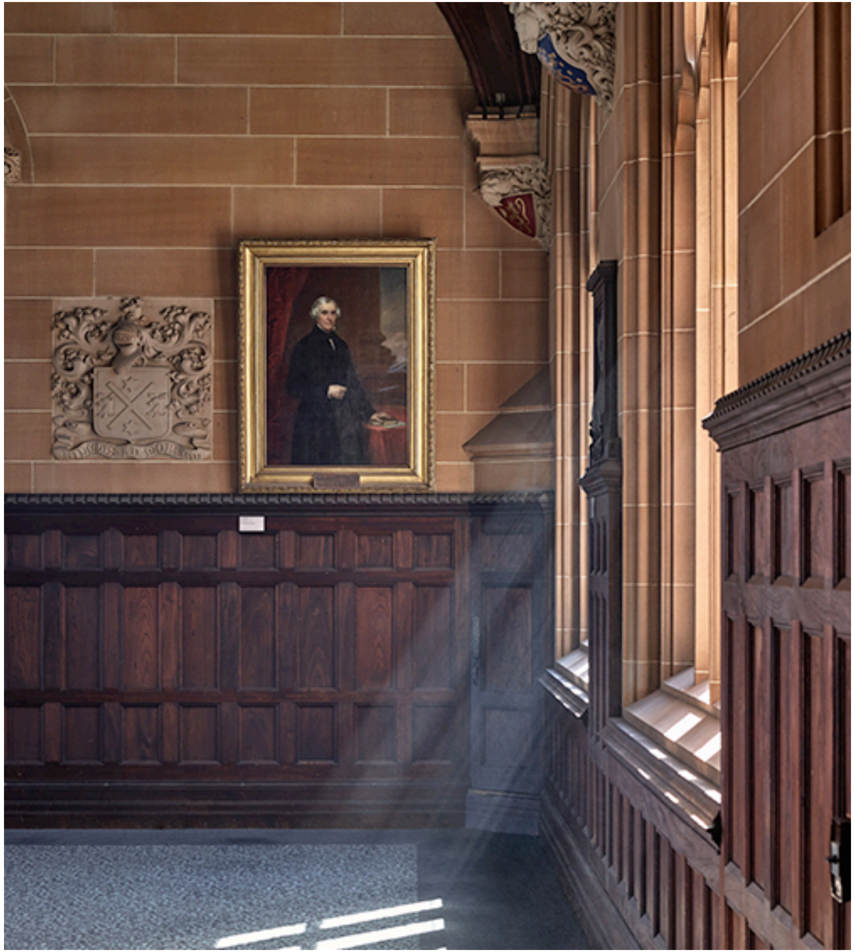
Single-cell analysis workshop

Yue Cao, Kevin Wang

Sydney Precision Bioinformatics Group
School of Mathematics and Statistics

THE UNIVERSITY OF
SYDNEY



Sydney Precision Bioinformatics Group

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

Our group consists of research leaders, research associates, PhD candidates, Honours and TSP students.

A/Prof. John Ormerod; Prof. Jean Yang; Prof. Samuel Mueller; Dr. Garth Tarr; Dr. Rachel Wang

Dr. Ellis Patrick; Dr. Pengyi Yang

Find out more:

http://www.maths.usyd.edu.au/bioinformatics/

Shiny apps:	http://shiny.maths.usyd.edu.au/
GitHub:	https://github.com/SydneyBioX

Roadmap for the workshop

12:30 – 12:40: Google cloud set up

12:40 – 13:00 Overview and Quality Control slides

13:45 – 14:00 scMerge data integration

14:45 – 15:00 Cell type identification via clustering, marker genes and composition

Scheduled to finish at 15:30

Setting up

- https://sydneybiox.github.io/cornell_sc_workshop/

- Go to address: http://34.68.240.36/

- Type code into the console

Overview of single-cell technology

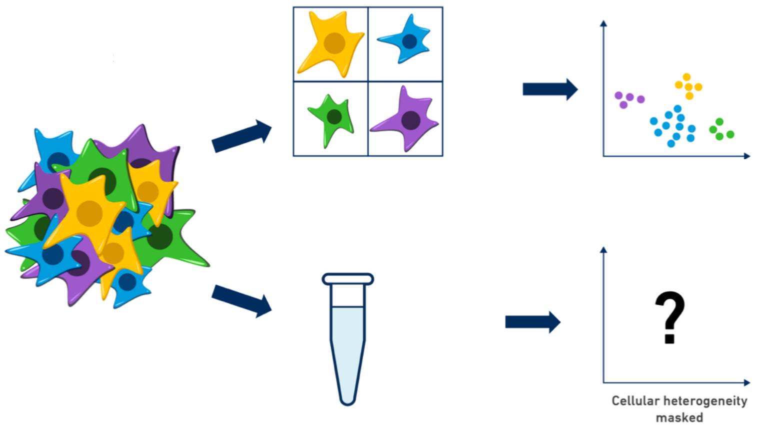
Single cell technology

- Resolving tissue and cellular heterogeneity

- Bulk RNA-Seq measures averaged signals from millions of cells

- scRNA-Seq measures individual cells

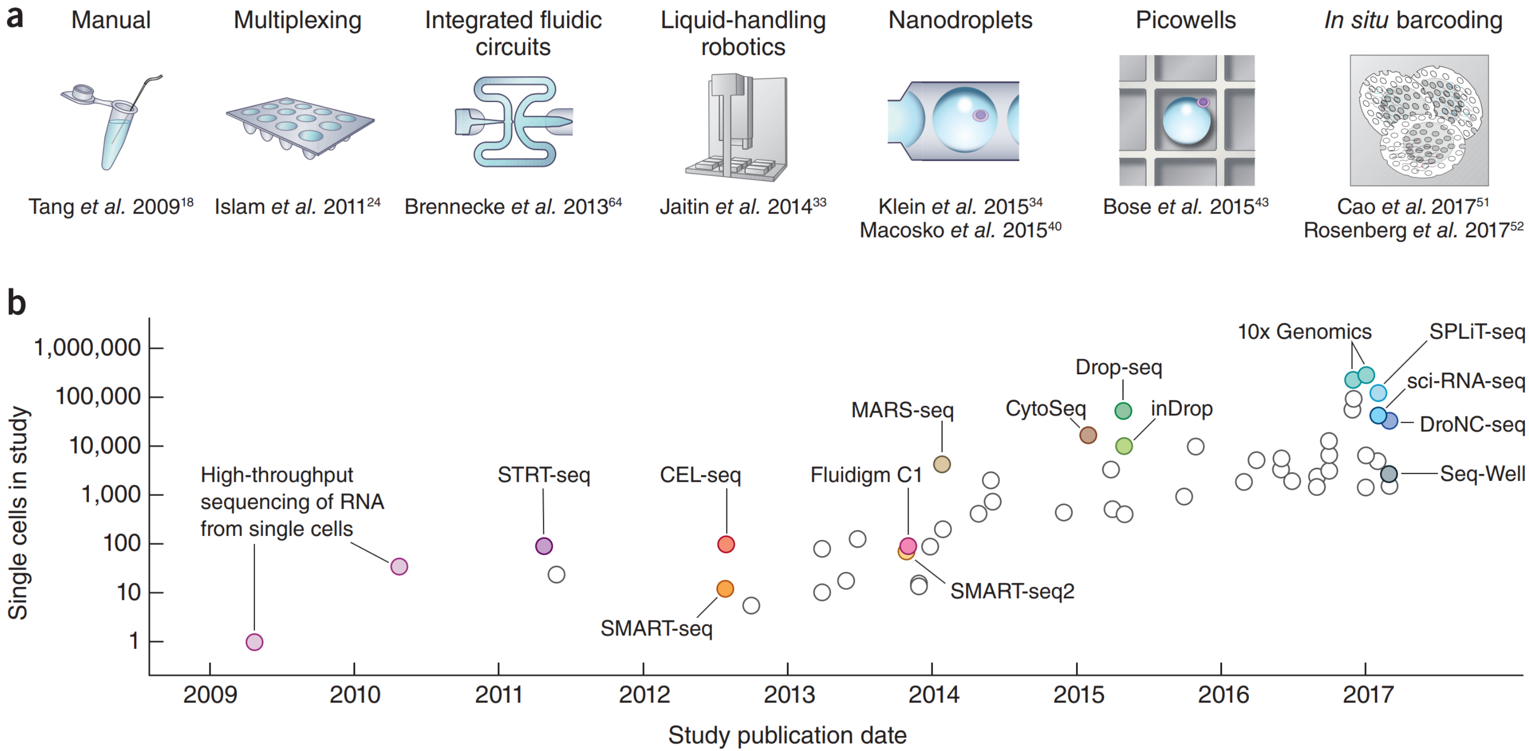
Single cells



Bulk cells

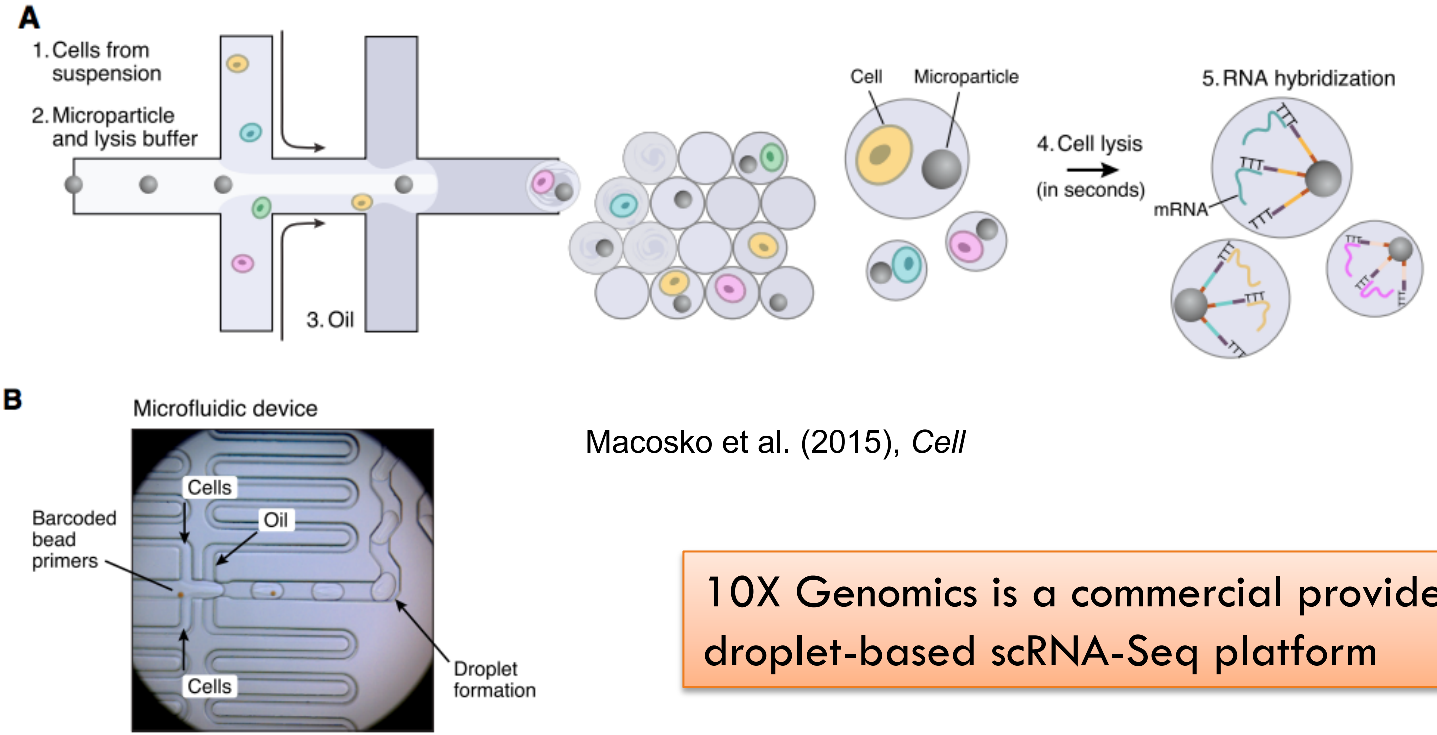
Goldman, S. L., MacKay, M., Afshinnekoo, E., Melnick, A. M., Wu, S., & Mason, C. E. (2019). The Impact of Heterogeneity on Single-Cell Sequencing. *Frontiers in Genetics*, *10*.
https://community.10xgenomics.com/t5/10x-Blog/Single-Cell-RNA-Seq-An-Introductory-Overview-and-Tools-for/ba-p/547

