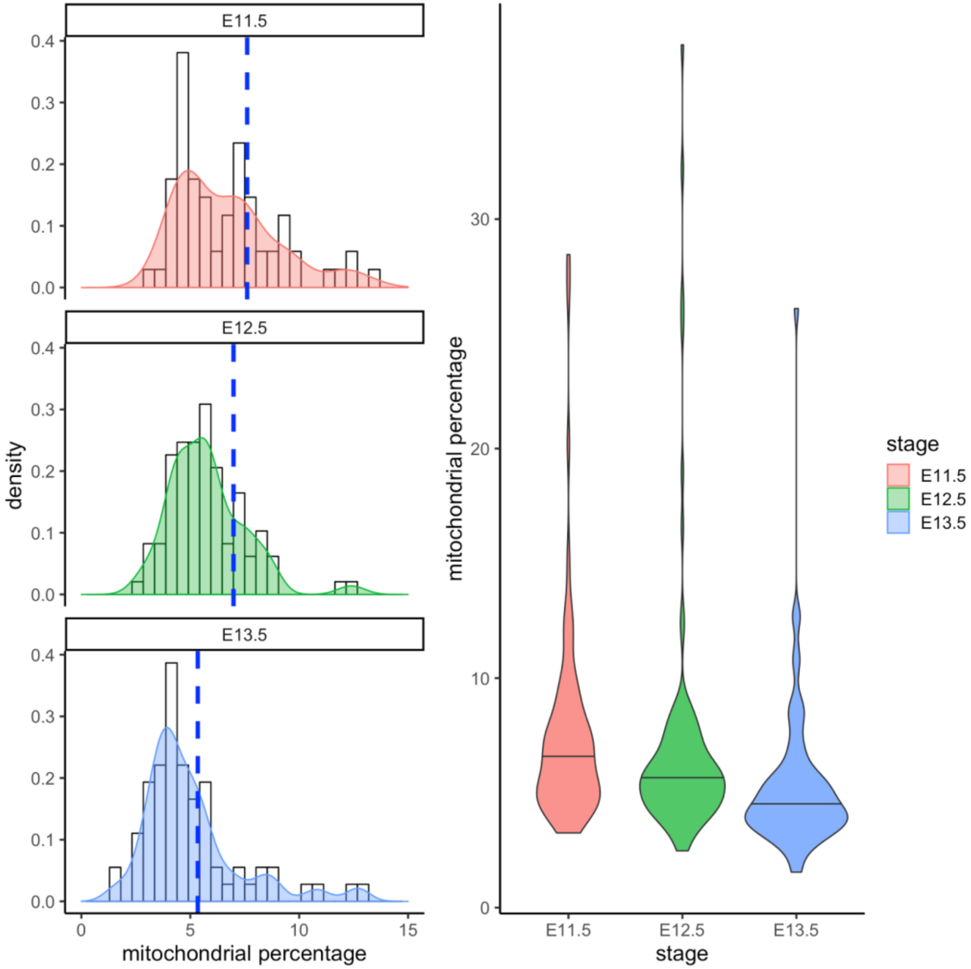
Software

- Seurat (all-purpose single cell R package)
- Scater
- DropletUtils (R package with a number of handy utility functions)
- Your own custom scripts

Considerations

- Filter out droplets with doublets – may be difficult to find. Can estimate expected rate by doing species mixture experiment
- Filter out droplets with no cells

Component 2: Data preprocessing – Quality control



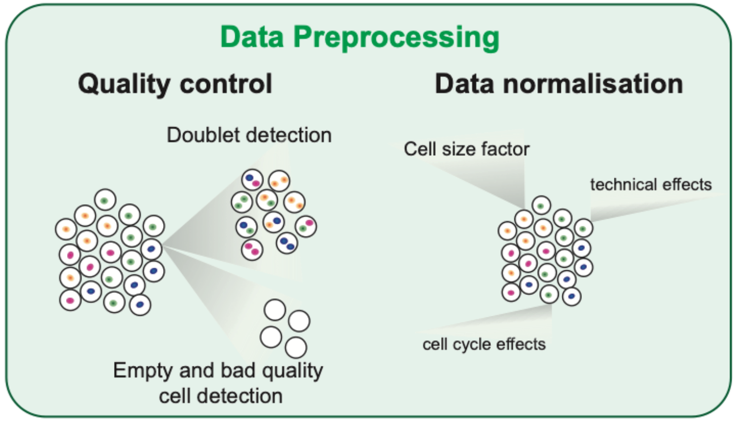
Software

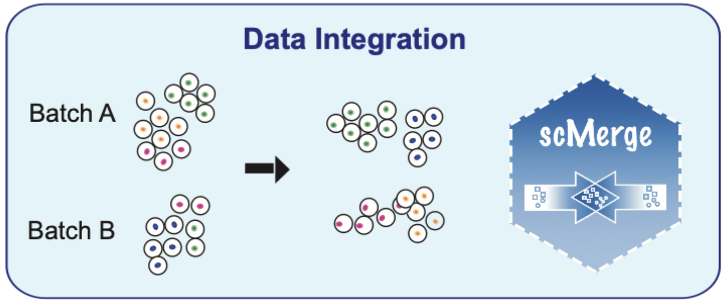
- Seurat (all-purpose single cell R package)
- Scater
- DropletUtils (R package with a number of handy utility functions)
- Your own custom scripts

Considerations

- Filter out droplets with doublets – may be difficult to find. Can estimate expected rate by doing species mixture experiment
- Filter out droplets with no cells
- Filter out droplets with damaged cells – look for high mitochondrial gene content or high spike-in

Component 3: Data integration



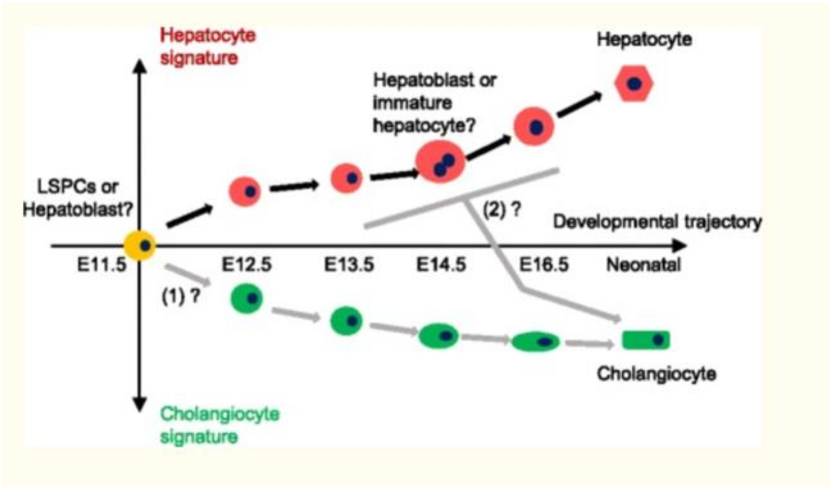


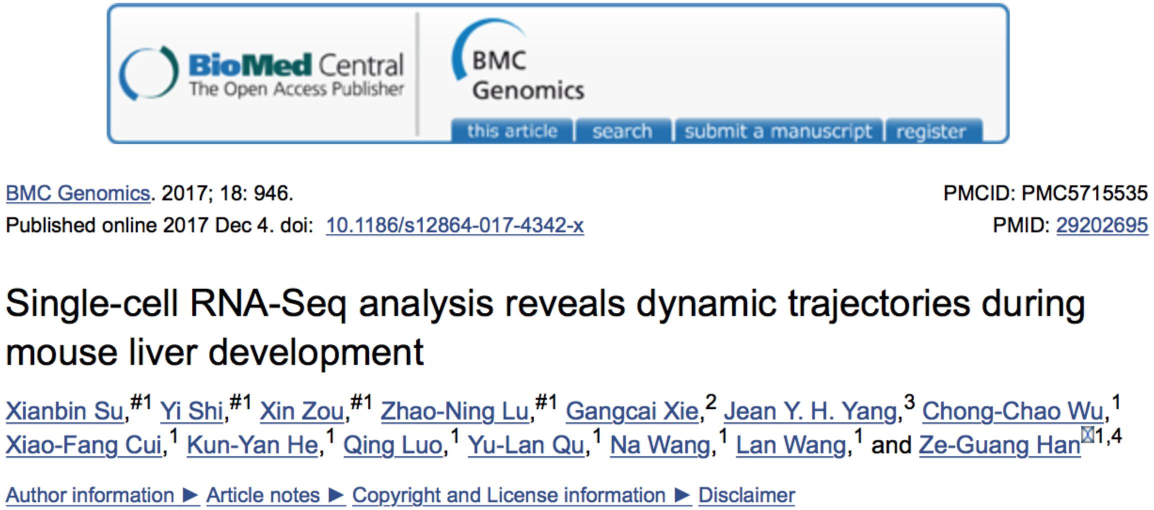
Software

- Seurat (all-purpose single cell R package) for very basic normalization
- Batch effect correction
  - mnnCorrect
  - ZINB-Wave
  - **scMerge**

scMerge motivation - Liver fetal development time course dataset

	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	E16.5	E17.5
GSE87795 Su et al.			●	●	●	●		●	





BMC Genomics. 2017; 18: 946.
Published online 2017 Dec 4. doi: 10.1186/s12864-017-4342-x

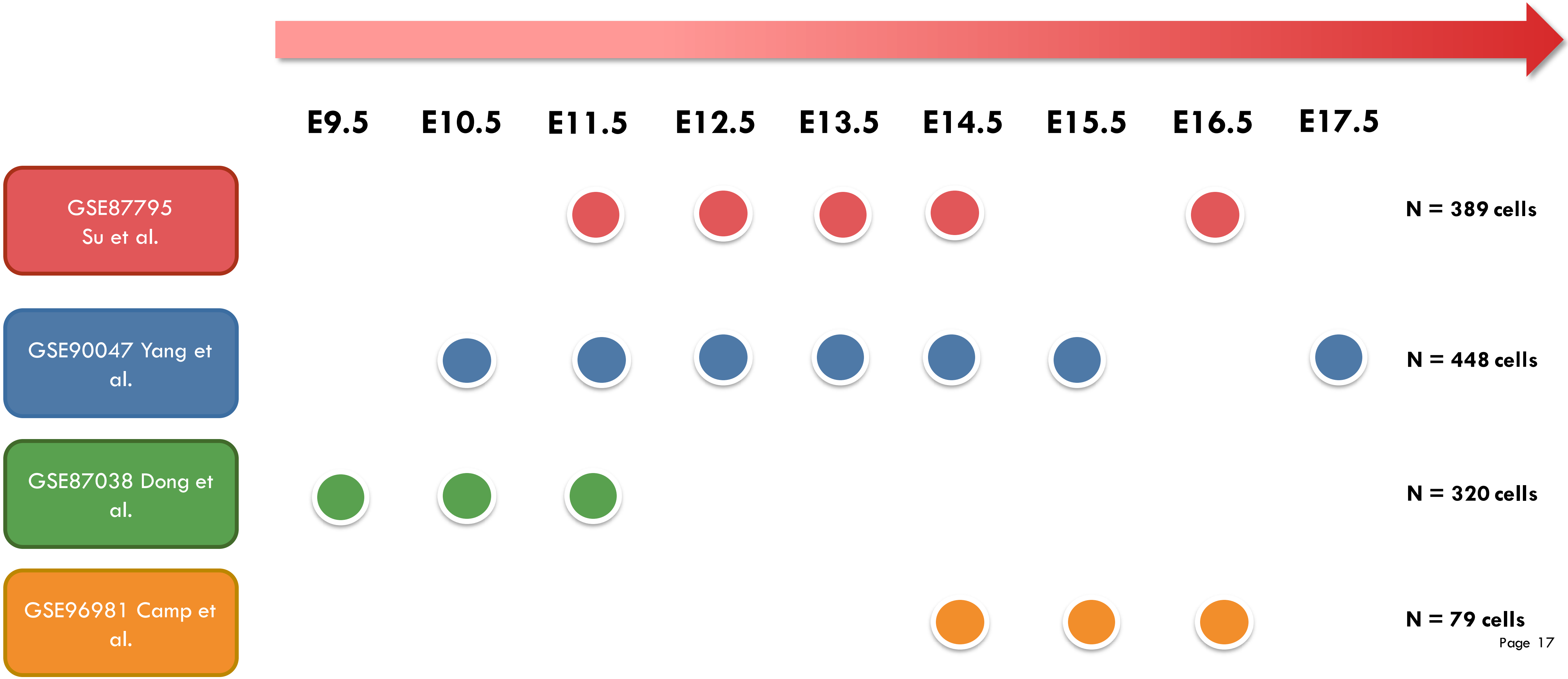
PMCID: PMC5715535
PMID: 29202695

Single-cell RNA-Seq analysis reveals dynamic trajectories during mouse liver development

Xianbin Su,[#1] Yi Shi,[#1] Xin Zou,[#1] Zhao-Ning Lu,[#1] Gangcai Xie,[2] Jean Y. H. Yang,[3] Chong-Chao Wu,[1] Xiao-Fang Cui,[1] Kun-Yan He,[1] Qing Luo,[1] Yu-Lan Qu,[1] Na Wang,[1] Lan Wang,[1] and Ze-Guang Han✉[1,4]

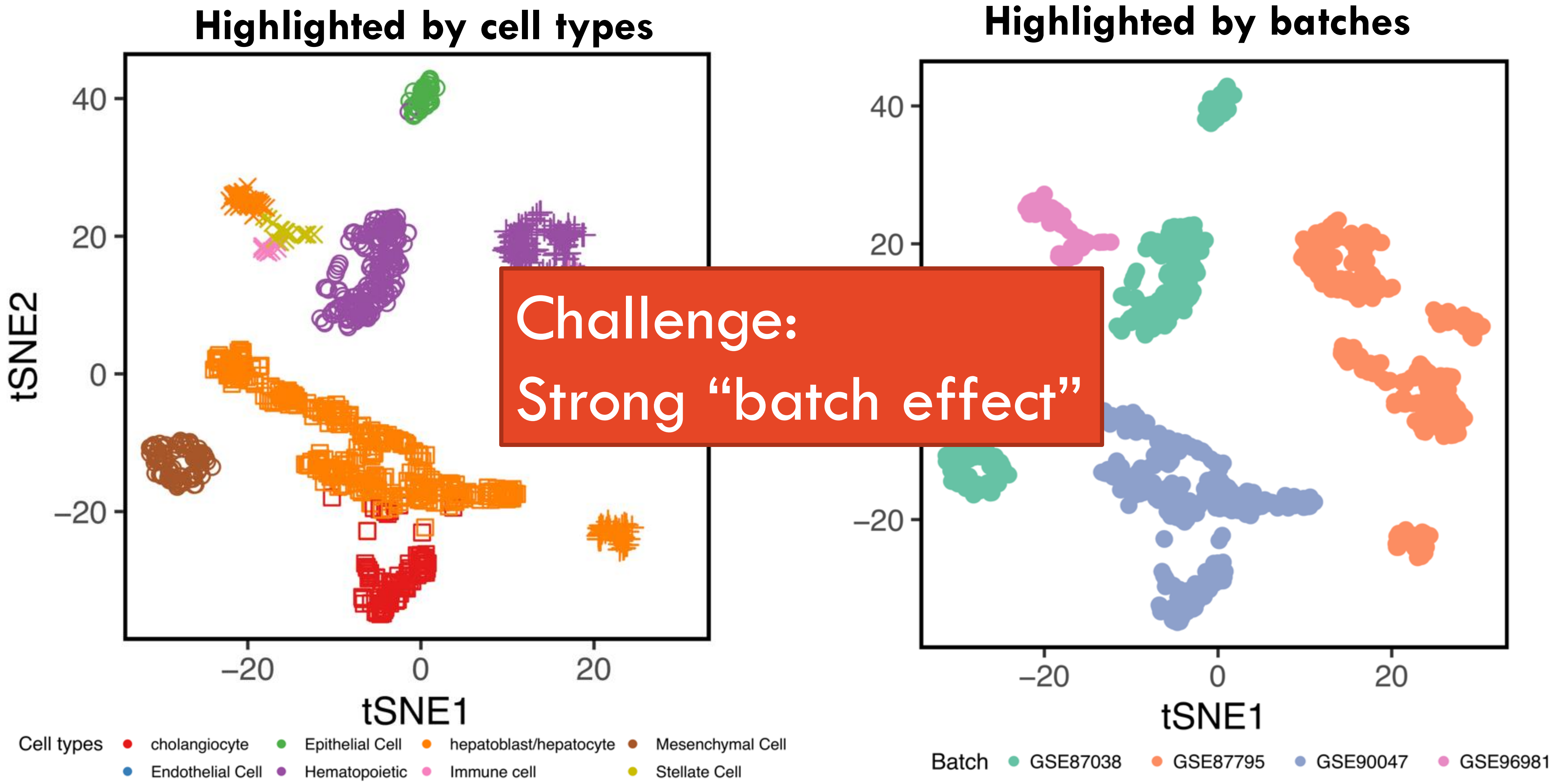
Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Liver fetal development time course datasets



