

scMerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo- replication

Presented by

Dr Shila Ghazanfar

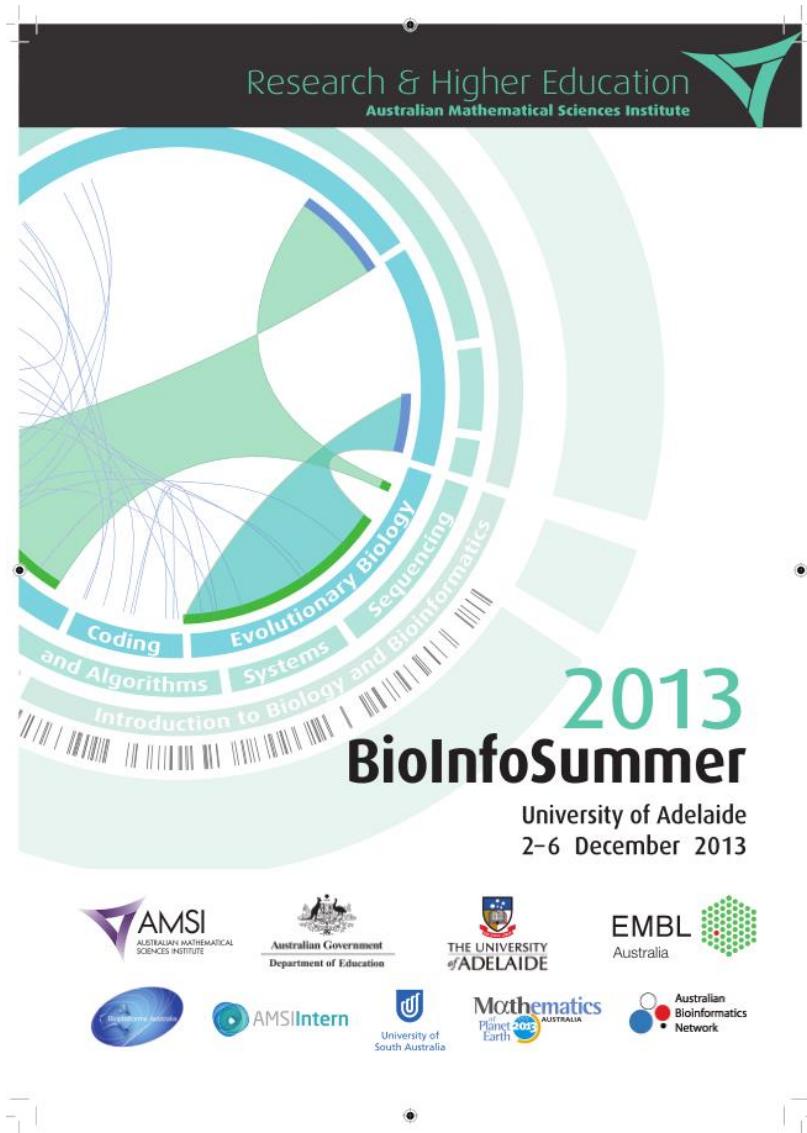
Judith and David Coffey LifeLab
Charles Perkins Centre

School of Mathematics and Statistics
The University of Sydney

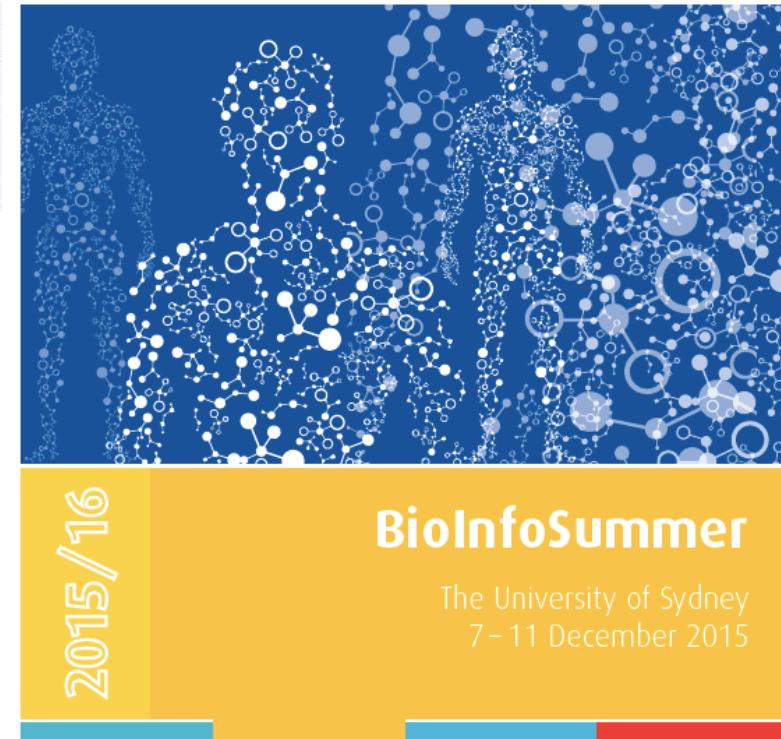
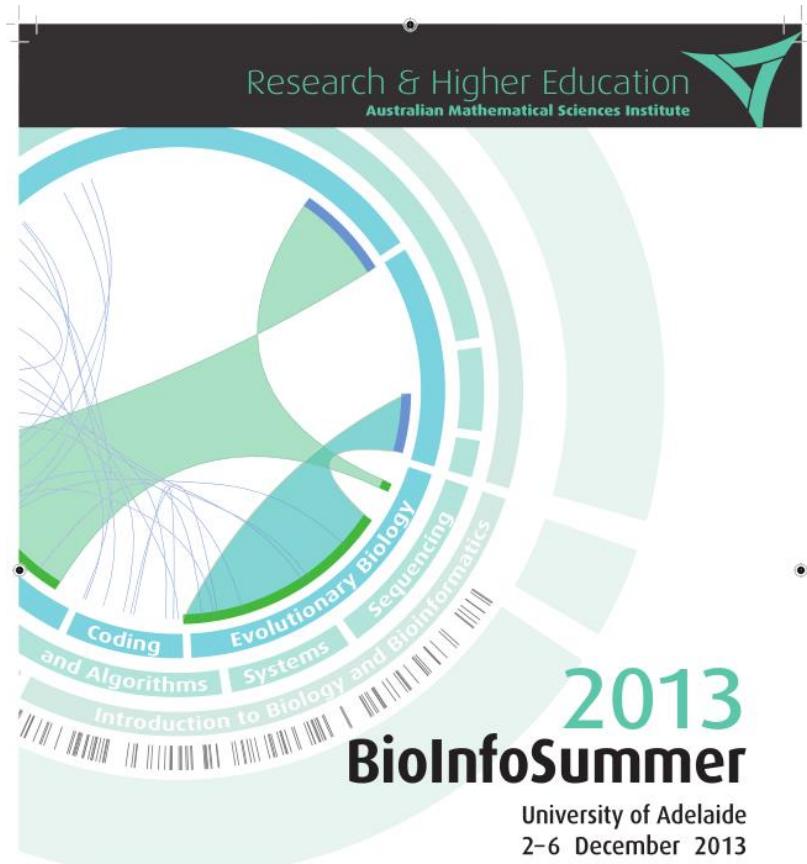
@shazanfar



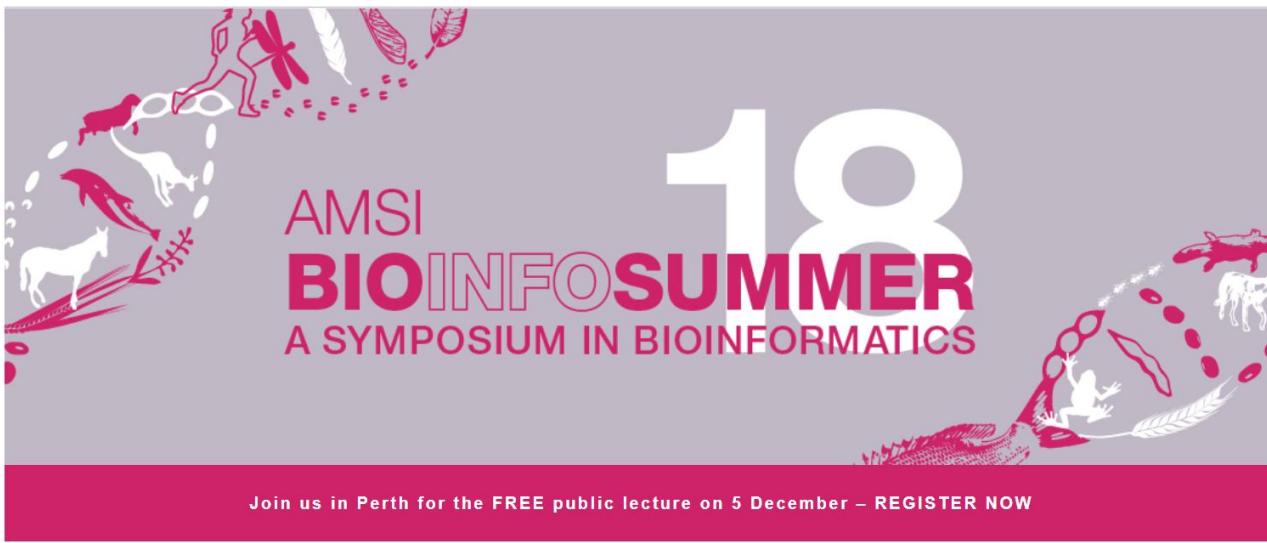
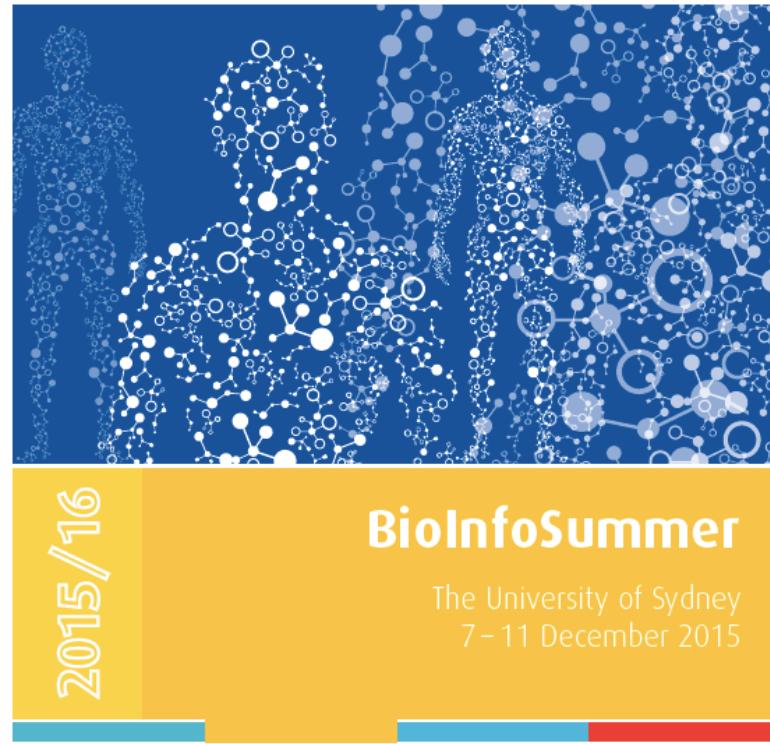
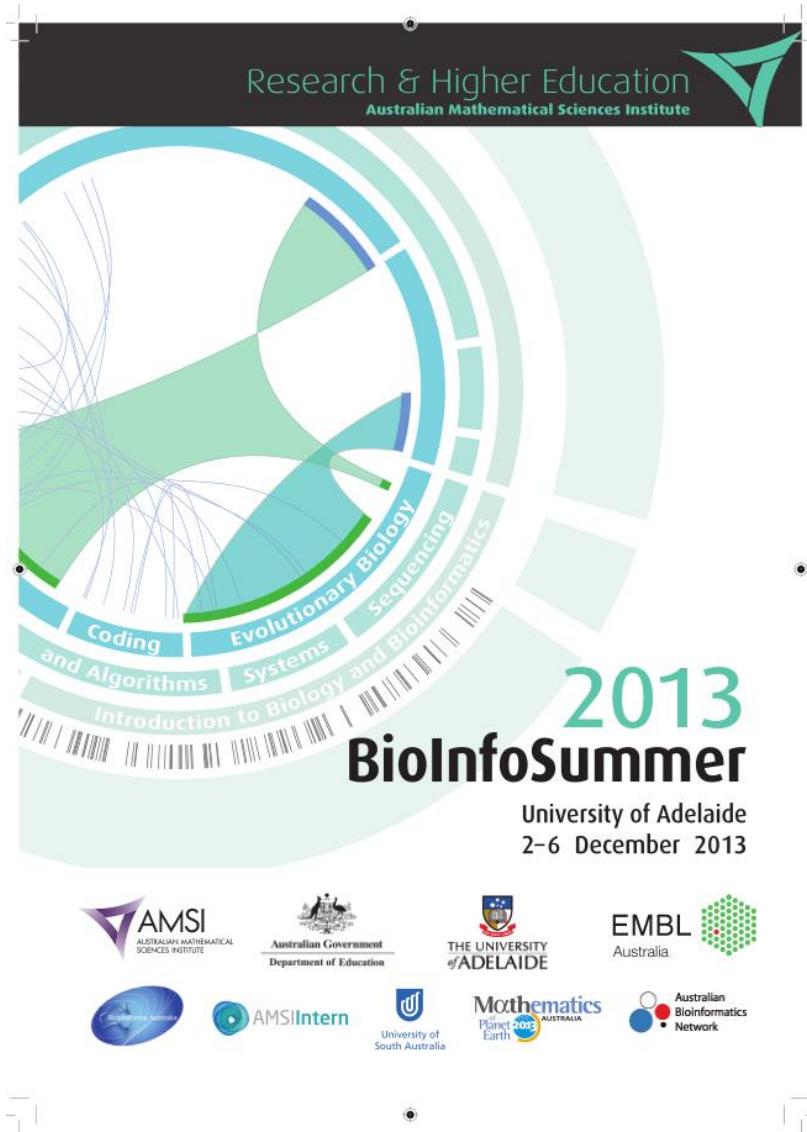
A brief history



A brief history



A brief history



Sydney Precision Bioinformatics Research Group



THE UNIVERSITY OF
SYDNEY

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

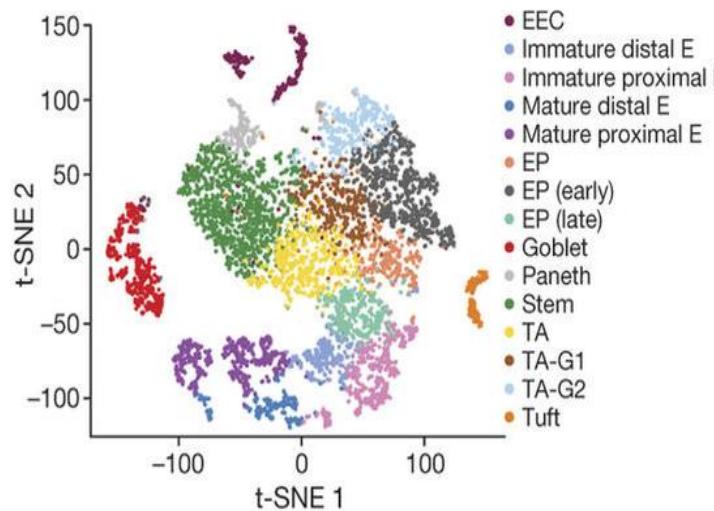
Meet our senior and junior research leaders ...



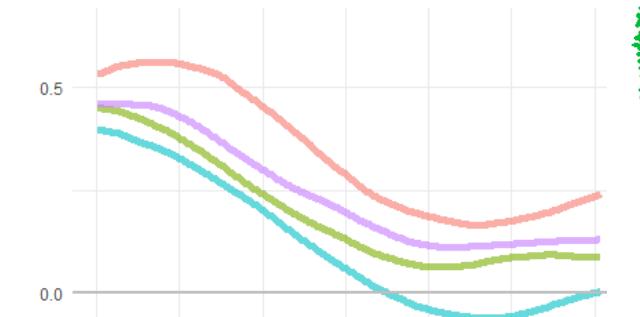
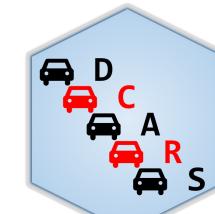
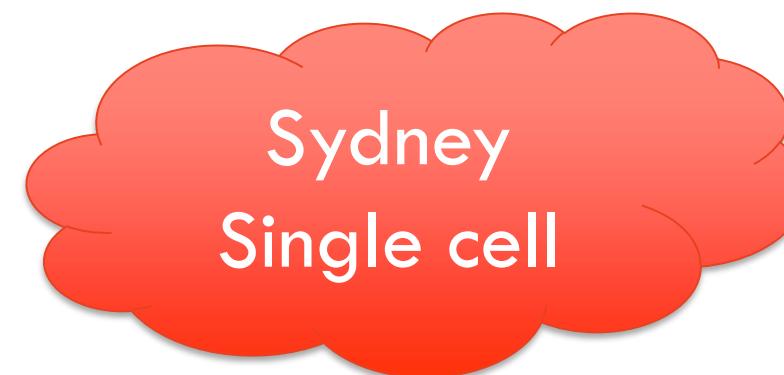
... and senior research associates: 4; PhD candidates: 20; Honours and TSP students: 8

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

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

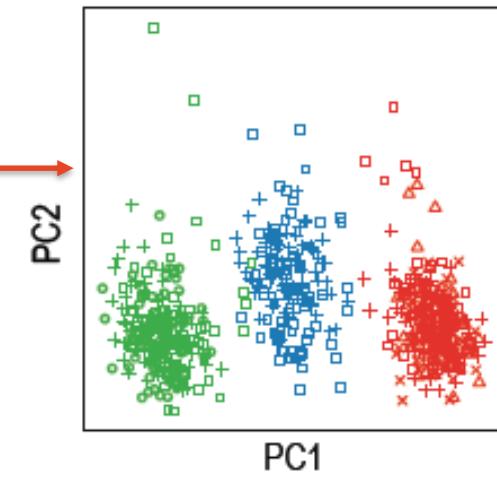
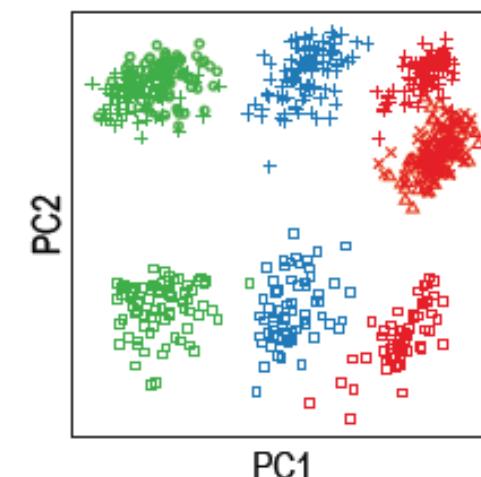
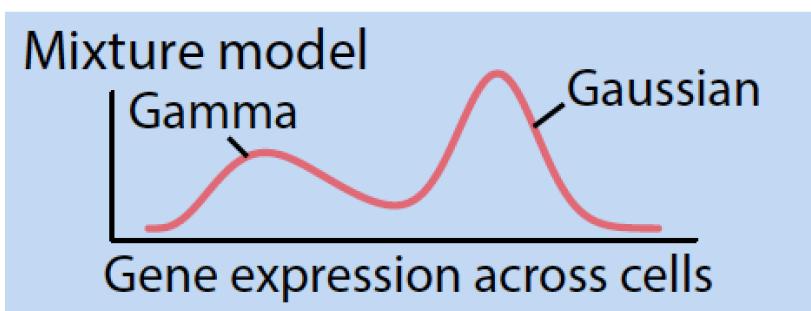


Clustering metrics

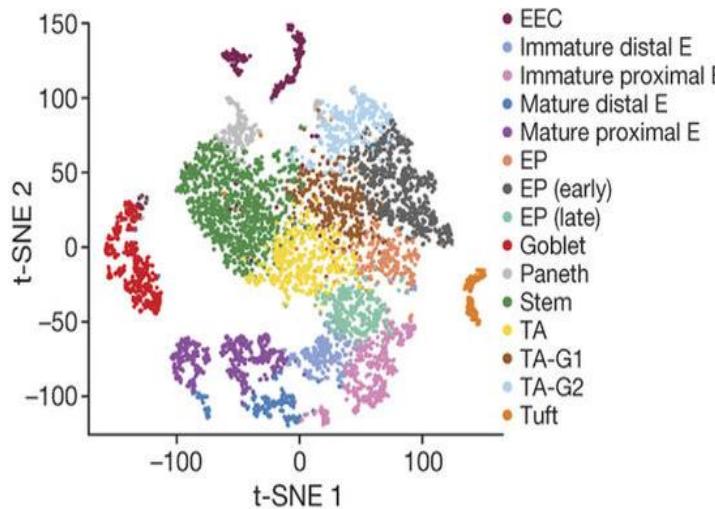


Differential correlation

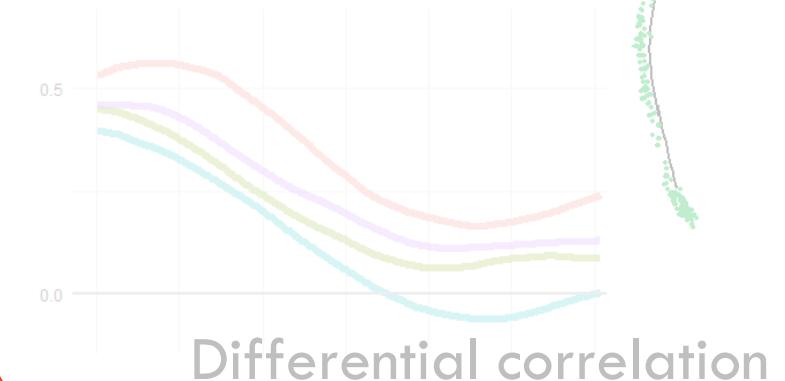
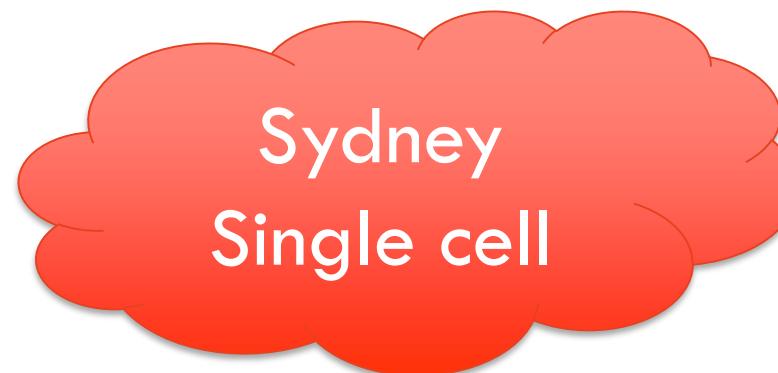
Finding stably expressed genes



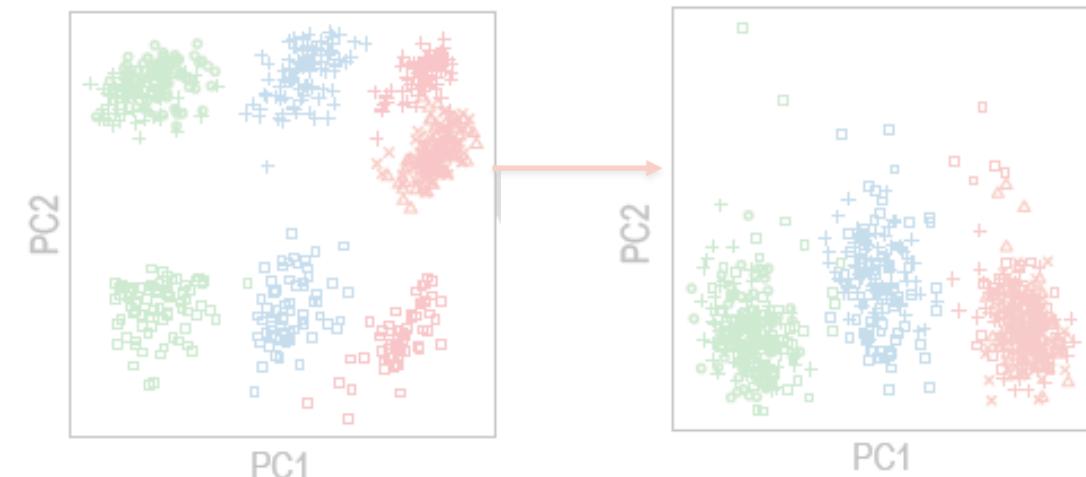
scMerge



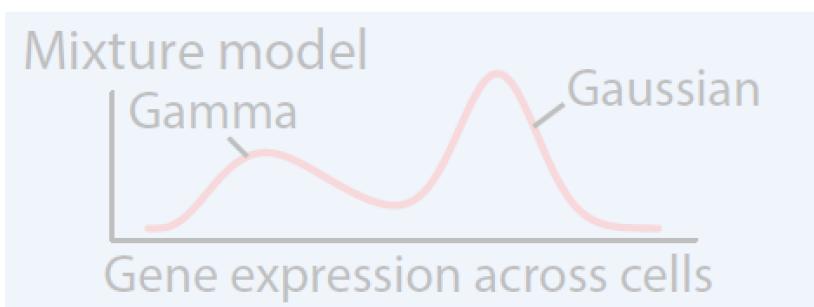
Clustering metrics



scMerge



Finding stably expressed genes



Key underlying aspect of clustering? Similarity metrics

k-means

Hierarchical

RacelD

SC3

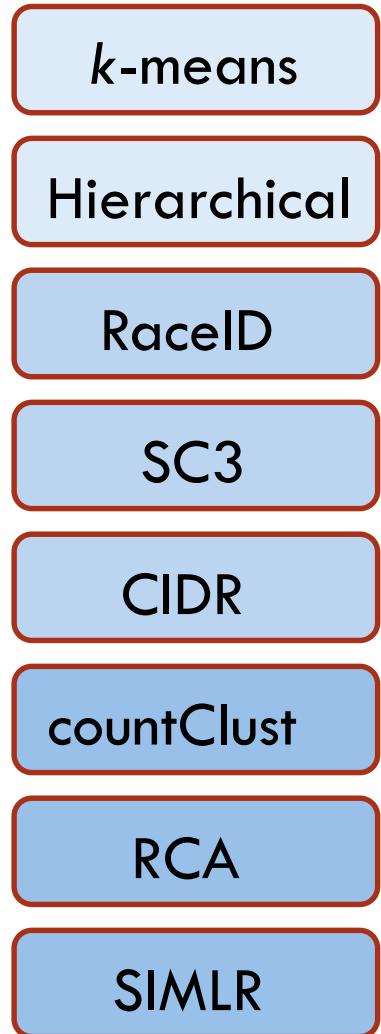
CIDR

countClust

RCA

SIMLR

Key underlying aspect of clustering? Similarity metrics



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

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

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

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

Key underlying aspect of clustering? Similarity metrics

- k-means
- Hierarchical
- RacelID
- 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