Exponential growth in single cell RNA-Seq technologies



Svensson et al. *Nature Protocols* (2018)

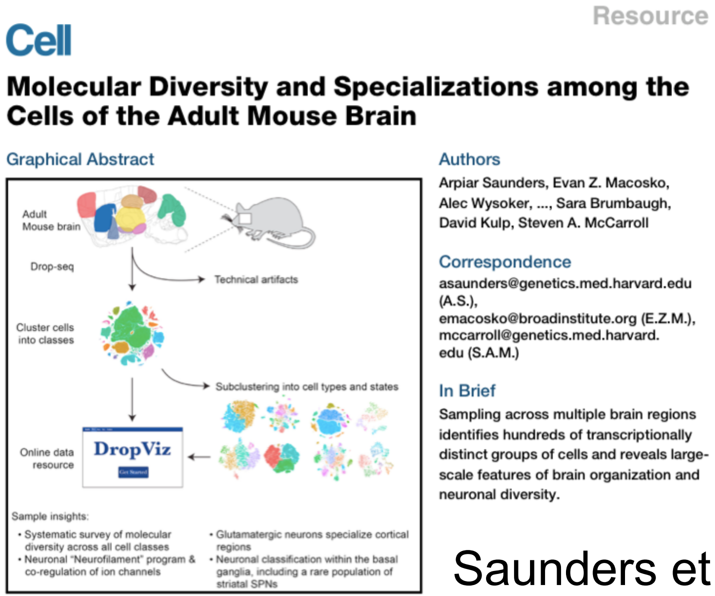
Droplet based technologies are now dominating



Macosko et al. (2015), *Cell*

10X Genomics is a commercial provider of droplet-based scRNA-Seq platform

scRNA-Seq experiments approaching 1million 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

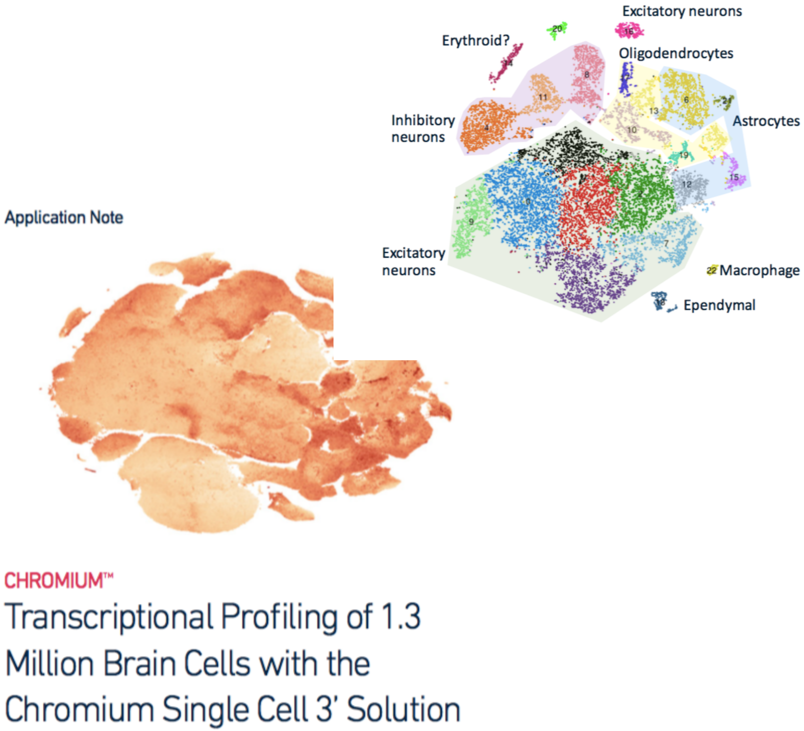
asaunders@genetics.med.harvard.edu (A.S.), emacosko@broadinstitute.org (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

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.

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

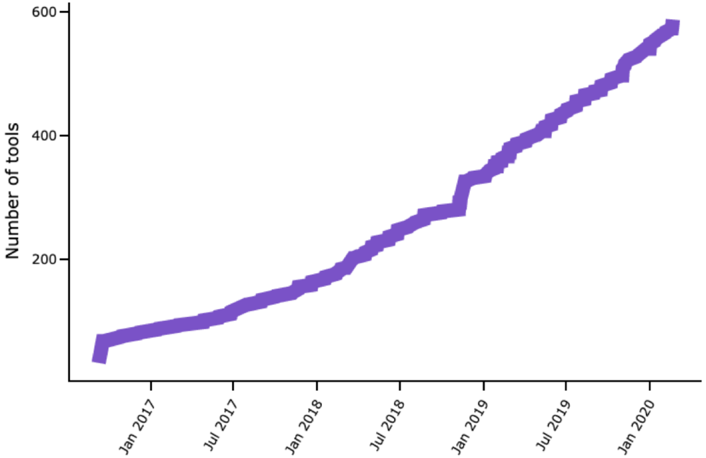
Single-cell RNA-Seq analysis

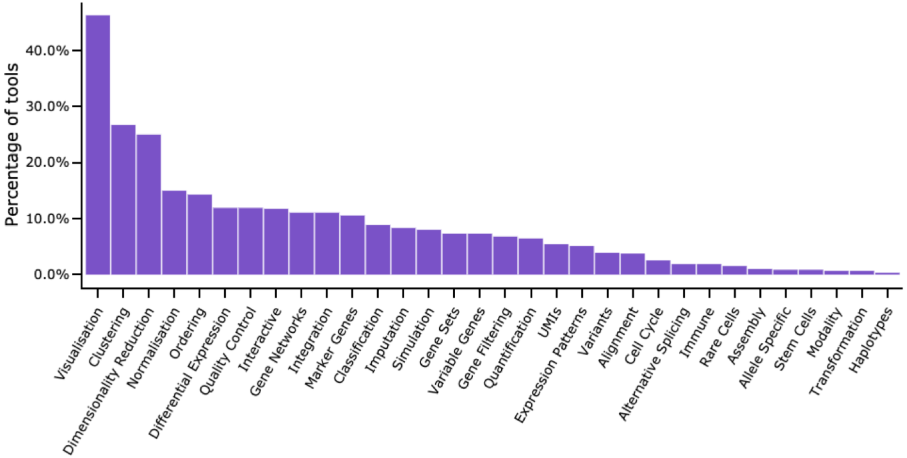
THE UNIVERSITY OF SYDNEY

Differences between single-cell and bulk RNA-Seq

- In scRNA-Seq, abundant genes are either highly expressed or undetected
- Biological (transcriptional bursts)
- Technical (drop-outs due to low capture efficiency)
  - An abundance of zeroes
  - Bimodal distribution of genes
- Many methods have been proposed to deal with drop-outs

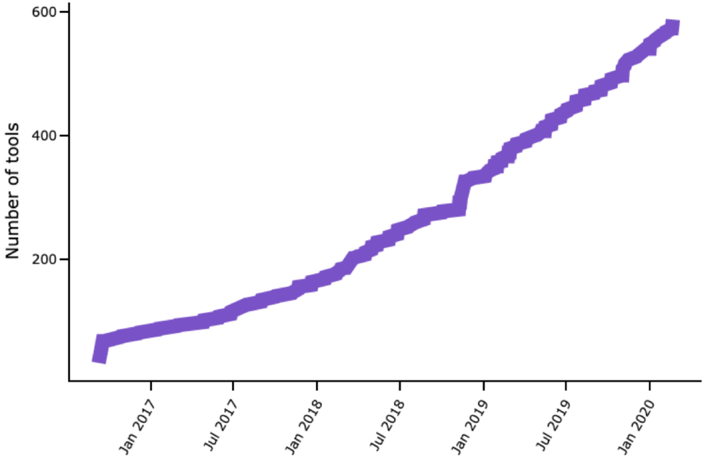
Rapid increase of scRNA-Seq tools

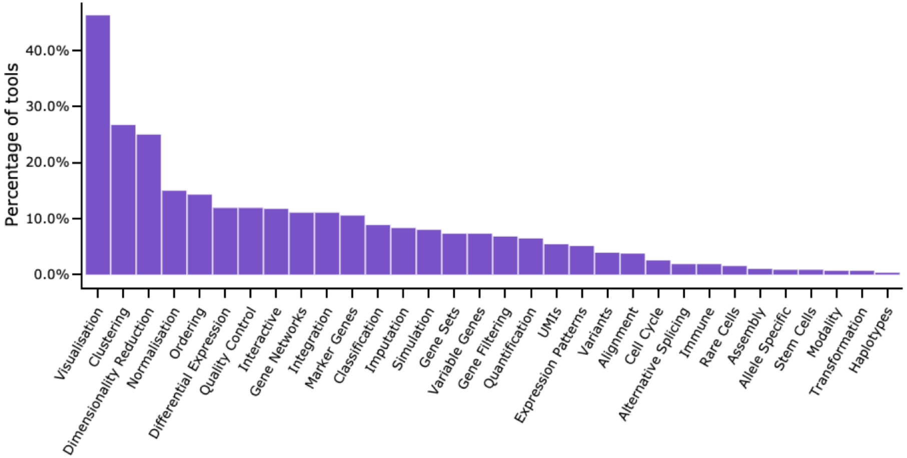




www.scrna-tools.org

Which tool should you use?





www.scrna-tools.org

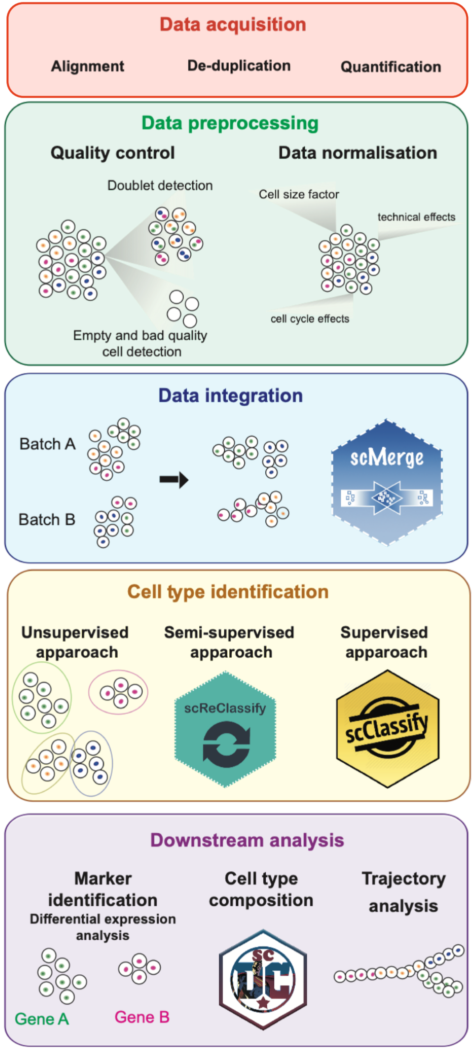
What biological questions are you trying to answer?