	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	E16.5	E17.5	
GSE87795 Su et al.			●	●	●	●		●		N = 389 cells
GSE90047 Yang et al.		●	●	●	●	●	●		●	N = 448 cells
GSE87038 Dong et al.	●	●	●							N = 320 cells
GSE96981 Camp et al.						●	●	●		N = 79 cells

tSNE of liver fetal development time course datasets



Challenge:
Strong "batch effect"

Breaking observed data into components

For *n* cells with data collected for *m* genes

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{W\alpha} + \boldsymbol{\epsilon}$$

The data we observe

Biologically relevant variation
cell types
p *wanted* variables

Unwanted variation
batch and technical effects
k *unwanted* variables

Random noise

scMerge algorithm

Estimated by **stably expressed genes** by factor analysis

$$Y = X\beta + W\alpha + \epsilon$$

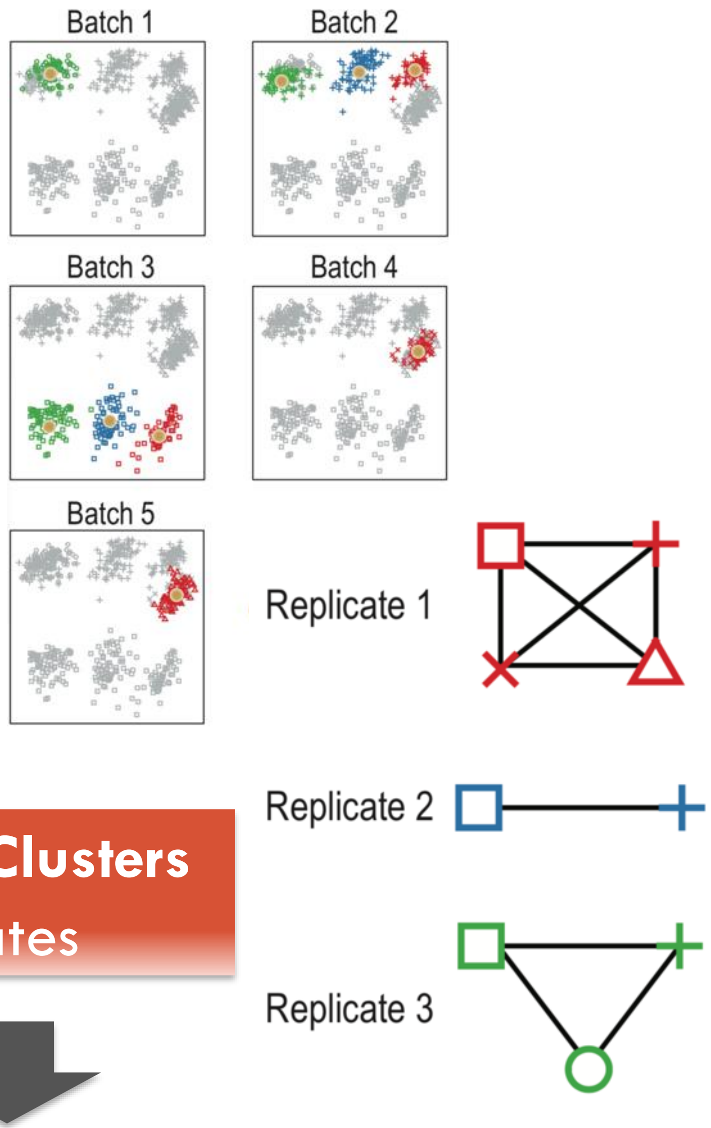
Estimated with **replicates** by factor analysis

RUVIII algorithm Molania et al. (2019), Nuclei Acids Res

scMerge algorithm

Clustering for each batch

(k-means by default)



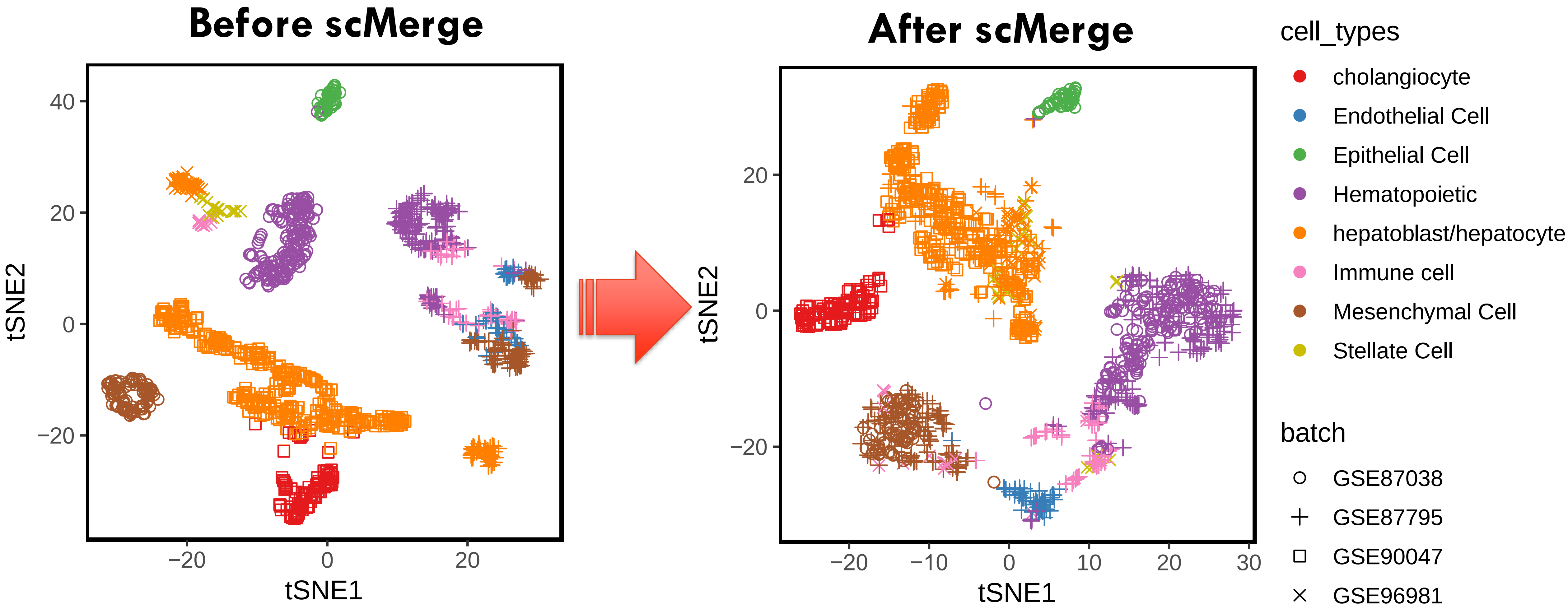
Find **Mutual Nearest Clusters** as pseudo-replicates

Frame as pseudo-replicate information

Pseudo-replicates

$$\begin{array}{c|ccc|}
\text{Cell 1} & 1 & 0 & 0 \\
\text{Cell 2} & 1 & 0 & 0 \\
\text{Cell 3} & 0 & 1 & 0 \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\text{Cell C} & 0 & 0 & 1 \\
 & \text{Replicate 1} & \text{Replicate 2} & \text{Replicate 3}
\end{array}$$

Coming back to our motivational data – Liver fetal development time course datasets



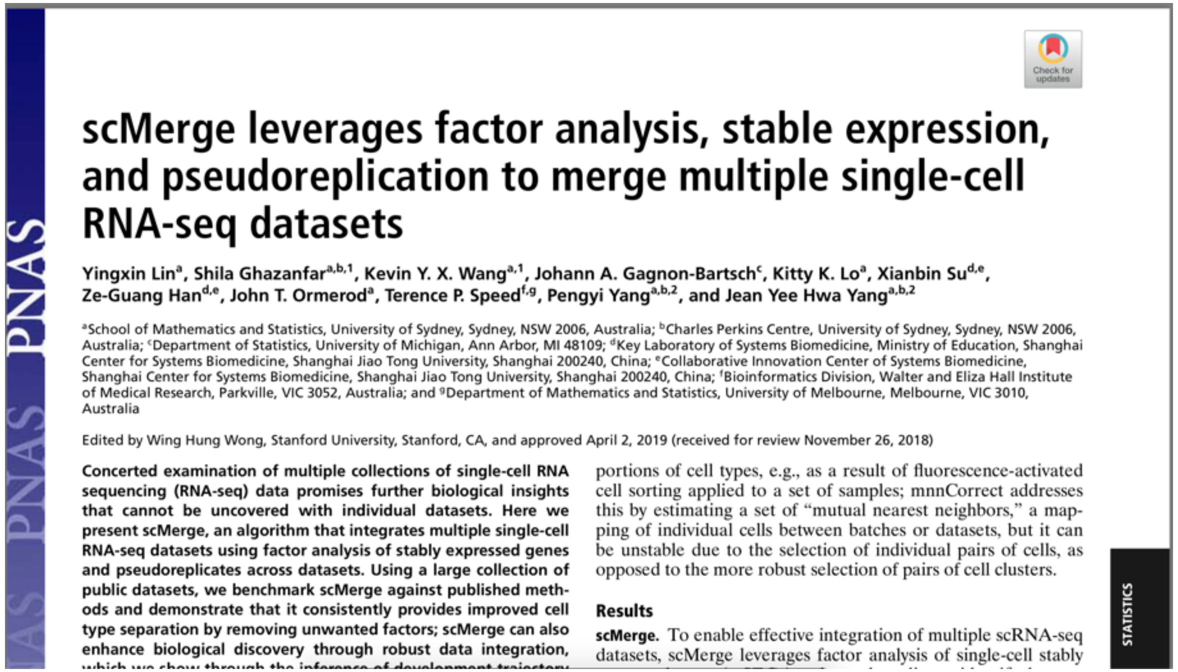
Before scMerge

After scMerge

More information

PNAS:

https://doi.org/10.1073/pnas.1820006116



Check for updates

STATISTICS

scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets

Yingxin Lin[a], Shila Ghazanfar[a,b,1], Kevin Y. X. Wang[a,1], Johann A. Gagnon-Bartsch[c], Kitty K. Lo[a], Xianbin Su[d,e], Ze-Guang Han[d,e], John T. Ormerod[a], Terence P. Speed[f,g], Pengyi Yang[a,b,2], and Jean Yee Hwa Yang[a,b,2]

[a]School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia; [b]Charles Perkins Centre, University of Sydney, Sydney, NSW 2006, Australia; [c]Department of Statistics, University of Michigan, Ann Arbor, MI 48109; [d]Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China; [e]Collaborative Innovation Center of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China; [f]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; and [g]Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved April 2, 2019 (received for review November 26, 2018)

Concerted examination of multiple collections of single-cell RNA sequencing (RNA-seq) data promises further biological insights that cannot be uncovered with individual datasets. Here we present scMerge, an algorithm that integrates multiple single-cell RNA-seq datasets using factor analysis of stably expressed genes and pseudoreplicates across datasets. Using a large collection of public datasets, we benchmark scMerge against published methods and demonstrate that it consistently provides improved cell type separation by removing unwanted factors; scMerge can also enhance biological discovery through robust data integration, which we show through the inference of development trajectory

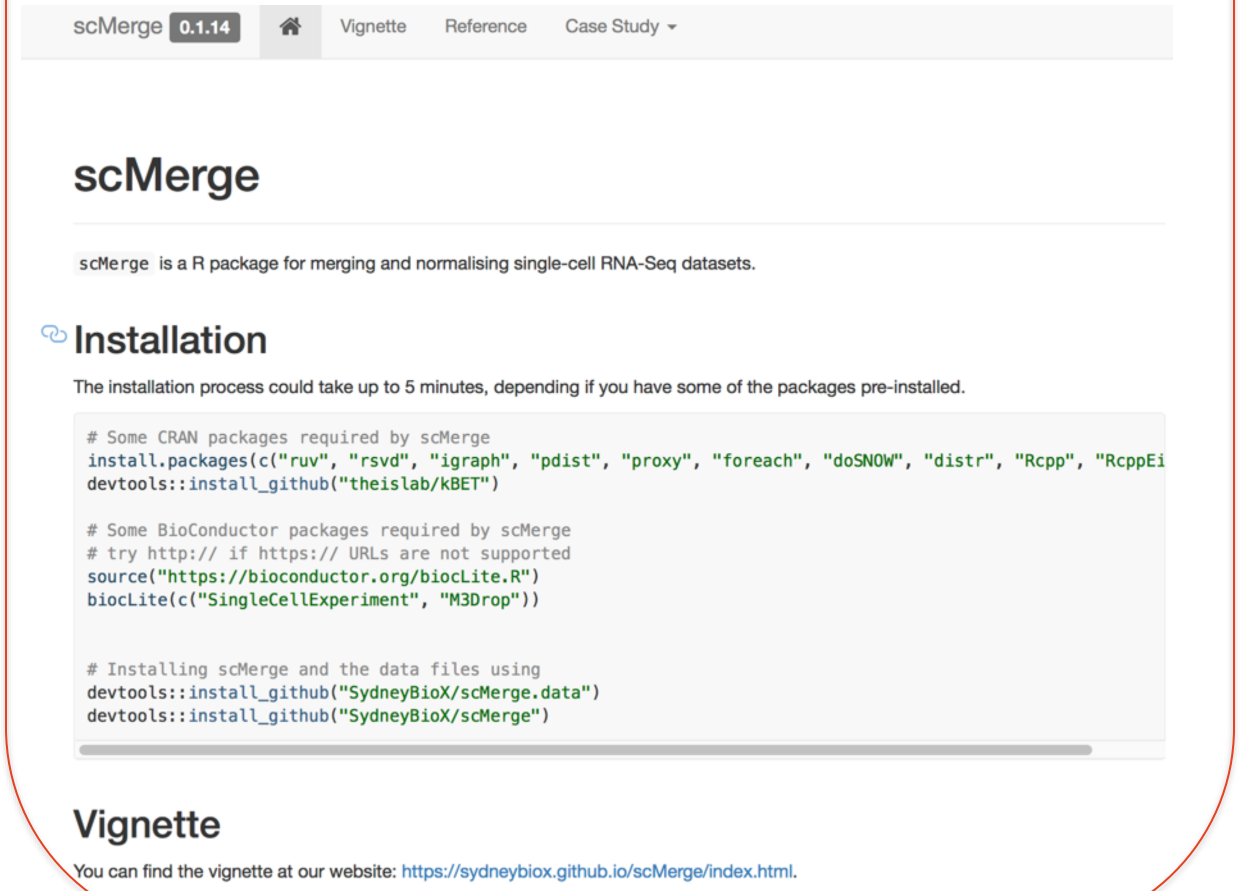
portions of cell types, e.g., as a result of fluorescence-activated cell sorting applied to a set of samples; mnnCorrect addresses this by estimating a set of "mutual nearest neighbors," a mapping of individual cells between batches or datasets, but it can be unstable due to the selection of individual pairs of cells, as opposed to the more robust selection of pairs of cell clusters.

Results

scMerge. To enable effective integration of multiple scRNA-seq datasets, scMerge leverages factor analysis of single-cell stably

scMerge R package and website:

https://sydneybiox.github.io/scMerge/



scMerge

scMerge is a R package for merging and normalising single-cell RNA-Seq datasets.

Installation

The installation process could take up to 5 minutes, depending if you have some of the packages pre-installed.

```
# Some CRAN packages required by scMerge
install.packages(c("ruv", "rsvd", "igraph", "pdist", "proxy", "foreach", "doSNOW", "distr", "Rcpp", "RcppEi
devtools::install_github("theislab/kBET")

# Some BioConductor packages required by scMerge
# try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite(c("SingleCellExperiment", "M3Drop"))


# Installing scMerge and the data files using
devtools::install_github("SydneyBioX/scMerge.data")
devtools::install_github("SydneyBioX/scMerge")
```

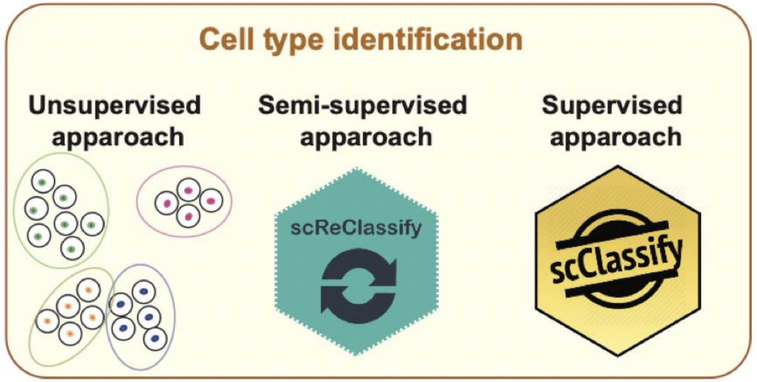
Vignette

You can find the vignette at our website: https://sydneybiox.github.io/scMerge/index.html.