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

Pearson

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

Spearmann

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

Correlation-based

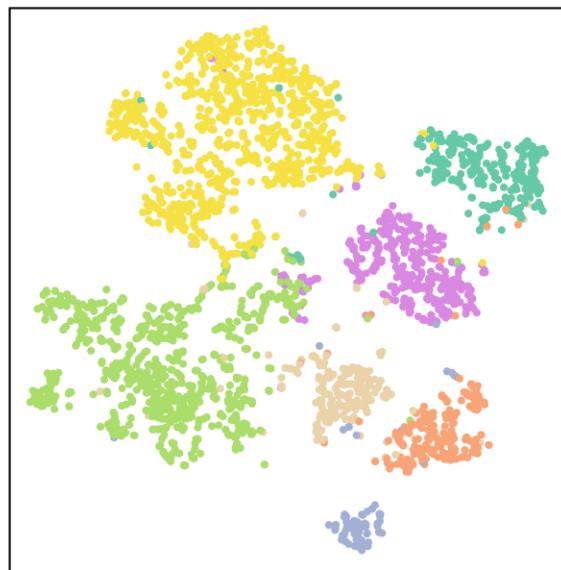
Distance-based

k-means Clustering on GSE60361

k-means

(a)

Annotated cells (GSE60361)

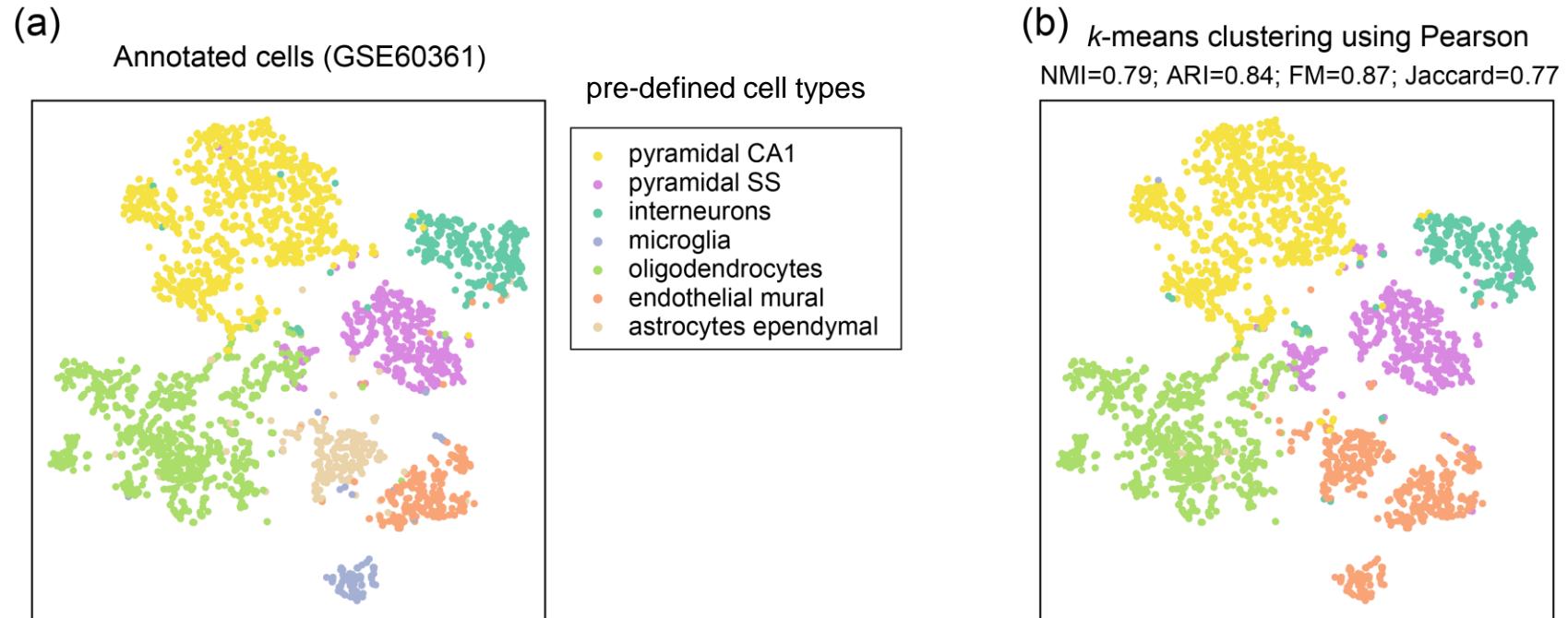


pre-defined cell types

- pyramidal CA1
- pyramidal SS
- interneurons
- microglia
- oligodendrocytes
- endothelial mural
- astrocytes ependymal

k-means Clustering on GSE60361

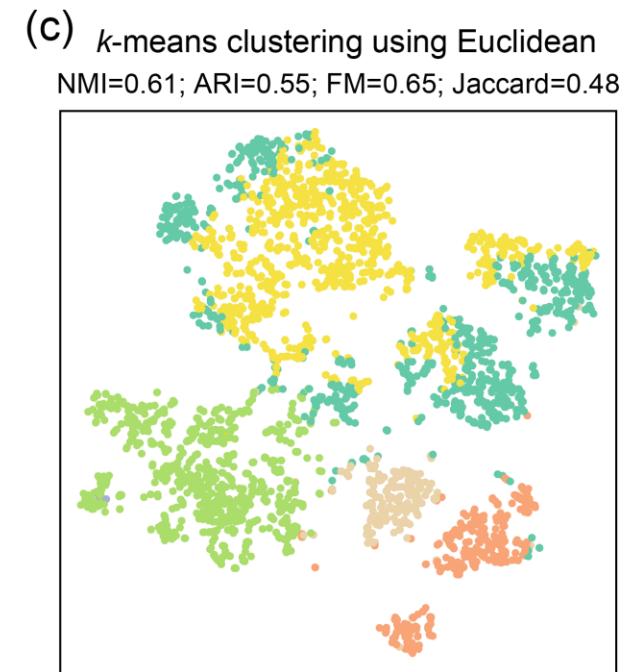
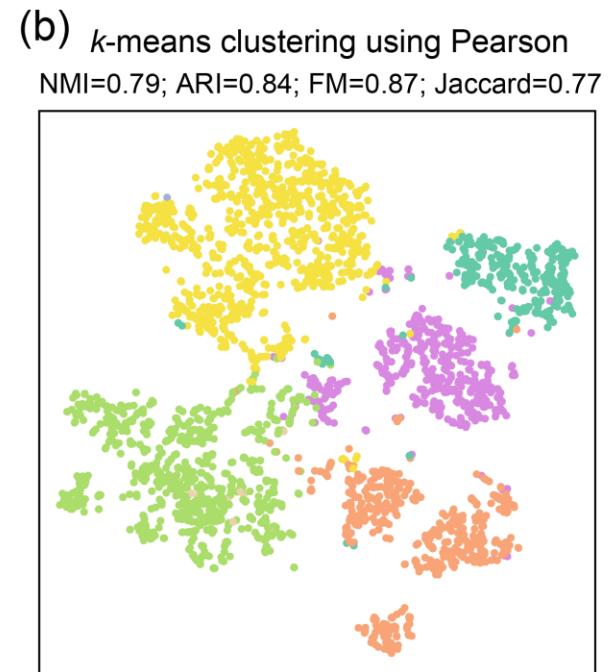
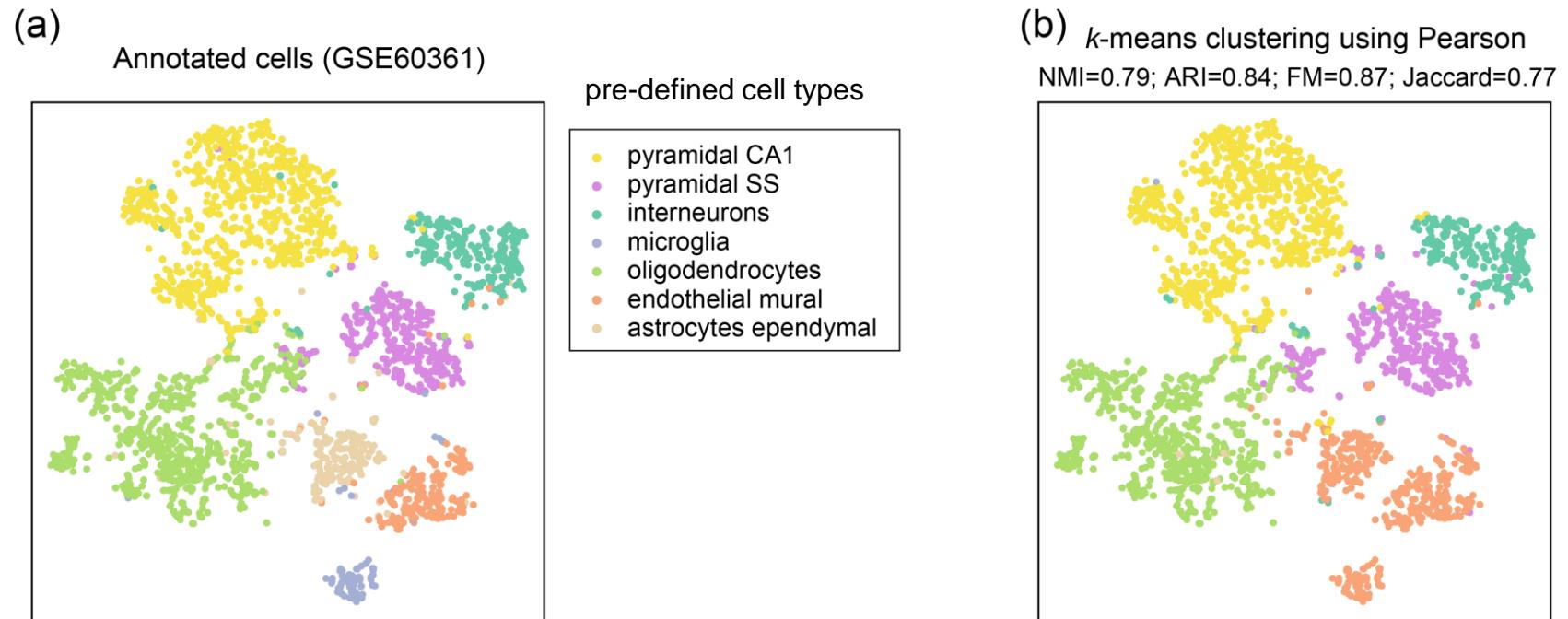
k-means



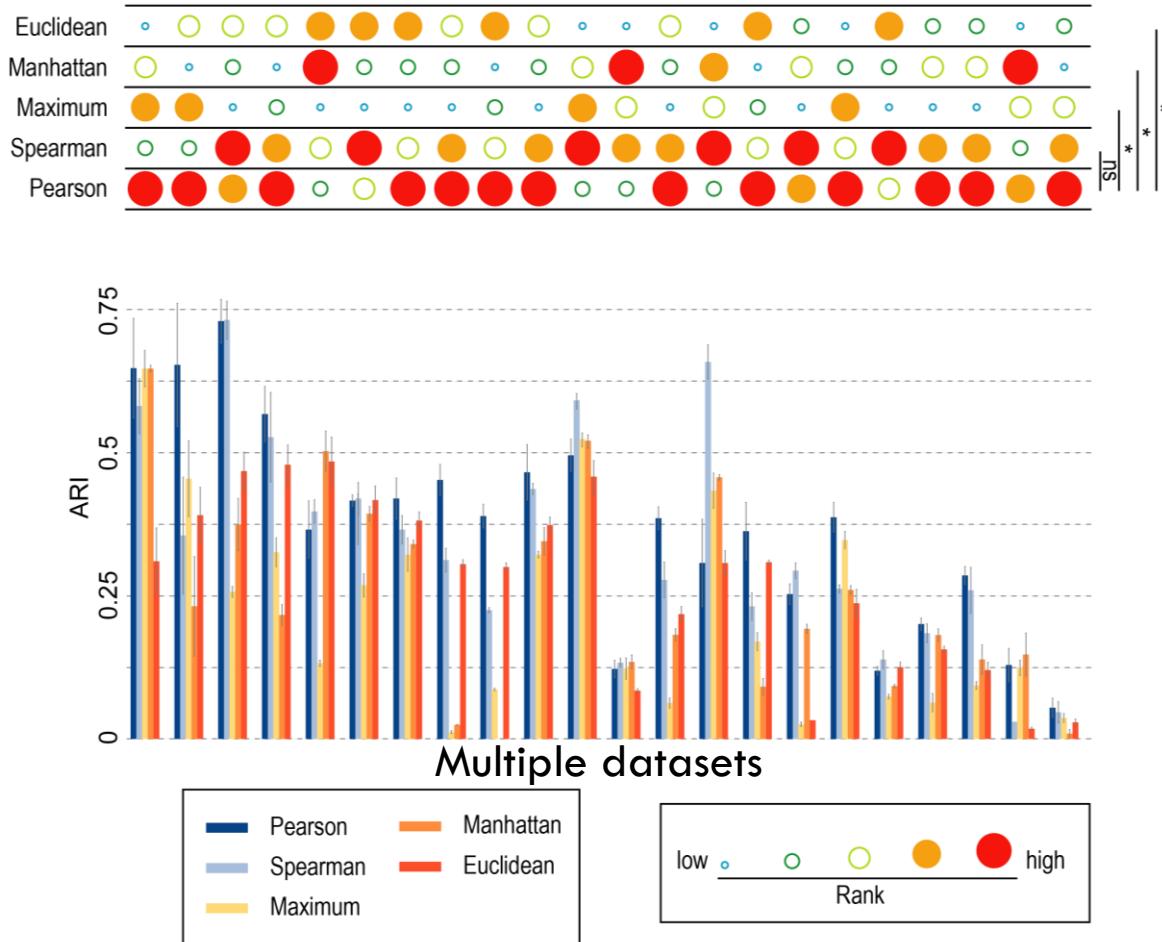
Zeisel A, et al. *Science* 2015

k-means Clustering on GSE60361

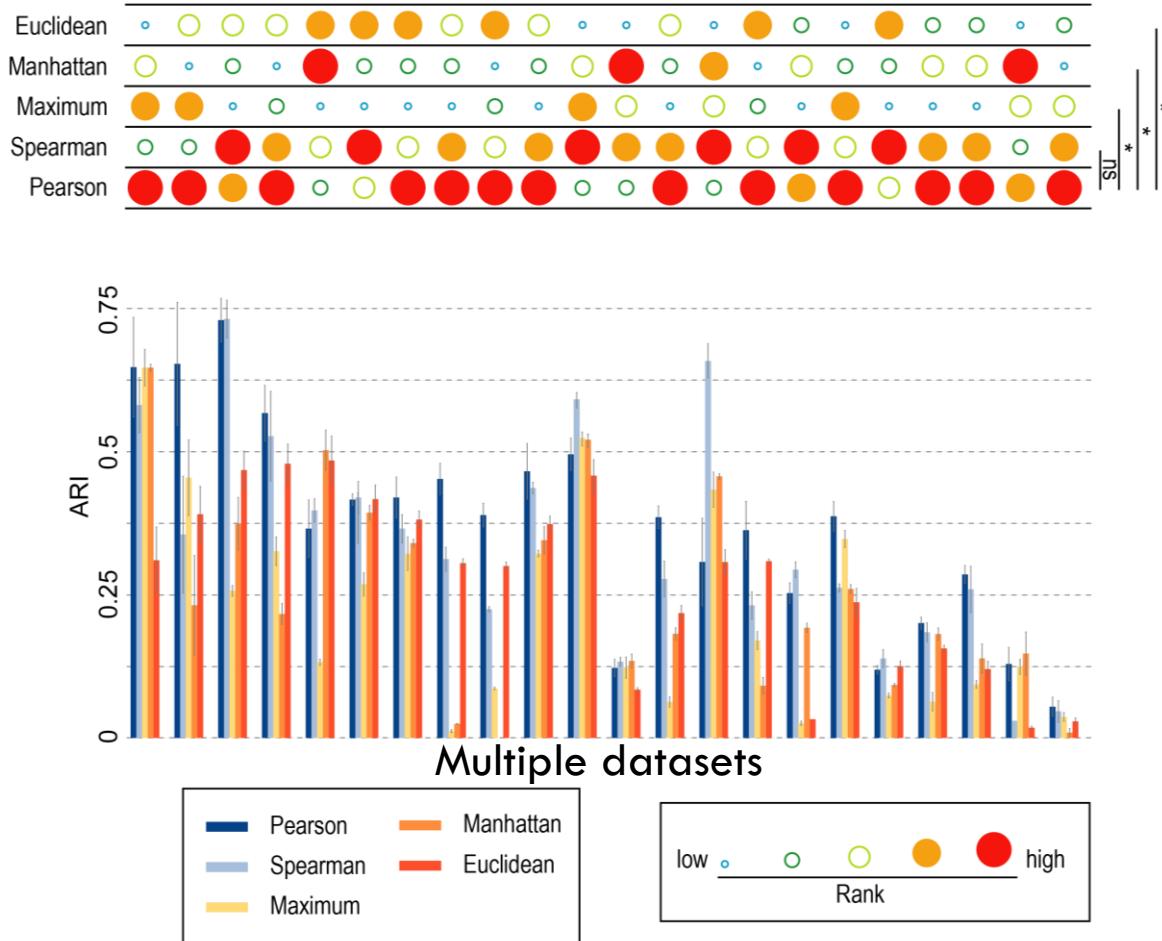
k-means



Evaluation results (against the pre-defined cell types)



Evaluation results (against the pre-defined cell types)



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

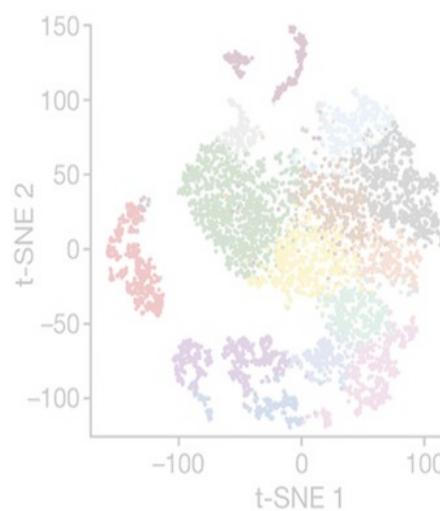
Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang,
Jean Yee Hwa Yang, Pengyi Yang

Briefings in Bioinformatics, bby076,



Pengyi Yang

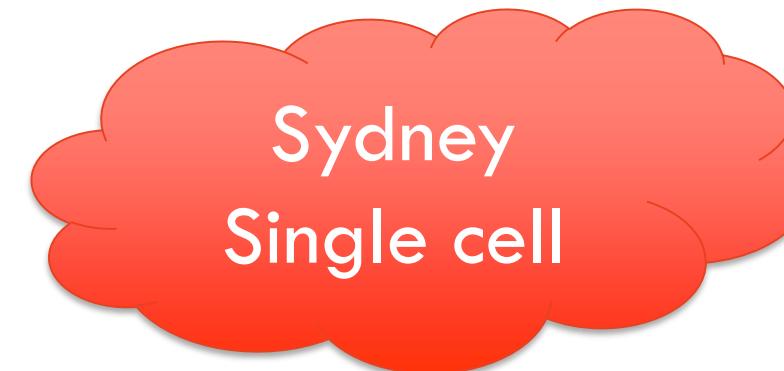
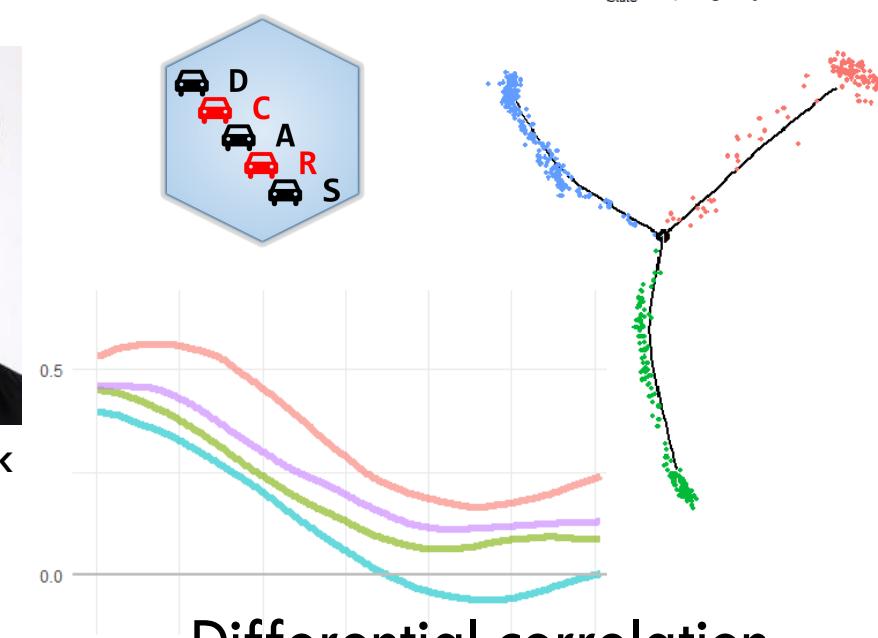
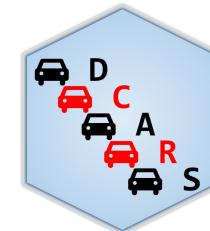
Taiyun Kim



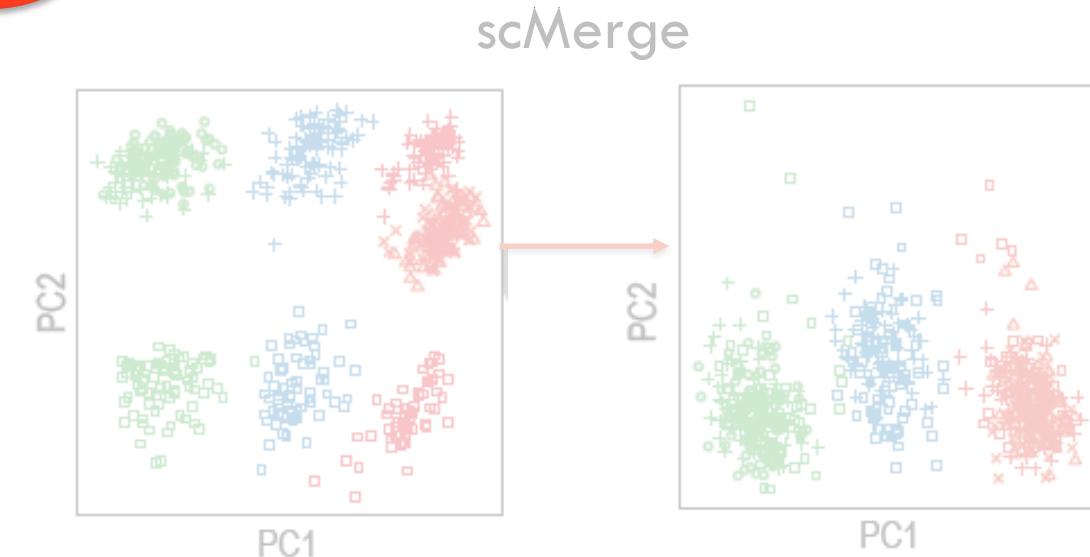
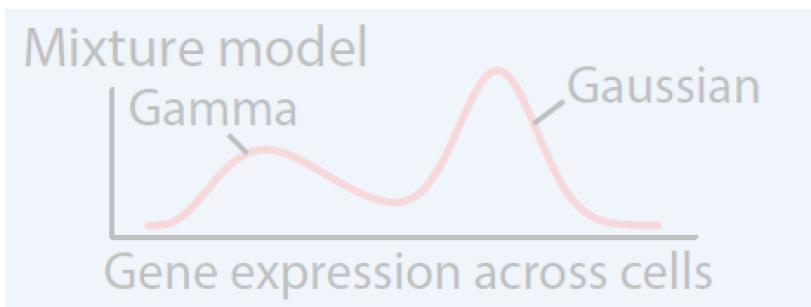
Clustering metrics



Ellis Patrick

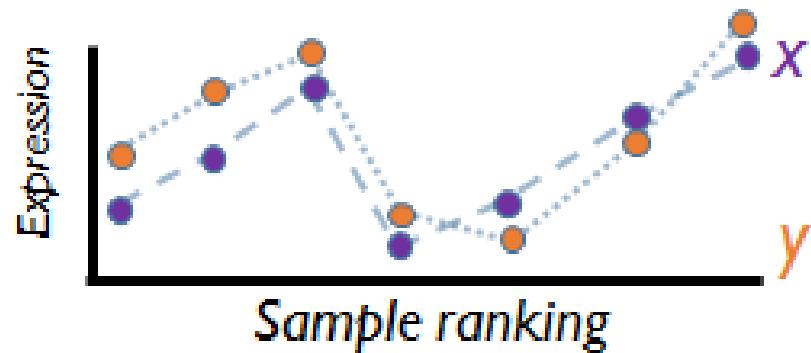


Finding stably expressed genes



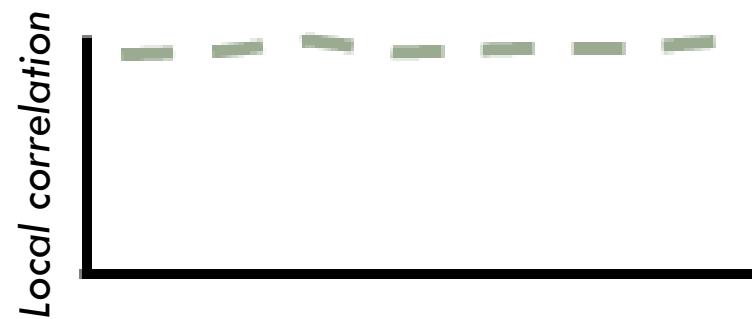
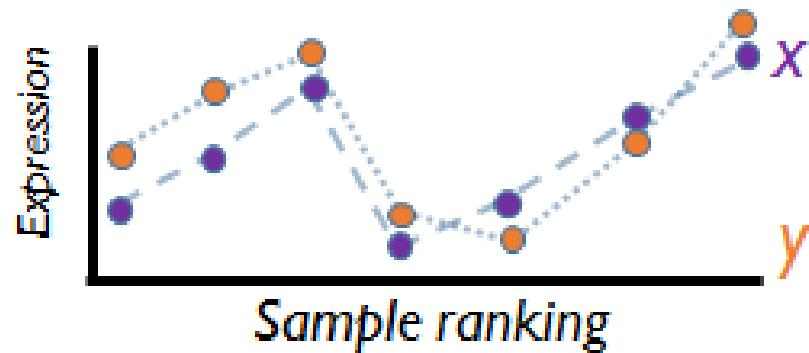
What is differential correlation?