- Can I get there using special modelling or just simple visualisation?

- Follow a well-established pipeline from Bioconductor https://osca.bioconductor.org/ or find suitable tools from https://www.scrna-tools.org/

- Use our tools and pipeline!

Components of a typical scRNA-Seq analysis



Component 1: Data acquisition

Data acquisition
Alignment
De-duplication
Quantification

Input

- BCL or FASTQ file from the sequencer

Output

- Gene-by-cell counts matrix

	Cell 1	Cell 2	Cell 3
ACTB	1	4	6
GAPDH	5	0	2
LBR	0	3	0
HIF1A	0	1	0

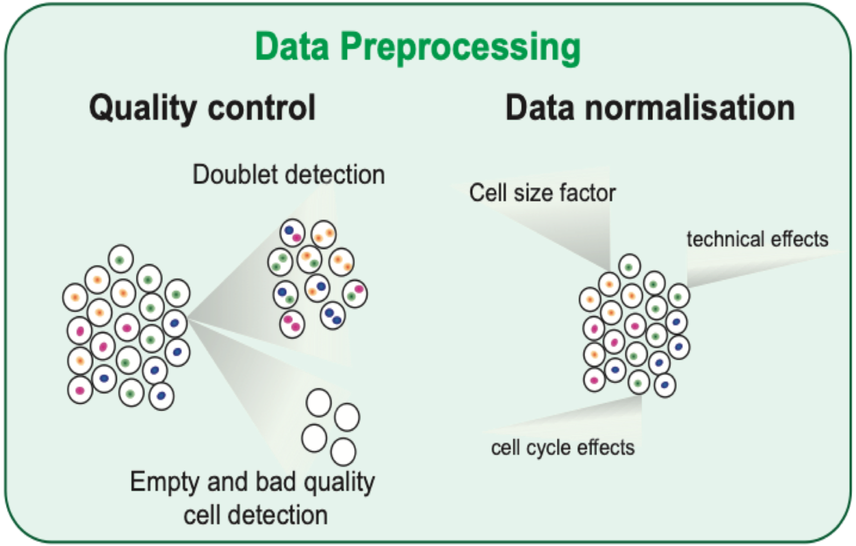
Software

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

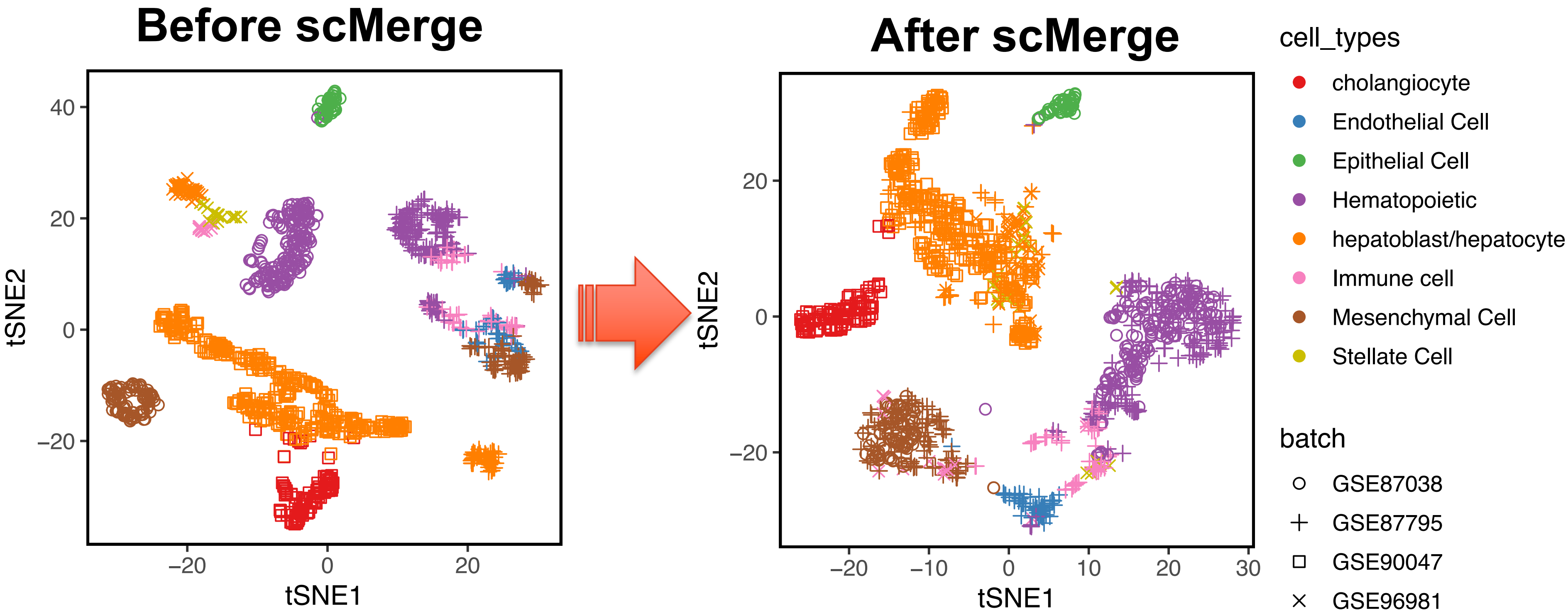
Considerations

- 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



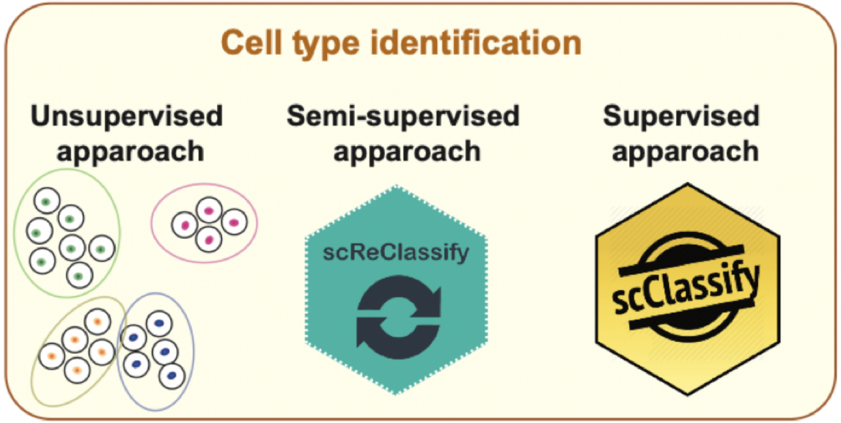
Component 3: Data integration



Before scMerge

After scMerge

Component 4: Cell type identification

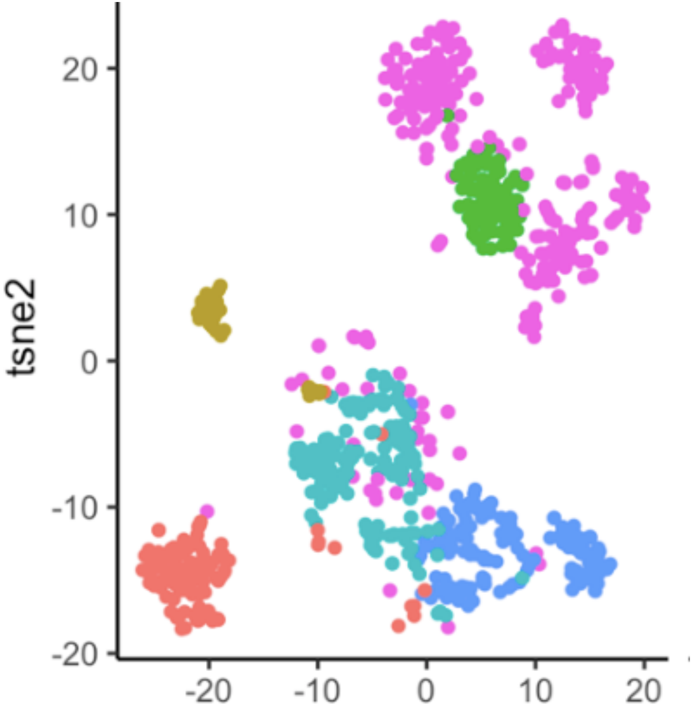


Cell type identification

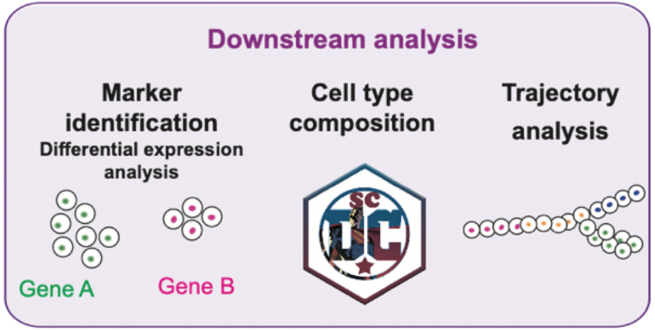
Unsupervised apparaoach

Semi-supervised apparaoach

Supervised apparaoach



Component 5: Downstream analysis

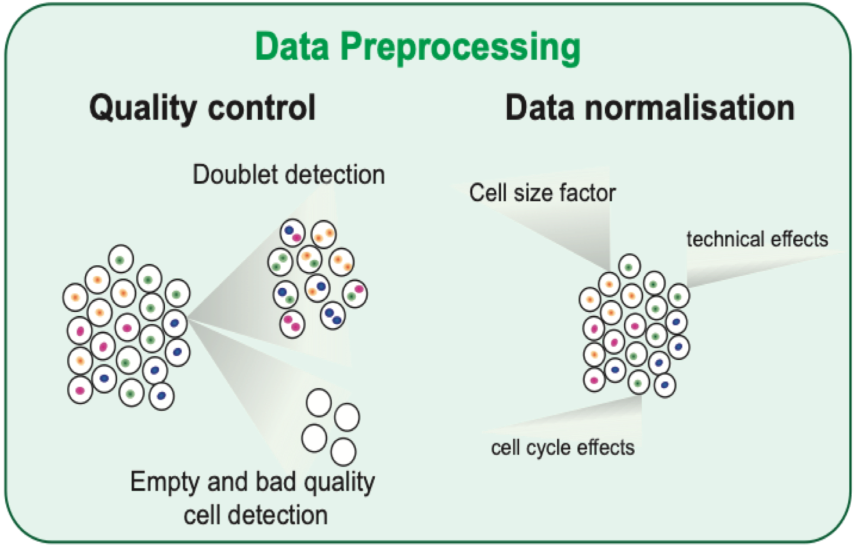


Science questions

- 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?