We will try this soon …

2:00 – 2:45 Quality control and data integration

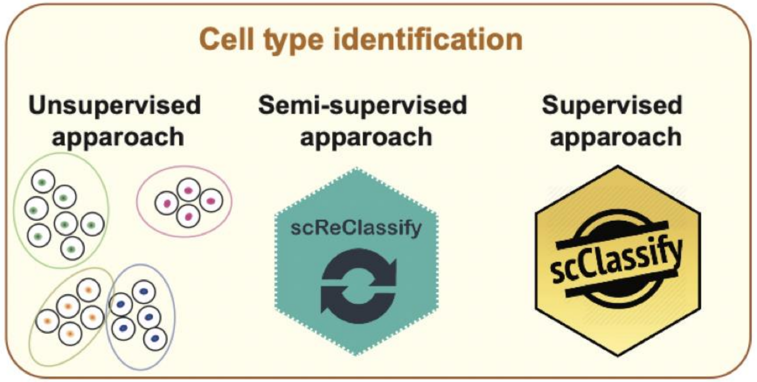
Component 4: Cell type identification



Science questions

- What cell types are present in the dataset?
- Can we identify the cell types?

Phase 3: Cell assignment



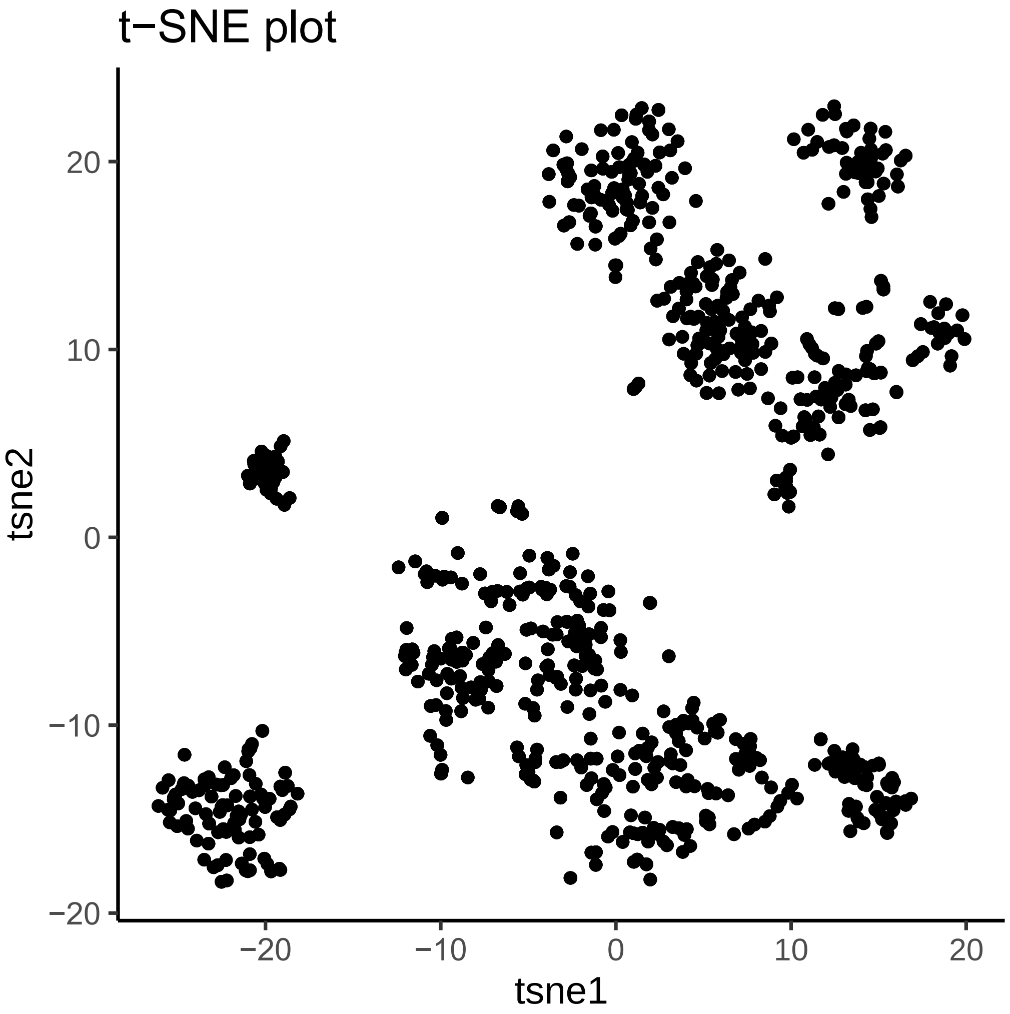
Science questions

- What cell types are present in the dataset?
- Can we identify the cell types?

Analysis techniques

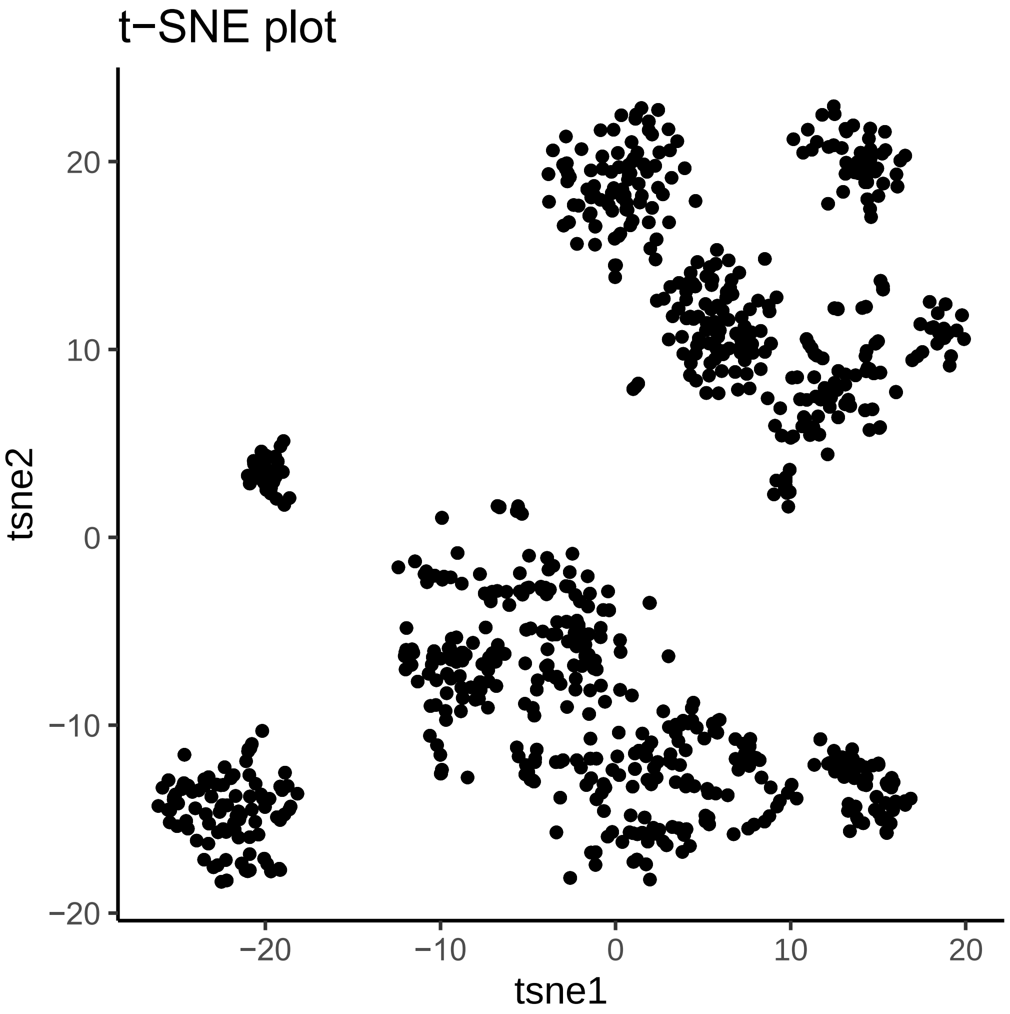
- Visualization (dimension reduction)
- Clustering (unsupervised learning)
- Classification (supervised learning)

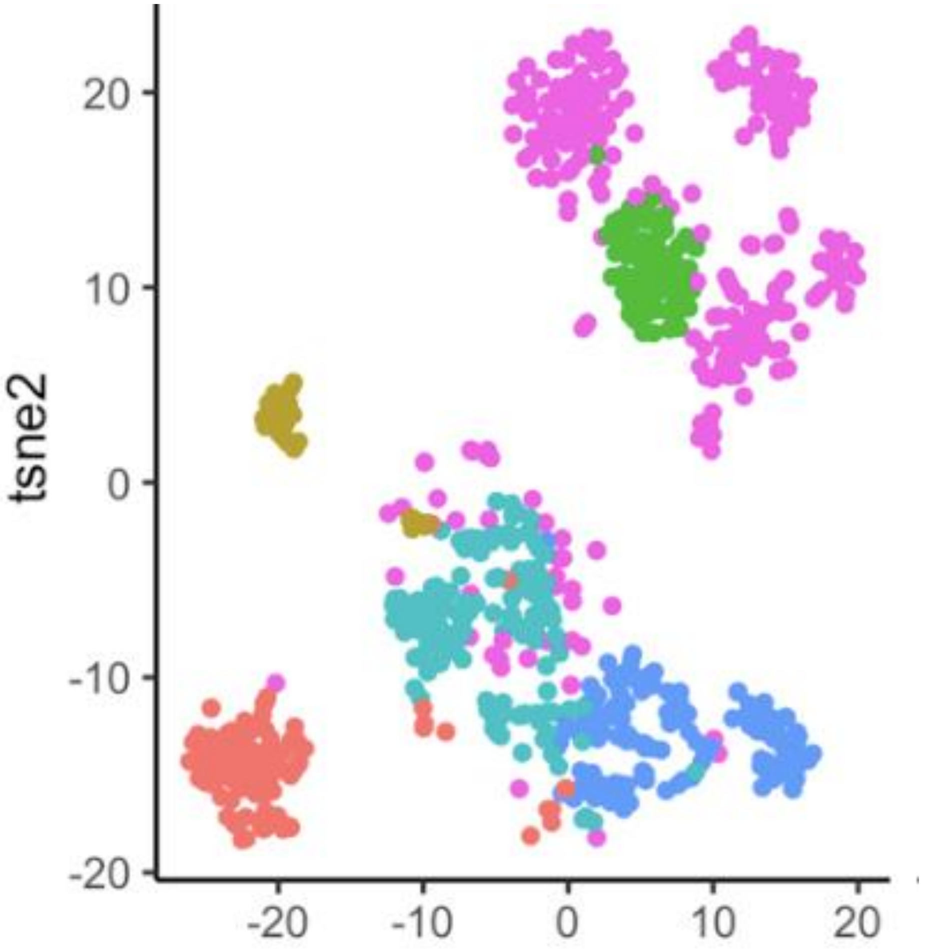
Dimension reduced plot of our data (tSNE plot)



How many cell types are there?
What are the cell types?

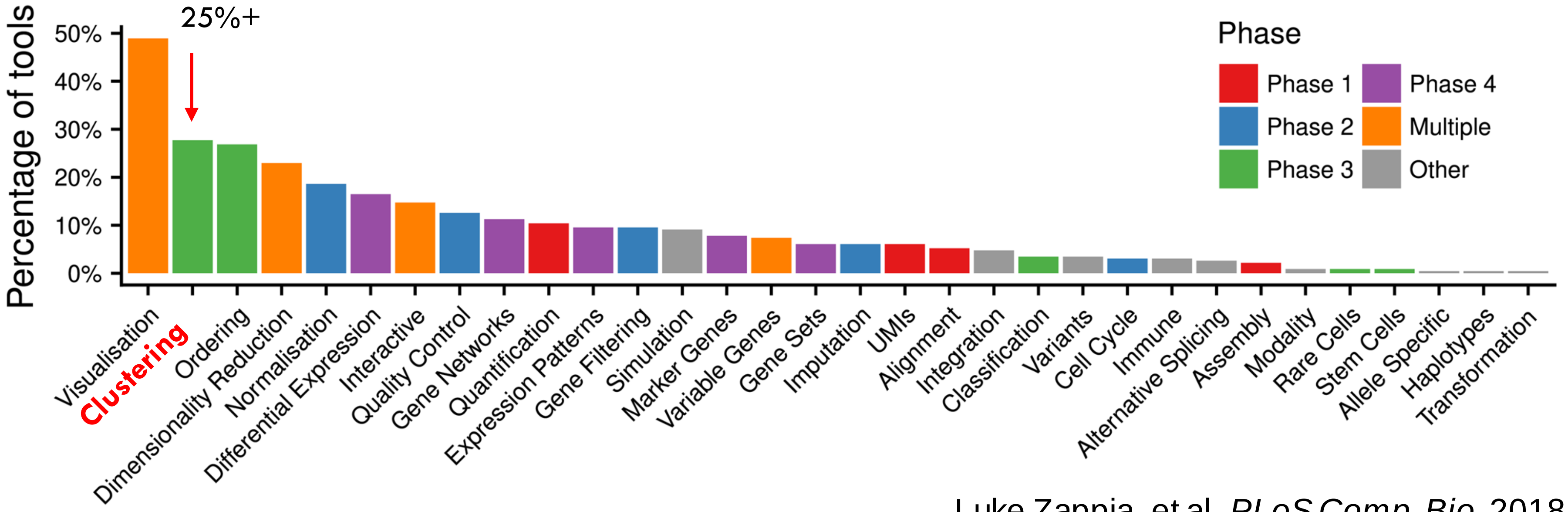
k-means clustering





Clustering algorithms for scRNA-seq

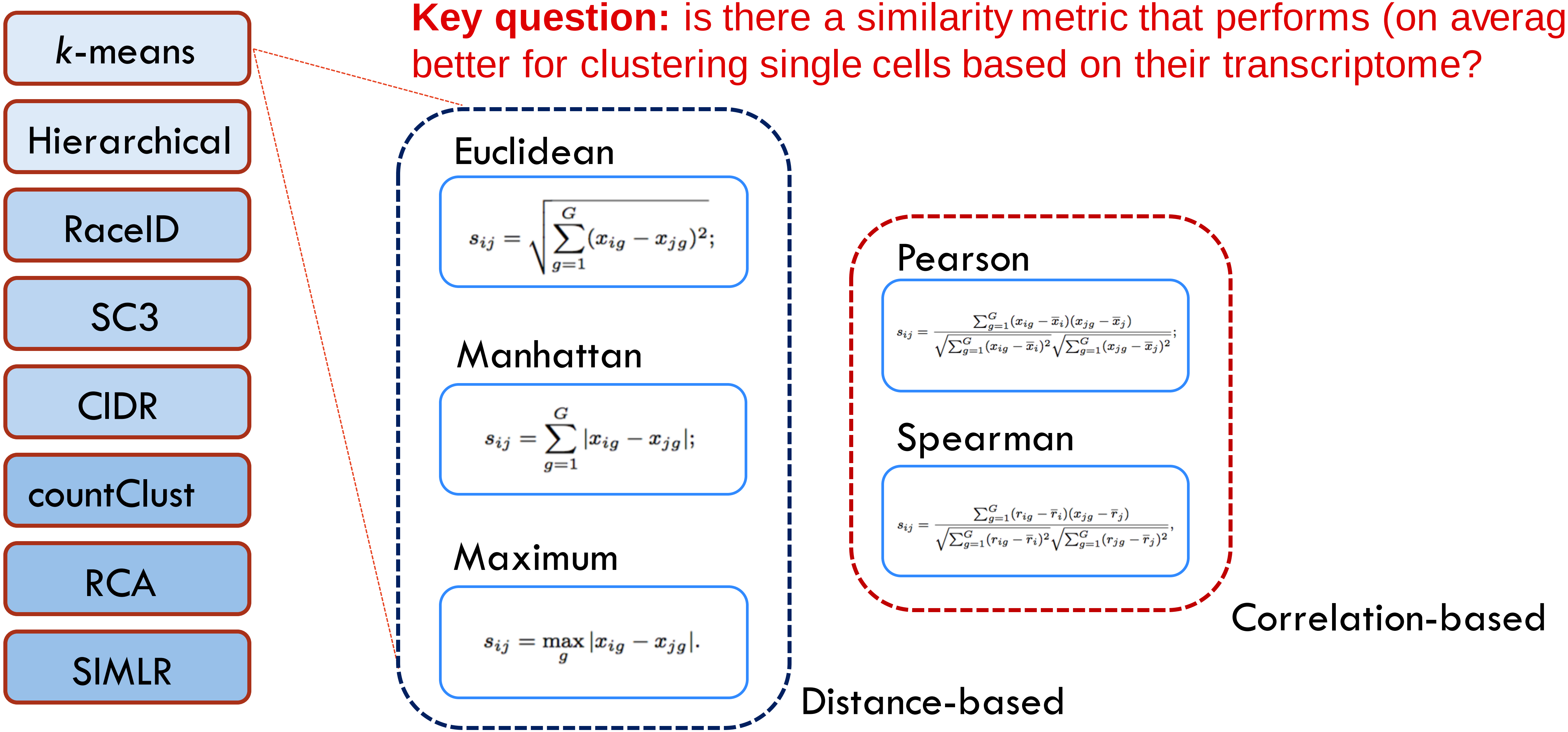
- *k*-means
- Hierarchical
- RaceID
- SC3
- CIDR
- countClust
- RCA
- SIMLR