Consider the expression of gene x and gene y



What is differential correlation?

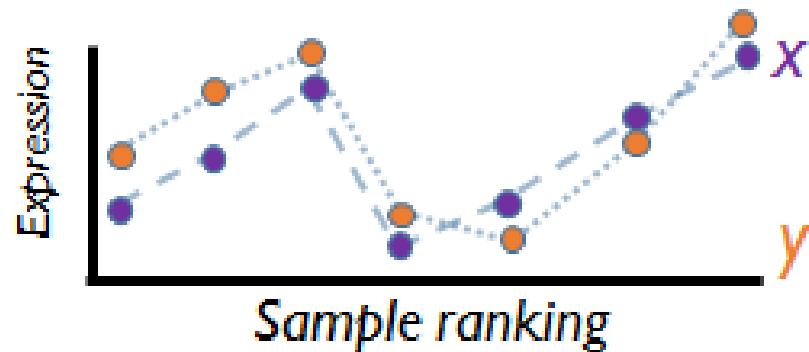
Consider the expression of gene x and gene y



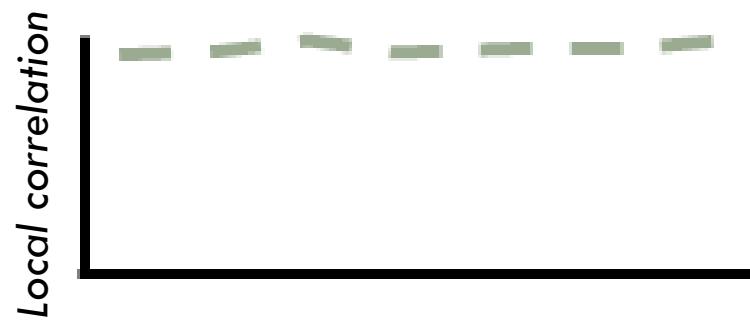
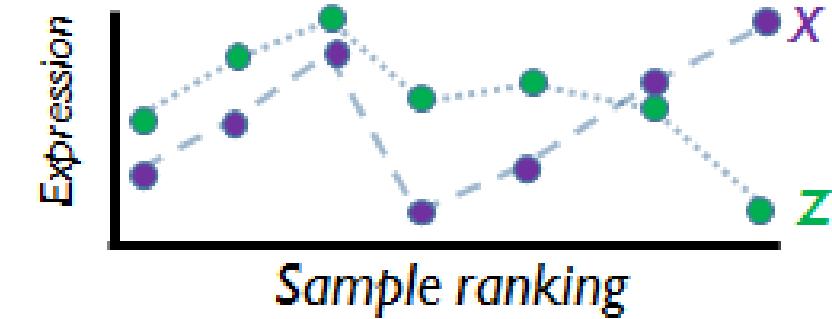
Constant correlation

What is differential correlation?

Consider the expression of gene x and gene y



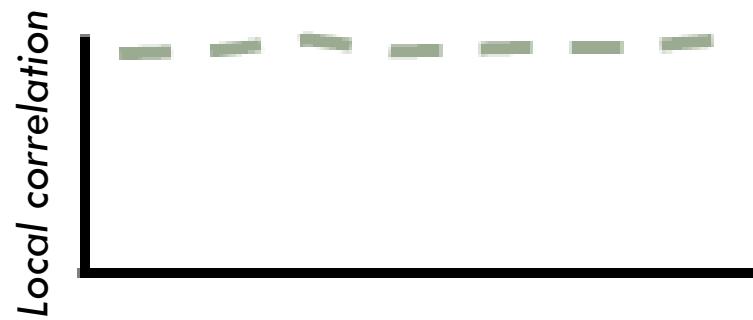
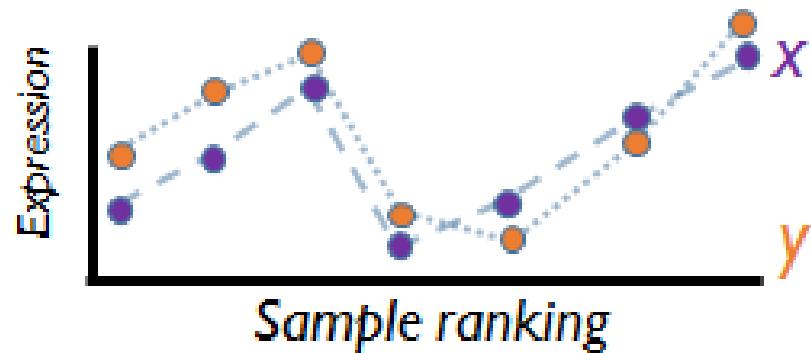
Now consider the expression of gene x and gene z



Constant correlation

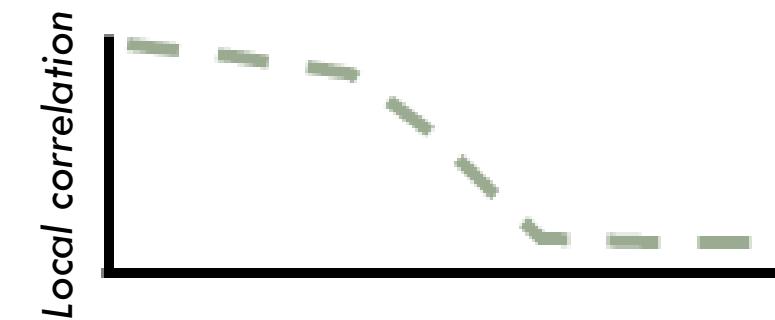
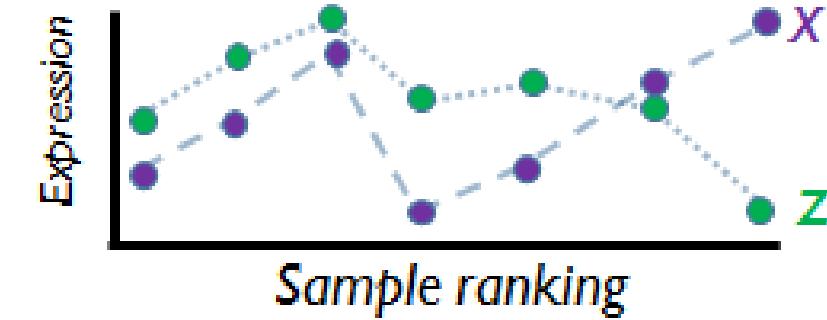
What is differential correlation?

Consider the expression of gene x and gene y



Constant correlation

Now consider the expression of gene x and gene z



Differential correlation

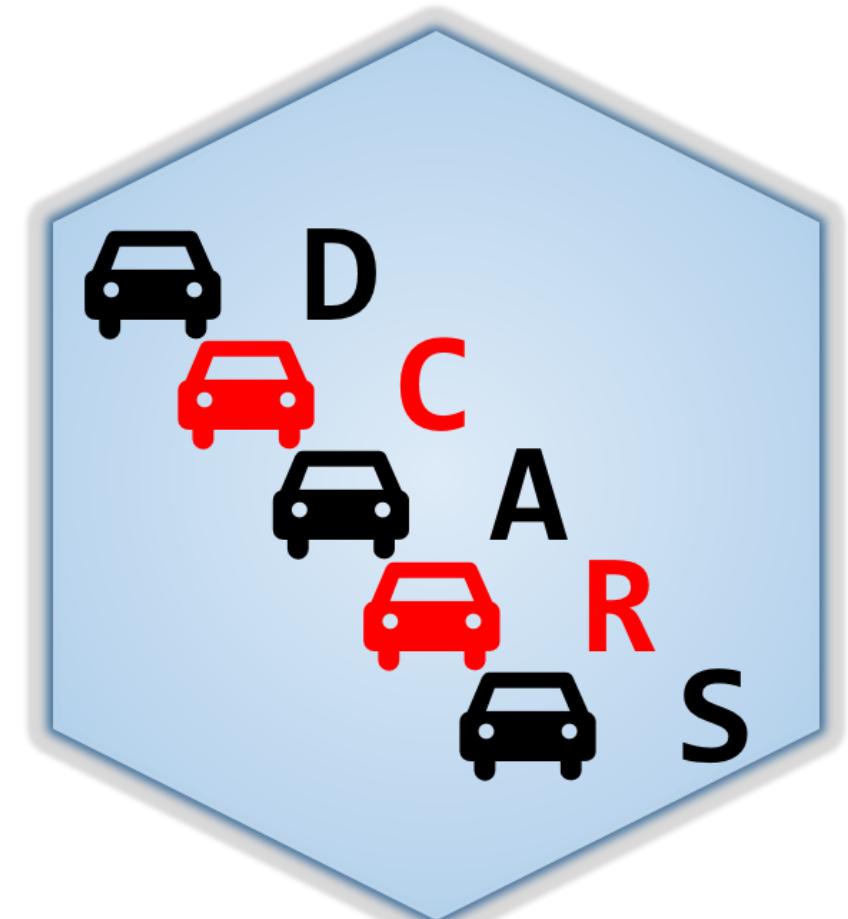
Systems biology

DCARS: differential correlation across ranked samples

Shila Ghazanfar  ^{1,2,*}, Dario Strbenac², John T. Ormerod^{2,3},
Jean Y. H. Yang^{2,1} and Ellis Patrick 

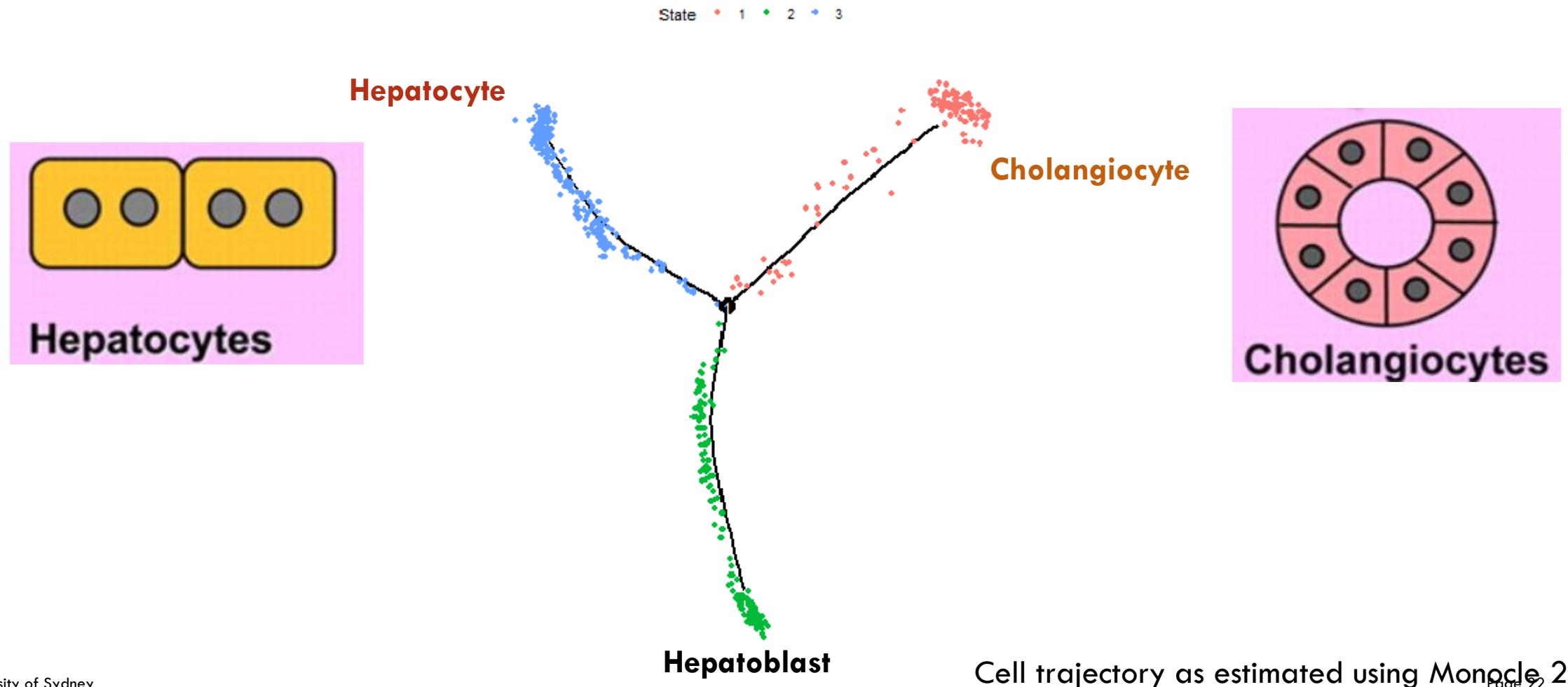
Installation

```
# Install the development version from GitHub:  
# install.packages("devtools")  
devtools::install_github("shazanfar/DCARS")  
library(DCARS)
```



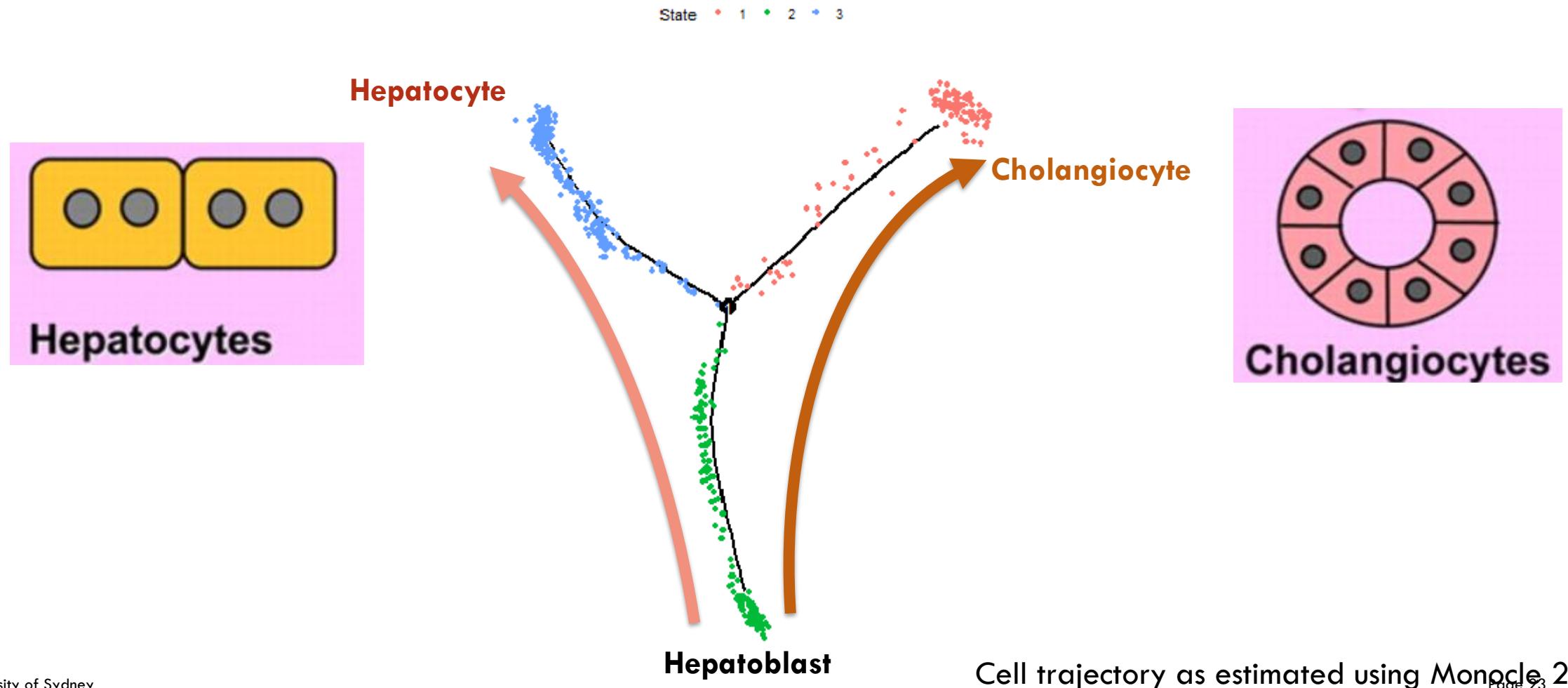


Differential correlation across pseudotime



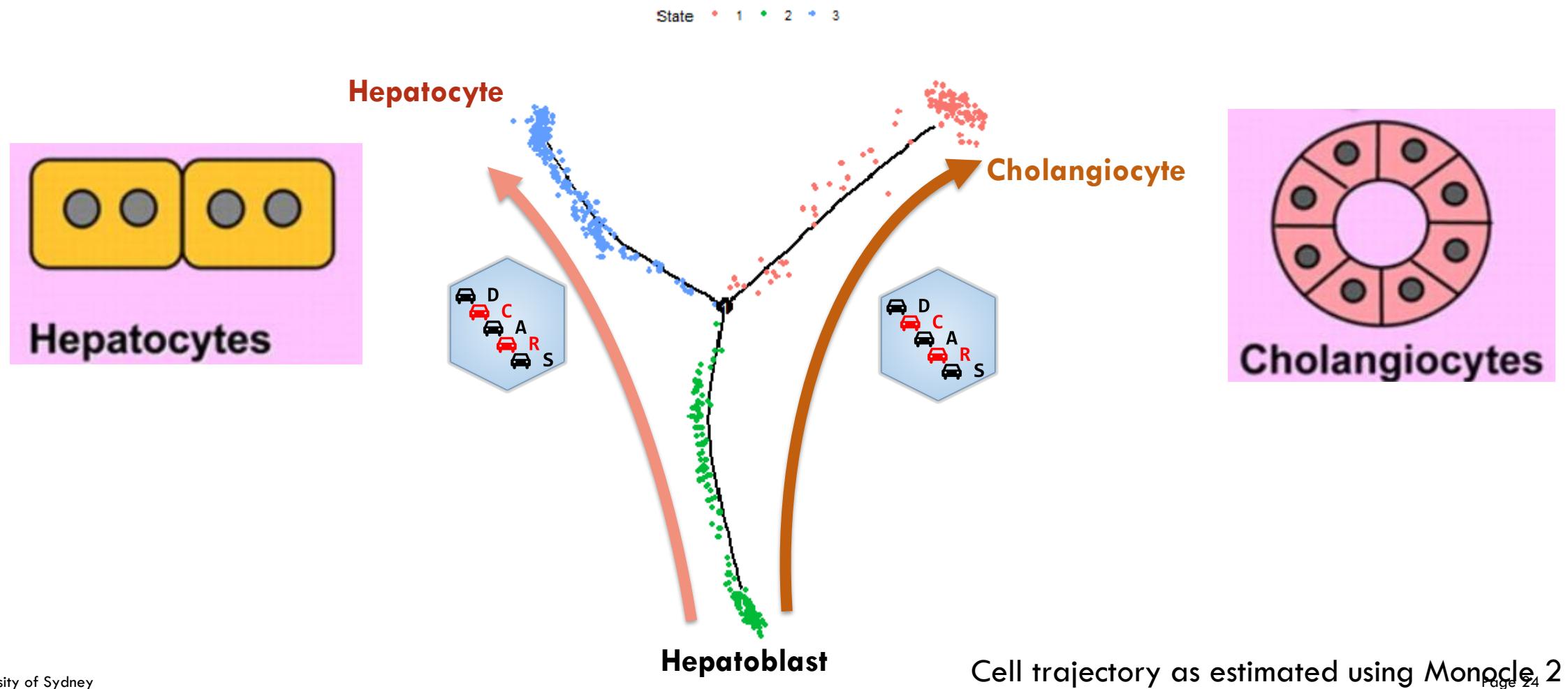


Differential correlation across pseudotime



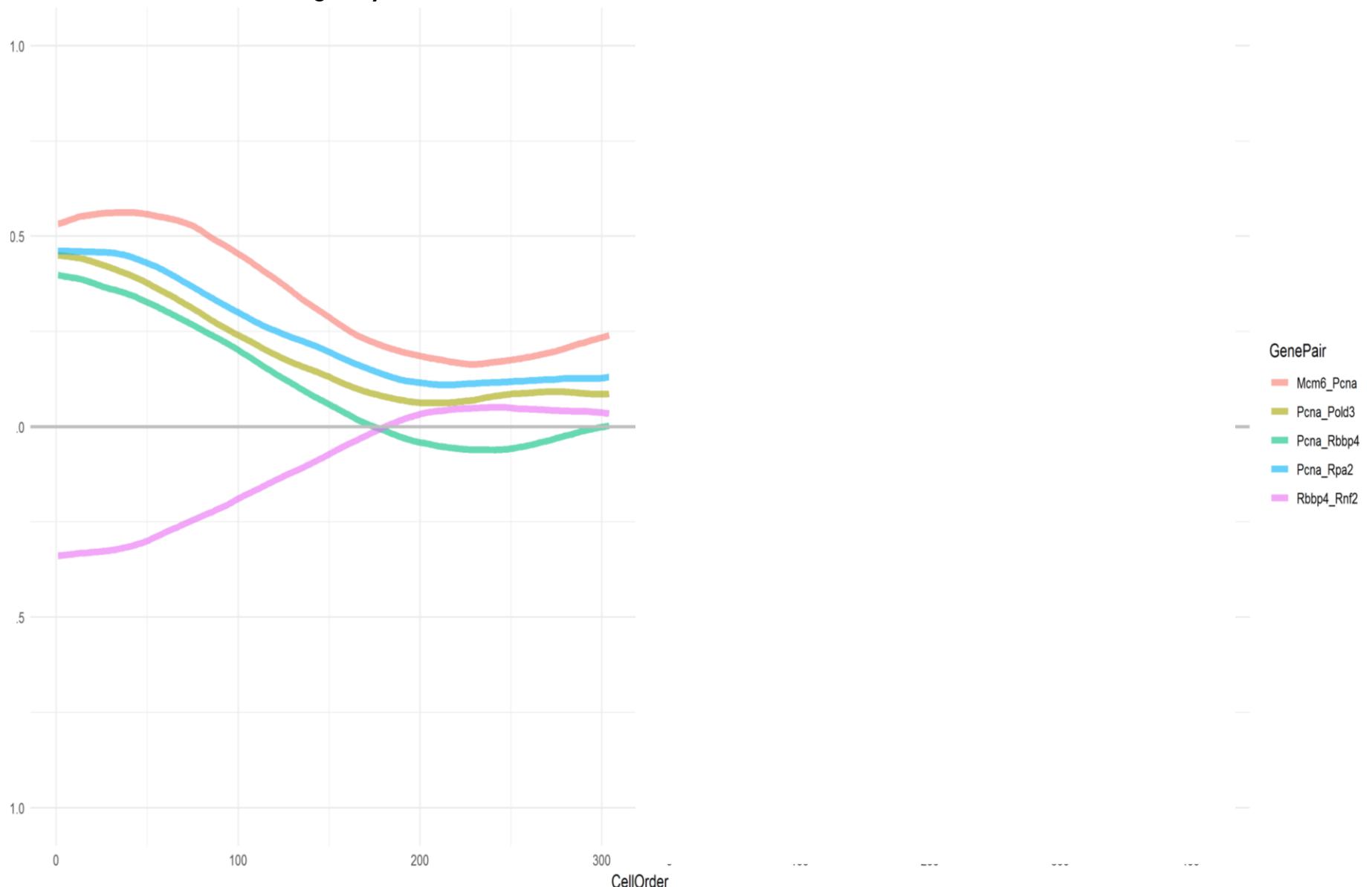
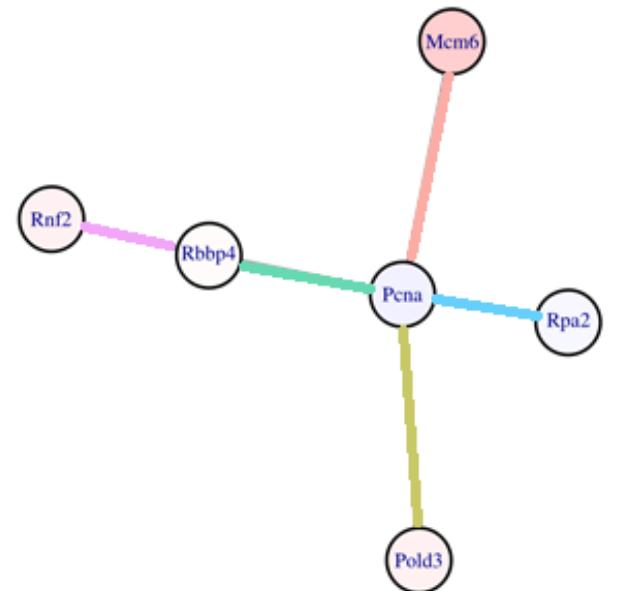