Quality control

Component 2: Data preprocessing – Quality control



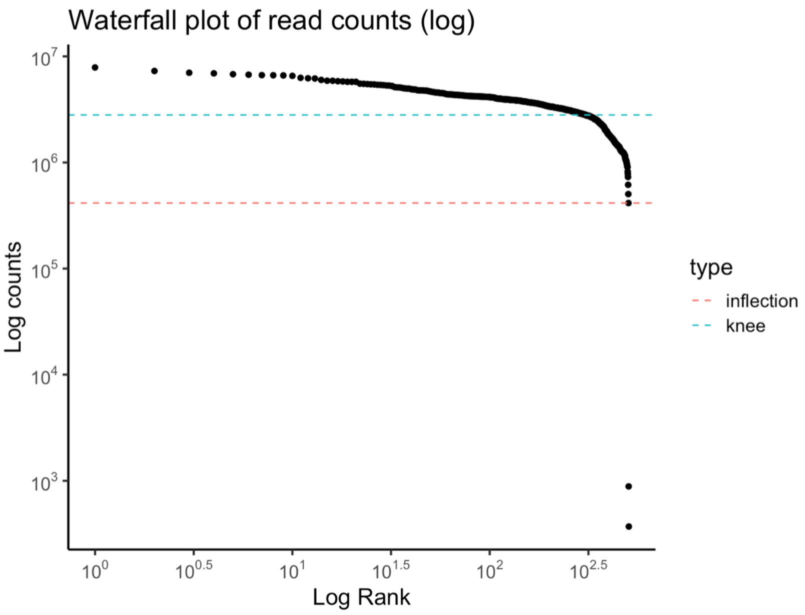
Software

- 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

Component 2: Data preprocessing – Quality control



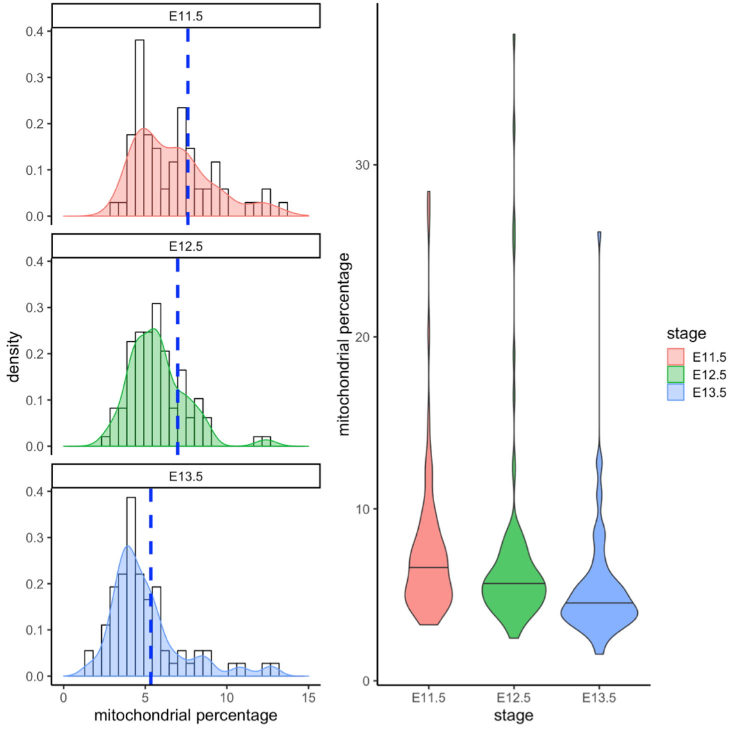
Software

- 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
- **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
- Filter out droplets with no cells
- **Filter out droplets with damaged cells – look for high mitochondrial gene content or high spike-in**

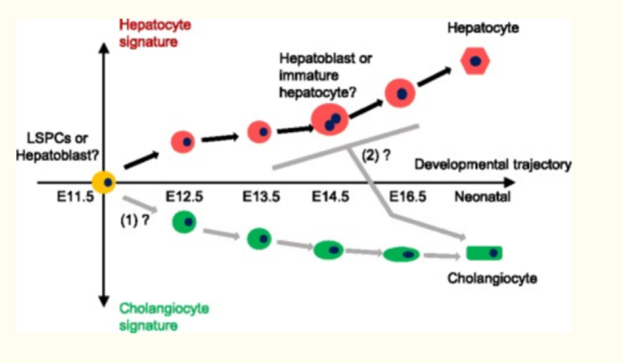
scMerge: merging scRNA-Seq data

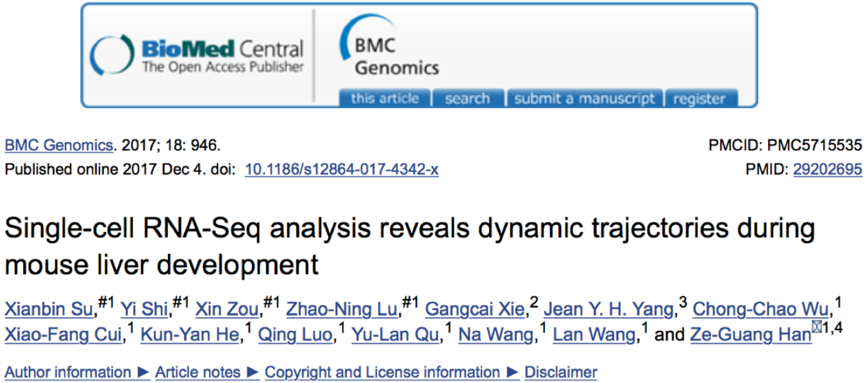
THE UNIVERSITY OF SYDNEY

Liver fetal development time course data

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]

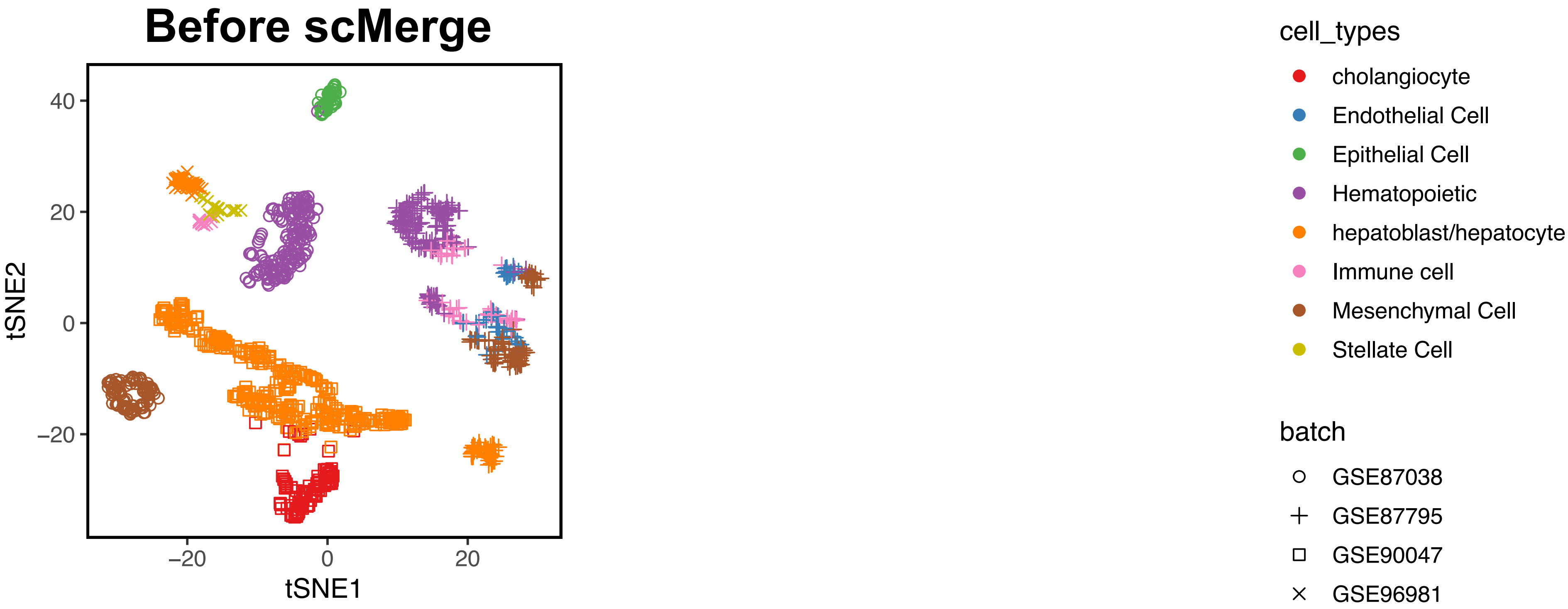
Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Liver fetal development time course data

https://sydneybiox.github.io/scMerge/articles/case_study/Mouse_Liver_Data.html

	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

Liver fetal development time course data



Breaking observed data into components

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

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

The data we observe

Biologically relevant variation

e.g. cell types

Unwanted variation

e.g. batch and technical effects

Random noise

Estimating unwanted variation

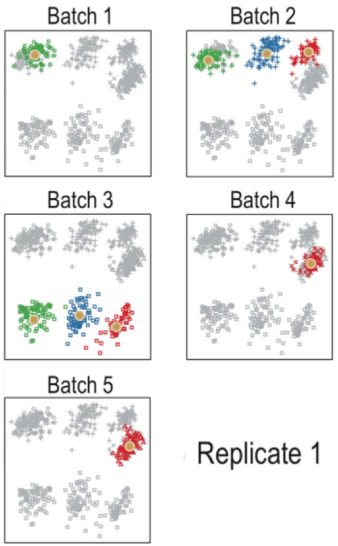
Estimated by **stably expressed genes** by factor analysis

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

Estimated with **replicates** by factor analysis

Molania et al. (2019), Nuclei Acids Res

scMerge algorithm



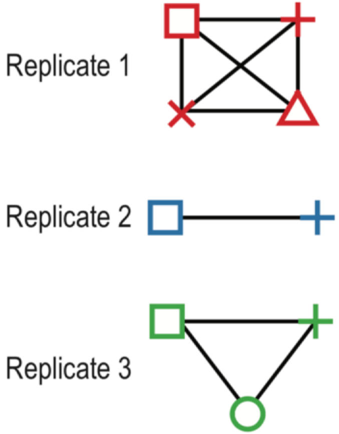
Clustering for each batch

(k-means by default)