Luke Zappia, et al. *PLoS Comp. Bio.* 2018

Similarity metric is the core of clustering algorithm

Key question: is there a similarity metric that performs (on average) better for clustering single cells based on their transcriptome?



k-means

Hierarchical

RaceID

SC3

CIDR

countClust

RCA

SIMLR

Euclidean

$$s_{ij} = \sqrt{\sum_{g=1}^{G} (x_{ig} - x_{jg})^2};$$

Manhattan

$$s_{ij} = \sum_{g=1}^{G} |x_{ig} - x_{jg}|;$$

Maximum

$$s_{ij} = \max_{g} |x_{ig} - x_{jg}|.$$

Distance-based

Pearson

$$s_{ij} = \frac{\sum_{g=1}^{G} (x_{ig} - \overline{x}_i)(x_{jg} - \overline{x}_j)}{\sqrt{\sum_{g=1}^{G} (x_{ig} - \overline{x}_i)^2} \sqrt{\sum_{g=1}^{G} (x_{jg} - \overline{x}_j)^2}};$$

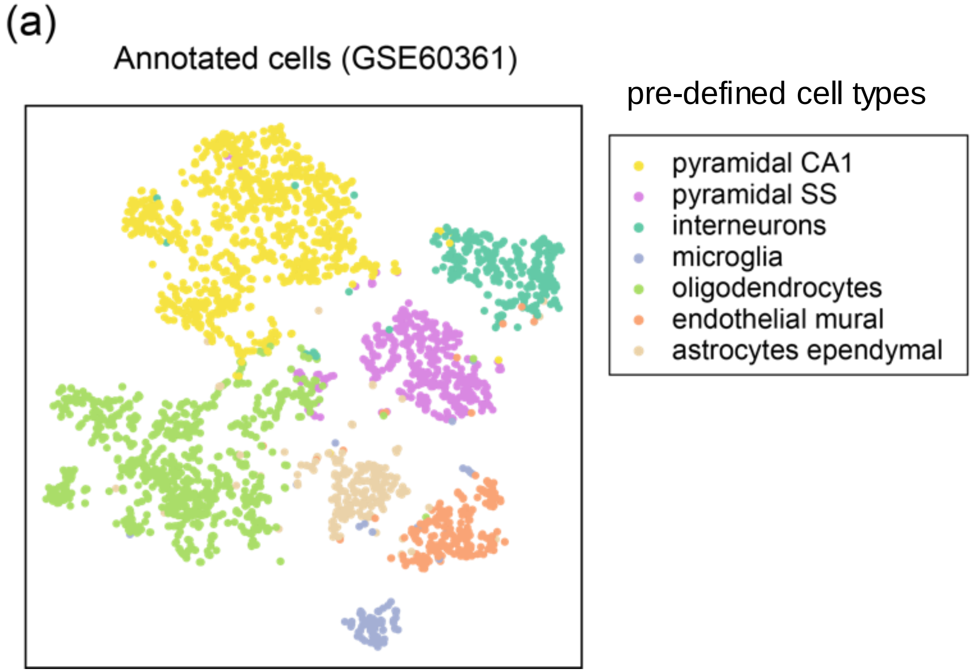
Spearman

$$s_{ij} = \frac{\sum_{g=1}^{G} (r_{ig} - \overline{r}_i)(x_{jg} - \overline{r}_j)}{\sqrt{\sum_{g=1}^{G} (r_{ig} - \overline{r}_i)^2} \sqrt{\sum_{g=1}^{G} (r_{jg} - \overline{r}_j)^2}},$$

Correlation-based

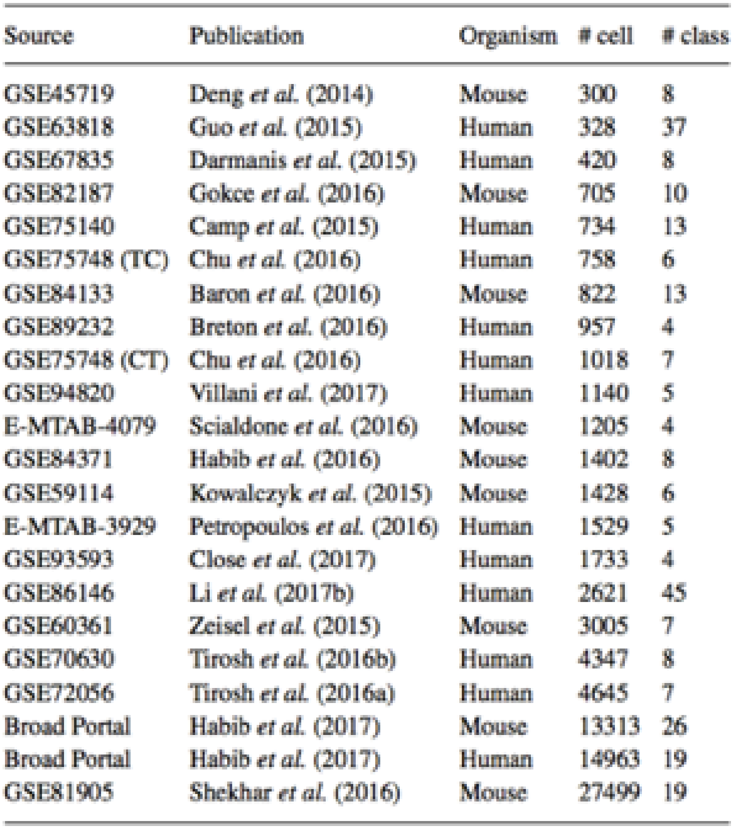
k-means Clustering on GSE60361

k-means

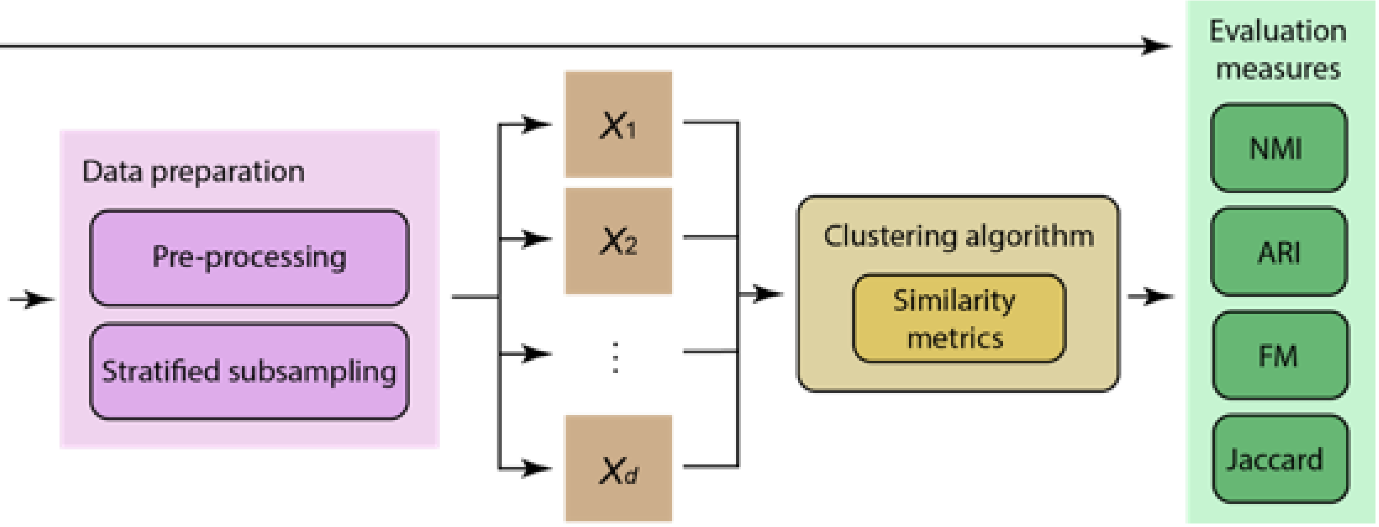


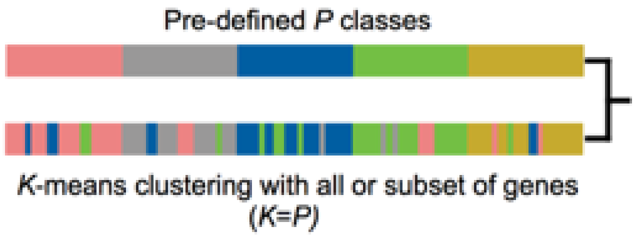
Zeisel A, et al. *Science* 2015

Evaluation framework



Source	Publication	Organism	# cell	# class
GSE45719	Deng *et al.* (2014)	Mouse	300	8
GSE63818	Guo *et al.* (2015)	Human	328	37
GSE67835	Darmanis *et al.* (2015)	Human	420	8
GSE82187	Gokce *et al.* (2016)	Mouse	705	10
GSE75140	Camp *et al.* (2015)	Human	734	13
GSE75748 (TC)	Chu *et al.* (2016)	Human	758	6
GSE84133	Baron *et al.* (2016)	Mouse	822	13
GSE89232	Breton *et al.* (2016)	Human	957	4
GSE75748 (CT)	Chu *et al.* (2016)	Human	1018	7
GSE94820	Villani *et al.* (2017)	Human	1140	5
E-MTAB-4079	Scialdone *et al.* (2016)	Mouse	1205	4
GSE84371	Habib *et al.* (2016)	Mouse	1402	8
GSE59114	Kowalczyk *et al.* (2015)	Mouse	1428	6
E-MTAB-3929	Petropoulos *et al.* (2016)	Human	1529	5
GSE93593	Close *et al.* (2017)	Human	1733	4
GSE86146	Li *et al.* (2017b)	Human	2621	45
GSE60361	Zeisel *et al.* (2015)	Mouse	3005	7
GSE70630	Tirosh *et al.* (2016b)	Human	4347	8
GSE72056	Tirosh *et al.* (2016a)	Human	4645	7
Broad Portal	Habib *et al.* (2017)	Mouse	13313	26
Broad Portal	Habib *et al.* (2017)	Human	14963	19
GSE81905	Shekhar *et al.* (2016)	Mouse	27499	19

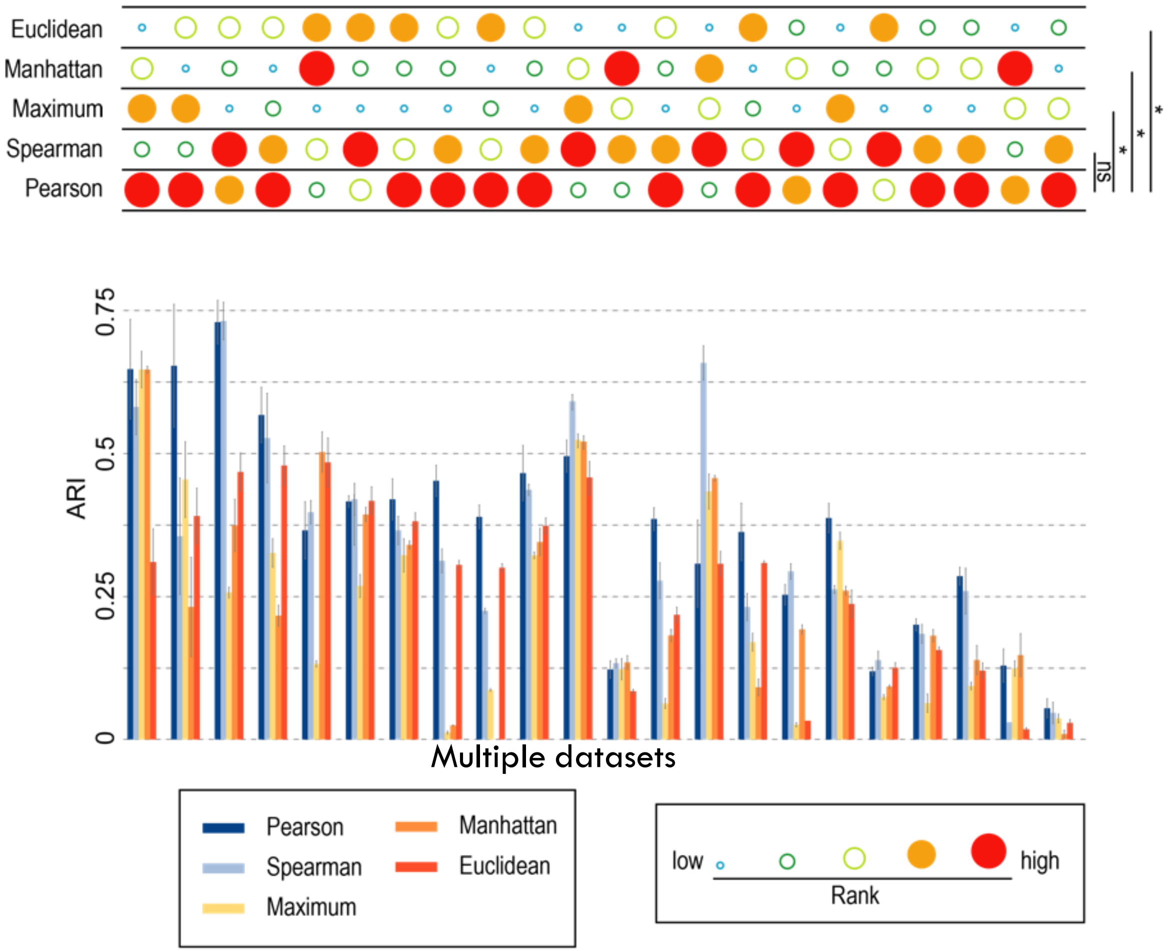




Agreement to pre-defined classes:
Normalized Mutual Information (NMI)
Adjusted Rand Index (ARI)
Fowlkes-Mallows Index (FM)
Jaccard Index (Jaccard)

Taiyun Kim

Evaluation results (against the pre-defined cell types)



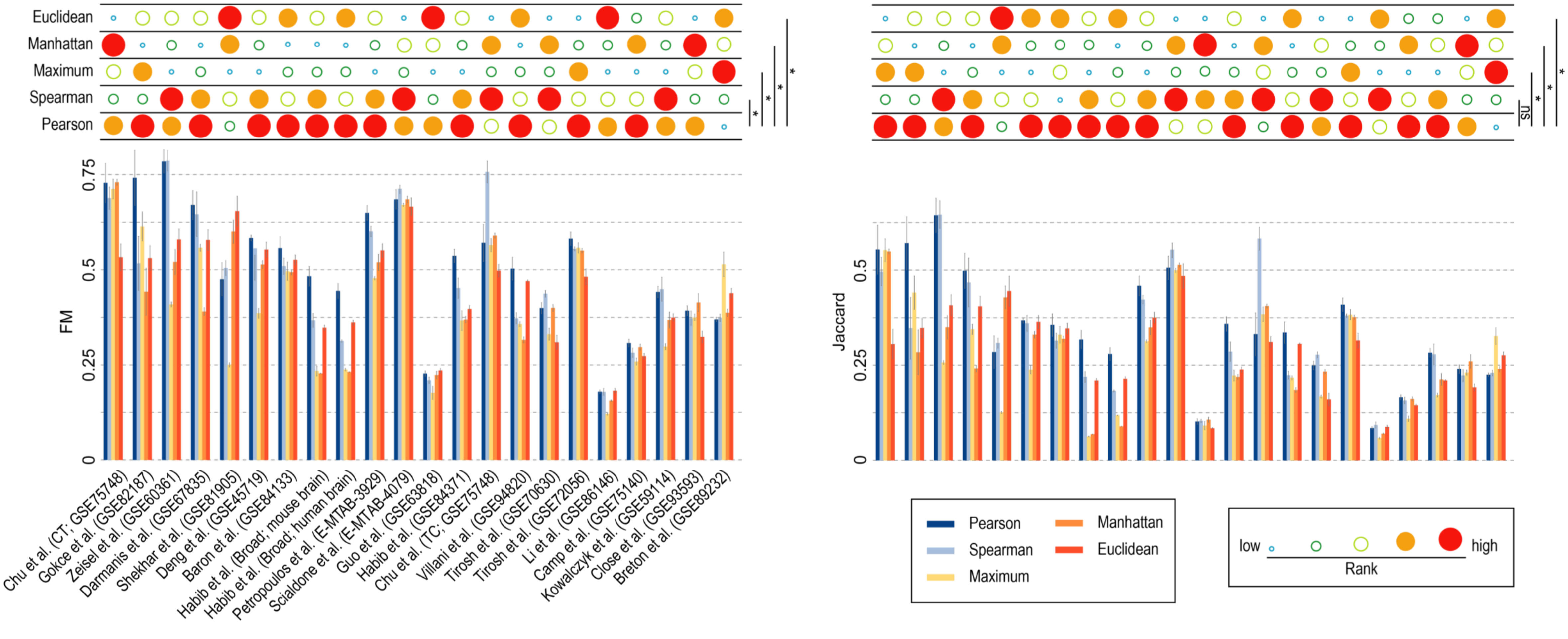
Impact of similarity metrics on single-cell RNA-seq data clustering

Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, Pengyi Yang

Briefings in Bioinformatics, bby076,

PhD student: Taiyun Kim

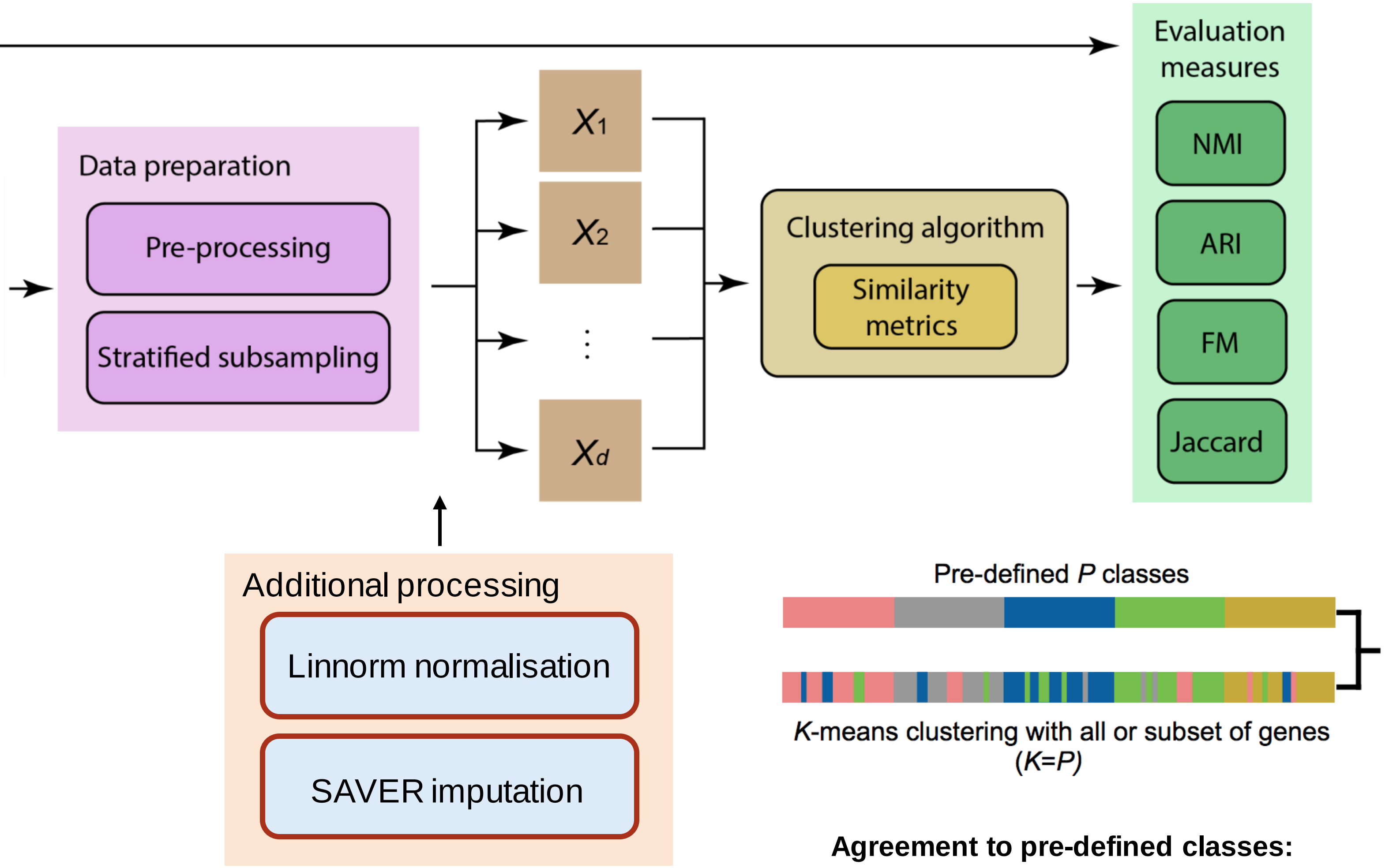
Evaluation results (against the pre-defined cell types) using other measures



On average, correlation-based metrics improved on distance-based metrics by 31.5% (NMI), 39.6% (ARI), 16% (FM), 23% (Jaccard)

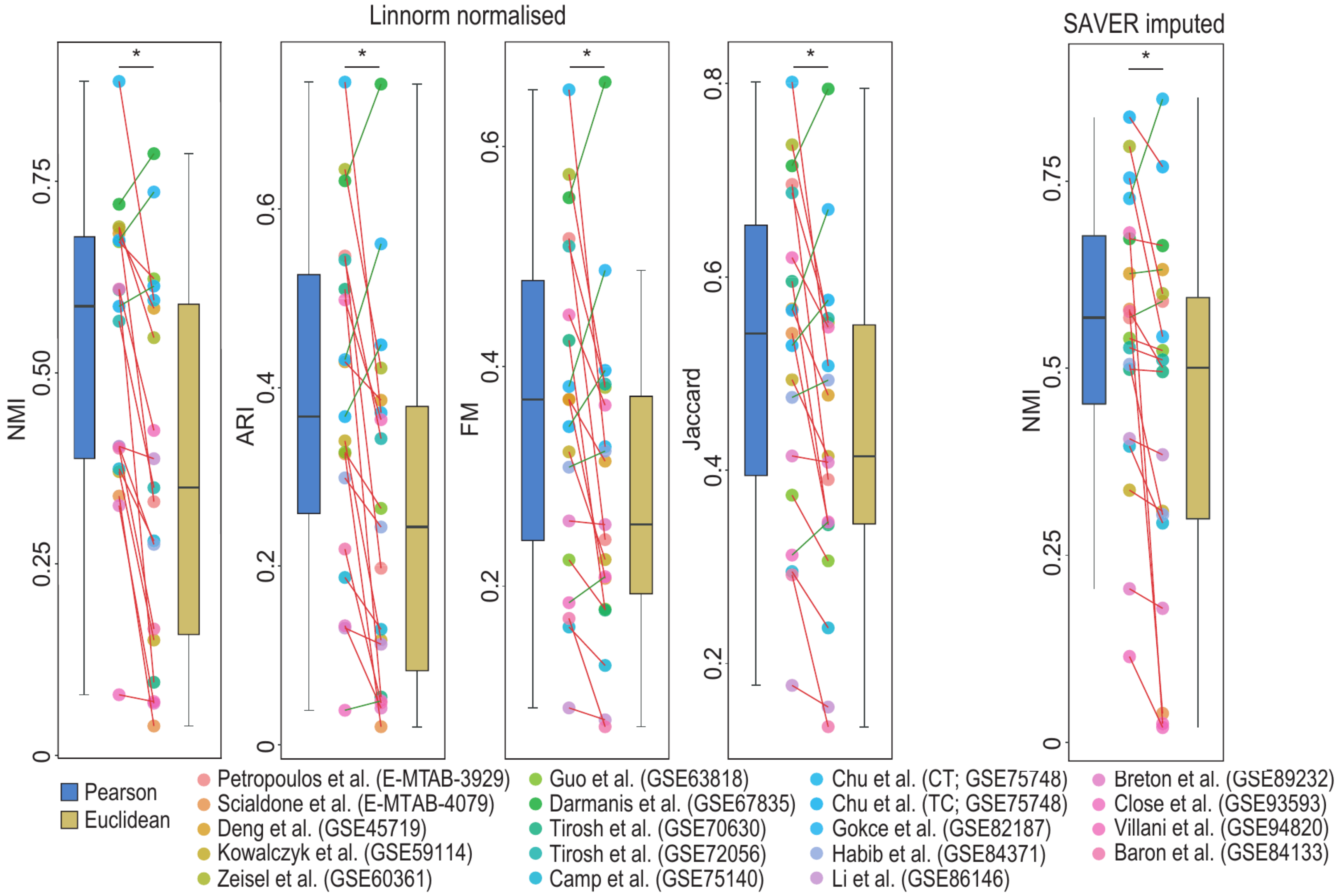
Account for data scaling and zero-counts

Source	Publication	Organism	# cell	# class
GSE45719	Deng *et al.* (2014)	Mouse	300	8
GSE63818	Guo *et al.* (2015)	Human	328	37
GSE67835	Darmanis *et al.* (2015)	Human	420	8
GSE82187	Gokce *et al.* (2016)	Mouse	705	10
GSE75140	Camp *et al.* (2015)	Human	734	13
GSE75748 (TC)	Chu *et al.* (2016)	Human	758	6
GSE84133	Baron *et al.* (2016)	Mouse	822	13
GSE89232	Breton *et al.* (2016)	Human	957	4
GSE75748 (CT)	Chu *et al.* (2016)	Human	1018	7
GSE94820	Villani *et al.* (2017)	Human	1140	5
E-MTAB-4079	Scialdone *et al.* (2016)	Mouse	1205	4
GSE84371	Habib *et al.* (2016)	Mouse	1402	8
GSE59114	Kowalczyk *et al.* (2015)	Mouse	1428	6
E-MTAB-3929	Petropoulos *et al.* (2016)	Human	1529	5
GSE93593	Close *et al.* (2017)	Human	1733	4
GSE86146	Li *et al.* (2017b)	Human	2621	45
GSE60361	Zeisel *et al.* (2015)	Mouse	3005	7
GSE70630	Tirosh *et al.* (2016b)	Human	4347	8
GSE72056	Tirosh *et al.* (2016a)	Human	4645	7
Broad Portal	Habib *et al.* (2017)	Mouse	13313	26
Broad Portal	Habib *et al.* (2017)	Human	14963	19
GSE81905	Shekhar *et al.* (2016)	Mouse	27499	19



Agreement to pre-defined classes:
Normalized Mutual Information (NMI)
Adjusted Rand Index (ARI)
Fowlkes-Mallows Index (FM)
Jaccard Index (Jaccard)

Account for normalisation and imputation

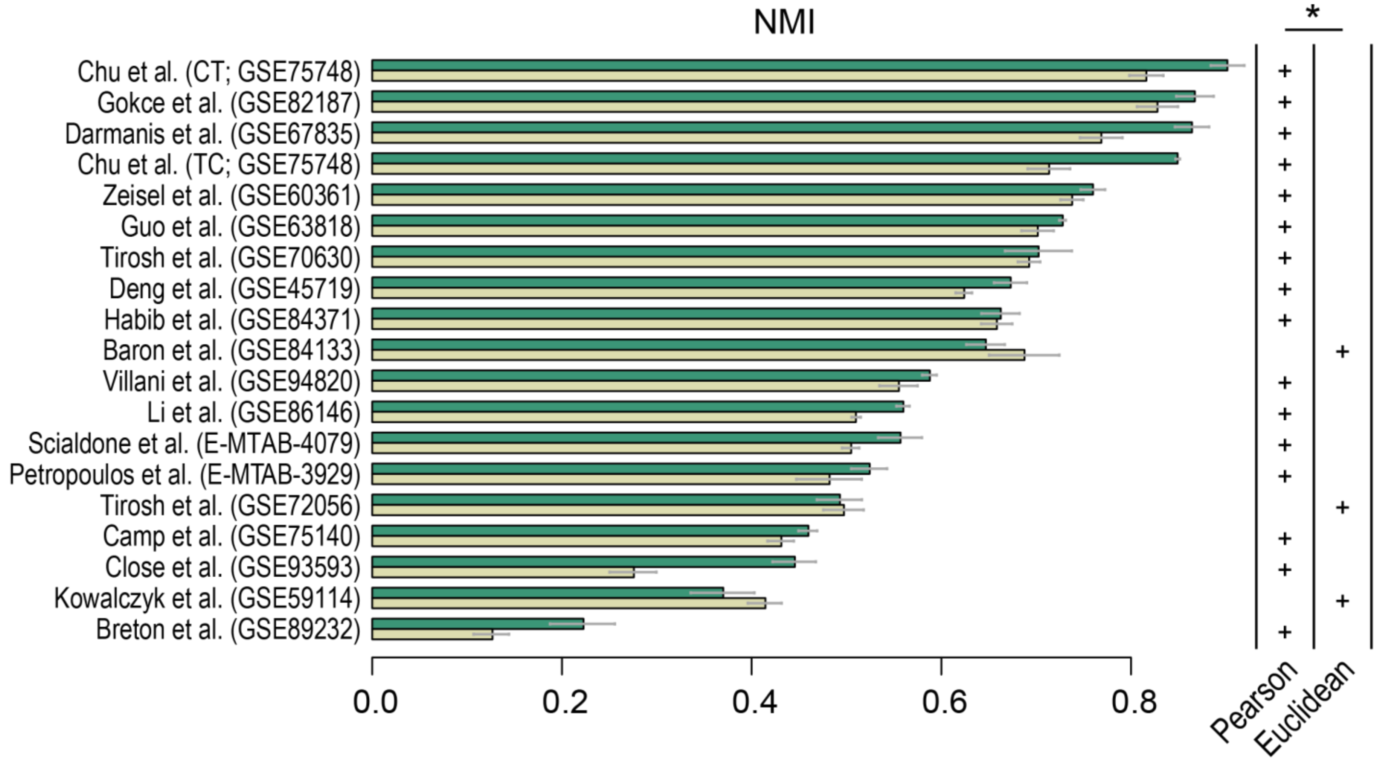


Improving the state-of-the-art clustering method using correlation metric

SIMLR

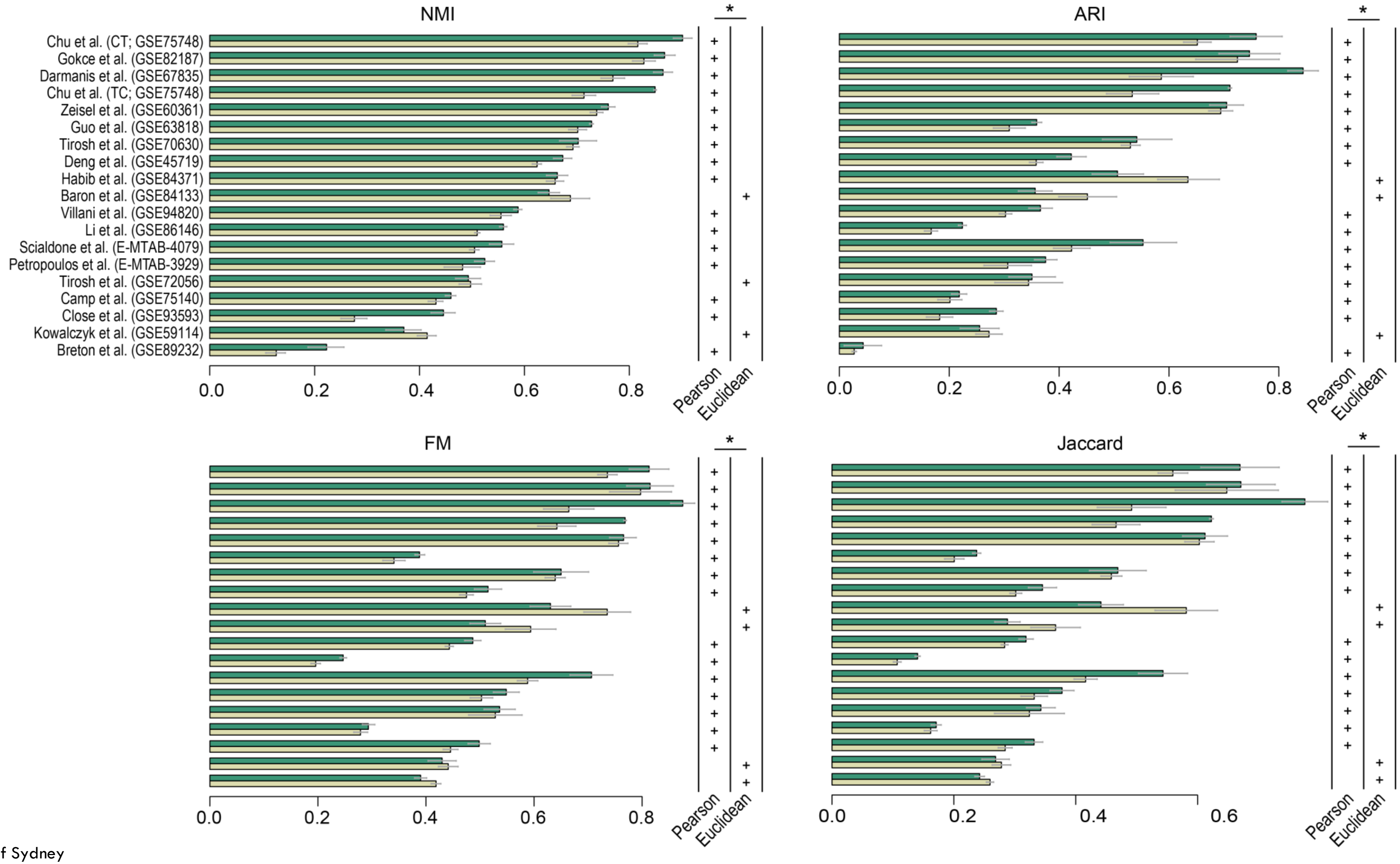
$$K(x_i, x_j) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} \exp\left(-\frac{\qquad\qquad}{2\epsilon_{ij}^2}\right)$$

$$s_{ij} = \frac{\sum_{g=1}^{G}(x_{ig}-\overline{x}_i)(x_{jg}-\overline{x}_j)}{\sqrt{\sum_{g=1}^{G}(x_{ig}-\overline{x}_i)^2}\sqrt{\sum_{g=1}^{G}(x_{jg}-\overline{x}_j)^2}};$$