Differential correlation across pseudotime

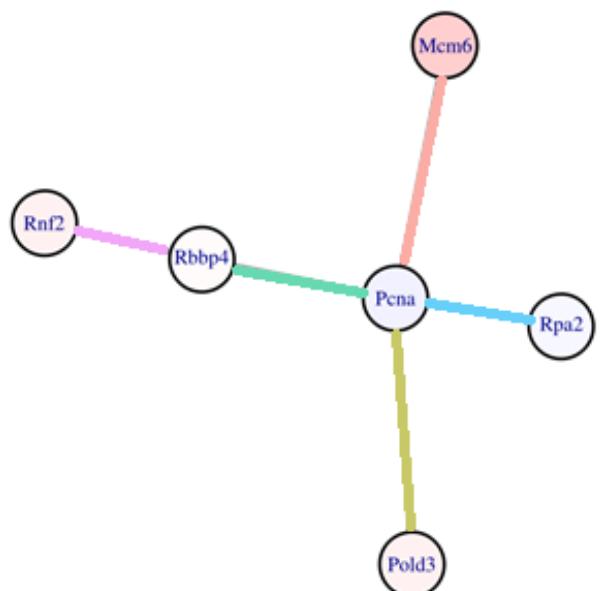


Pcna: DNA replication, cell cycle. Correlation lost during differentiation

Cholangiocyte

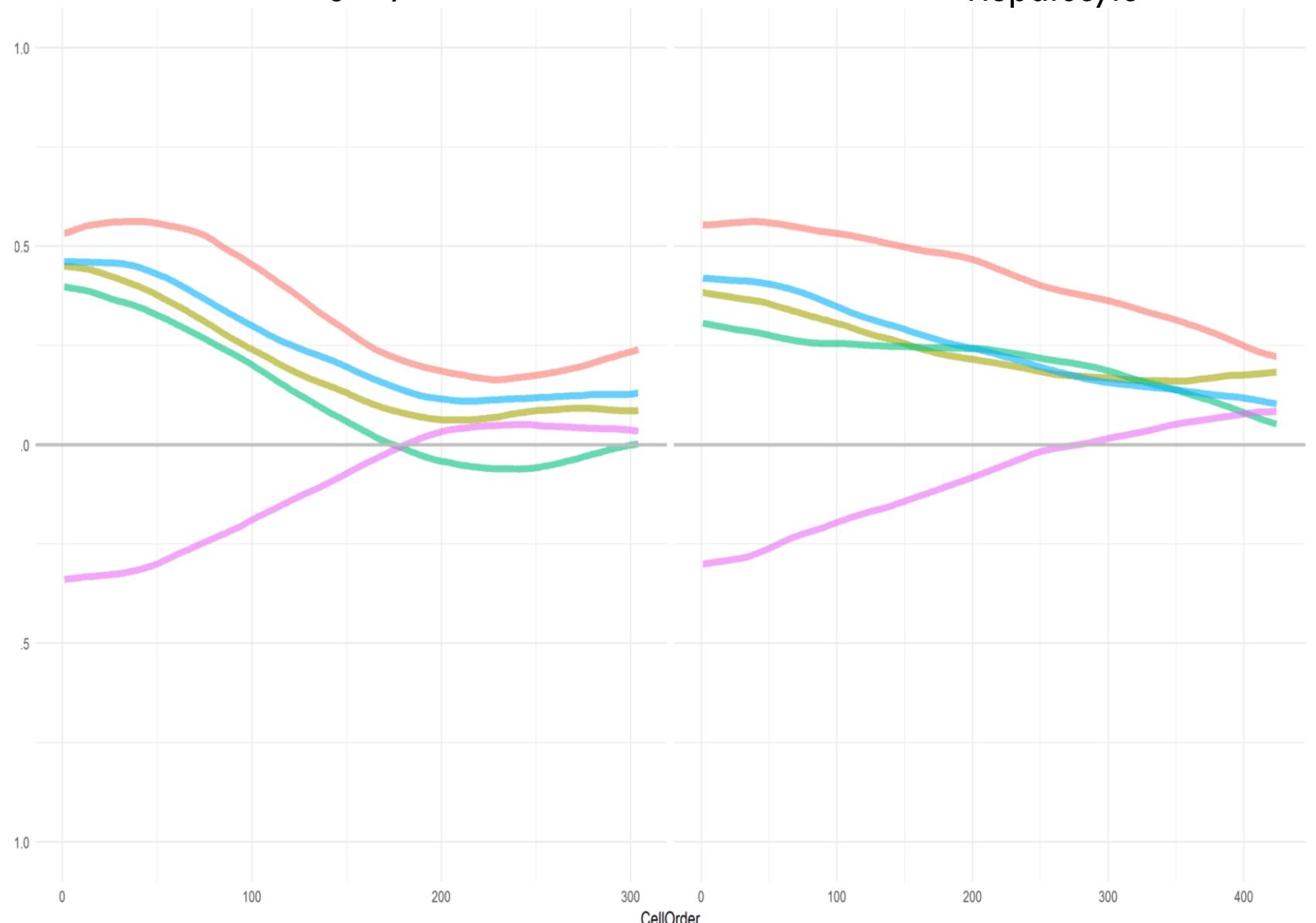


Pcna: DNA replication, cell cycle. Correlation lost during differentiation

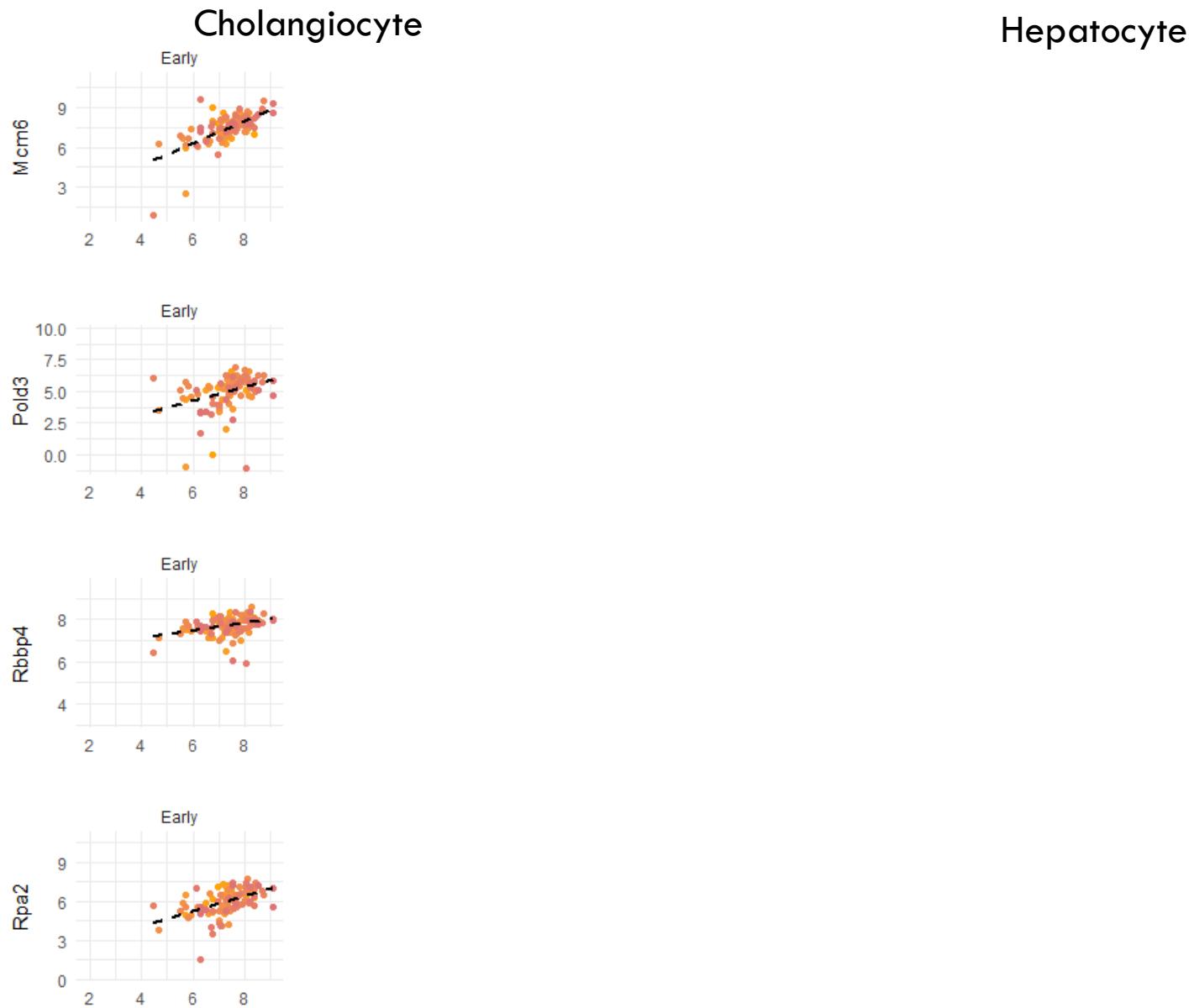
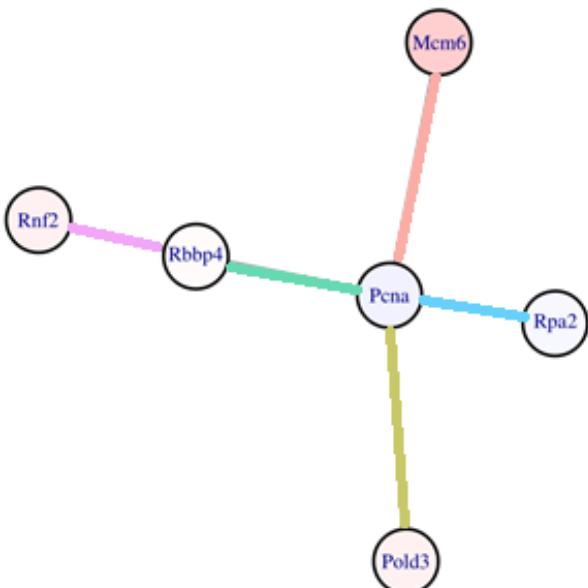


Cholangiocyte

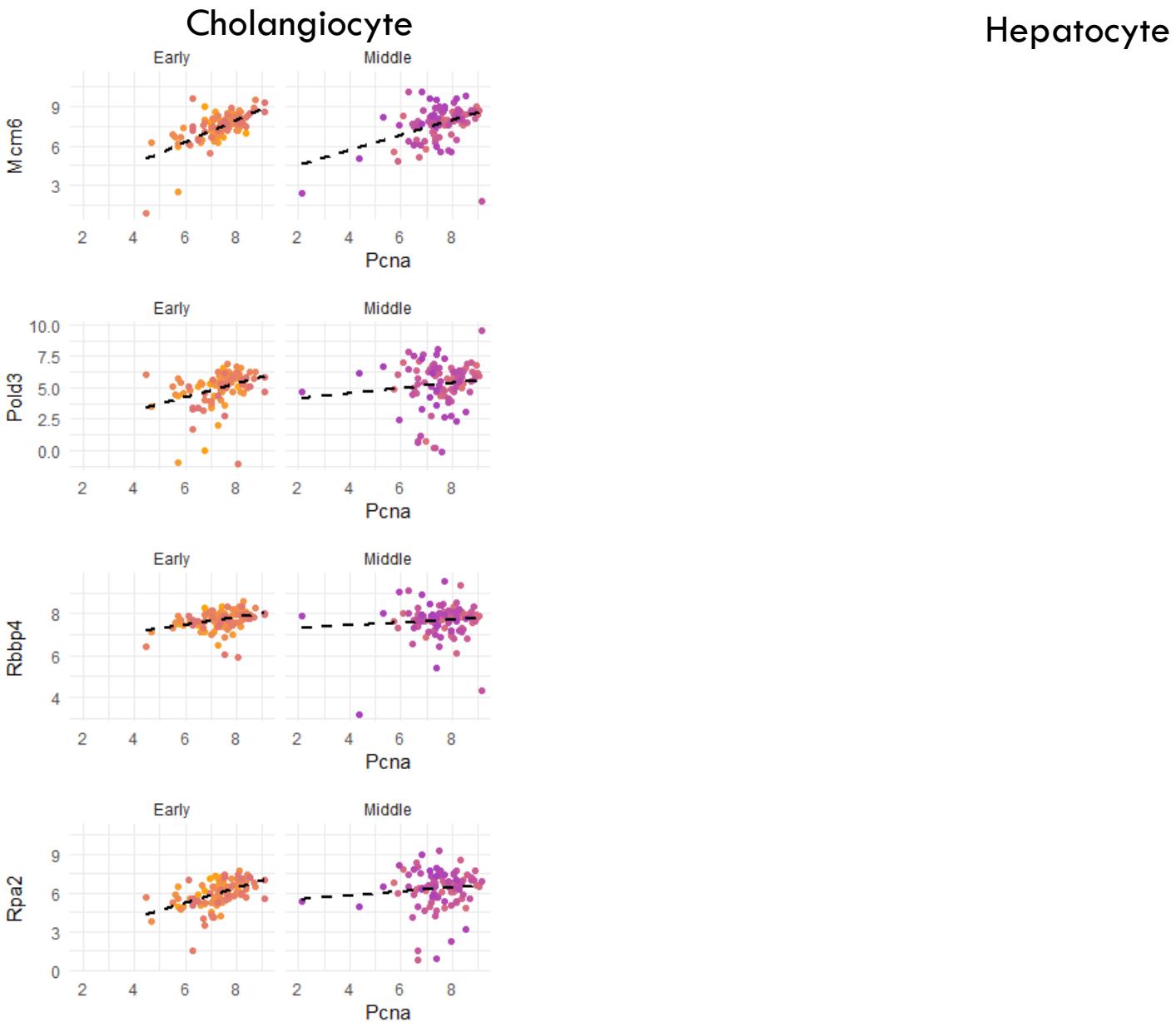
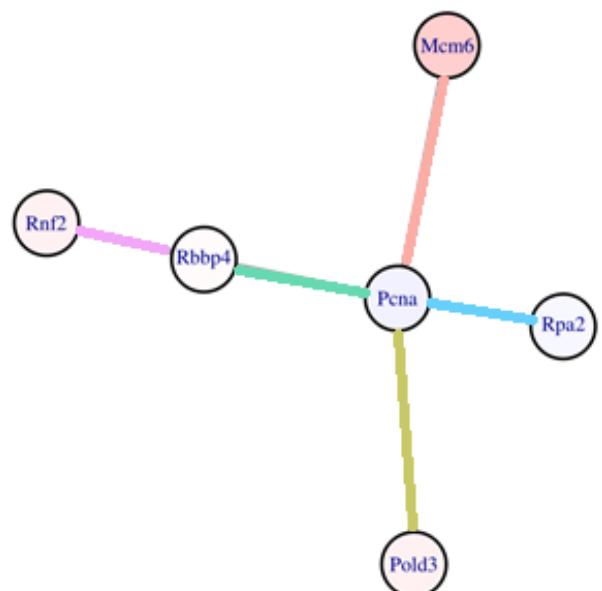
Hepatocyte



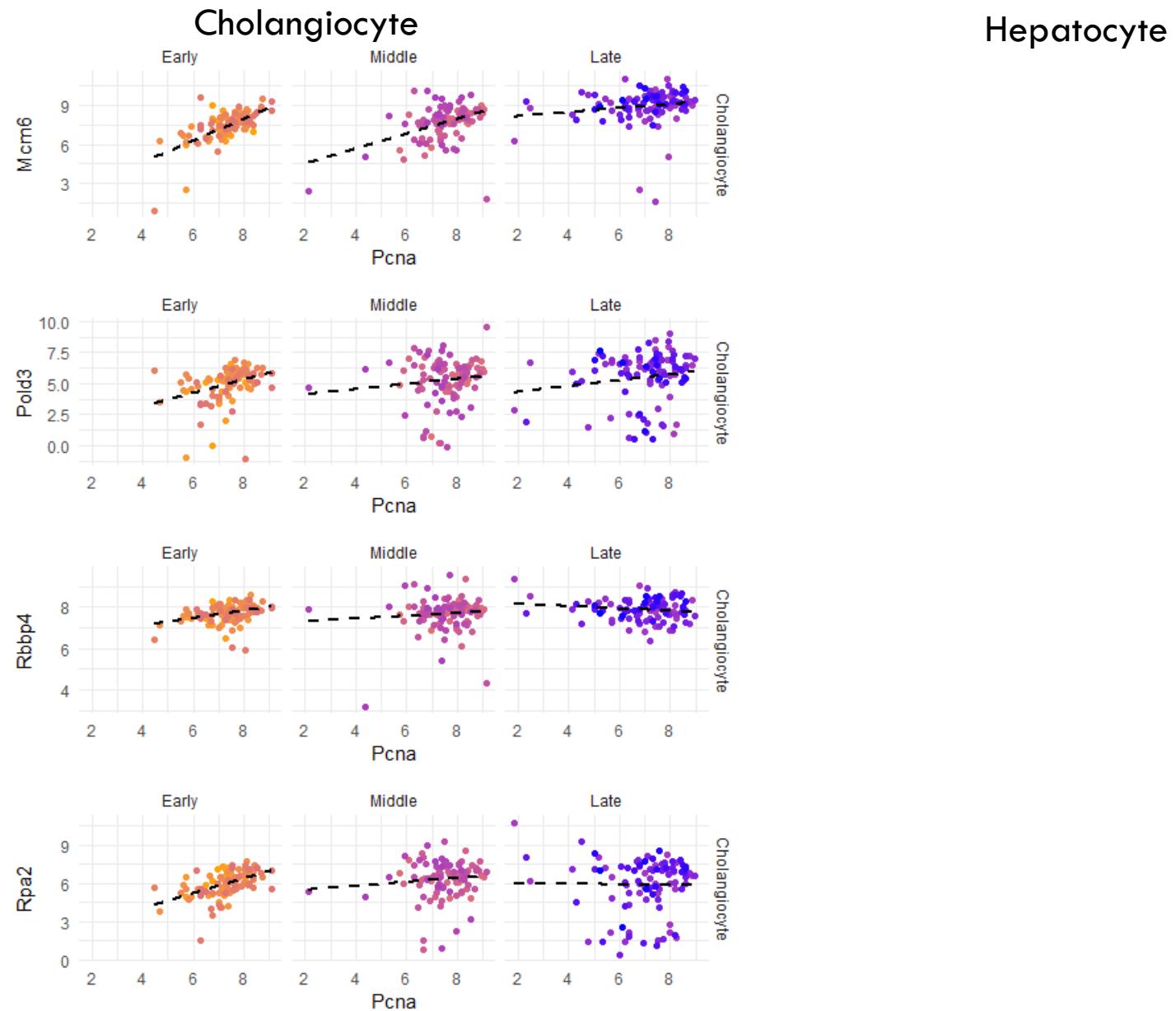
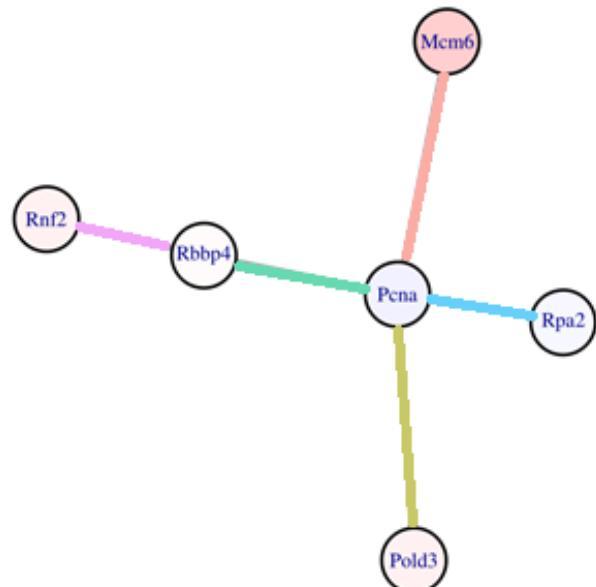
Pcna: DNA replication, cell cycle. Correlation lost during differentiation



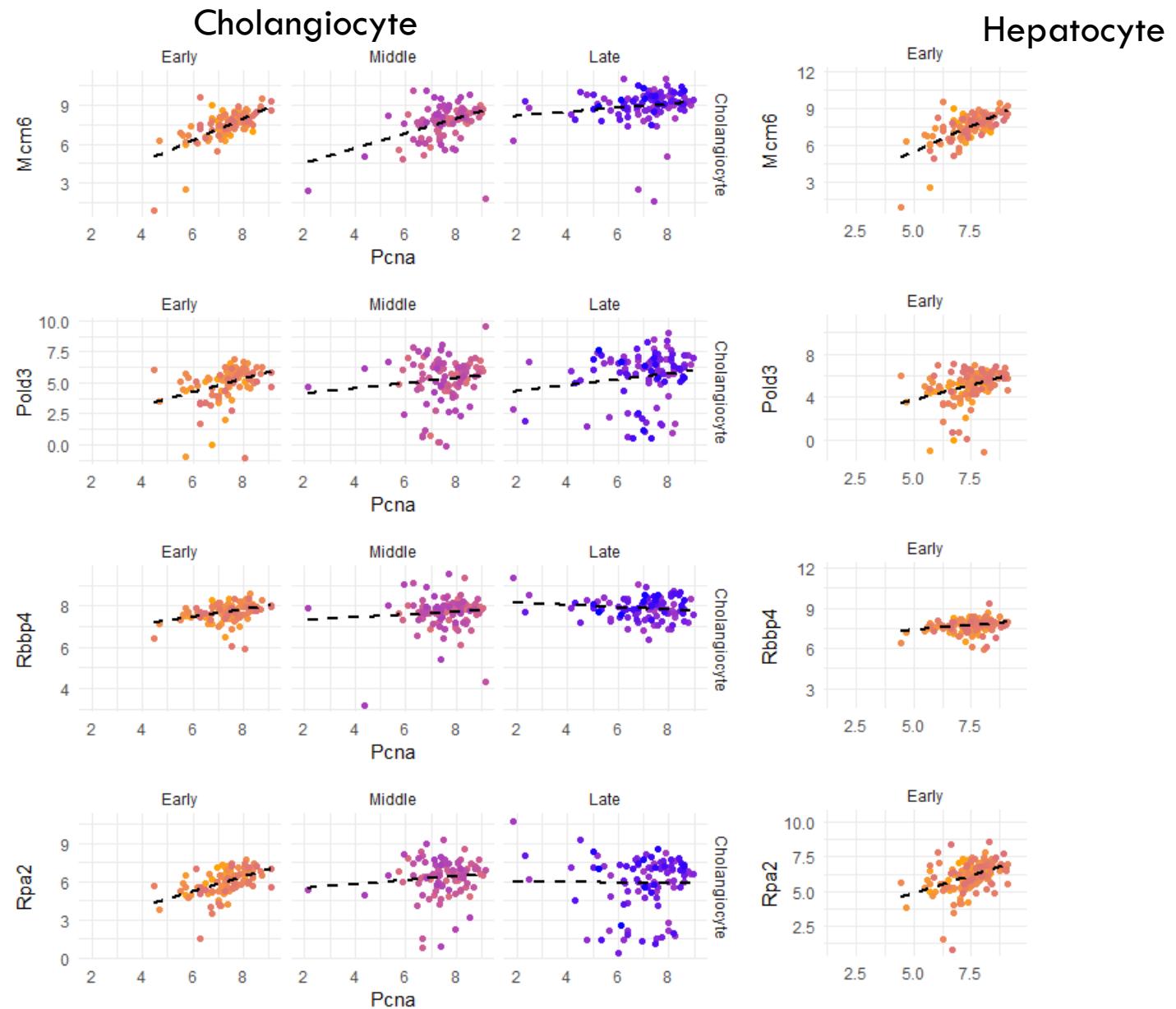
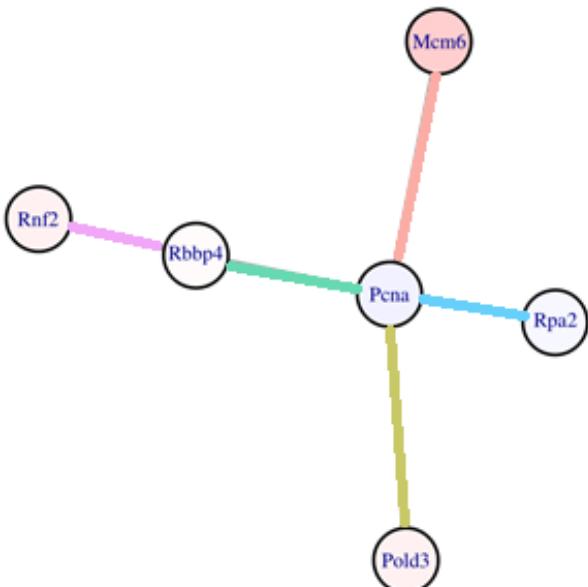
Pcna: DNA replication, cell cycle. Correlation lost during differentiation



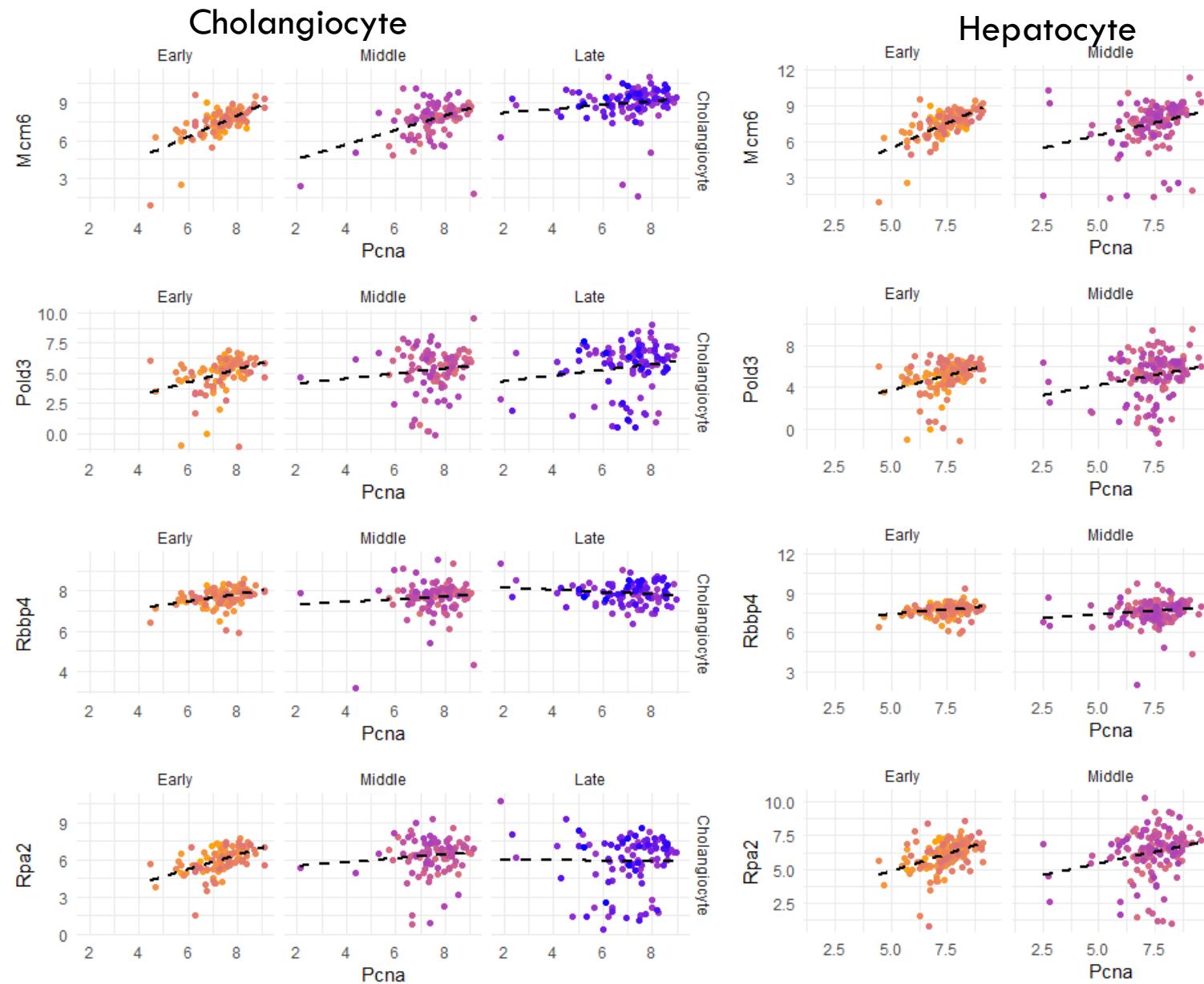
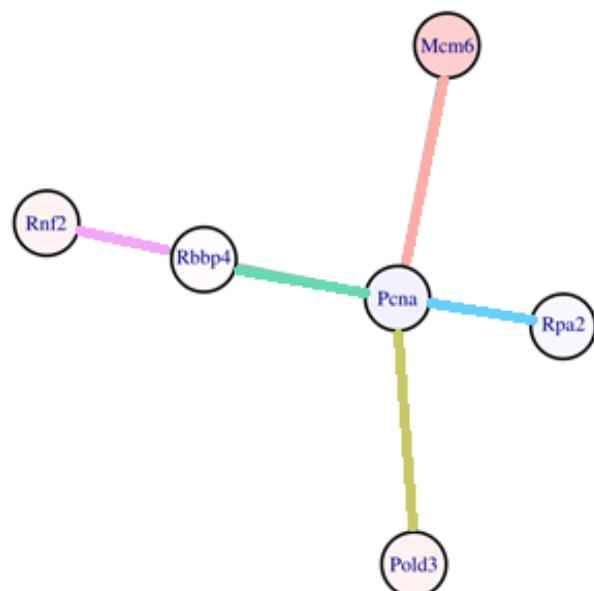
Pcna: DNA replication, cell cycle. Correlation lost during differentiation