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

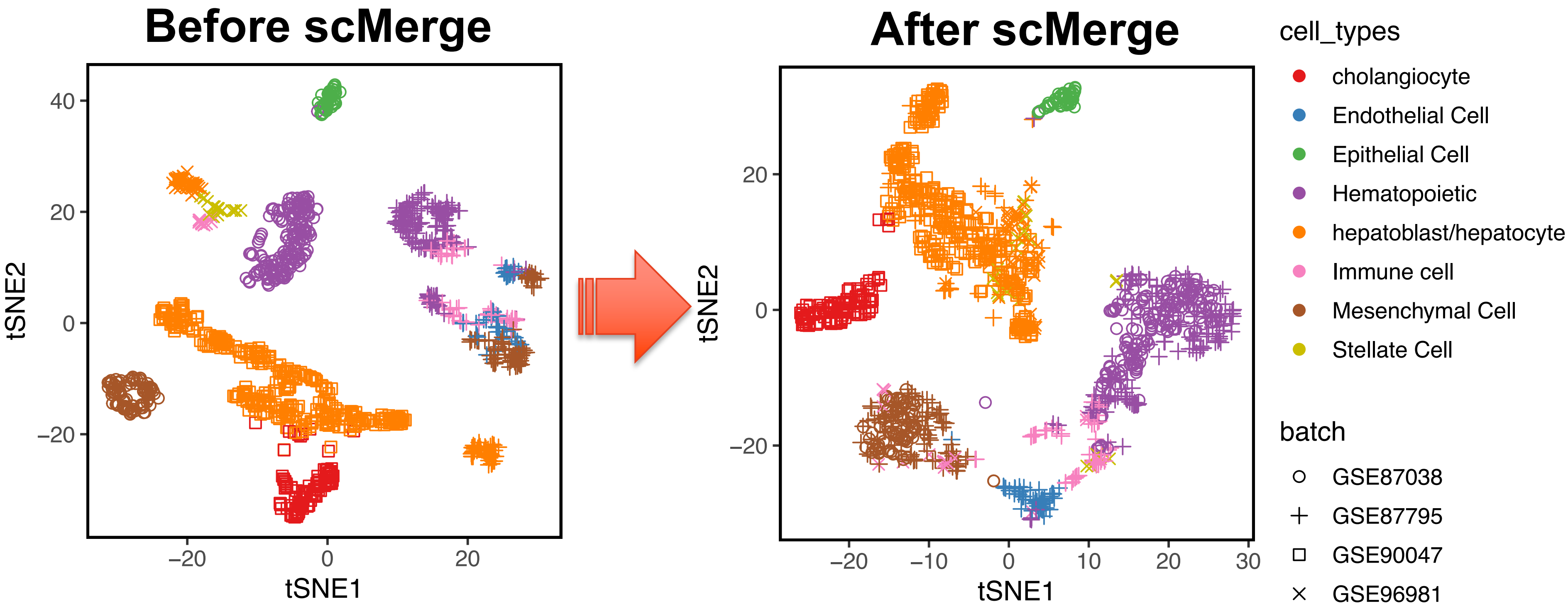
Frame as pseudo-replicate information

Pseudo-replicates



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

Liver fetal development time course data



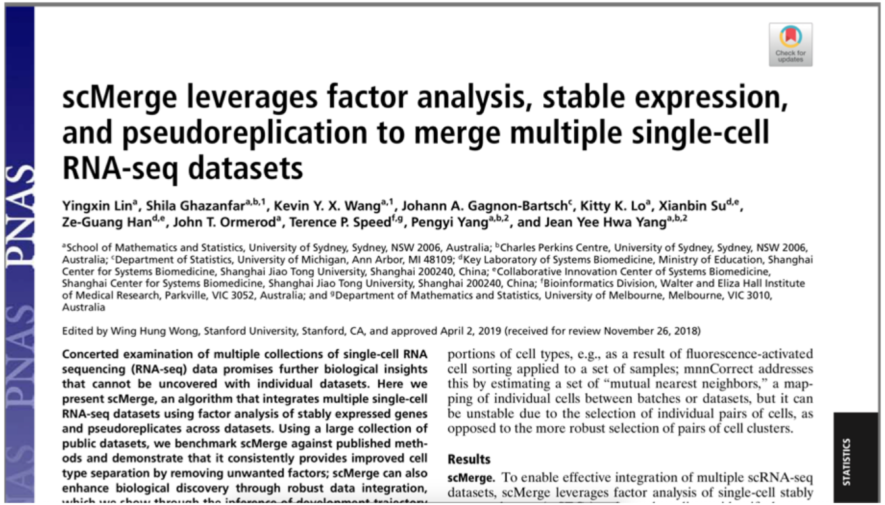
Before scMerge

After scMerge

More information

PNAS:

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



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

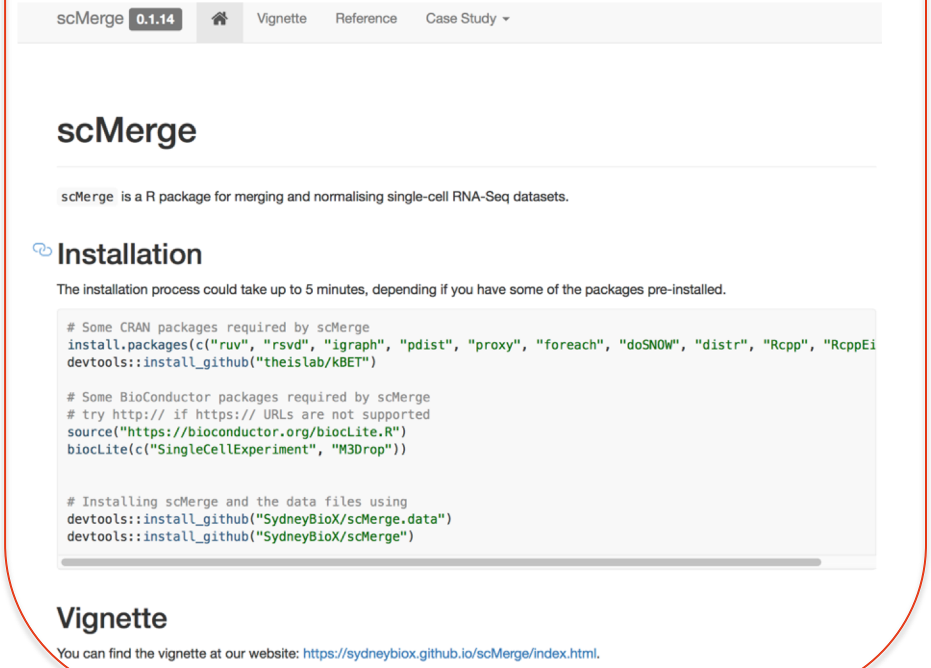
portions of cell types, e.g., as a result of fluorescence-activated cell sorting applied to a set of samples; mnnCorrect addresses this by estimating a set of "mutual nearest neighbors," a mapping of individual cells between batches or datasets, but it can be unstable due to the selection of individual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

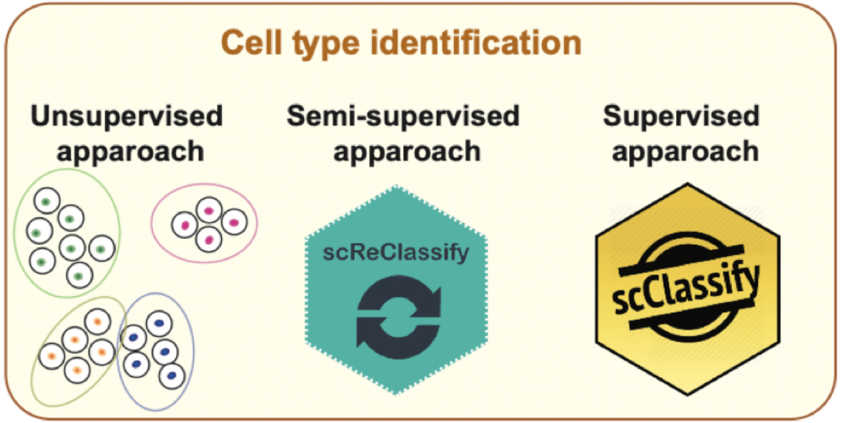
Vignette

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

Cell type identification - clustering

THE UNIVERSITY OF
SYDNEY

Component 4: Cell type identification



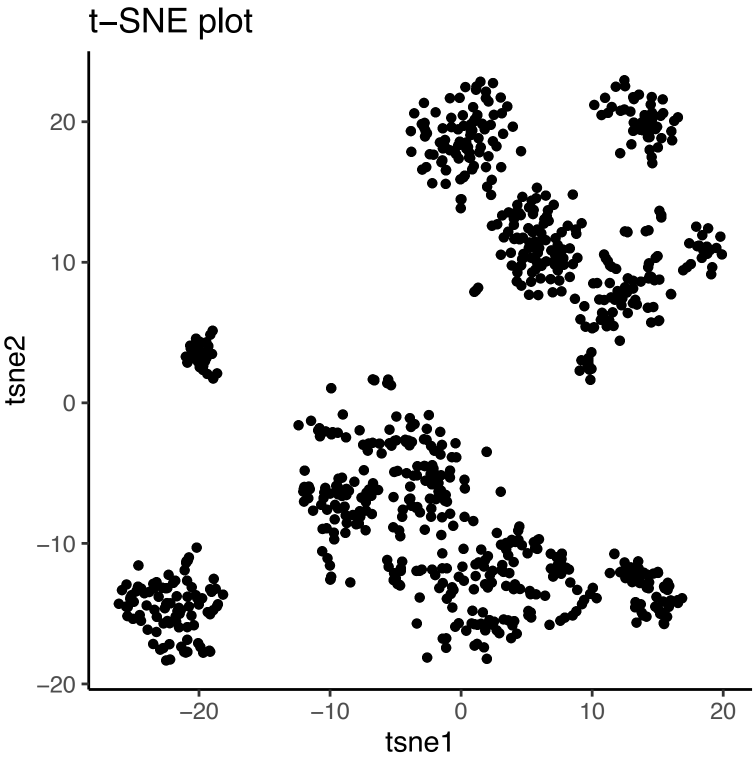
Science questions

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

Analysis techniques

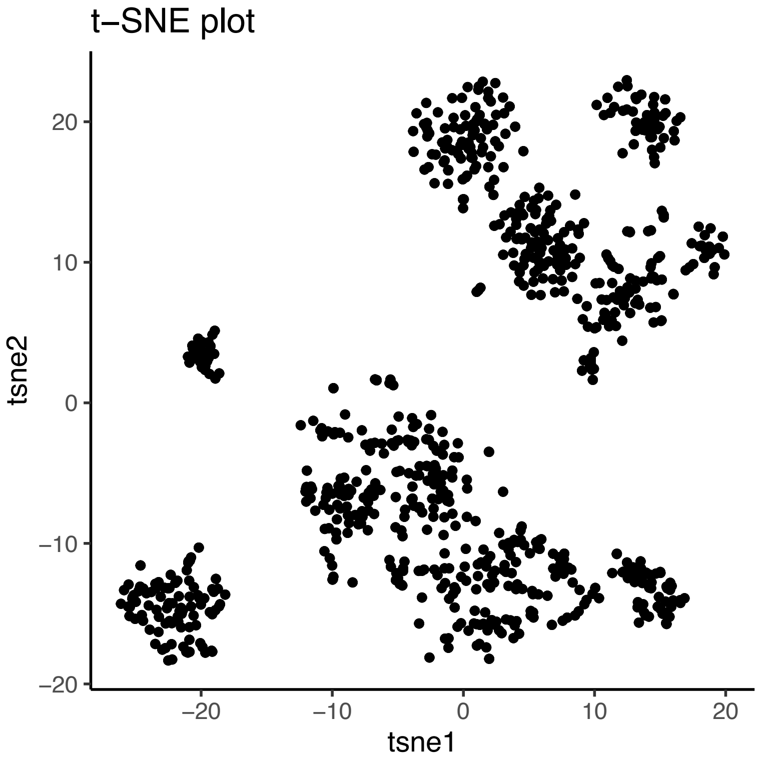
- Visualization (dimension reduction)
- **Clustering (unsupervised learning)**
- Classification (supervised learning)