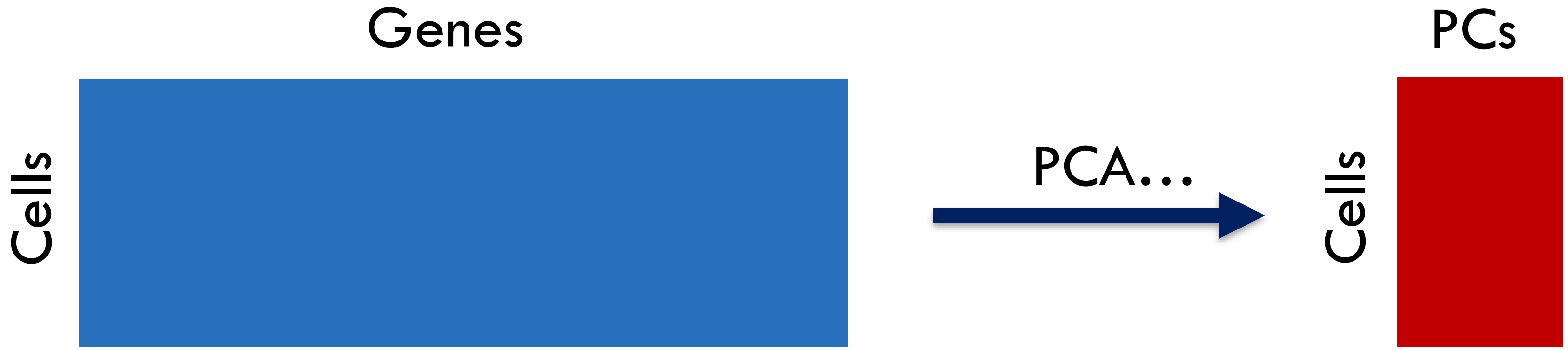


Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, **14**(4), 414.

Evaluation results of SIMLR with Pearson or Euclidean metrics



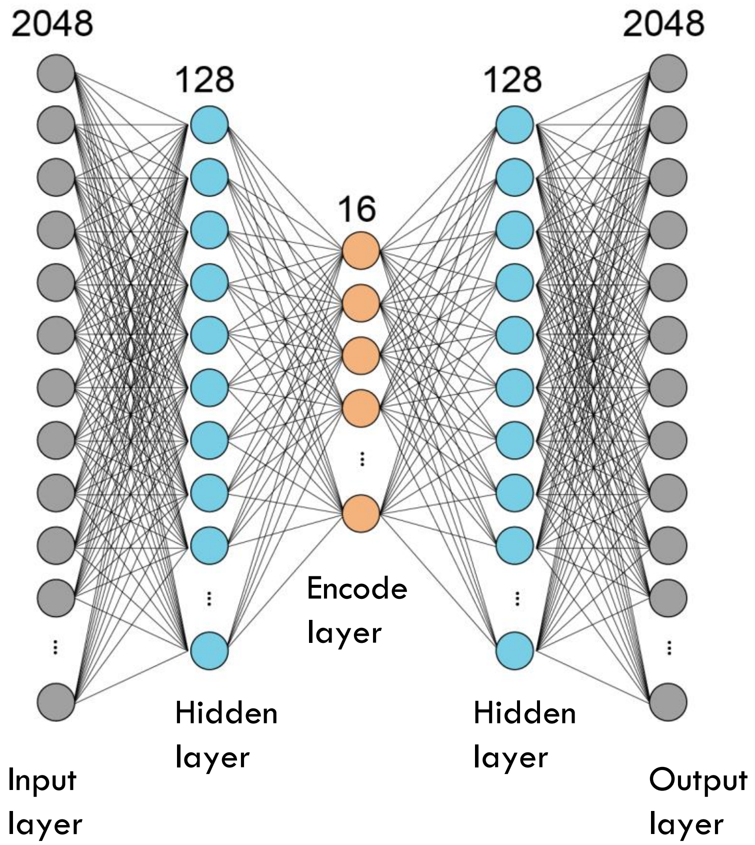
Extension: Methods for accounting high-dimensionality of scRNA-seq



Problem of PCA is that PCs can only be linear combination of genes:

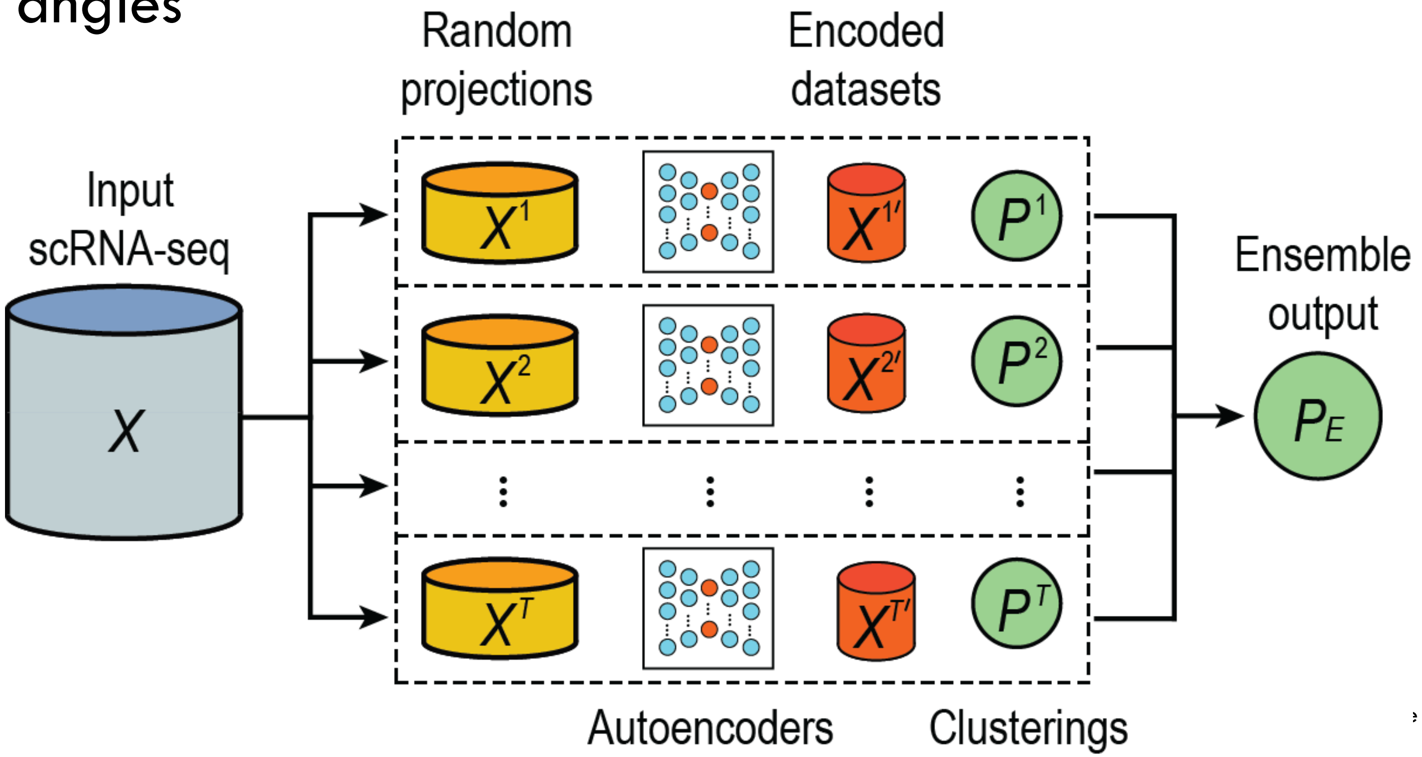
$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

Dimension reduction using an ensemble of autoencoders

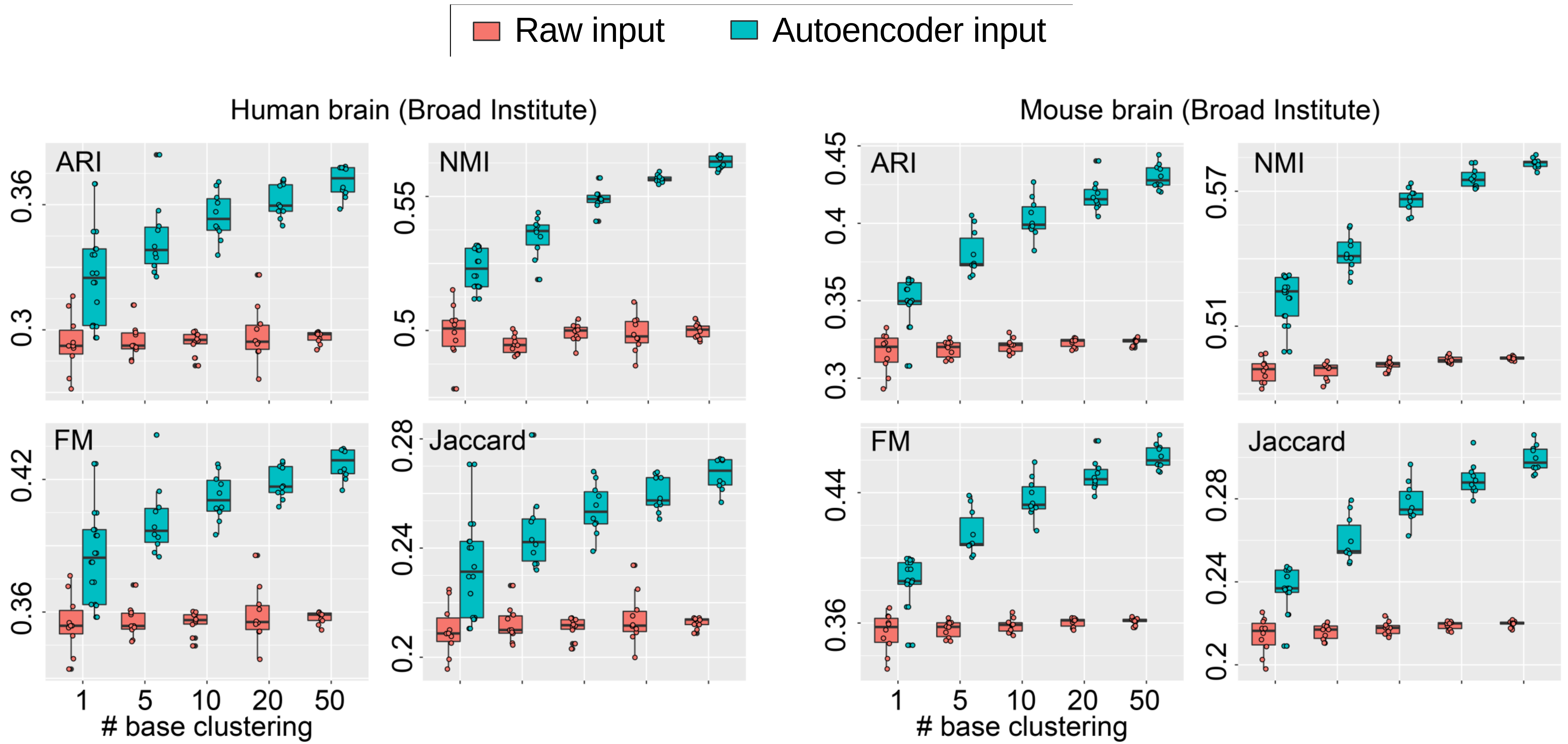


Autoencoder, a deep learning model, allows nonlinear dimension reduction

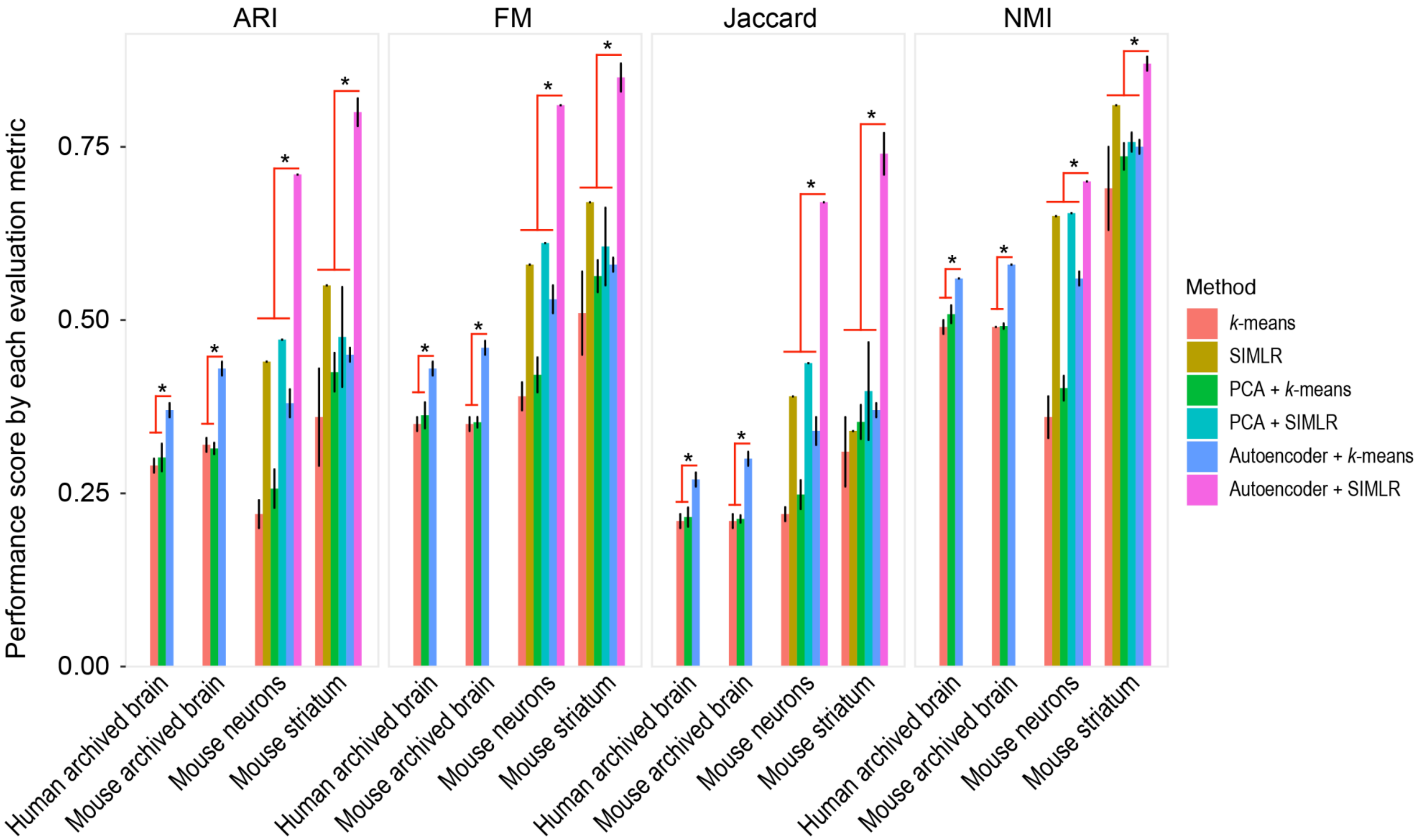
Random projection based ensemble of autoencoders allow multiple views of the scRNA-seq data from different “angles”



Ensemble of autoencoders – does it work (with k-means)?