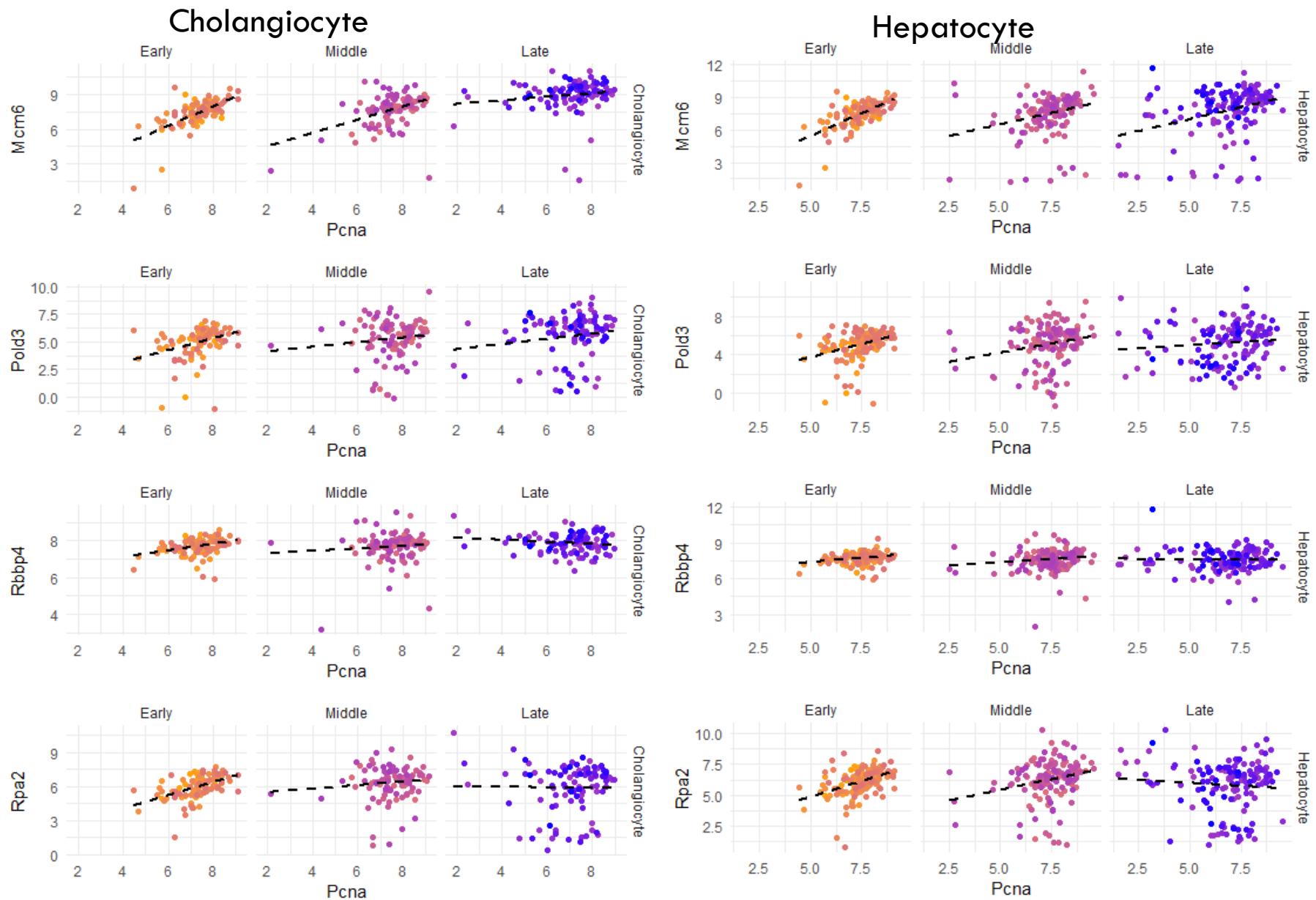
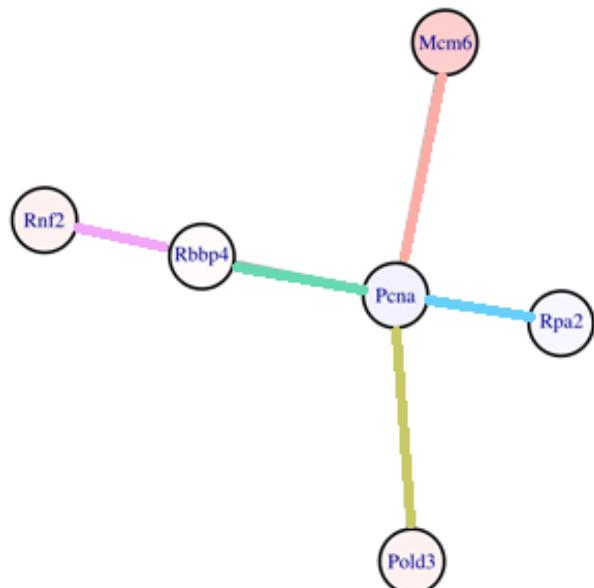
Pcna: DNA replication, cell cycle. Correlation lost during differentiation

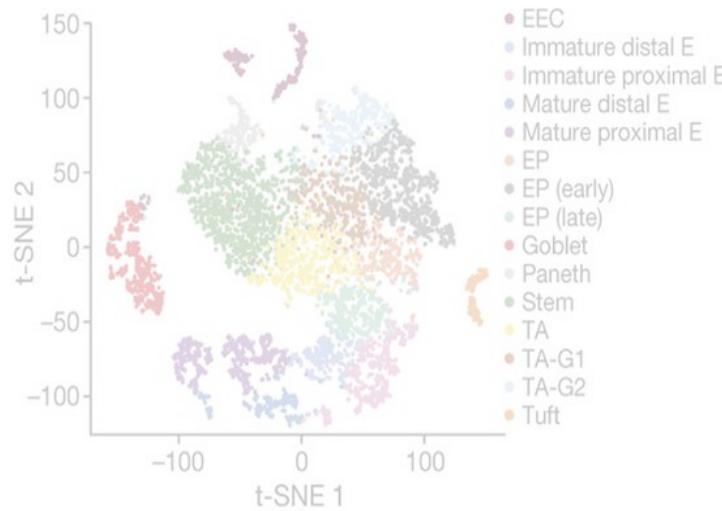


Pcna: DNA replication, cell cycle. Correlation lost during differentiation

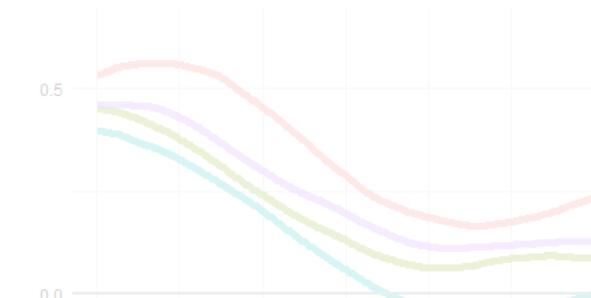


Pcna: DNA replication, cell cycle. Correlation lost during differentiation





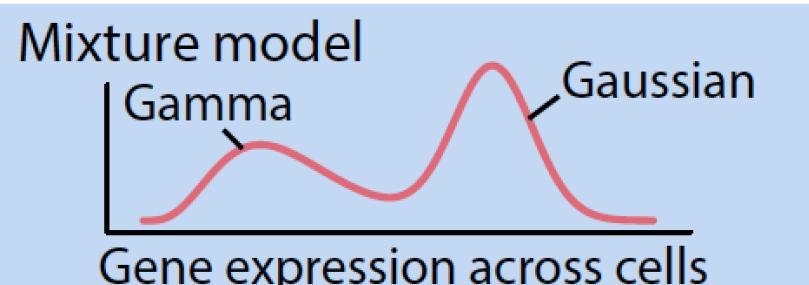
Clustering metrics



Differential correlation

Sydney
Single cell

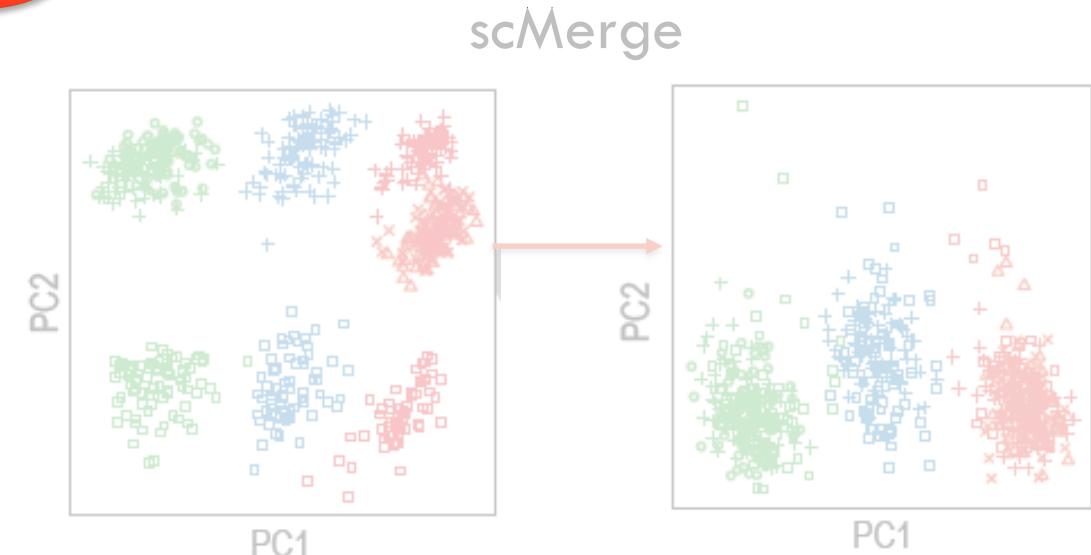
Finding stably expressed genes



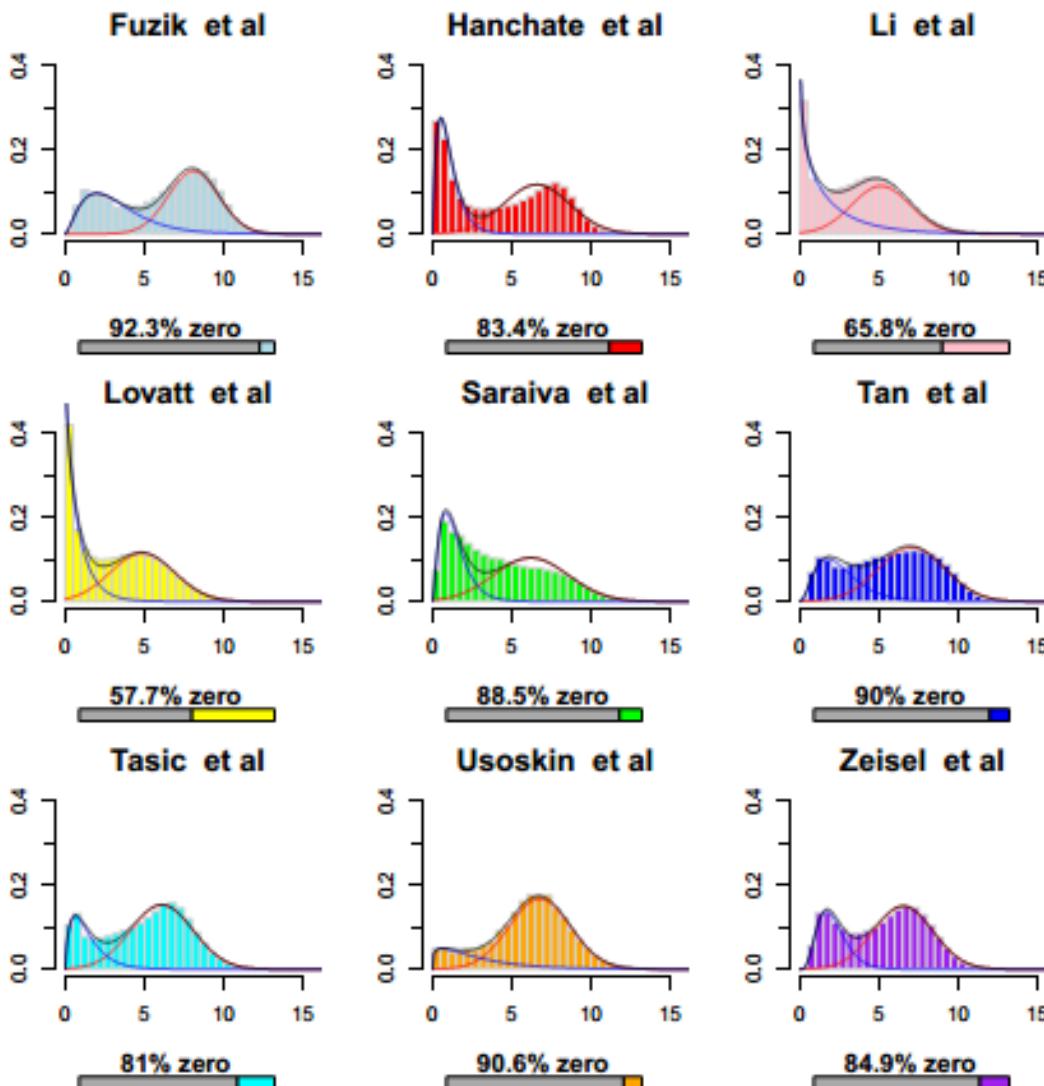
The University of Sydney



Yingxin Lin



Previous work on mixture modelling



BMC Syst Biol. 2016; 10(Suppl 5): 127.

Published online 2016 Dec 5. doi: [10.1186/s12918-016-0370-4](https://doi.org/10.1186/s12918-016-0370-4)

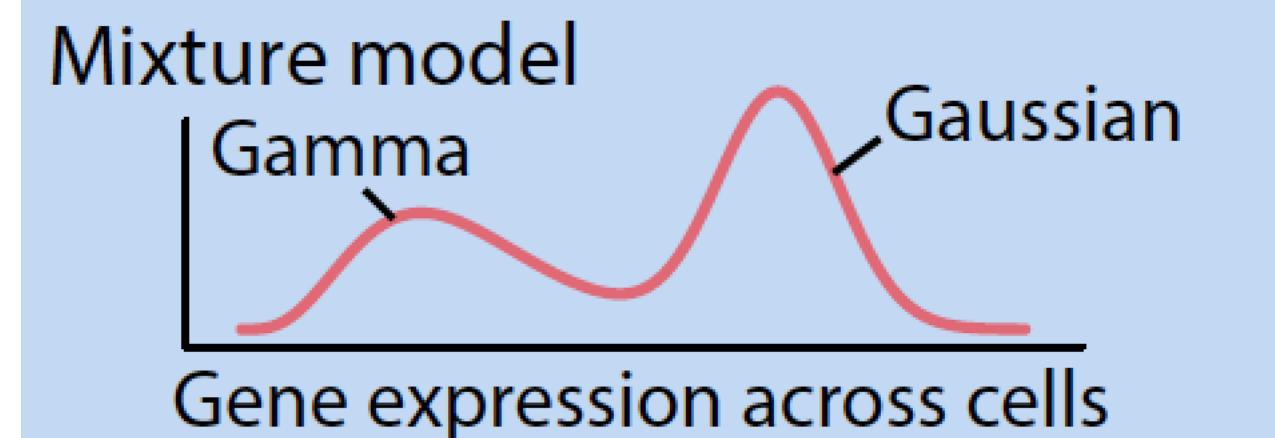
PMCID: PMC5249008

PMID: [278105940](https://pubmed.ncbi.nlm.nih.gov/278105940/)

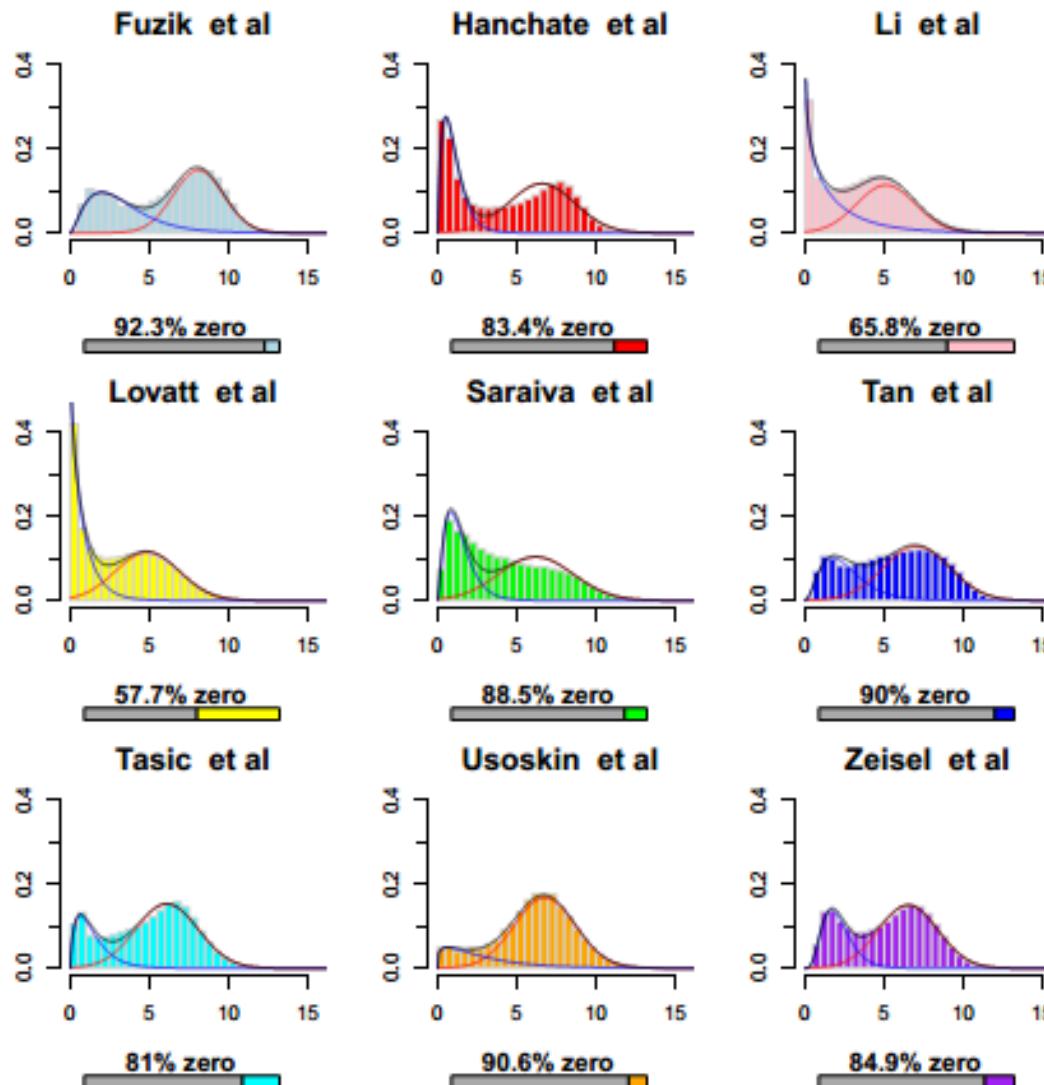
Integrated single cell data analysis reveals cell specific networks and novel coactivation markers

Shila Ghazanfar,^{✉1} Adam J. Bisogni,² John T. Ormerod,^{1,3} David M. Lin,² and Jean Y. H. Yang¹

[Author information](#) ► [Article notes](#) ► [Copyright and License information](#) ► [Disclaimer](#)



Previous work on mixture modelling



BMC Syst Biol. 2016; 10(Suppl 5): 127.

Published online 2016 Dec 5. doi: [10.1186/s12918-016-0370-4](https://doi.org/10.1186/s12918-016-0370-4)

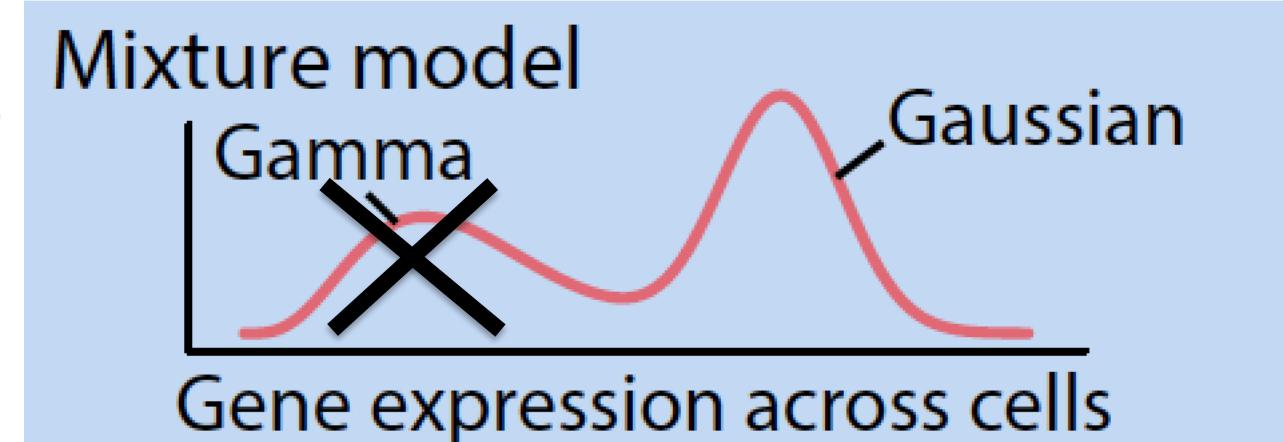
PMCID: PMC5249008

PMID: [278105940](https://pubmed.ncbi.nlm.nih.gov/278105940/)

Integrated single cell data analysis reveals cell specific networks and novel coactivation markers

Shila Ghazanfar,^{✉1} Adam J. Bisogni,² John T. Ormerod,^{1,3} David M. Lin,² and Jean Y. H. Yang¹

[Author information](#) ► [Article notes](#) ► [Copyright and License information](#) ► [Disclaimer](#)

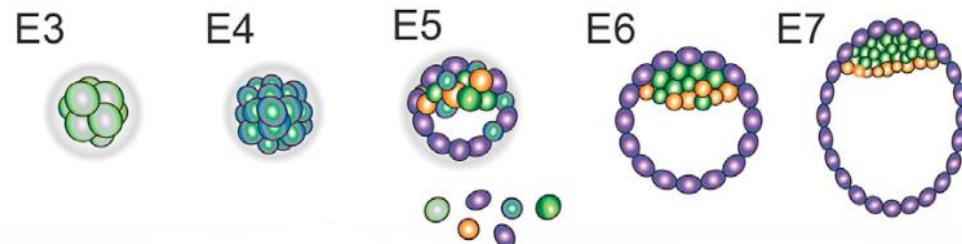


Establishing a good set of stably expressed genes

Datasets with wide ranges of cell types

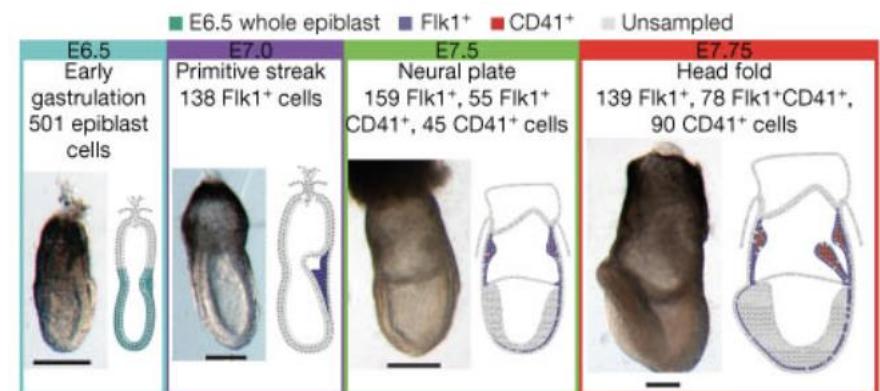
Human early developmental dataset
(Petropoulos et al.):

1529 cells from five timepoints:
E3, E4, E5, E6, E7



Mouse early developmental dataset
(Scialdone et al.):

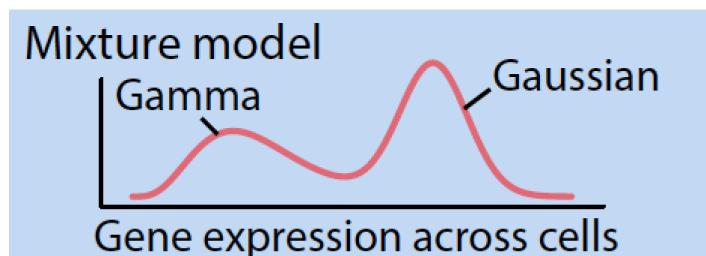
1205 cells from four timepoints:
E6.5, E7.0, E7.5, E7.75



Building features for stably expressed genes

- Mixture Model:
 - ▶ **Mixing Proportion** of the second component (λ_i)
 - ▶ **Standard deviation** of the second component (σ_i)
- **Zero Proportion** of each gene across all the cells (ω_i)
- **F-statistics** from one-way ANOVA (F_i):