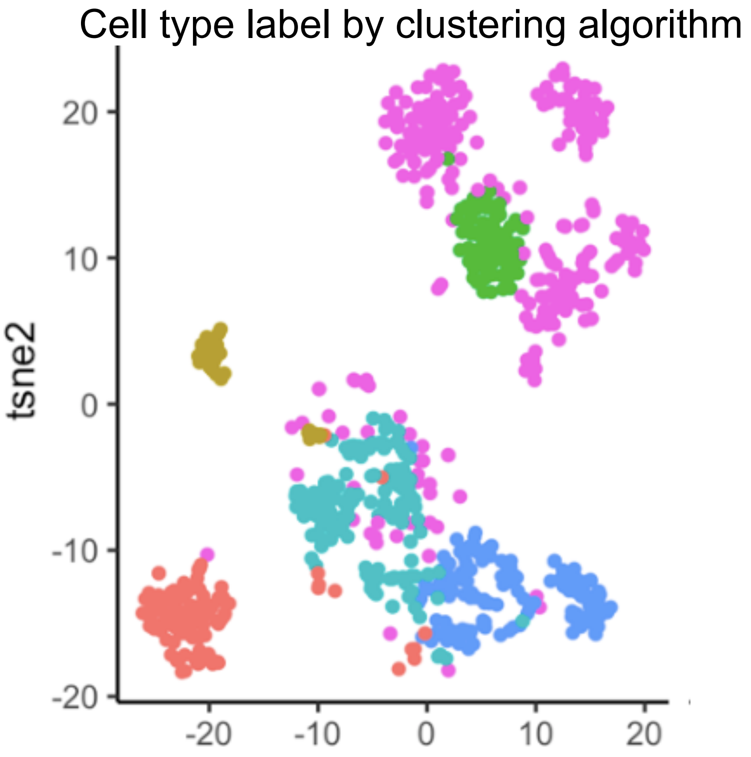
tSNE dimension reduction



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

tSNE dimension reduction + clustering





Clustering algorithms for scRNA-seq

k-means

Hierarchical

RaceID

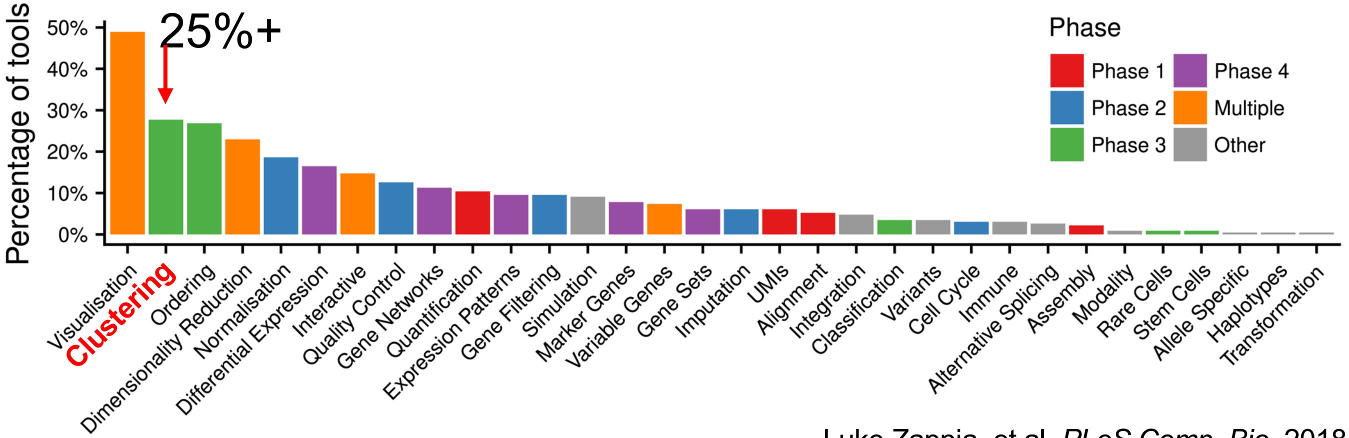
SC3

CIDR

countClust

RCA

SIMLR



25%+

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

Which clustering method should I pick?

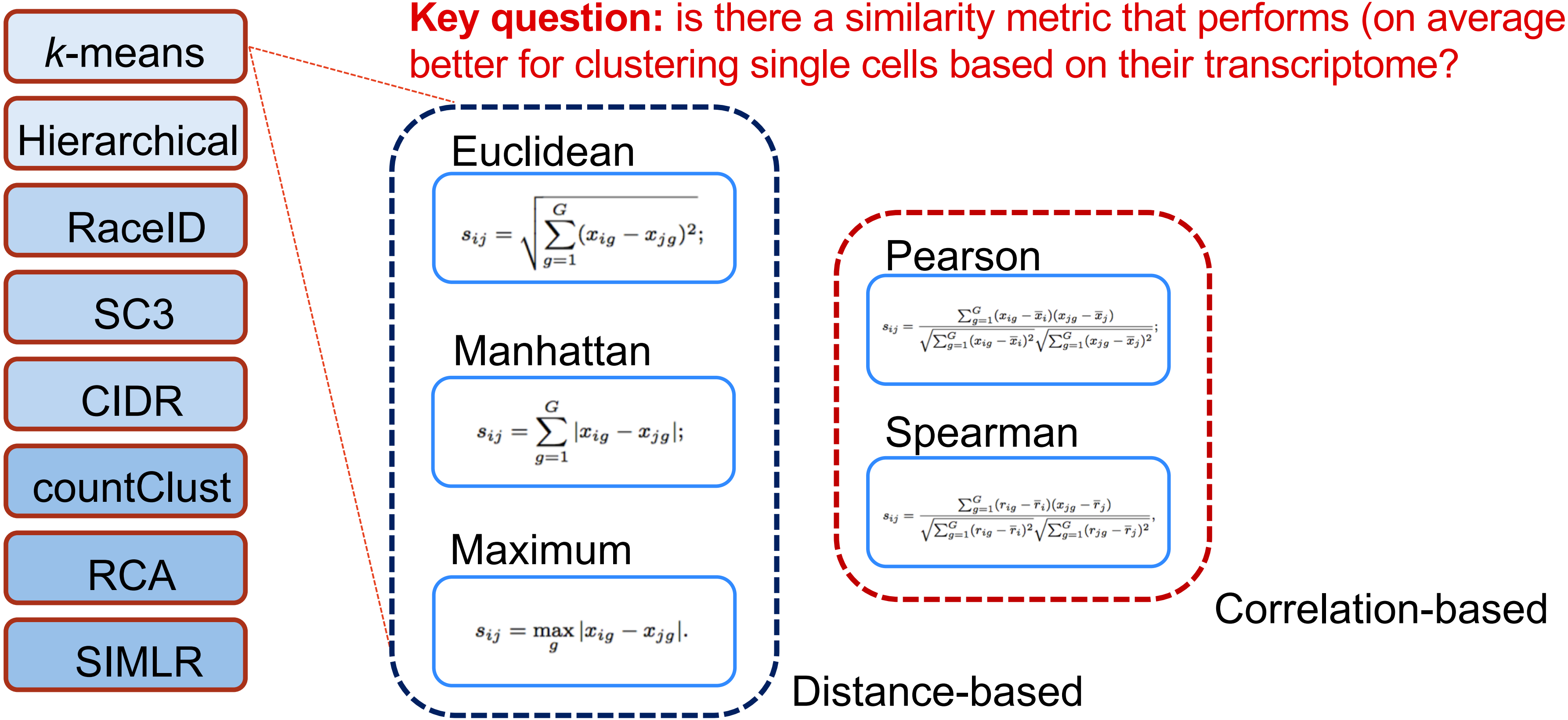
- Different methods make different assumptions, which may or may not be satisfied by your data

- Try a few different ones to understand what makes a method work well for your own data

- We did the same and found **similarity metrics** has a huge impact on performance of methods

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

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

Spearman

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

Correlation-based

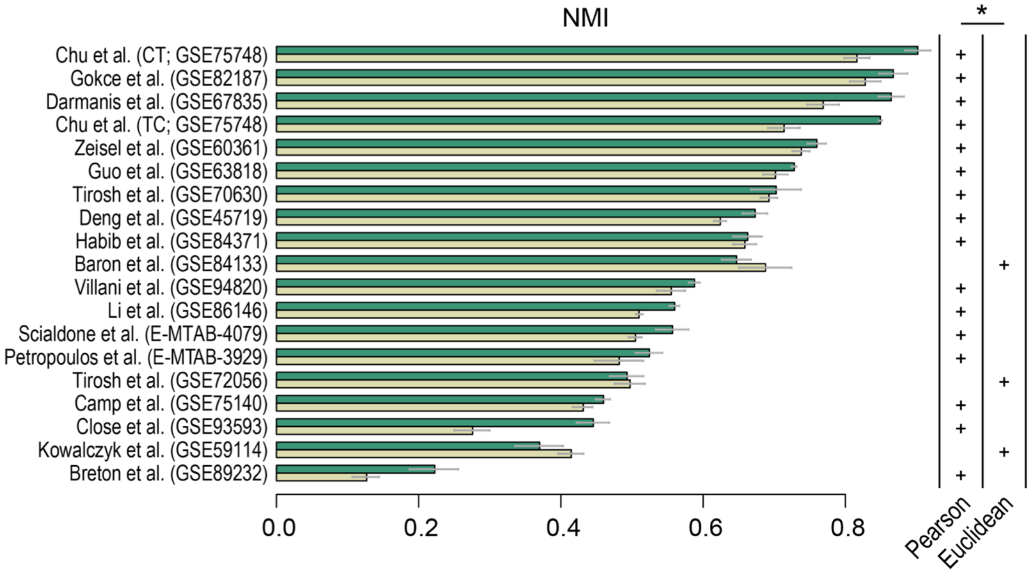
scClust: improved clustering methods using correlation metrics

SIMLR

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

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

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.



PhD student: Taiyun Kim

scClassify

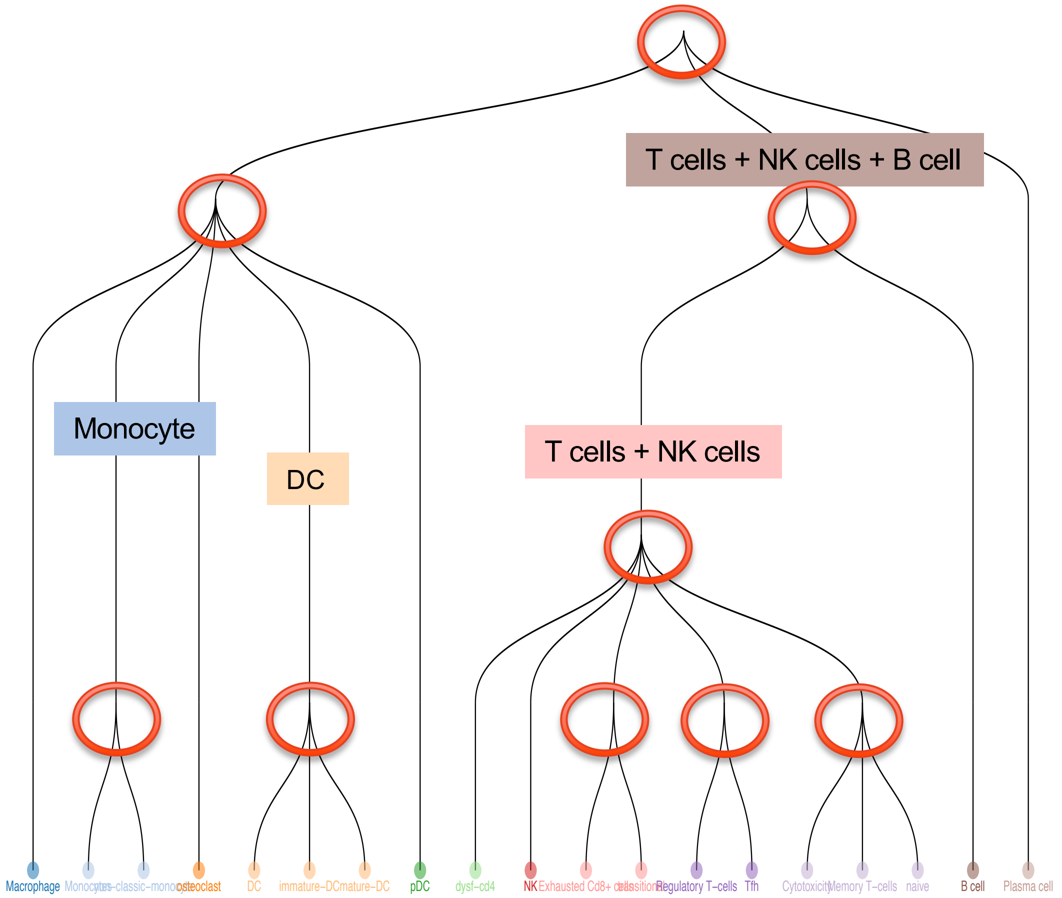
Feature selection at each branch point.