More benchmark of autoencoder ensemble with PCA using k-means & SIMLR



Geddes T *et al.*, Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis, ***BMC Bioinformatics*** (2019)

We will try this soon…

2:45 – 3:45 Cell type identification via clustering analysis (scClust)

scClassify: Algorithm

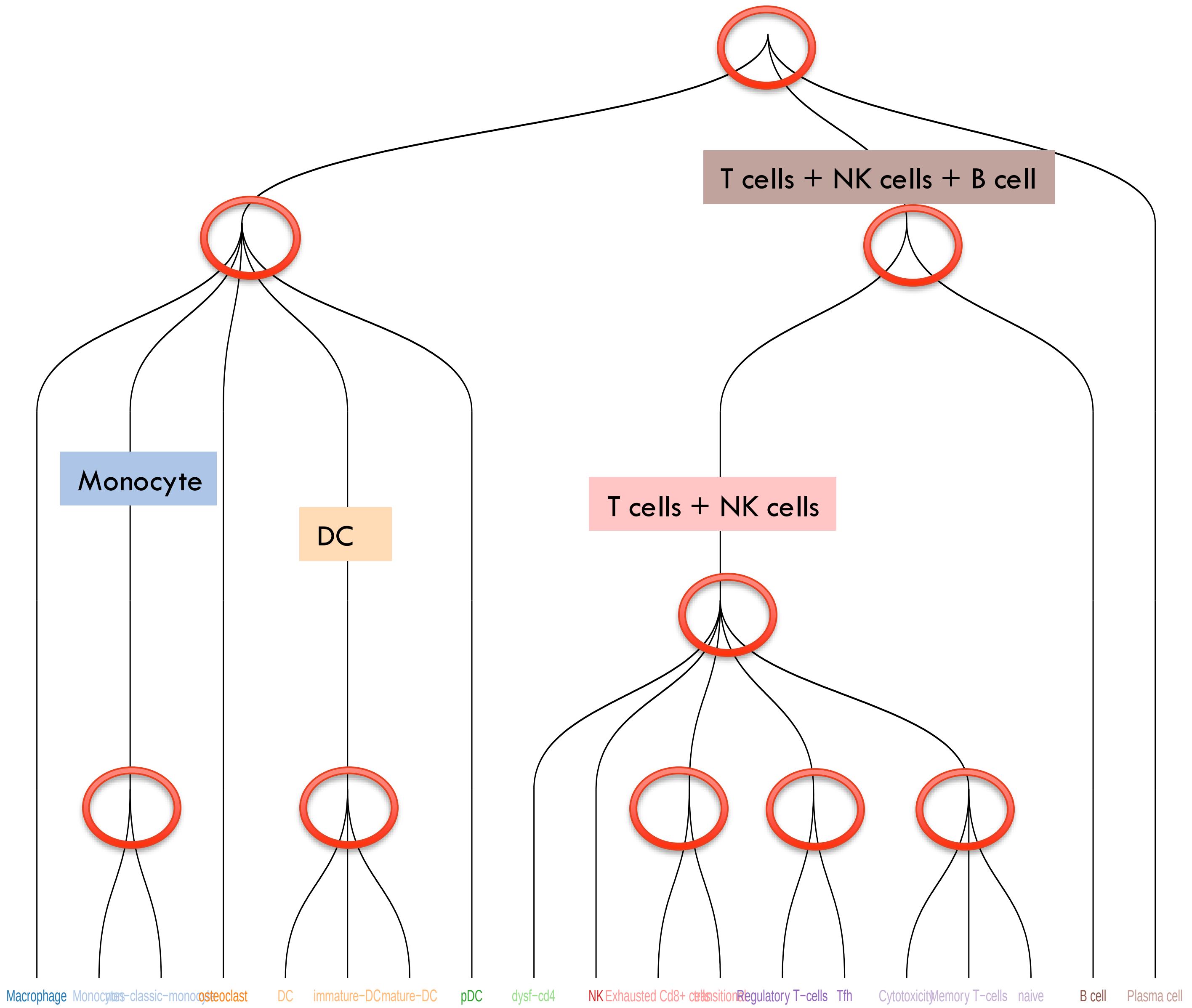
Feature selection at each branch point.

Features are selected from :

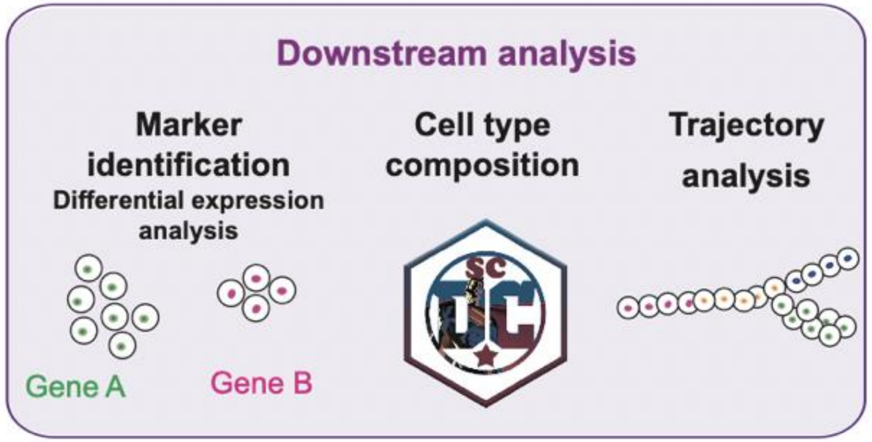
- *Differential expression analysis;*
- *Differential variability analysis;*
- *Differential distribution analysis;*
- *Chi-squared test,*

……

PhD student: Yingxin Lin



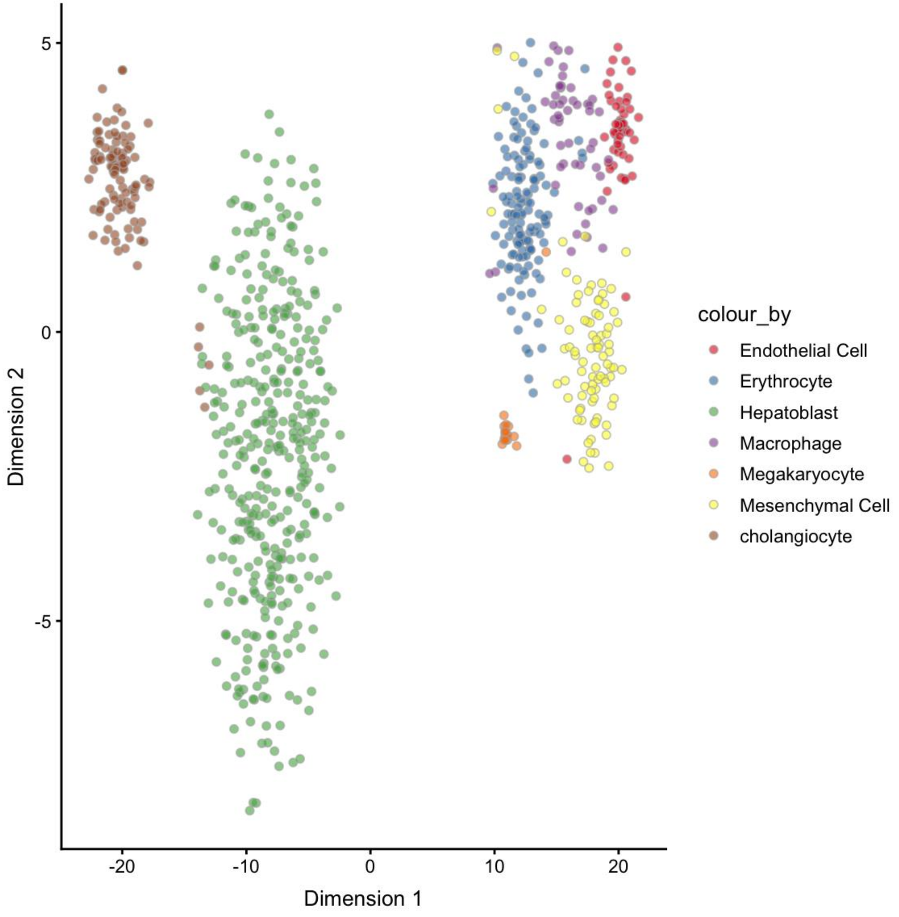
Component 5: Downstream analysis

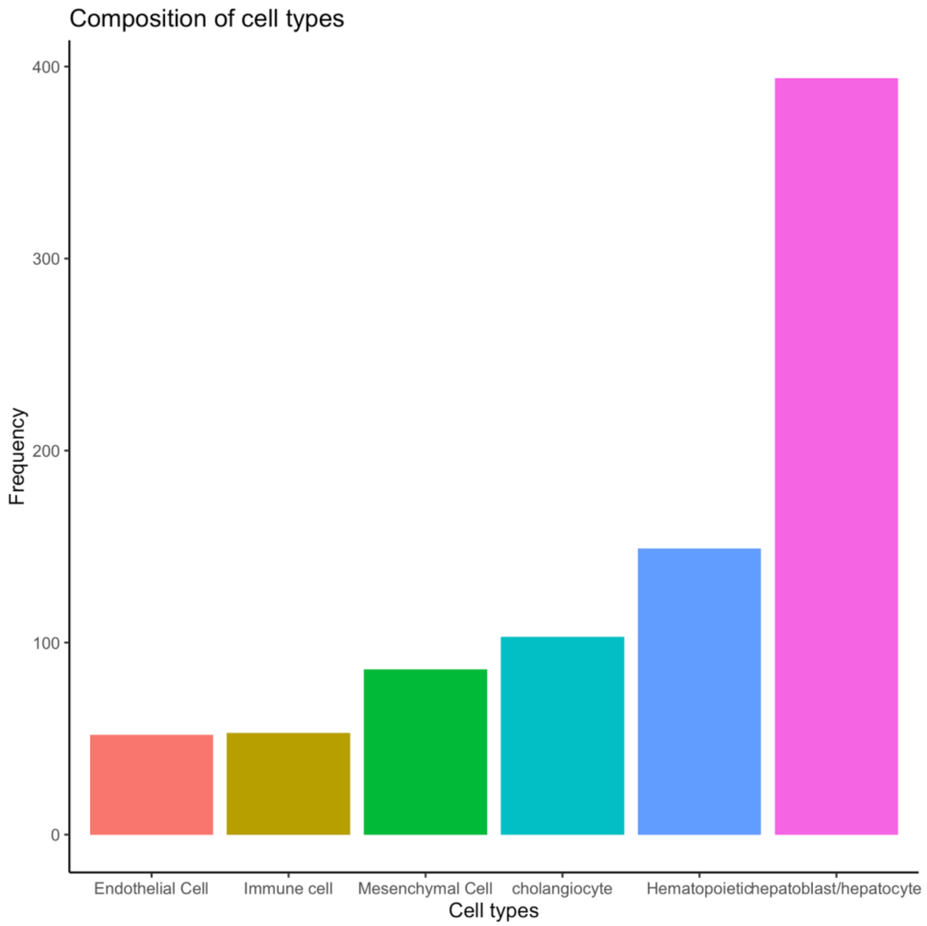


Science questions

- Which genes are differentially expressed between cell types?
- What are the marker genes for each cell type?
- What is the cell type composition?
- Are the cells transitioning from one state to another?

Cell type proportions





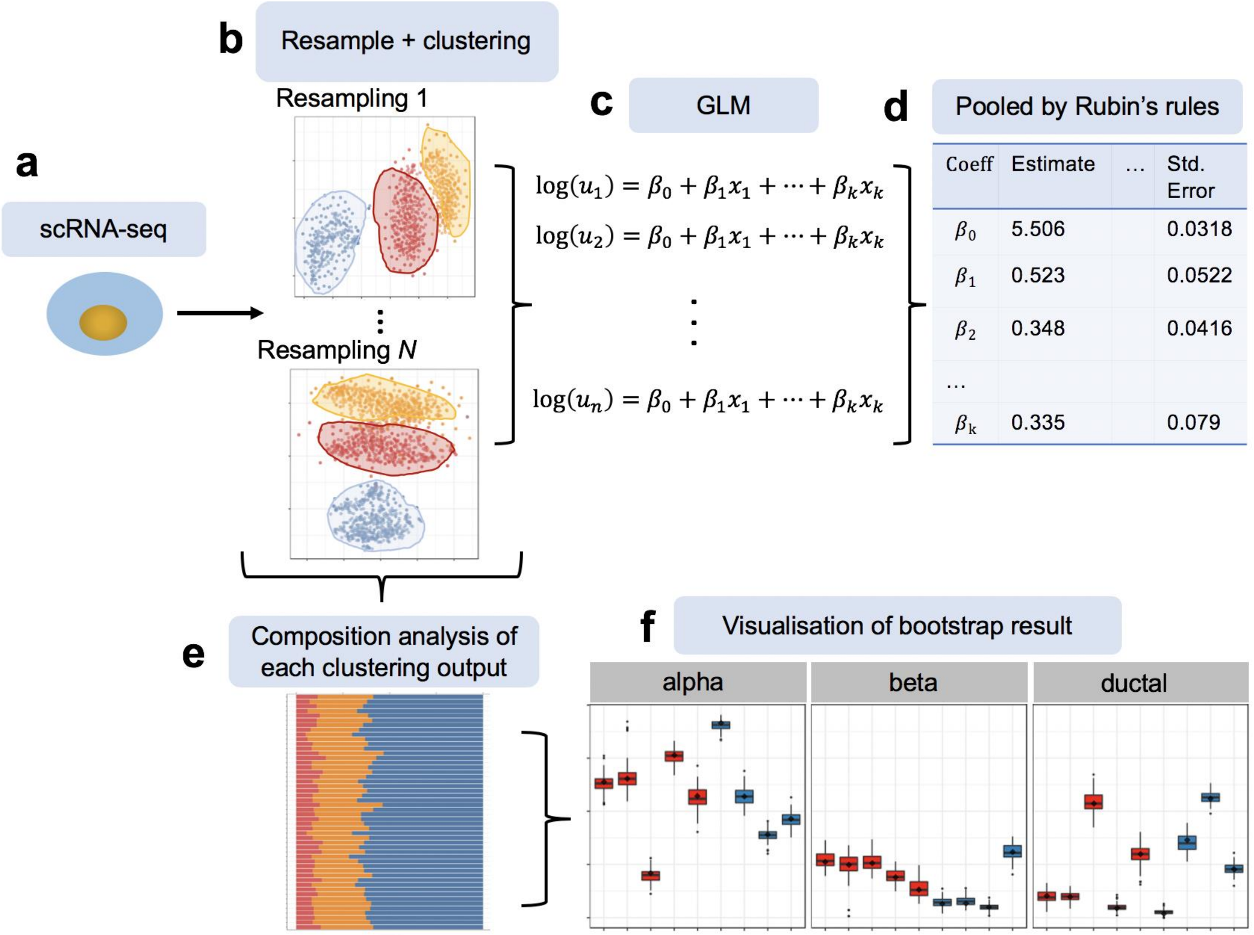
Can we conclude that there are more cholangiocytes than mesenchymal cells?