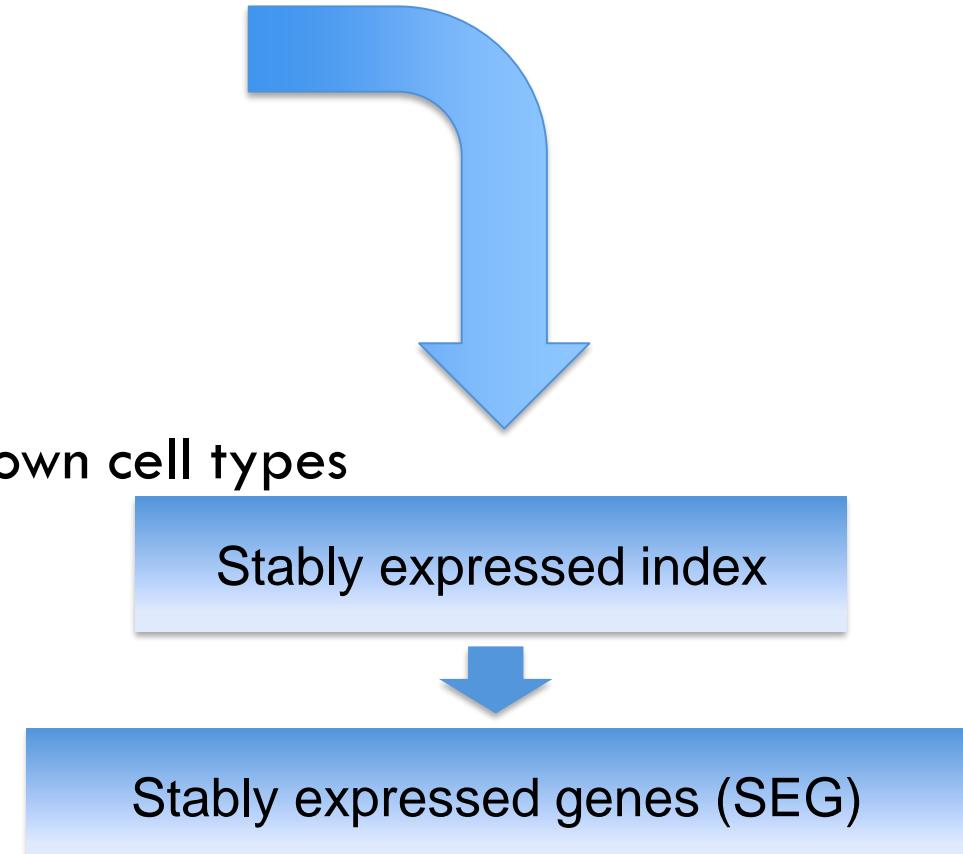
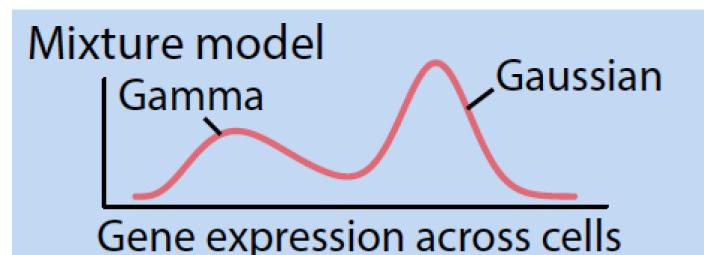
$\log_2 \text{FPKM} \sim \text{Any conditions or known cell types}$



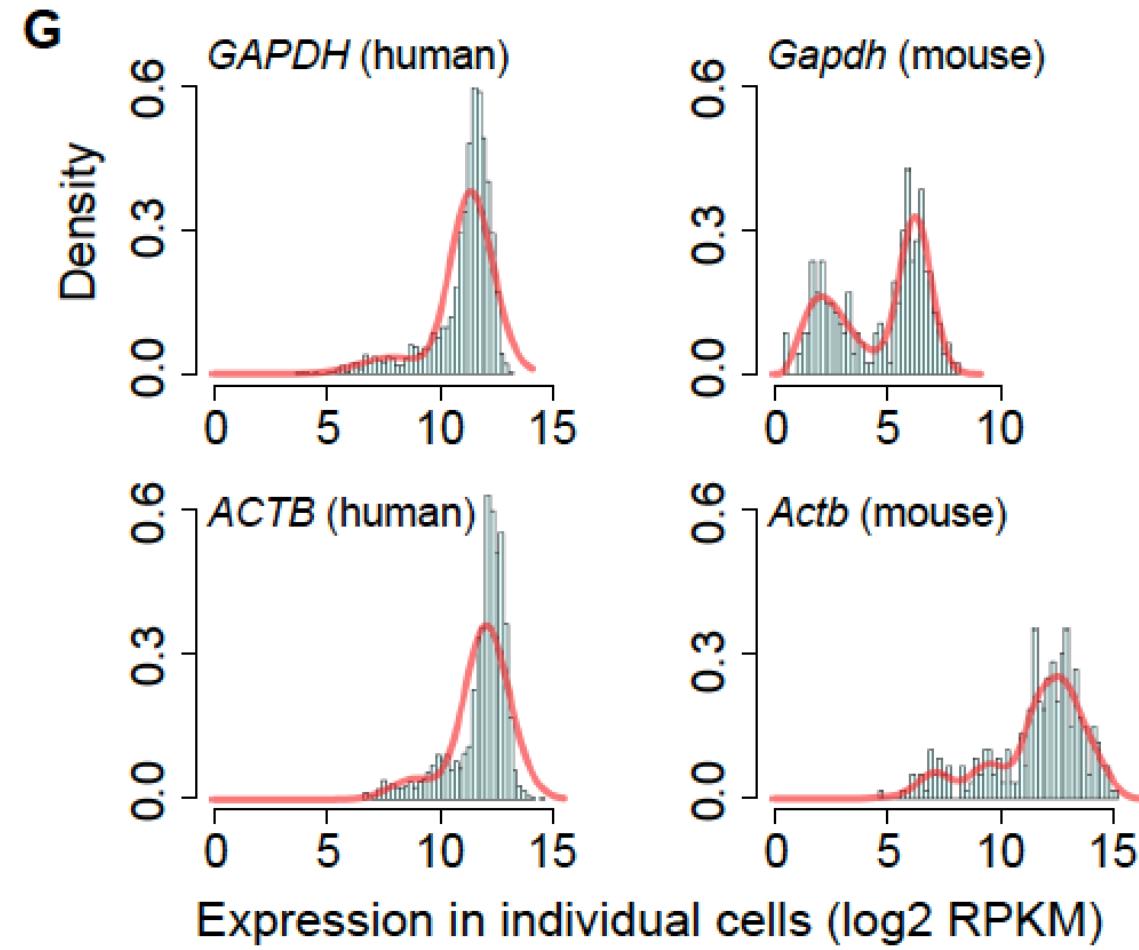
Building features for stably expressed genes

- Mixture Model:
 - ▶ **Mixing Proportion** of the second component (λ_i)
 - ▶ **Standard deviation** of the second component (σ_i)
- **Zero Proportion** of each gene across all the cells (ω_i)
- **F-statistics** from one-way ANOVA (F_i):

$\log_2 \text{FPKM} \sim \text{Any conditions or known cell types}$

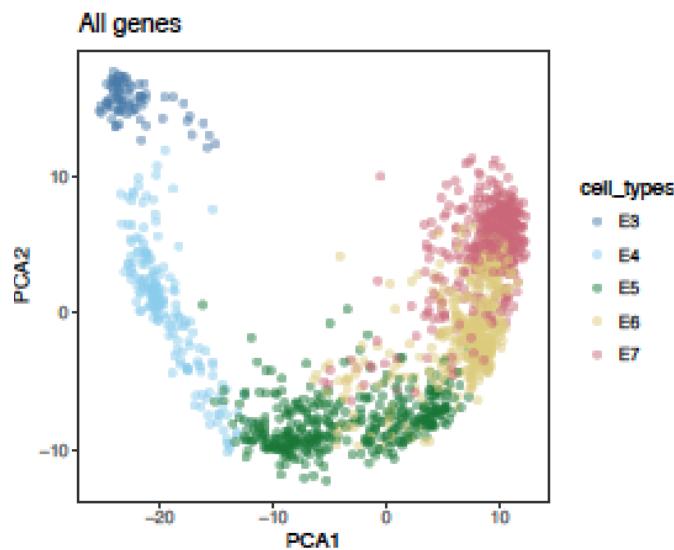


What about GAPDH?



Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



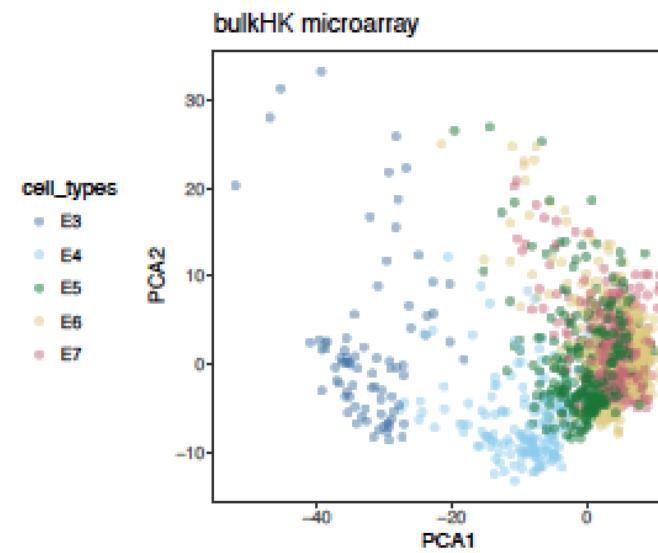
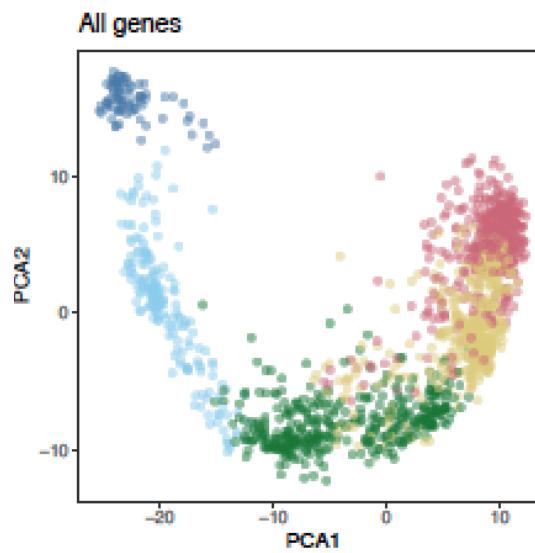
**House-keeping
genes based on
microarray data**

**House keeping
genes based on
bulk RNA-seq**

**Proposed
scSEG**

Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



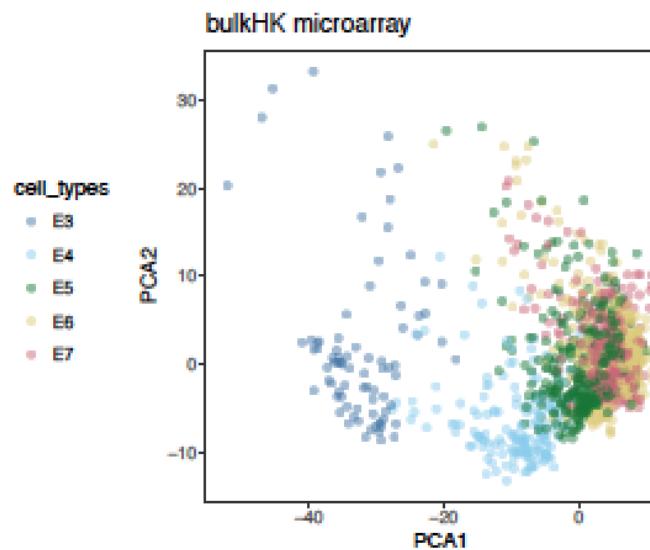
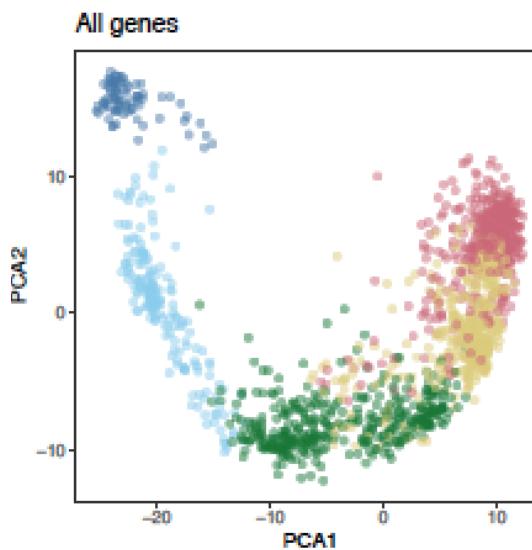
**House-keeping
genes based on
microarray data**

**House keeping
genes based on
bulk RNA-seq**

**Proposed
scSEG**

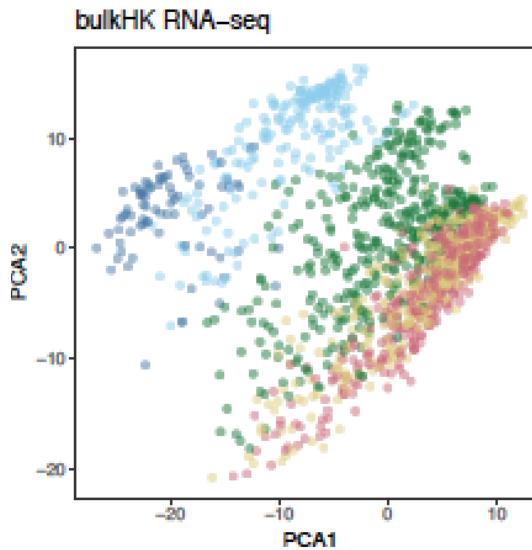
Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



**House-keeping
genes based on
microarray data**

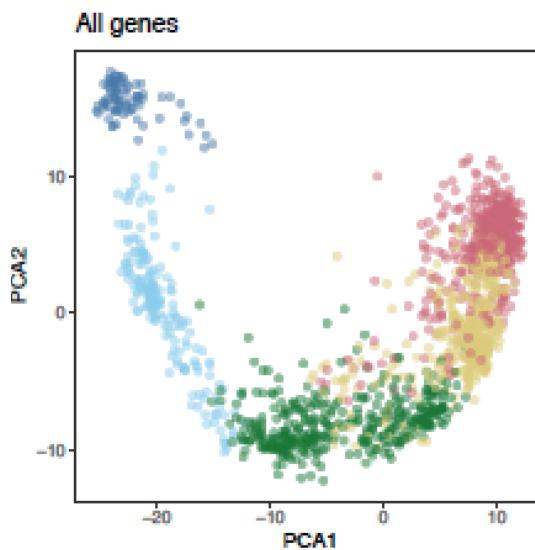
**House keeping
genes based on
bulk RNA-seq**



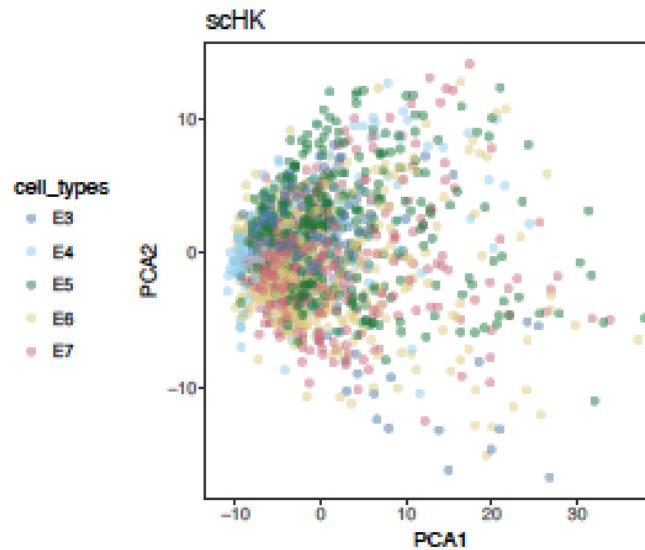
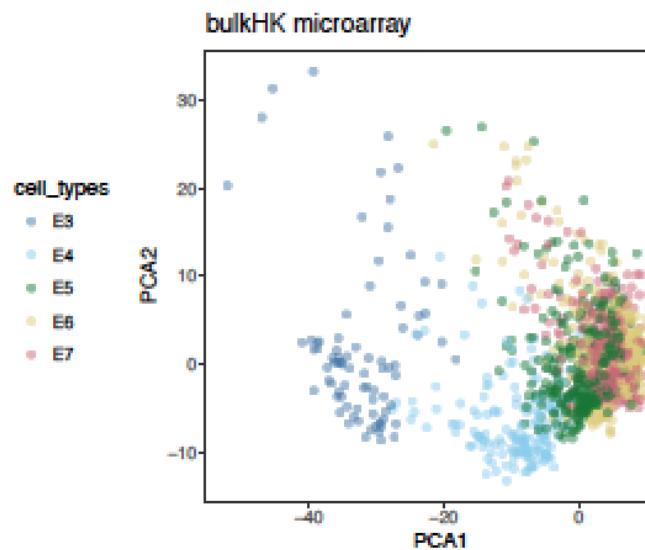
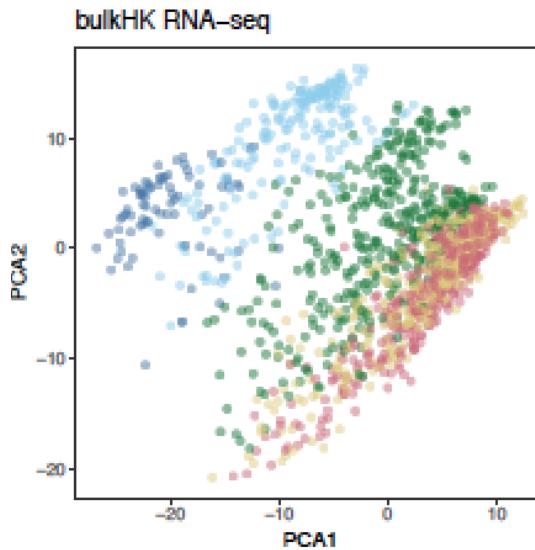
**Proposed
scSEG**

Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



**House keeping
genes based on
bulk RNA-seq**

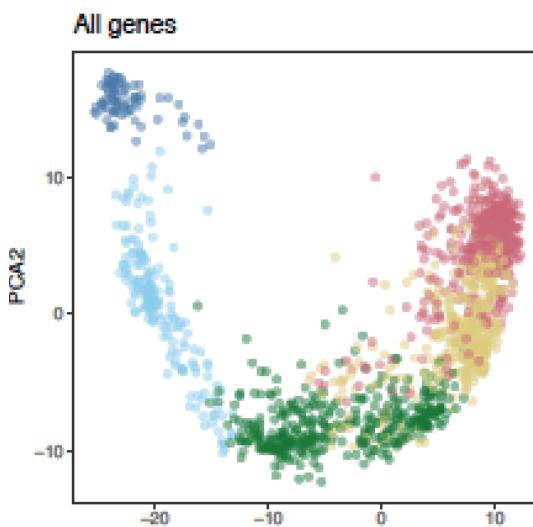


**House-keeping
genes based on
microarray data**

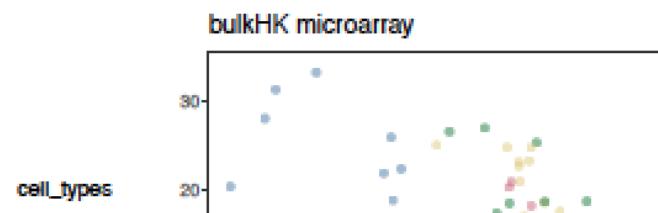
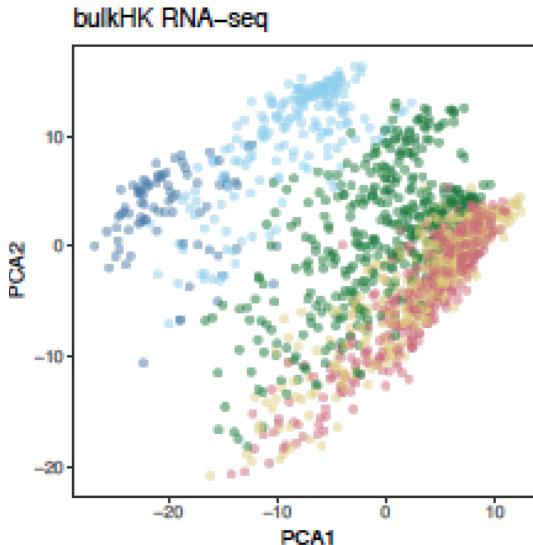
**Proposed
scSEG**

Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



**House keeping
genes based on
bulk RNA-seq**



**House-keeping
genes based on
microarray data**

New Results

Evaluating stably expressed genes in single cells

Yingxin Lin, Shila Ghazanfar, Dario Strbenac, Andy Wang, Ellis Patrick, Dave Lin, Terence Speed, Jean Yang, Pengyi Yang

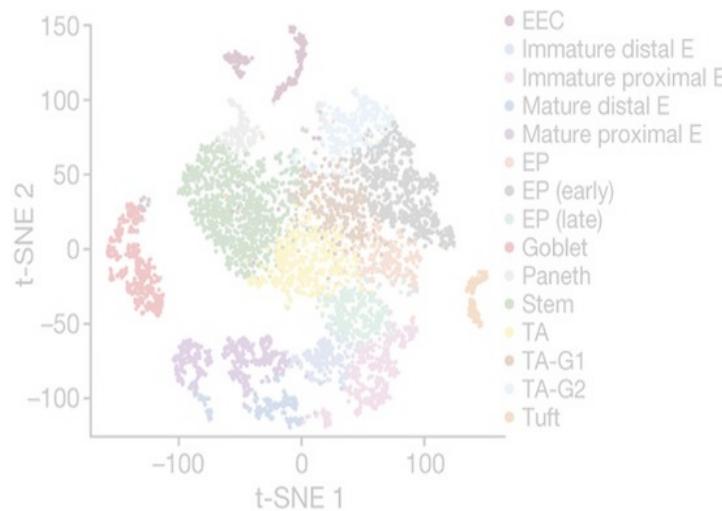
doi: <https://doi.org/10.1101/229815>

This article is a preprint and has not been peer-reviewed [what does this mean?].

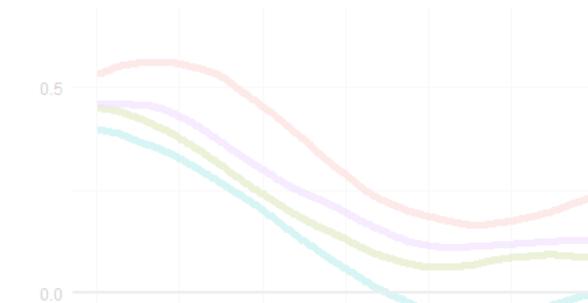
The CSHL logo is displayed next to the bioRxiv logo.

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

SEG



Clustering metrics



Differential correlation

Sydney
Single cell

Finding stably expressed genes



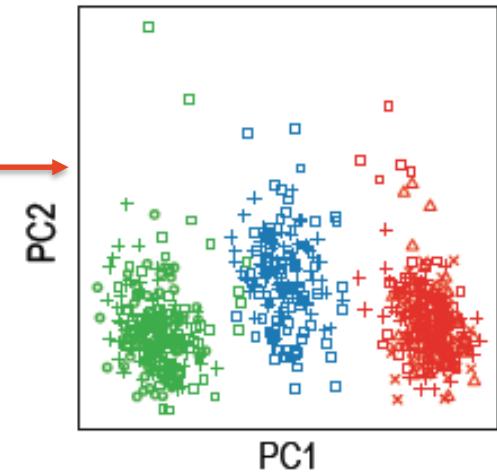
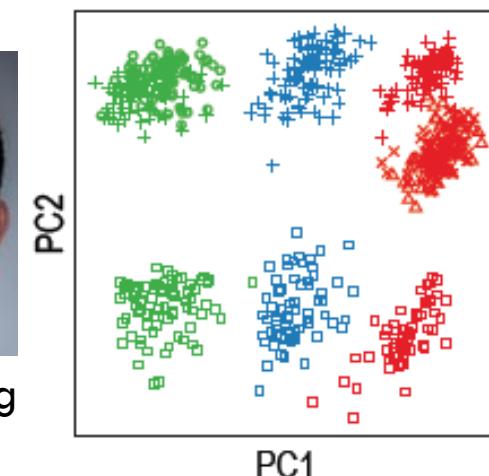
The University of Sydney



Yingxin Lin



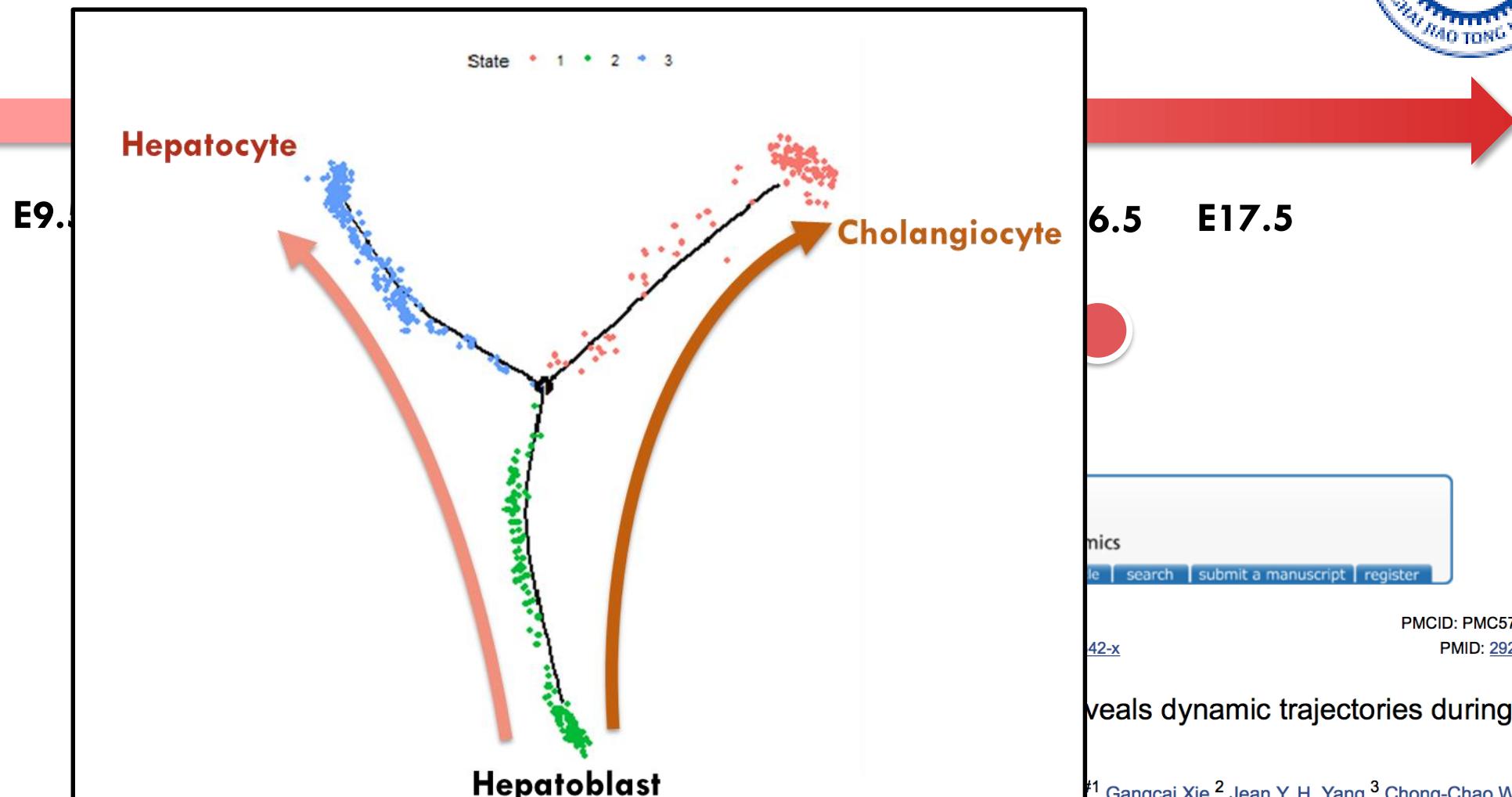
Kevin Wang



Liver fetal development time course datasets – integrating multiple datasets



GSE87795
Su et al.