Features are selected from :

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

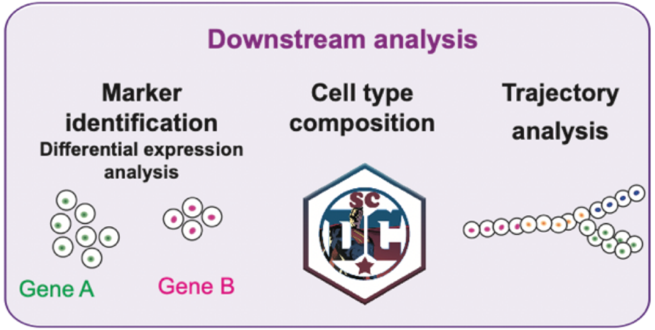
……

PhD student: Yingxin Lin



Downstream analysis

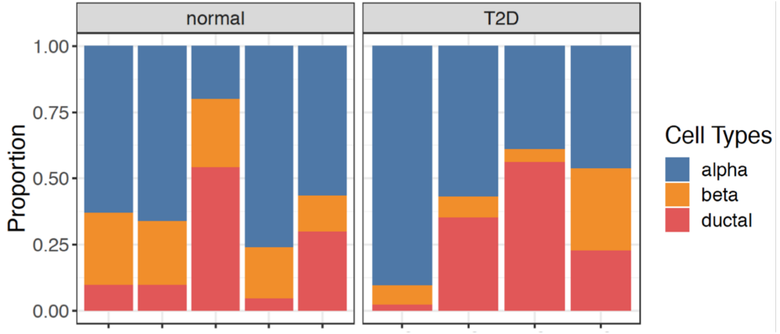
Component 5: Downstream analysis

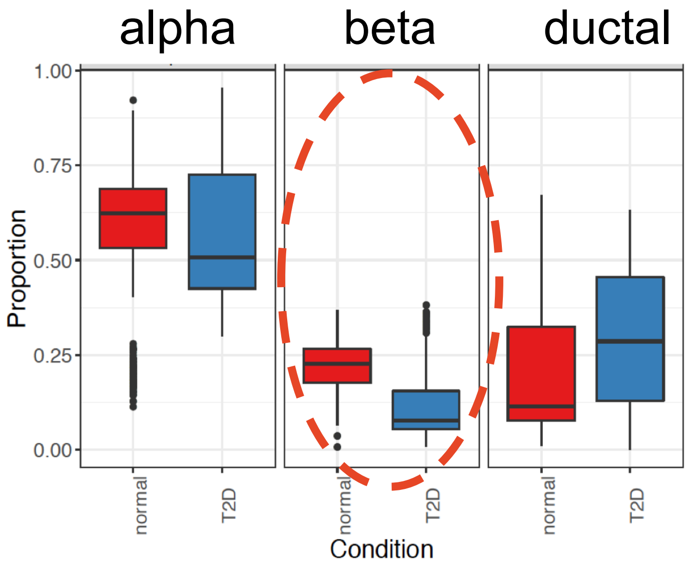


Science questions

- 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?

Compare these proportions





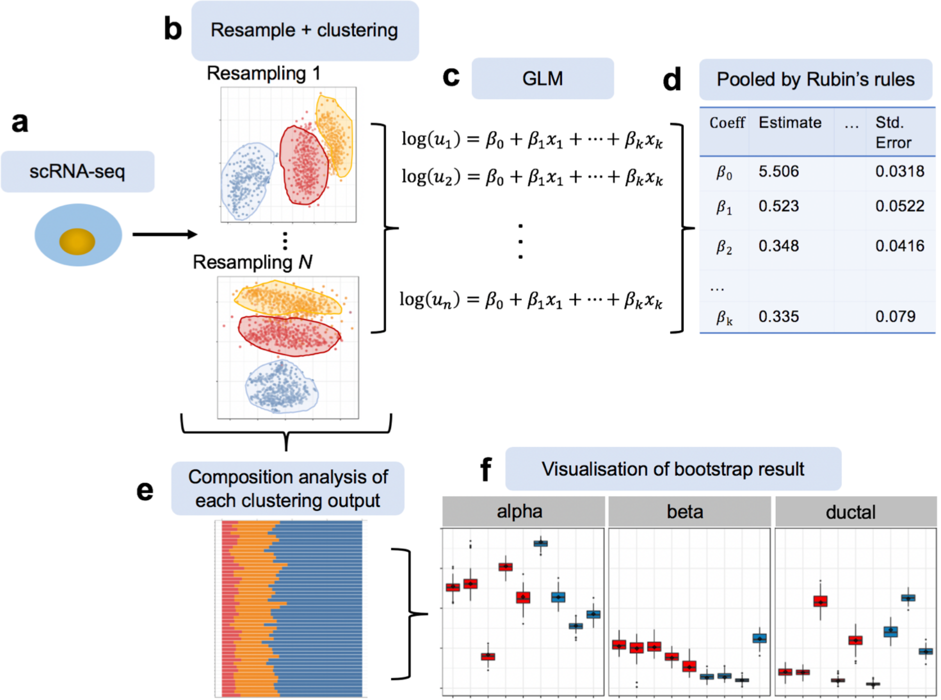
Single cell Differential Composition (scDC)

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)
- Cell – type proportions standard error from bootstrap samples
- Calculation of pooled log-linear model using Rubin's pooled estimate

PhD student: Yue Cao



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

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

$$\vdots$$

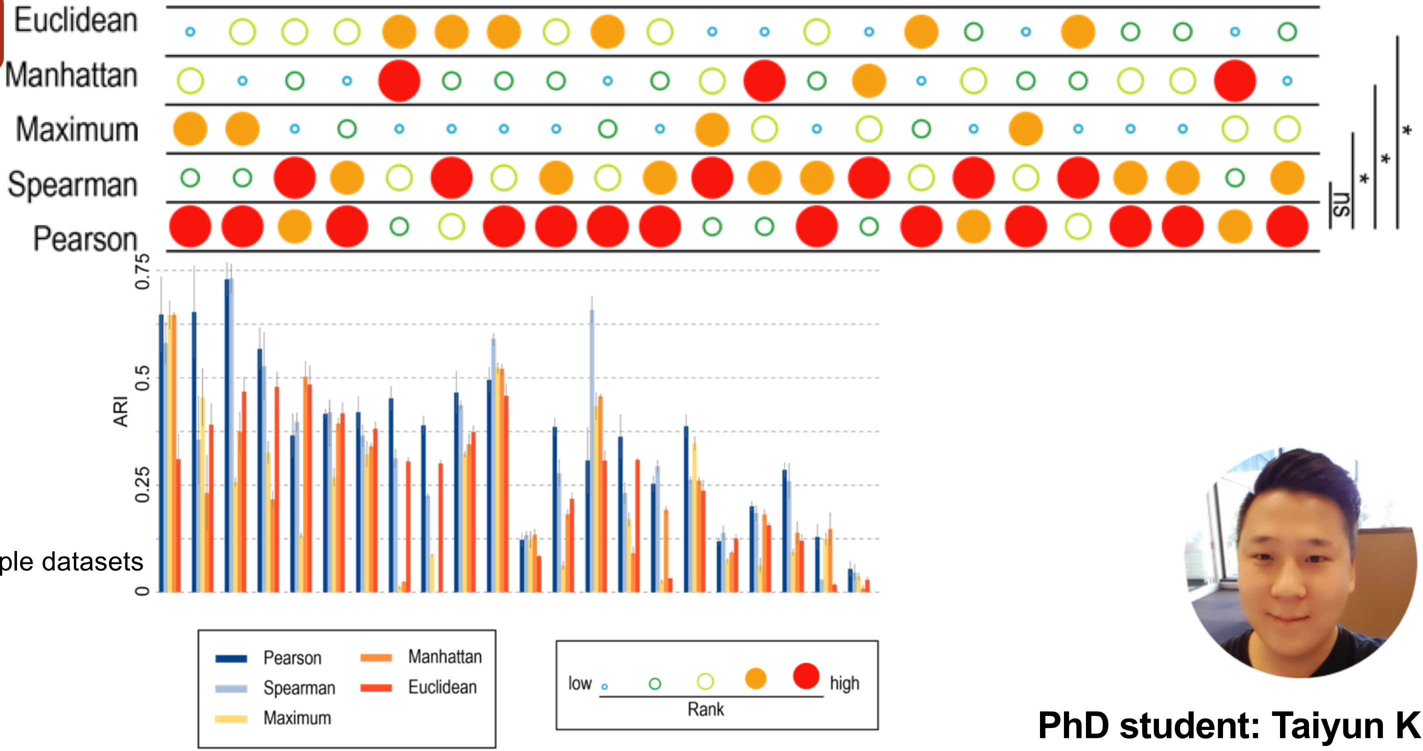
$$\log(u_n) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

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

Additional slides

Evaluation results (against the pre-defined cell types)