Single cell Differential Composition (scDC)

scDC simulates ***uncertainty*** in cell-type proportions via bootstrapping

Main components:

- Sample with replacement from count matrix, stratified by patient
- Cell type identification via clustering (PCA -> Kmeans (Pearson correlation)
- Calculations of cell – type proportions standard error from bootstrap samples
- Calculation of pooled log-linear model using Rubin's pooled estimate



$$\log(u_1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\log(u_2) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\vdots$$

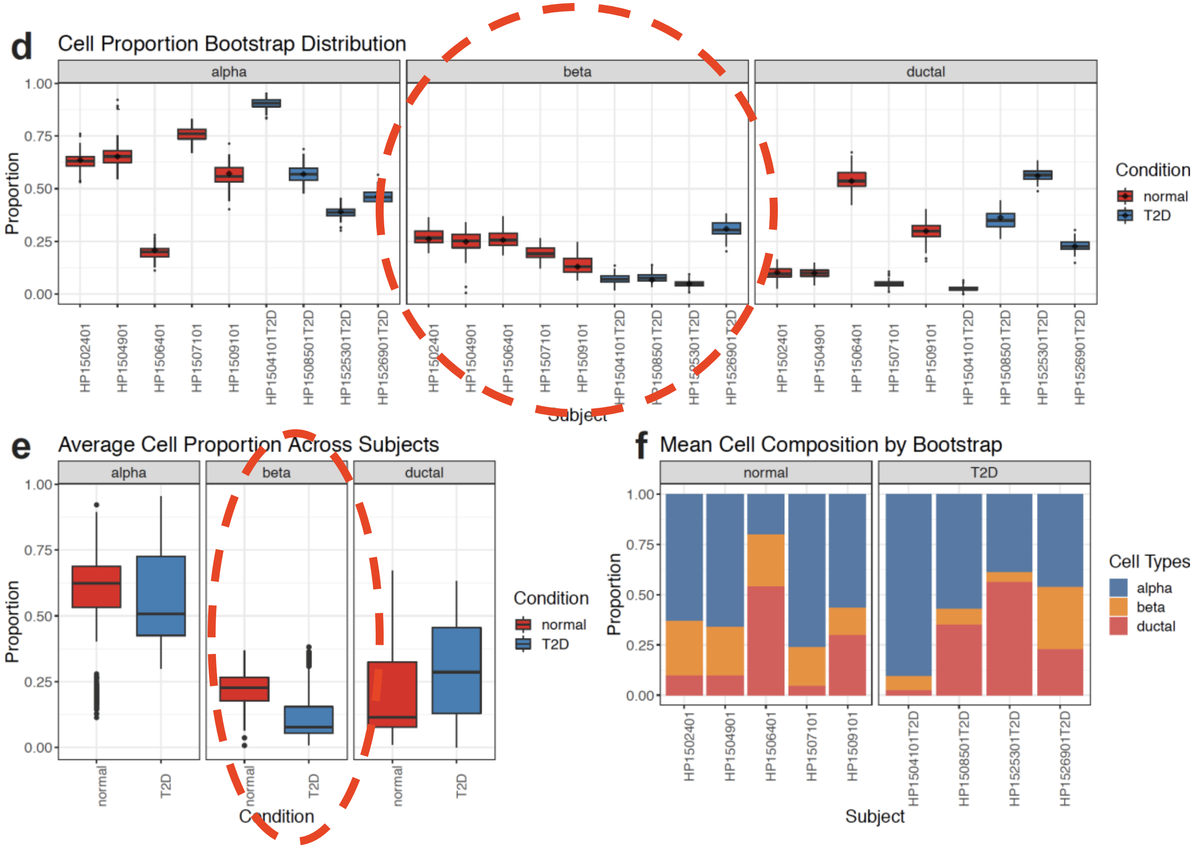
$$\log(u_n) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Coeff	Estimate	…	Std. Error
β_0	5.506		0.0318
β_1	0.523		0.0522
β_2	0.348		0.0416
…			
β_k	0.335		0.079

PhD student: Yue Cao

Single cell Differential Composition (scDC)

- Examined two synthetic datasets constructed from two sets of real experimental data — Pancreas (T2D vs healthy) and Neuronal (developing mouse)
- In pancreas dataset
  - confirmed the original finding that 1 of the 4 subjects has a higher beta cell value, as IQR non overlap
- In neuronal dataset
  - Revealed new finding that progenitor cells percentage increase over time



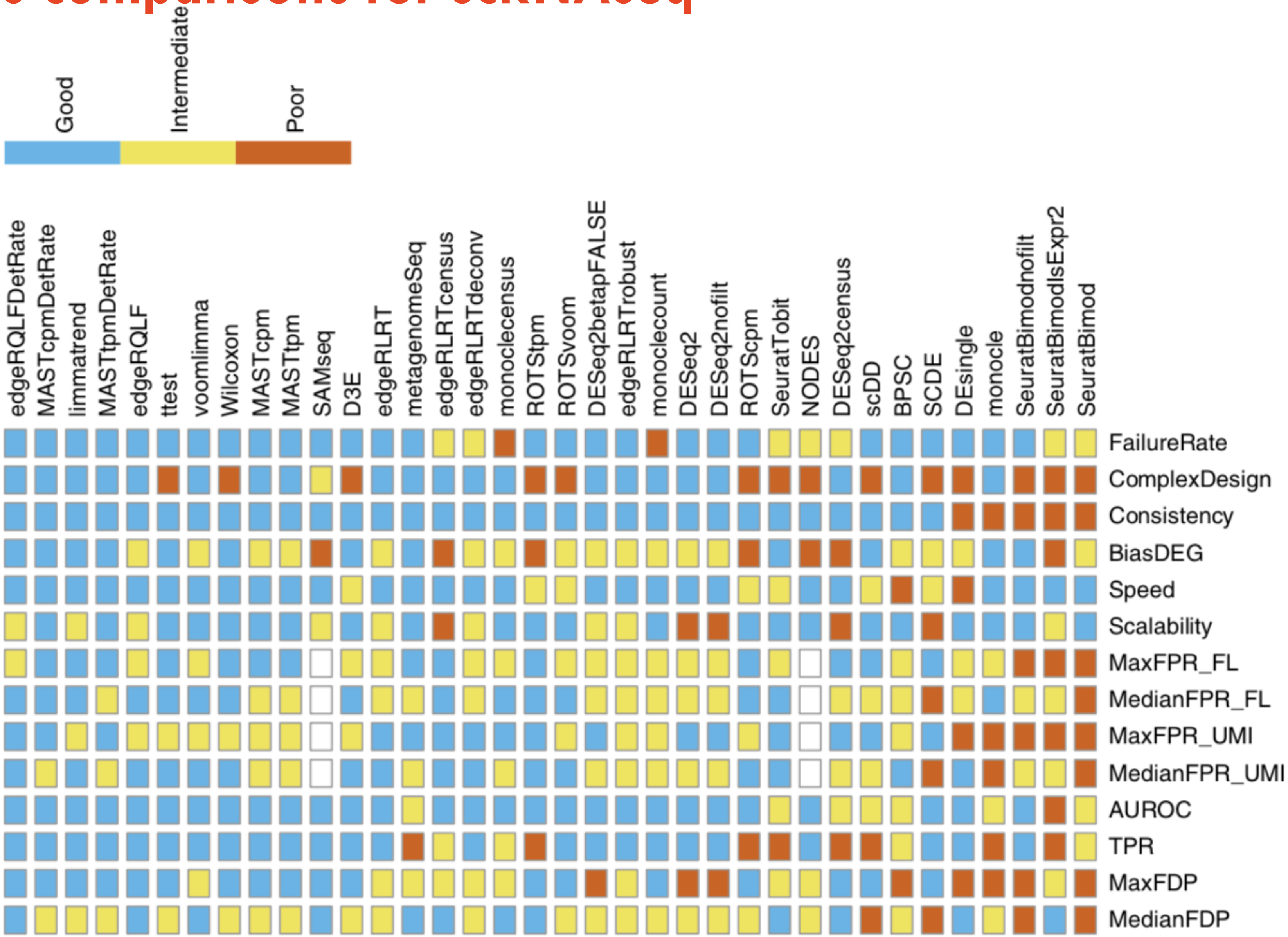
Differences between single cell and bulk RNAseq

- Single cell gene expressions show a bimodal expression pattern – abundant genes are either highly expressed or undetected.
- This can be technical (drop-outs) or biological (transcriptional bursts).
- Drop-outs lead to technical zeroes in the data.
- Technical zeroes are due to low capture efficiency in scRNAseq experiments.
- Many methods have been proposed to deal with drop-outs

Differential expression analysis

- Simple statistical test
  - Wilcoxon rank test, t-test
- Methods developed for bulk RNAseq DE
- DESeq2
  - EdgeR
  - Voom-Limma
- scRNA specific
  - MAST
  - DECENT
  - D3E
  - …. many more!

DE methods comparisons for scRNAseq



Soneson and Robinson (2018) Nature methods

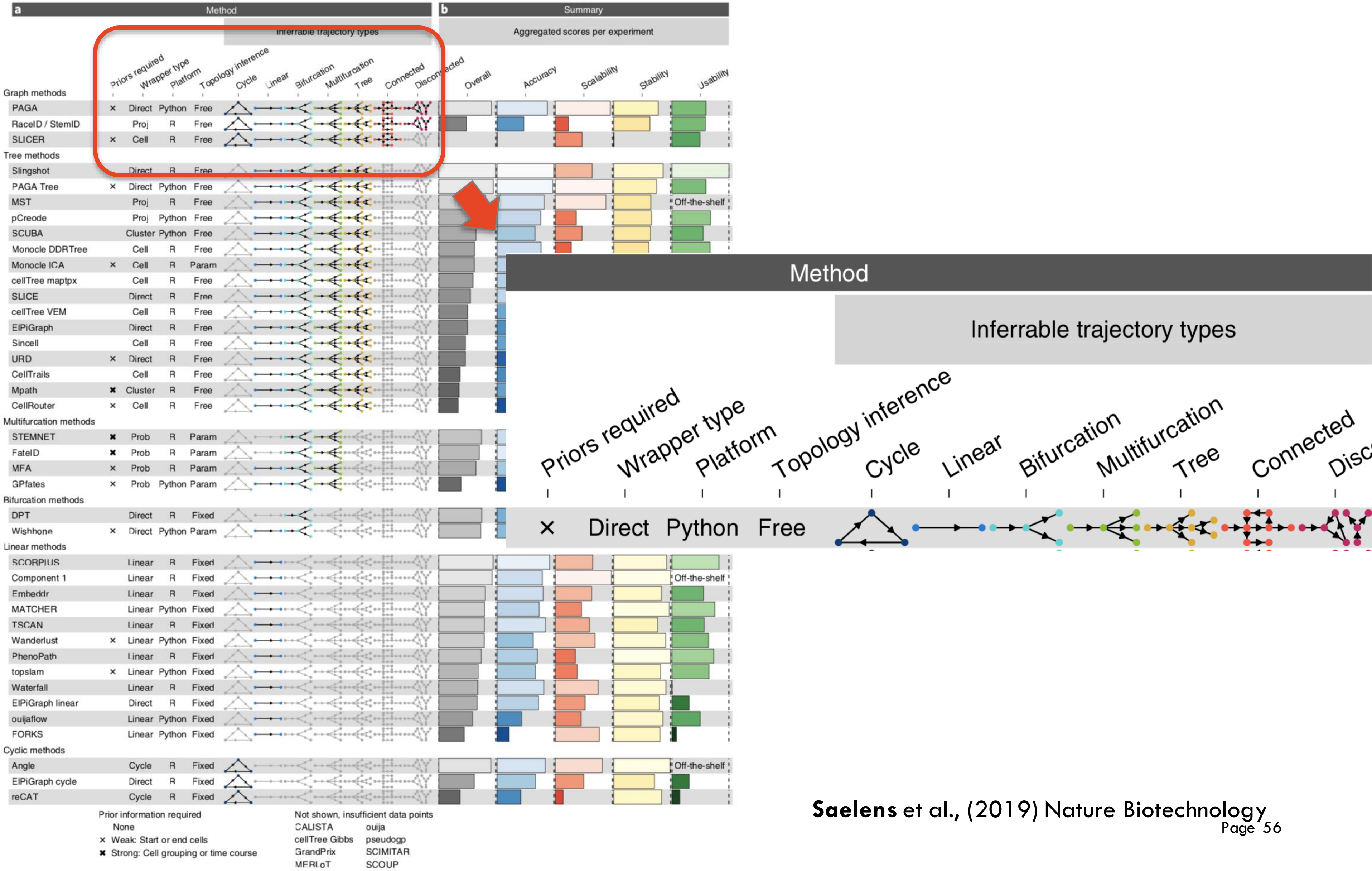
Pseudotime inference

- Why pseudotime?
  - Sometimes cells do not occupy discrete states, rather cell states may follow a smooth trajectory
  - Example: stem cell differentiation

- What is pseudotime?
  - Abstract unit of progress along some trajectory

- Typical steps involved in pseudotime inference:
  - Reduce the dimensionality of the data
  - Build some kind of lineage structure
  - Order the cells in pseudotime

Comparisons of pseudotime inference methods



	Priors required	Wrapper type	Platform	Topology inference
Graph methods				
PAGA	×	Direct	Python	Free
RaceID / StemID		Proj	R	Free
SLICER	×	Cell	R	Free
Tree methods				
Slingshot		Direct	R	Free
PAGA Tree	×	Direct	Python	Free
MST		Proj	R	Free
pCreode		Proj	Python	Free
SCUBA		Cluster	Python	Free
Monocle DDRTree		Cell	R	Free
Monocle ICA	×	Cell	R	Param
cellTree maptpx		Cell	R	Free
SLICE		Direct	R	Free
cellTree VEM		Cell	R	Free
ElPiGraph		Direct	R	Free
Sincell		Cell	R	Free
URD	×	Direct	R	Free
CellTrails		Cell	R	Free
Mpath	✖	Cluster	R	Free
CellRouter	×	Cell	R	Free
Multifurcation methods				
STEMNET	✖	Prob	R	Param
FateID	✖	Prob	R	Param
MFA	×	Prob	R	Param
GPfates	×	Prob	Python	Param
Bifurcation methods				
DPT		Direct	R	Fixed
Wishbone	×	Direct	Python	Param
Linear methods				
SCORPIUS		Linear	R	Fixed
Component 1		Linear	R	Fixed
Embeddr		Linear	R	Fixed
MATCHER		Linear	Python	Fixed
TSCAN		Linear	R	Fixed
Wanderlust	×	Linear	Python	Fixed
PhenoPath		Linear	R	Fixed
topslam	×	Linear	Python	Fixed
Waterfall		Linear	R	Fixed
ElPiGraph linear		Direct	R	Fixed
ouijaflow		Linear	Python	Fixed
FORKS		Linear	Python	Fixed
Cyclic methods				
Angle		Cycle	R	Fixed
ElPiGraph cycle		Direct	R	Fixed
reCAT		Cycle	R	Fixed

Prior information required
None
× Weak: Start or end cells
✖ Strong: Cell grouping or time course

Not shown, insufficient data points

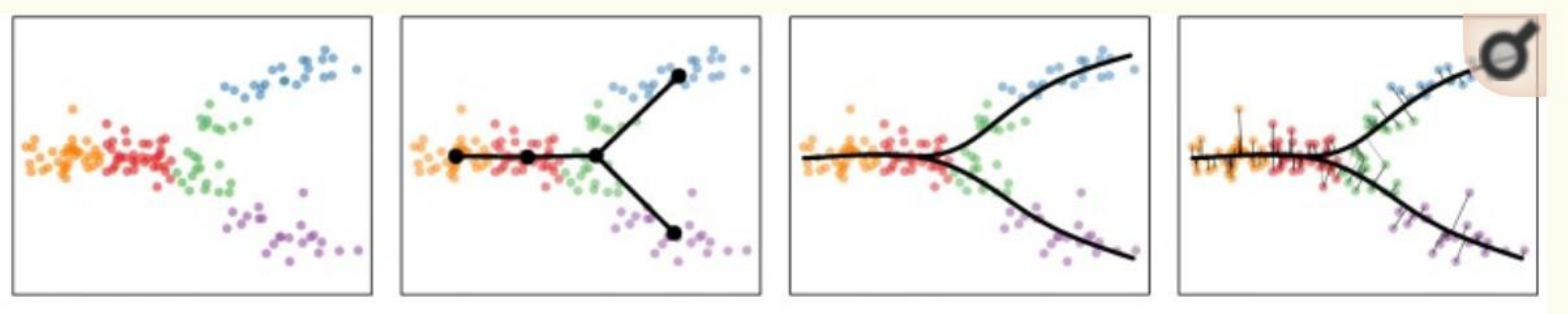
CALISTA	ouija
cellTree Gibbs	pseudogp
GrandPrix	SCIMITAR
MERLoT	SCOUP

Saelens et al., (2019) Nature Biotechnology

Slingshot example (Street et al., 2018)

Two stages:

1. Inference of the global lineage structure. Uses cluster-based minimum spanning tree
2. Inference of pseudotime variables for cells along each lineage. Fits simultaneous principal curves



-We will try this soon…

3:45 – 4:30 Downstream analysis: identify marker genes & cell type composition
Extension: cell type identification via supervised classification and single cell trajectory analysis

THE UNIVERSITY OF SYDNEY