#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 – integrating multiple datasets



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](https://doi.org/10.1186/s12864-017-4342-x)

PMCID: PMC5715535

PMID: [29202695](https://pubmed.ncbi.nlm.nih.gov/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 – integrating multiple datasets



LETTER

doi:10.1038/nature22796

BioMed Central
The Open Access Publisher

BMC Genomics

this article | search | submit a manuscript | register

© 2017; 18: 946.

© 2017 Dec 4. doi: 10.1186/s12864-017-4342-x

PMCID: PMC5715535

PMID: 29202695

RNA-Seq analysis reveals dynamic trajectories during liver development

Yi Shi,¹ Xin Zou,¹ Zhao-Ning Lu,¹ Gangcai Xie,² Jean Y. H. Yang,³ Chong-Chao Wu,¹ ui,¹ Kun-Yan He,¹ Qing Luo,¹ Yu-Lan Qu,¹ Na Wang,¹ Lan Wang,¹ and Ze-Guang Han^{1,4}

[Introduction](#) ► [Article notes](#) ► [Copyright and License information](#) ► [Disclaimer](#)

Multilineage communication regulates human liver bud development from pluripotency

J. Gray Camp^{1*}, Keisuke Sekine^{2*}, Tobias Gerber¹, Henry Loeffler-Wirth³, Hans Binder³, Małgorzata Gac¹, Sabina Kanton¹, Jorge Kageyama¹, Georg Damm^{4,5}, Daniel Seehofer^{4,5}, Lenka Belicova⁶, Marc Bickle⁶, Rico Barsacchi⁶, Ryo Okuda², Emi Yoshizawa², Masaki Kimura², Hiroaki Ayabe², Hideki Taniguchi², Takanori Takebe^{2,7} & Barbara Treutlein^{1,6}

Liver fetal development time course datasets – integrating multiple datasets



E9.5 E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 E17.5

GSE87795
Su et al.

HEPATOLOGY

RAPID COMMUNICATION | HEPATOLOGY, VOL. 66, NO. 5, 2017



LET] A Single-Cell Transcriptomic Analysis Reveals Precise Pathways Multilin and Regulatory Mechanisms Underlying bud dev Hepatoblast Differentiation

J. Gray Camp^{1*}, Kei Jorge Kageyama¹, C Li Yang,^{1,2*} Wei-Hua Wang,^{1,2*} Wei-Lin Qiu,^{1,3*} Zhen Guo,¹ Erfei Bi,⁴ and Cheng-Ran Xu¹ Emi Yoshizawa,² M



946.
doi: [10.1186/s12864-017-4342-x](https://doi.org/10.1186/s12864-017-4342-x)

PMCID: PMC5715535
PMID: [29202695](https://pubmed.ncbi.nlm.nih.gov/29202695/)

A-Seq analysis reveals dynamic trajectories during development

¹ Xin Zou,^{#1} Zhao-Ning Lu,^{#1} Gangcai Xie,² Jean Y. H. Yang,³ Chong-Chao Wu,¹ Yan He,¹ Qing Luo,¹ Yu-Lan Qu,¹ Na Wang,¹ Lan Wang,¹ and Ze-Guang Han^{1,4}

Liver fetal development time course datasets – integrating multiple datasets



E9.5 E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 E17.5

GSE87795
Su et al.

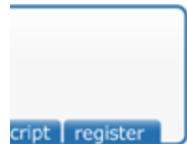
HEPATOLOGY

RAPID COMMUNICATION | HEPATOLOGY, VOL. 66, NC

Dong et al. *Genome Biology* (2018) 19:31
<https://doi.org/10.1186/s13059-018-1416-2>

RESEARCH ARTICLE

Open Access



PMCID: PMC5715535

PMID: [29202695](#)

LET] A Single-Cell Transcriptomic Analysis Reveals Precise Pathway and Regulatory Mechanisms in Hepatoblast Differentiation

Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis

Ji Dong^{1,2†}, Yuqiong Hu^{1,2,3†}, Xiaoying Fan^{1,2†}, Xinglong Wu^{1,2,3†}, Yunuo Mao^{1,2}, Boqiang Hu^{1,2}, Hongshan Guo^{1,2}, Lu Wen^{1,2} and Fuchou Tang^{1,2,3*}

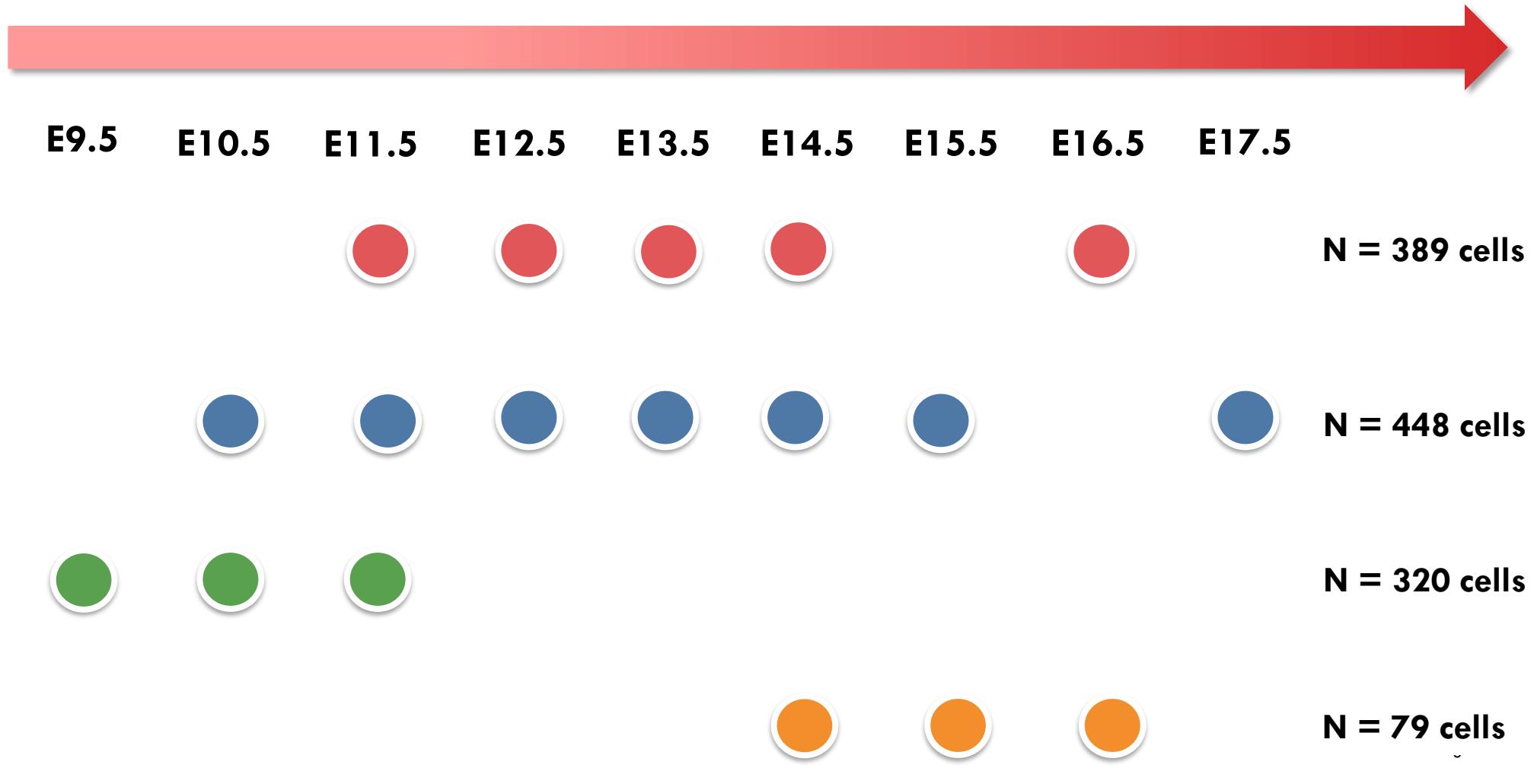
✉ tangfc@sjtu.edu.cn

¹ Xin Zou, ^{#1} Zhao-Ning Lu, ^{#1} Gangcai Xie, ² Jean Y. H. Yang, ³ Chong-Chao Wu, ¹ Yan He, ¹ Qing Luo, ¹ Yu-Lan Qu, ¹ Na Wang, ¹ Lan Wang, ¹ and Ze-Guang Han ^{✉1,4}

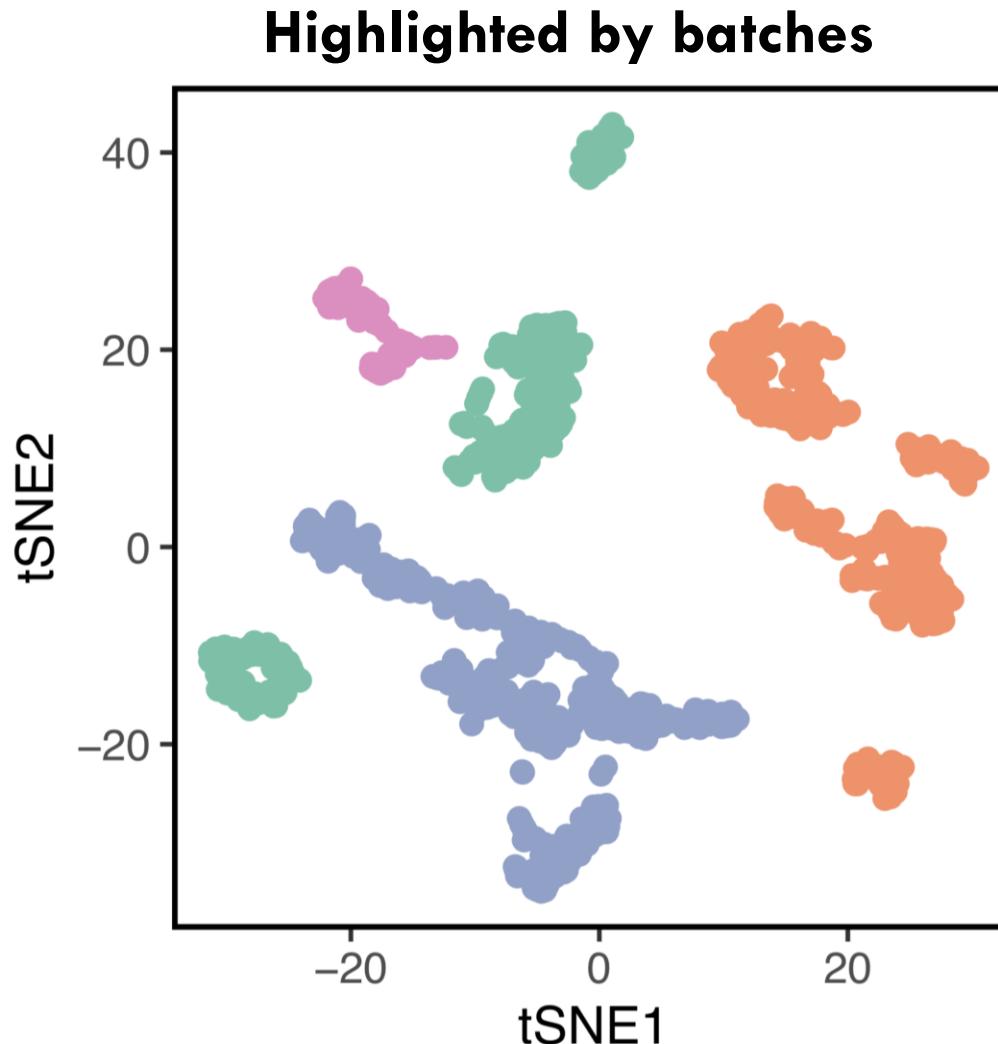
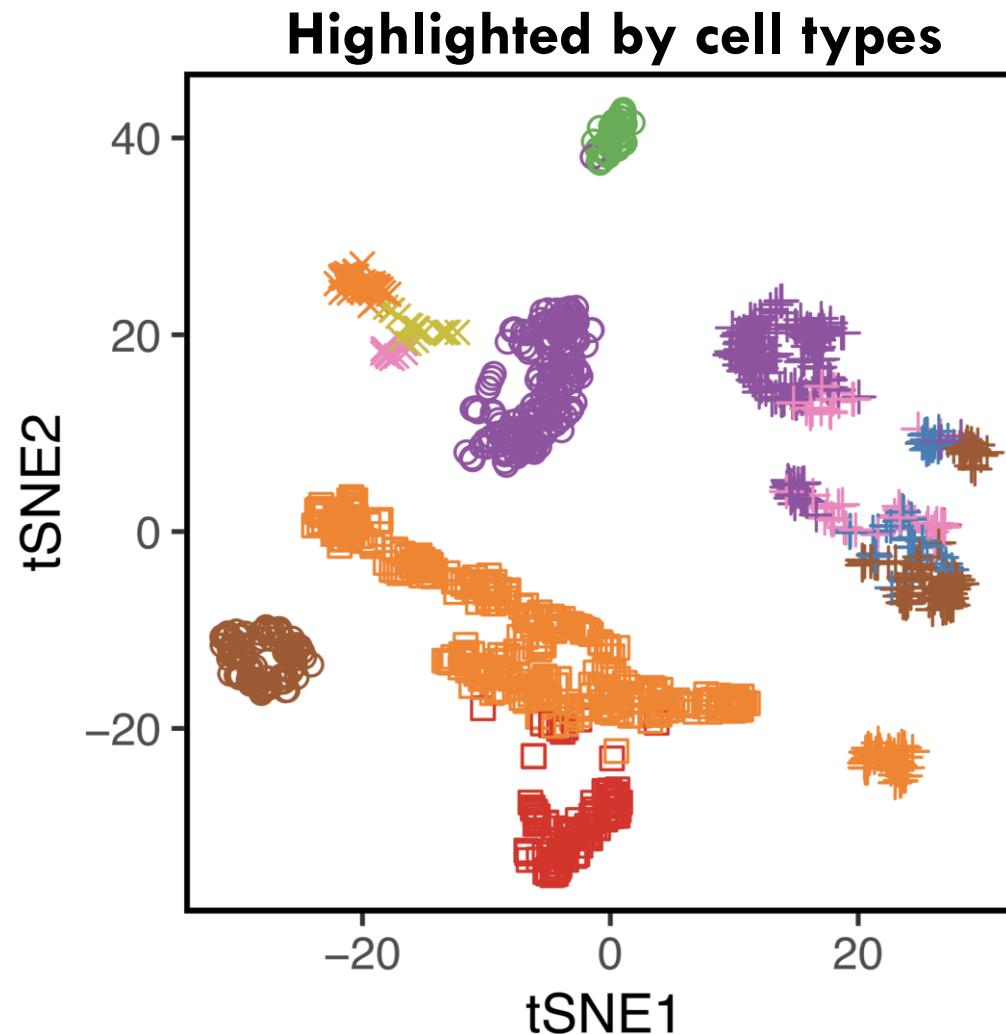
J. Gray Camp^{1*}, Kei Jorge Kageyama¹, C Li Yang,^{1,2*} Wei-Hua Wang,^{1,2*} Wei-Lin Qiu,^{1,3*} Zhen Guo,¹ Erfei Bi,⁴ and Cheng-Ran Xu¹ Emi Yoshizawa,² M

[Article notes](#) ▶ [Copyright and License information](#) ▶ [Disclaimer](#)

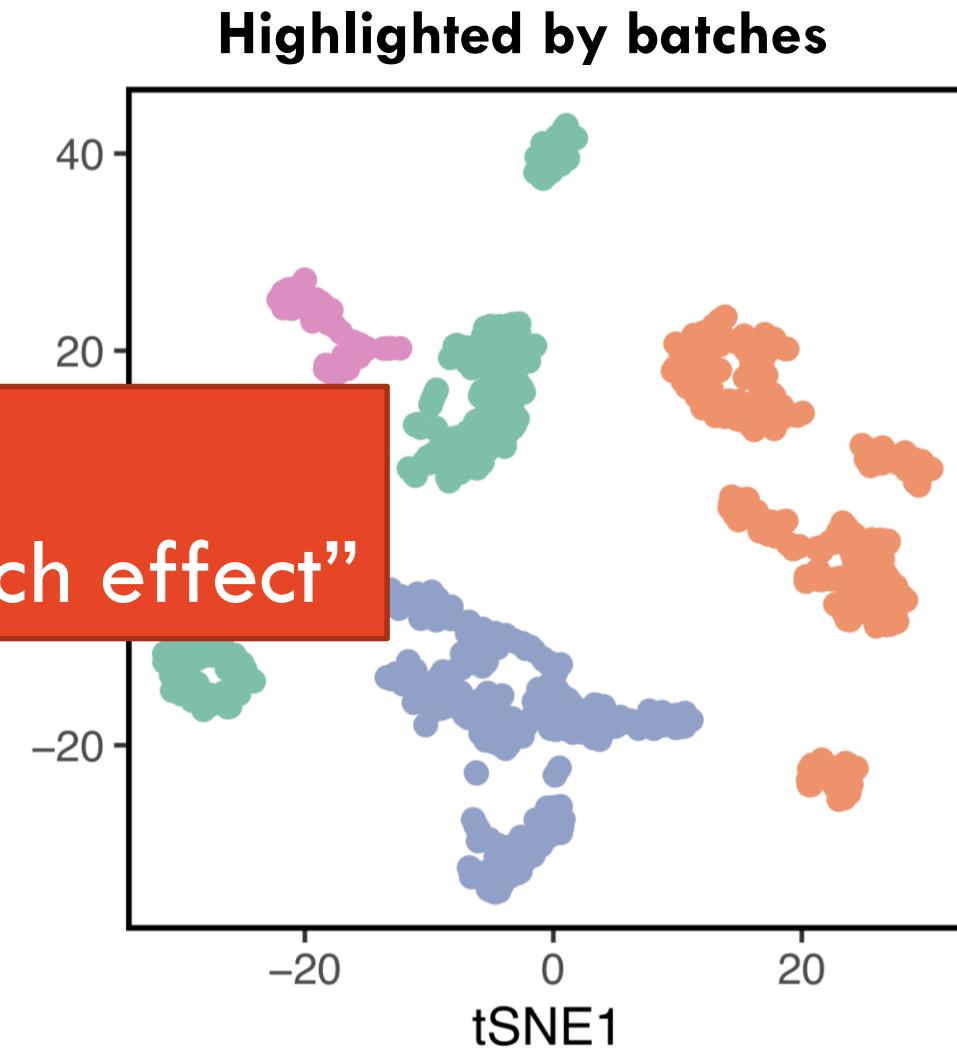
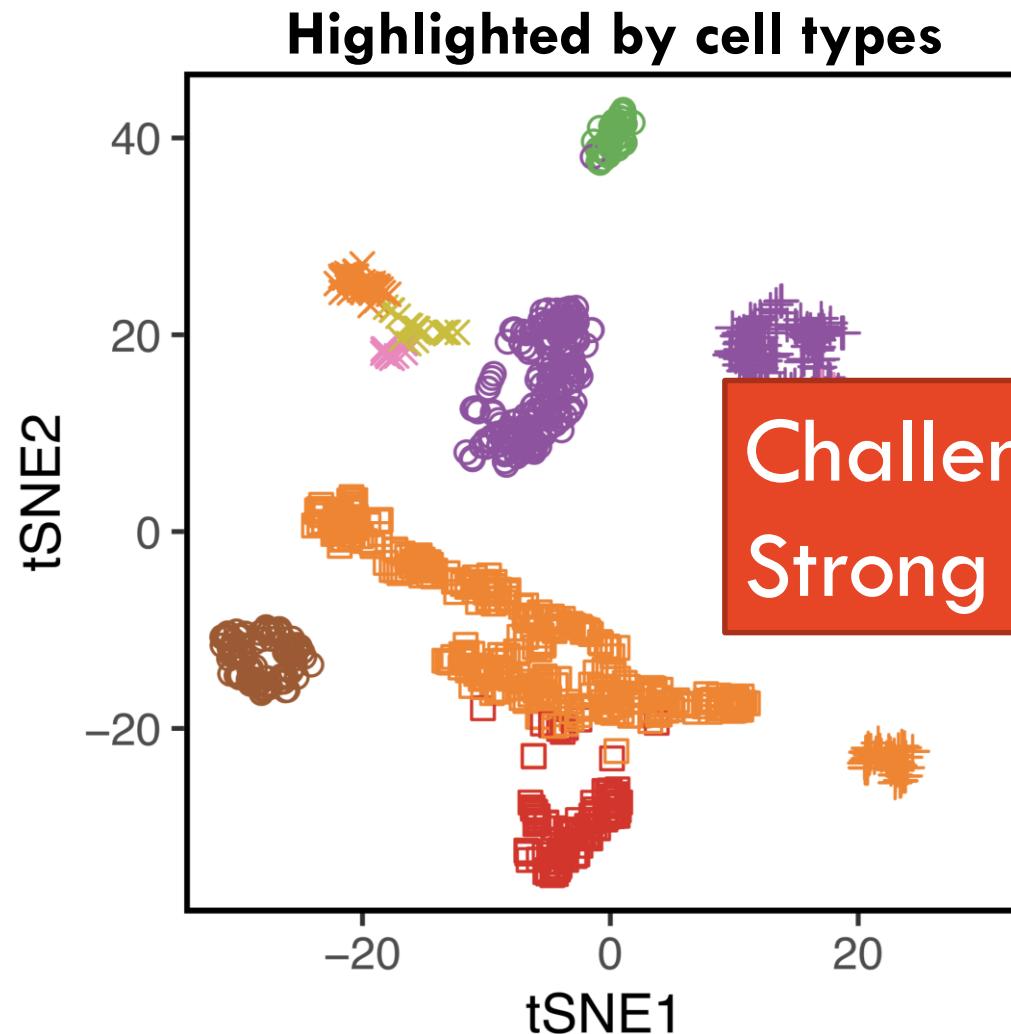
Liver fetal development time course datasets



tSNE of liver fetal development time course datasets



tSNE of liver fetal development time course datasets



Cell types

- cholangiocyte
- Epithelial Cell
- hepatoblast/hepatocyte
- Mesenchymal Cell
- Endothelial Cell
- Hematopoietic
- Immune cell
- Stellate Cell

Batch

- GSE87038
- GSE87795
- GSE90047
- GSE96981

Current approaches

Current approaches

ANALYSIS

_computational
BIOLOGY

Normalization of RNA-seq data using factor analysis
of control genes or samples

Davide Risso¹, John Ngai²⁻⁴, Terence P Speed^{1,5,6} & Sandrine Dudoit^{1,7}

Current approaches

ANALYSIS

_computational BIOLOGY

Biostatistics

Issues Advance articles Submit ▾ Purchase Alerts About ▾

Article Navigation

Adjusting batch effects in microarray expression data using empirical Bayes methods FREE

W. Evan Johnson, Cheng Li, Ariel Rabinovic

Biostatistics, Volume 8, Issue 1, 1 January 2007, Pages 118–127, <https://doi.org/10.1093/biostatistics/kxj037>

Published: 21 April 2006 Article history ▾

Normalization of RNA-seq data using factor analysis of control genes or samples

Davide Risso¹, John Ngai²⁻⁴, Terence P Speed^{1,5,6} & Sandrine Dudoit^{1,7}

Current approaches

The screenshot shows a research article from **Nature Biotechnology**. The article title is "Integrating single-cell transcriptomic data across different conditions, technologies, and species". It is authored by Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexis & Rahul Satija. The article was published in **Nature Biotechnology** 36, 411–420 (2018). The URL is <https://doi.org/10.1038/nbt.4082>. The article is categorized under **ANALYSIS** and **computational BIOLOGY**.

The screenshot shows a research article from **Biostatistics**. The article title is "Adjusting batch effects in microarray expression data using empirical Bayes methods". It is authored by W. Evan Johnson, Cheng Li, Ariel Rabinovic. The article was published in **Biostatistics**, Volume 8, Issue 1, 1 January 2007, Pages 118–127. The URL is <https://doi.org/10.1093/biostatistics/kxj037>. The article is categorized under **Article Navigation**.

Current approaches

The image displays three academic publications arranged diagonally from top-left to bottom-right:

- Top Left:** A purple banner with the text "ANALYSIS" and "computational BIOLOGY" above a white header for "nature biotechnology". The main title is "Integrating single-cell RNA-seq data across different technologies, and". Below it are author names: Andrew Butler, Paul Hoffman, Peter Smits, and Nature Biotechnology 36, 411–420 (2018). A red arrow points from this paper towards the center.
- Center:** A white header for "Biostatistics" with navigation links: Issues, Advance articles, Submit, Purchase, Alerts, and About. The main title is "Adjusting batch effects in microarray expression data using empirical Bayes methods" by W. Evan Johnson, Cheng Li, Ariel Rabinovic. It includes publication details: Biostatistics, Volume 8, Issue 1, 1 January 2007, Pages 118–127, and a DOI: 10.1093/biostatistics/kxj037. A blue arrow points from this paper towards the bottom-right.
- Bottom Right:** A green header for "nature biotechnology" with a dropdown menu. The main title is "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors" by Laleh Haghverdi, Aaron T L Lun, Michael D Morgan & John C Marioni. Below it are author names: Andrew Butler, Paul Hoffman, Peter Smits, and Nature Biotechnology 36, 411–420 (2018).

Current approaches

The image displays a collage of academic publications and a software interface, all related to the analysis of RNA-seq data, specifically addressing batch effects and signal extraction.

Top Left: A purple ribbon banner with the word "ANALYSIS" and "computational BIOLOGY" written on it.

Top Right: A screenshot of the **Biostatistics** journal website. The main title "Biostatistics" is in large blue letters. Below it is a navigation bar with links: Issues, Advance articles, Submit, Purchase, Alerts, and About. The main content area features an article titled "Adjusting batch effects in microarray expression data with respect to covariates" by W. Evan Johnson, Cheng Li, and Ariel Rabinovic. The article is marked as "FREE". It includes a "Article history" link and was published in Biostatistics, Volume 8, Issue 1, 1 January 2007, Pages 118–127. The DOI is 10.1093/biostatistics/kxj033.