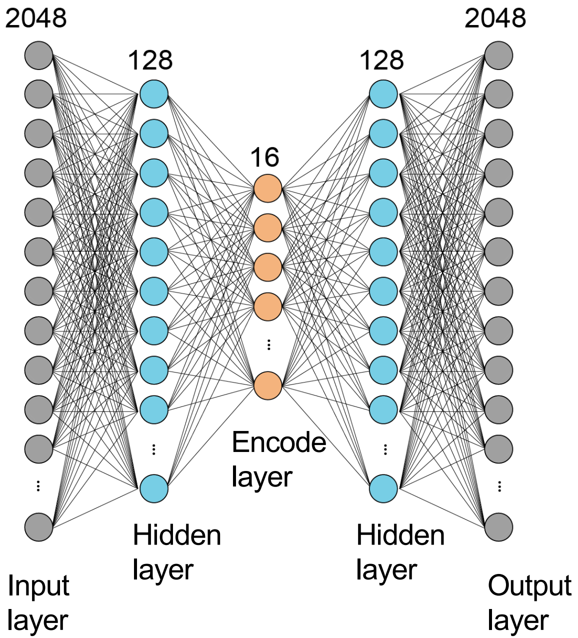
k-means



Multiple datasets

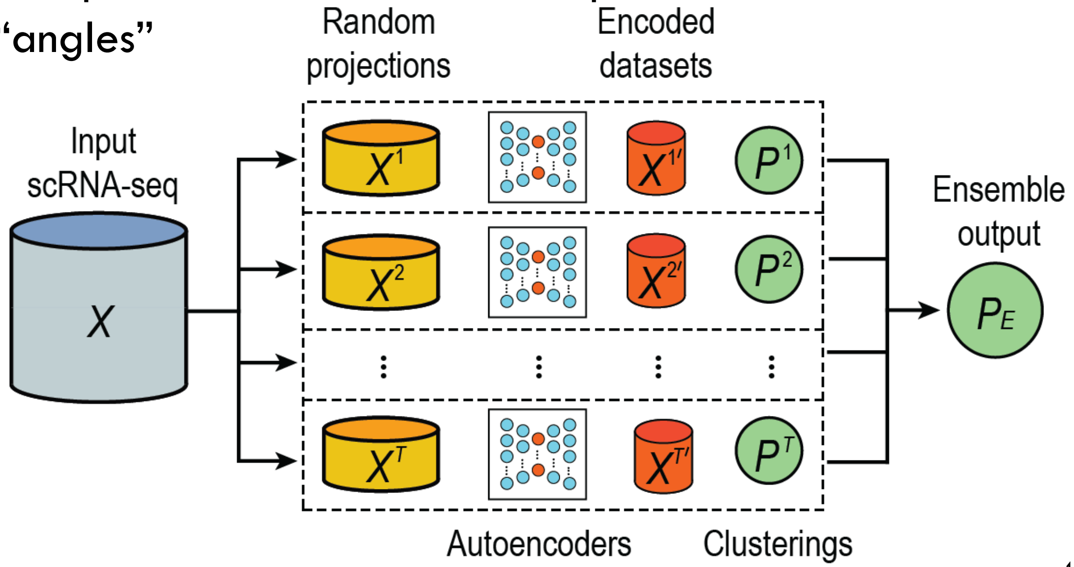
PhD student: Taiyun Kim

Dimension reduction using an ensemble of autoencoders



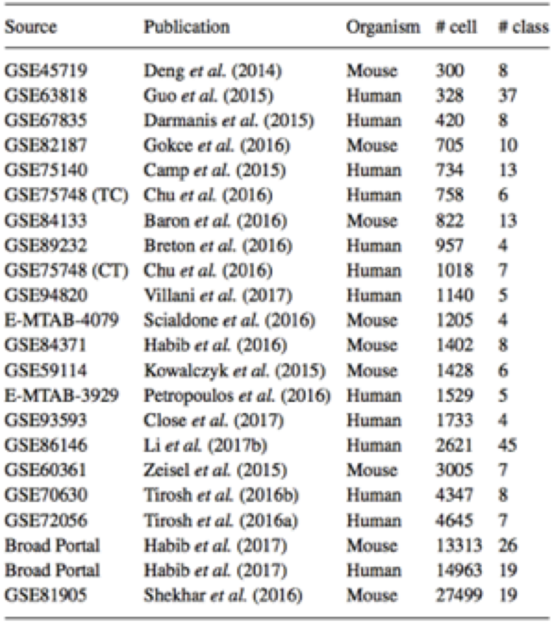
Autoencoder, a deep learning model, allows nonlinear dimension reduction

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

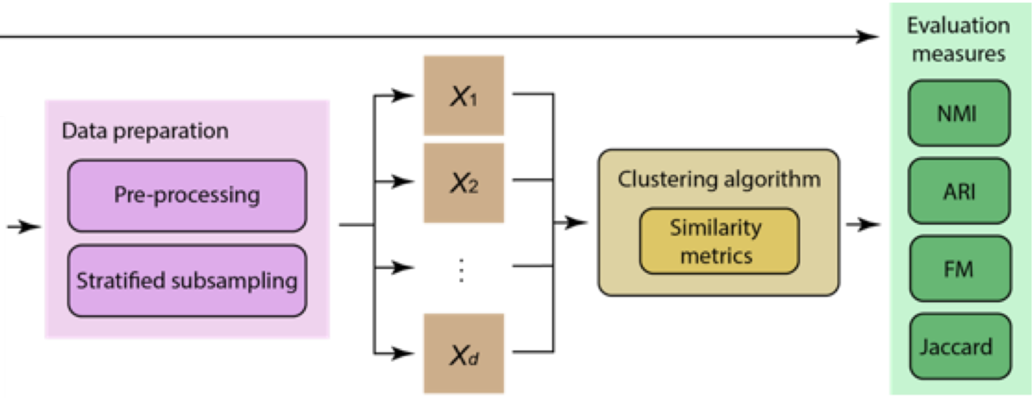


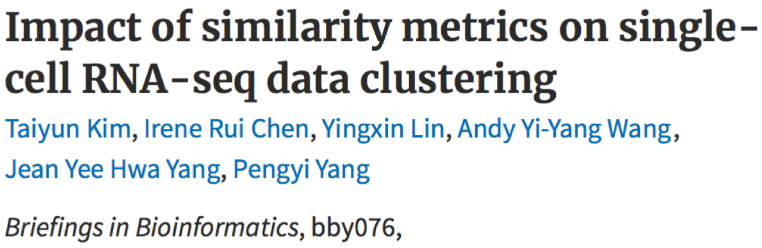
Geddes T *et al.*, Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis, ***BMC Bioinformatics*** (2019)

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





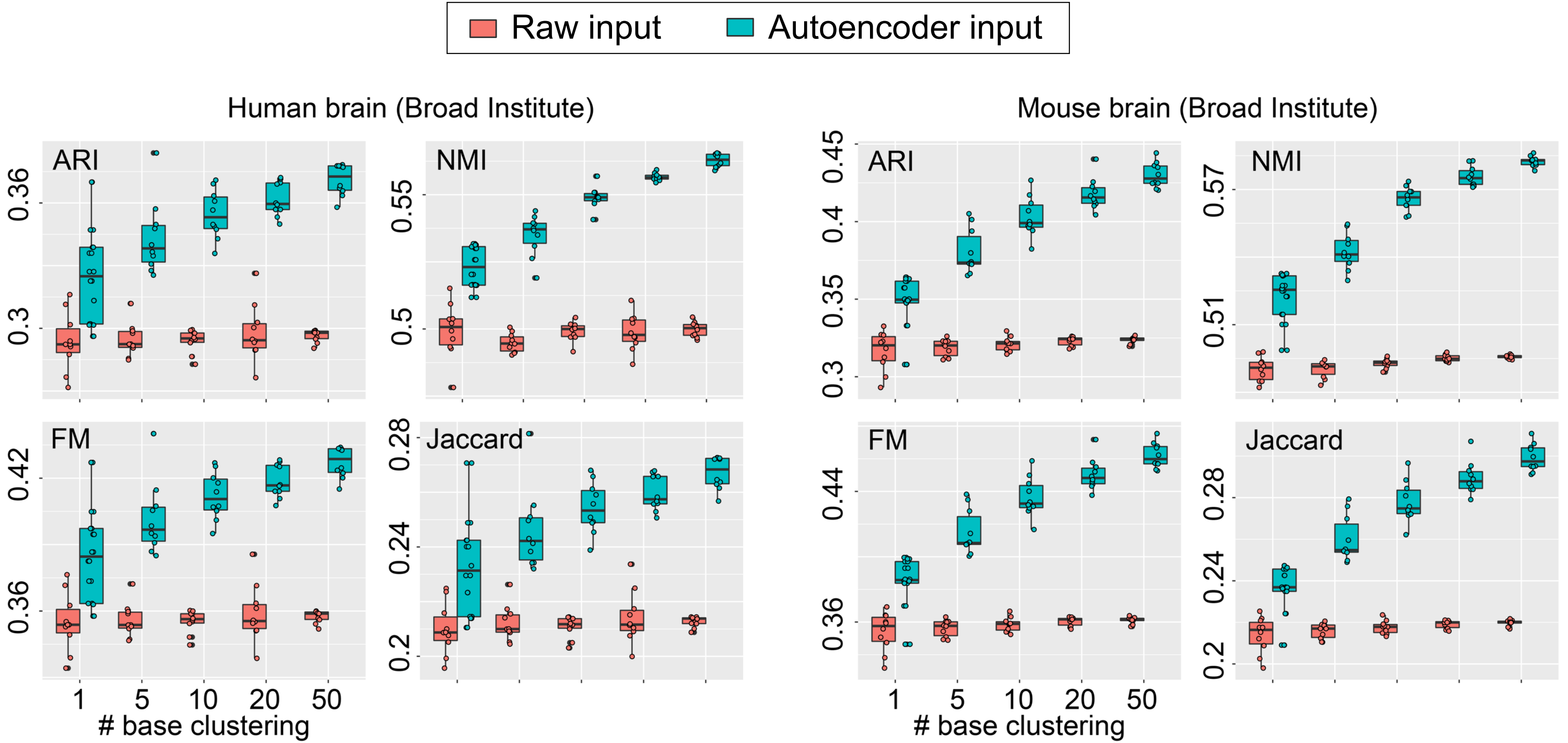
Impact of similarity metrics on single-cell RNA-seq data clustering

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

Briefings in Bioinformatics, bby076,

Taiyun Kim

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



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

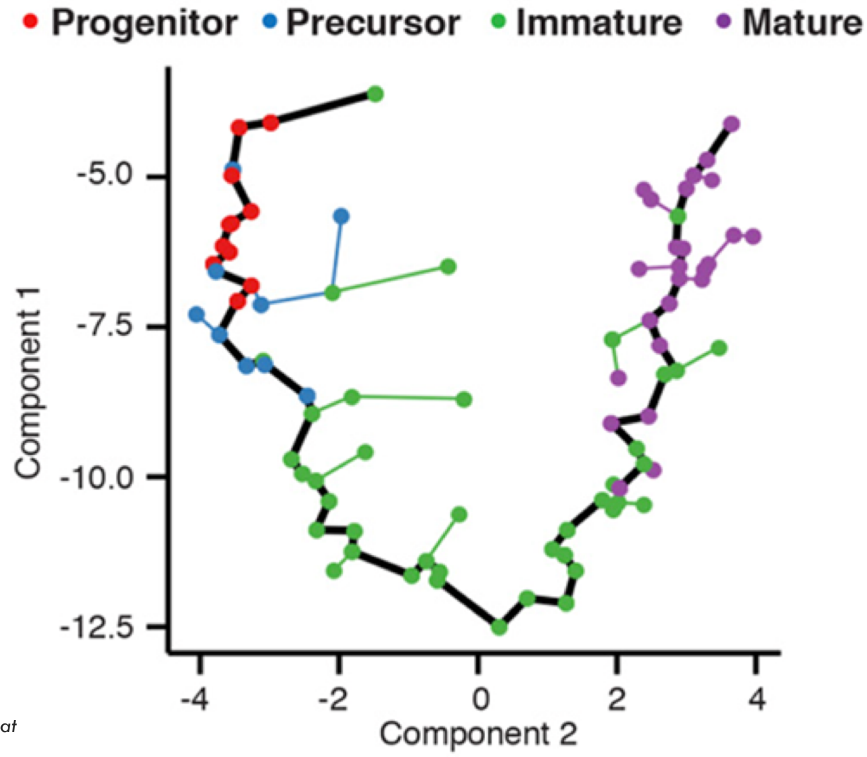
- Simple statistical test
  - Wilcoxon rank test, t-test
- Methods developed for bulk RNAseq DE
- DESeq2
  - EdgeR
  - Voom-Limma
- scRNA specific
  - MAST
  - DECENT
  - D3E
  - …. many more!

Trajectory analysis

- Inference on a dynamic process such as cell cycle/differentiation

- Dimensional reduction to learn the key genes

- Trees are then grown to connect the cell types



Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019). https://doi.org/10.1038/s41587-019-0071-9