Middle Left: A screenshot of the **nature biotechnology** journal website. The main title "nature biotechnology" is in white on a dark green background. Below it is a sub-section with the title "Integrating single data across different technologies, and" and authors Andrew Butler, Paul Hoffman, Peter Smit. It is published in Nature Biotechnology 36, 411–420 (2018). The date "Published: 02 April 2018" is also present.

Middle Center: A screenshot of the **nature biotechnology** journal website. The main title "nature biotechnology" is in white on a dark green background. Below it is a sub-section with the title "Batch effects in single sequencing data are correctly identified by matching mutual nearest neighbor distances" and authors Laleh Haghverdi, Aaron T L Lun, Michael D Morgan & others. It is published in Nature Biotechnology 36, 421–427 (2018). The date "Published: 02 April 2018" is also present.

Bottom Right: A screenshot of the **nature COMMUNICATIONS** journal website. The main title "nature COMMUNICATIONS" is in white on a grey background. Below it is a featured article titled "A general and flexible method for signal extraction from single-cell RNA-seq data" by Davide Risso, Fanny Perraudau, Svetlana Gribkova, Sandrine Dudoit, & Jean-Philippe Vert. The article is marked as "OPEN" and has a DOI of 10.1038/s41467-017-02554-5.

Bottom Left: The University of Sydney logo.

Bottom Center: The text "Nature Biotechnology 36, 421–427 (2018) | Download" is displayed.

scMerge



THE UNIVERSITY OF
SYDNEY

Breaking observed data into components

For n cells with data collected for m genes

$$Y_{m \times n} = X_{m \times p} \beta_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$

The data we observe	Biologically relevant variation cell types p wanted variables	Unwanted variation batch and technical effects k unwanted variables	Random noise
---------------------	---	---	--------------

New Results

A new normalization for the Nanostring nCounter gene expression assay

Ramyar Molania, Johann A Gagnon-Bartsch, Alexander Dobrovic, Terence P Speed
doi: <https://doi.org/10.1101/374173>

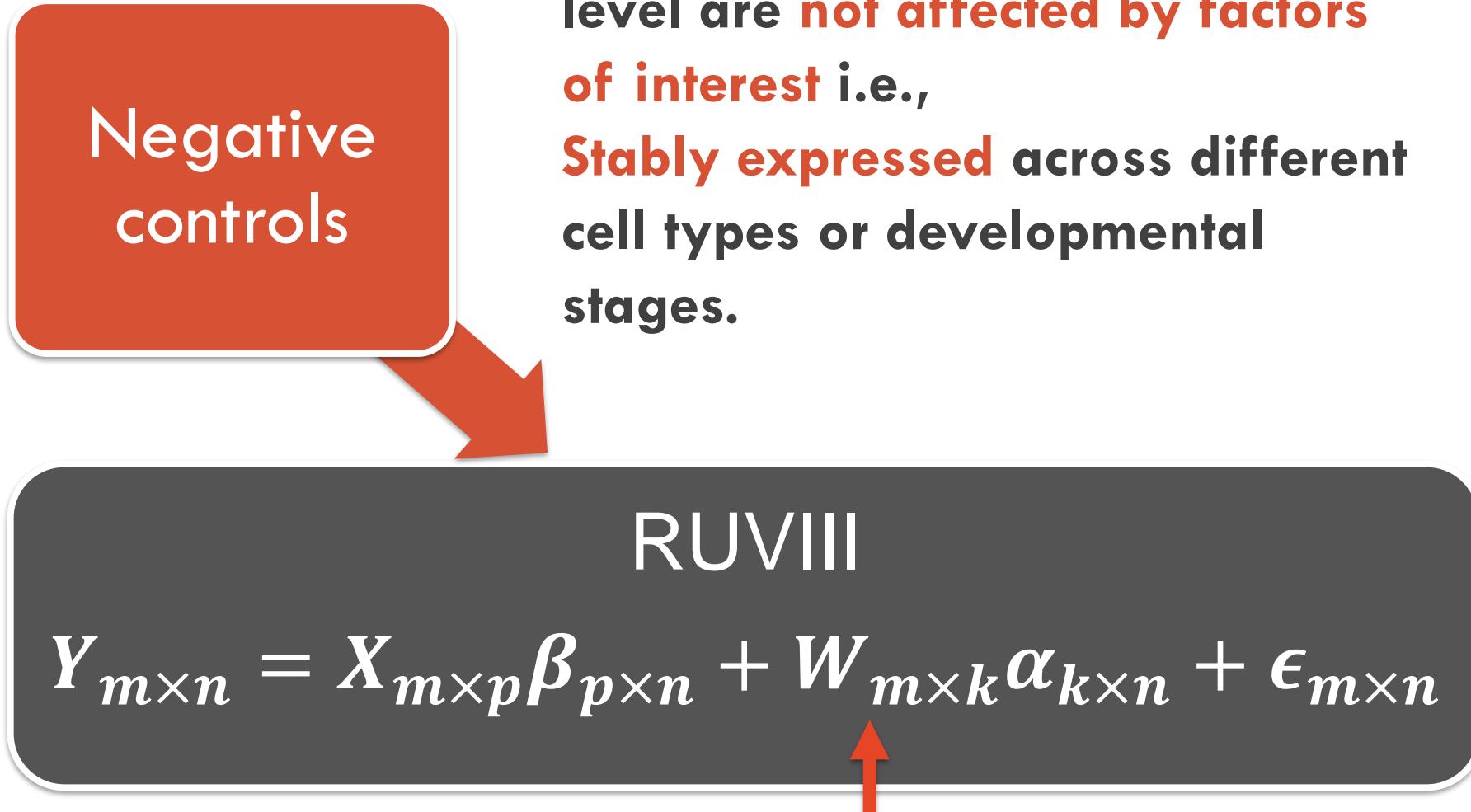
This article is a preprint and has not been peer-reviewed [what does this mean?].

RUVIII

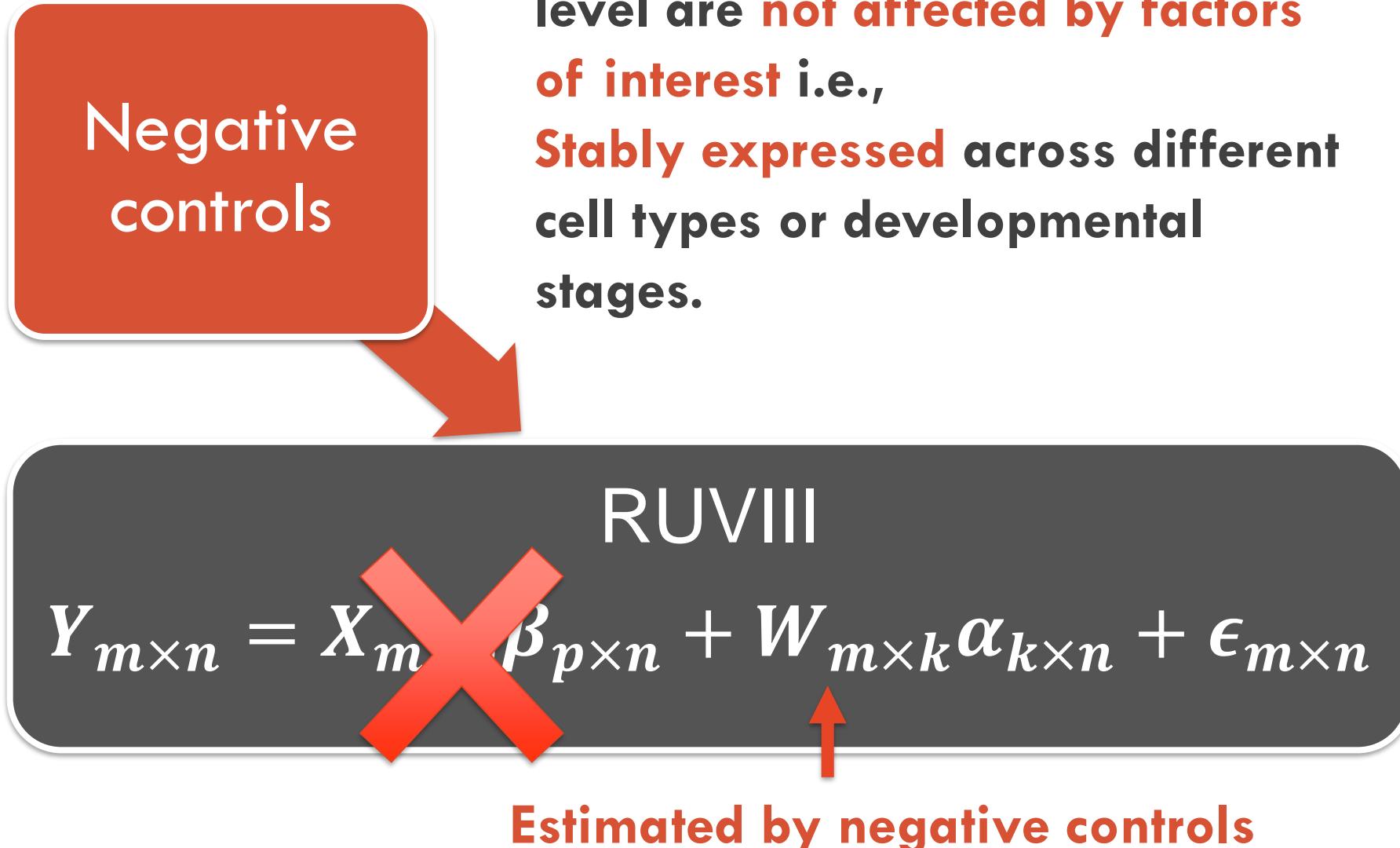
$$Y_{m \times n} = X_{m \times p} \beta_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$

Estimate unwanted variation

scMerge: algorithm



scMerge: algorithm



scMerge: algorithm

Single-cell
stably
expressed
genes (SEG)

Genes with stable expression level are not affected by factors of interest i.e., Stably expressed across different cell types or developmental stages.

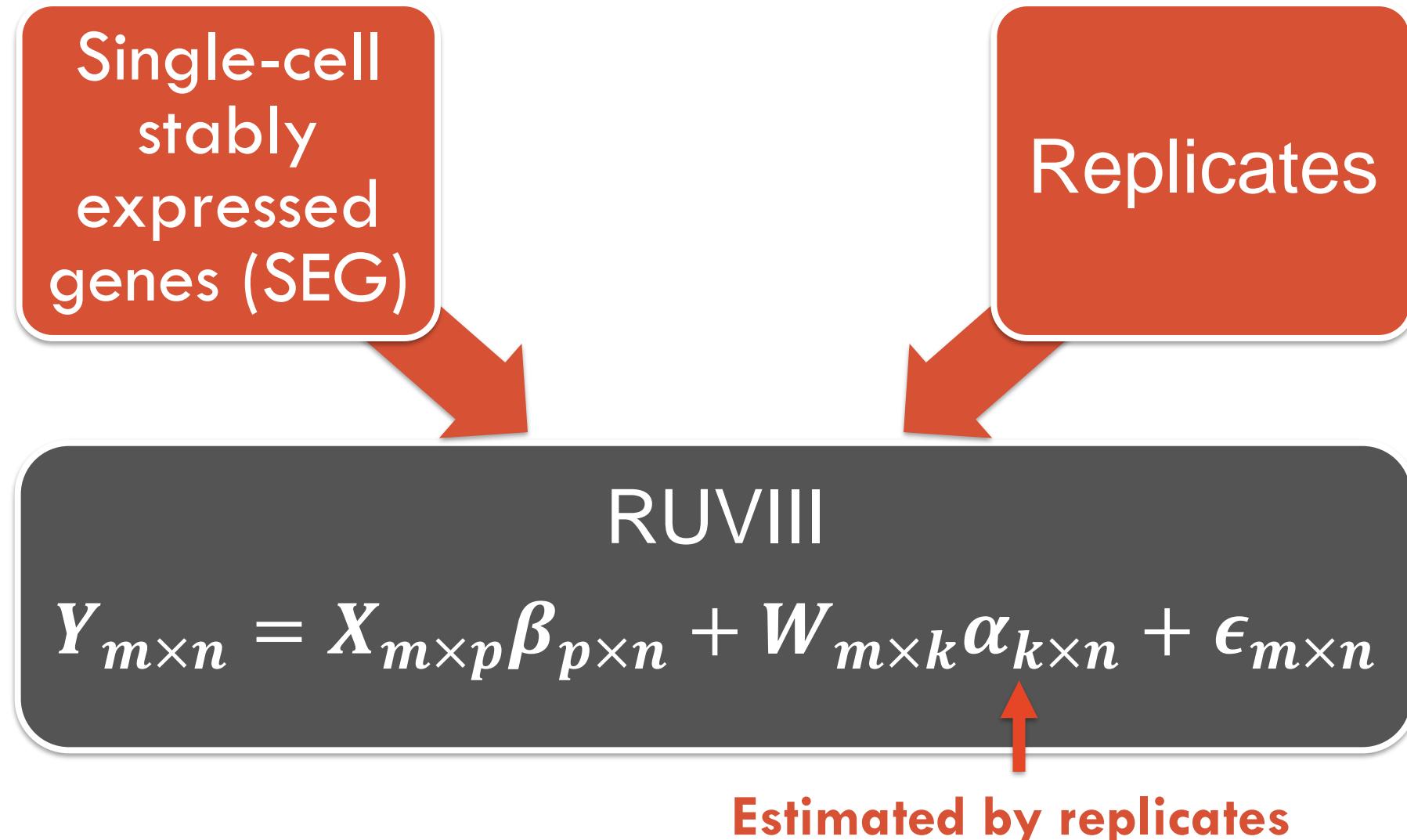
RUVIII

$$Y_{m \times n} = X_m \cancel{\beta}_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$



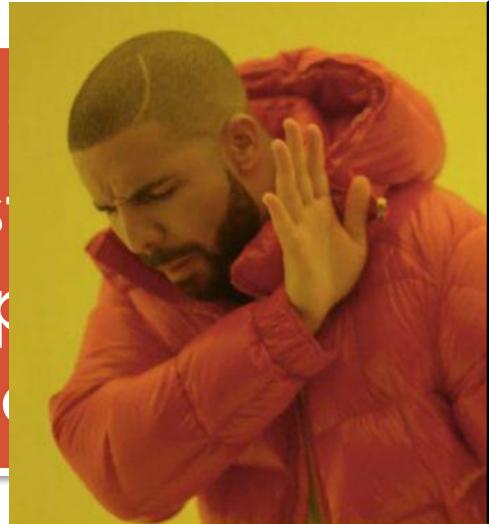
Estimated by negative controls

scMerge: algorithm



scMerge: algorithm

Sing
st
exp
geno



But cells don't have
replicates!

RUVIII

$$Y_{m \times n} = X_{m \times p} \beta_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$

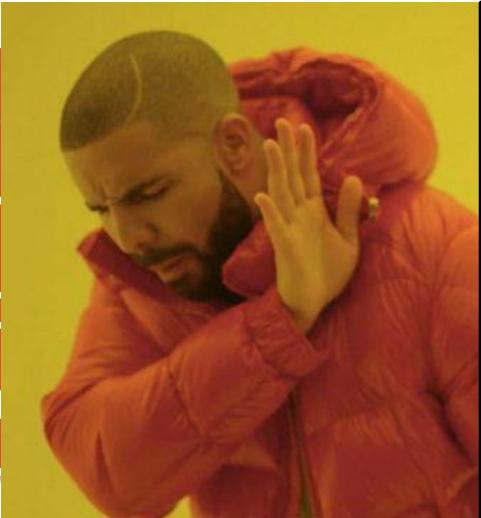


Estimated by replicates

scMerge: algorithm

Sing
st
exp
gen

Y_{mx}



But cells don't have
replicates!

Estimate
pseudo-replicates

Estimated by replicates

scMerge: algorithm

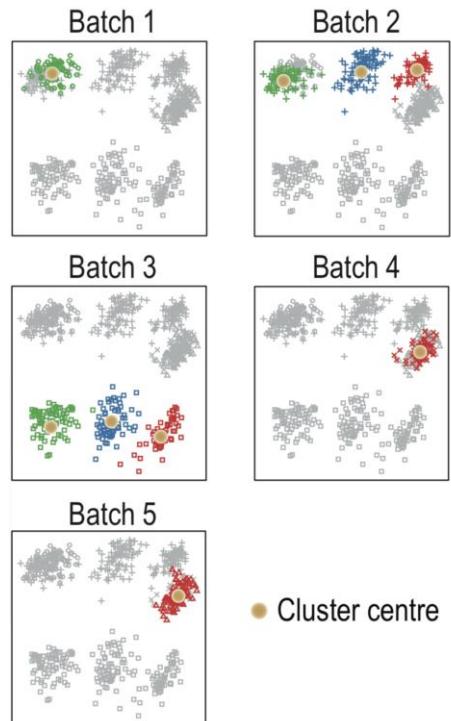
Idea:

- Consider cells from the same cell type as a set of pseudo-replicates
- Match the cell types across different batches

Pseudo-replicates

scMerge: algorithm

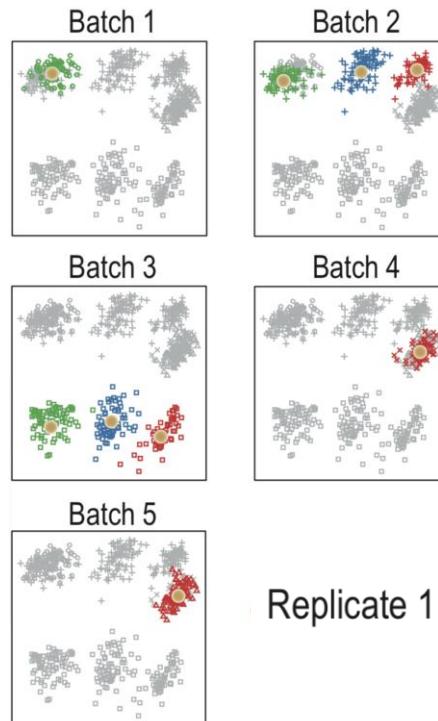
Clustering for each batch
(k-means by default)



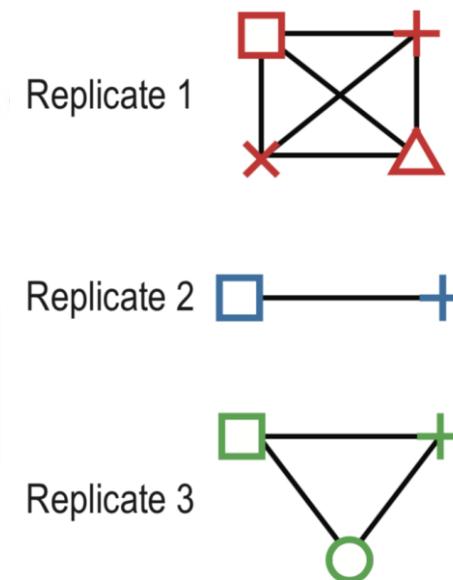
Pseudo-replicates

scMerge: algorithm

Clustering for each batch
(k-means by default)



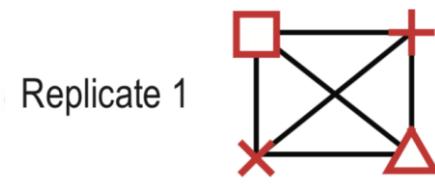
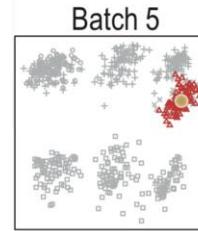
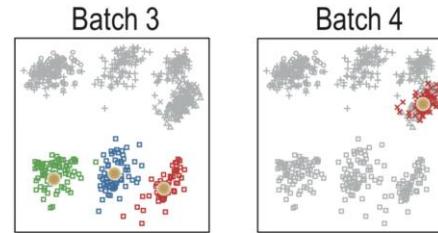
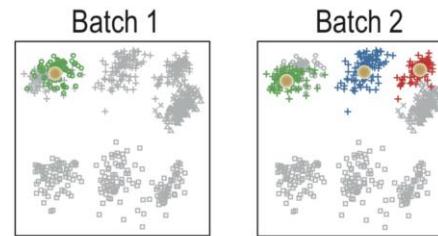
Find **Mutual Nearest Clusters**
as pseudo-replicates



Pseudo-
replicates

scMerge: algorithm

Clustering for each batch
(k-means by default)



Pseudo-replicates

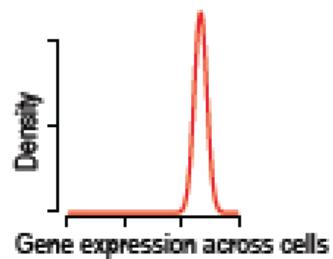
Find Mutual Nearest Clusters
as pseudo-replicates



Frame as pseudo-replicate
information

	Replicate 1	Replicate 2	Replicate 3
Cell 1	1	0	0
Cell 2	1	0	0
Cell 3	0	1	0
.	.	.	.
.	.	.	.
Cell C	0	0	1

scMerge: algorithm



Single-cell
stably
expressed
genes (SEG)

Pseudo-
replicates

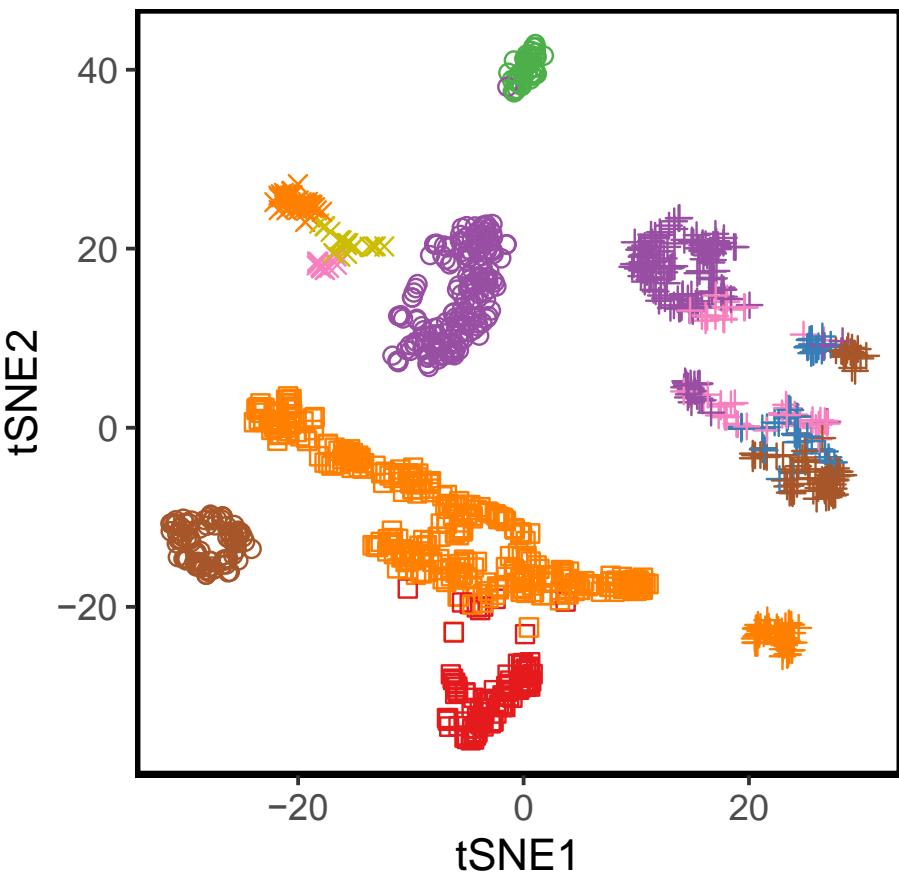
RUVIII

$$Y_{m \times n} = X_{m \times p} \beta_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$

	Cell 1	Cell 2	Cell 3	
Cell 1	1	0	0	
Cell 2	1	0	0	
Cell 3	0	1	0	
.
.
.
Cell C	0	0	1	
	Replicate 1	Replicate 2	Replicate 3	

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

Before scMerge



cell_types

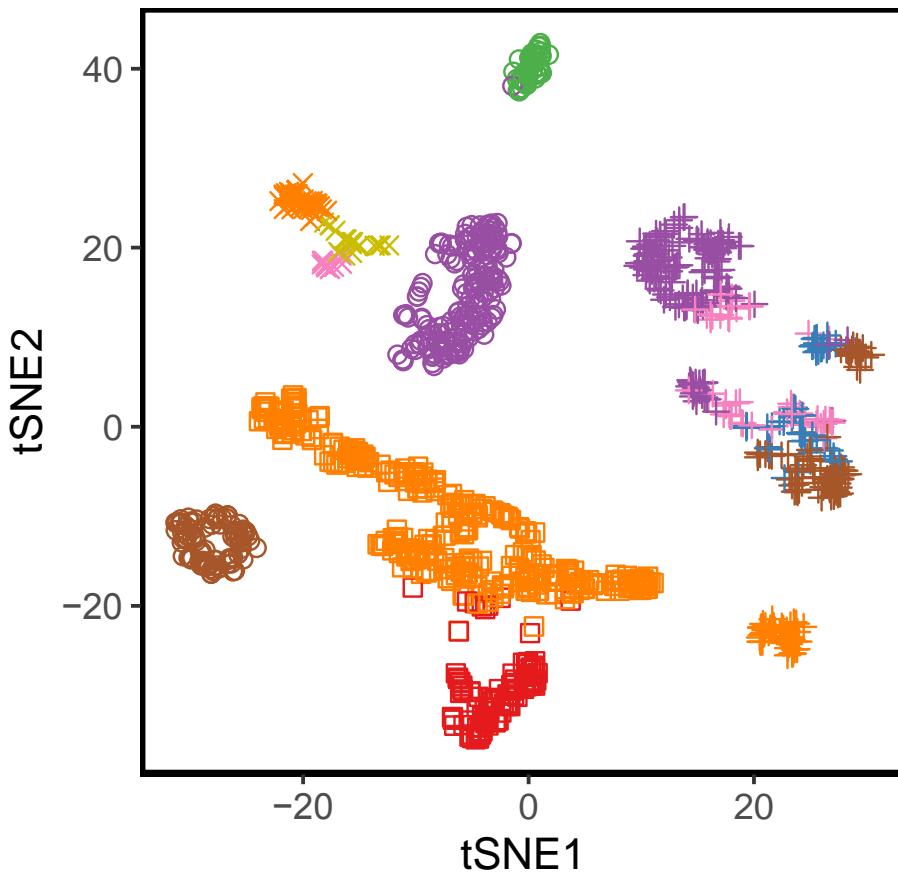
- cholangiocyte
- Endothelial Cell
- Epithelial Cell
- Hematopoietic
- hepatoblast/hepatocyte
- Immune cell
- Mesenchymal Cell
- Stellate Cell

batch

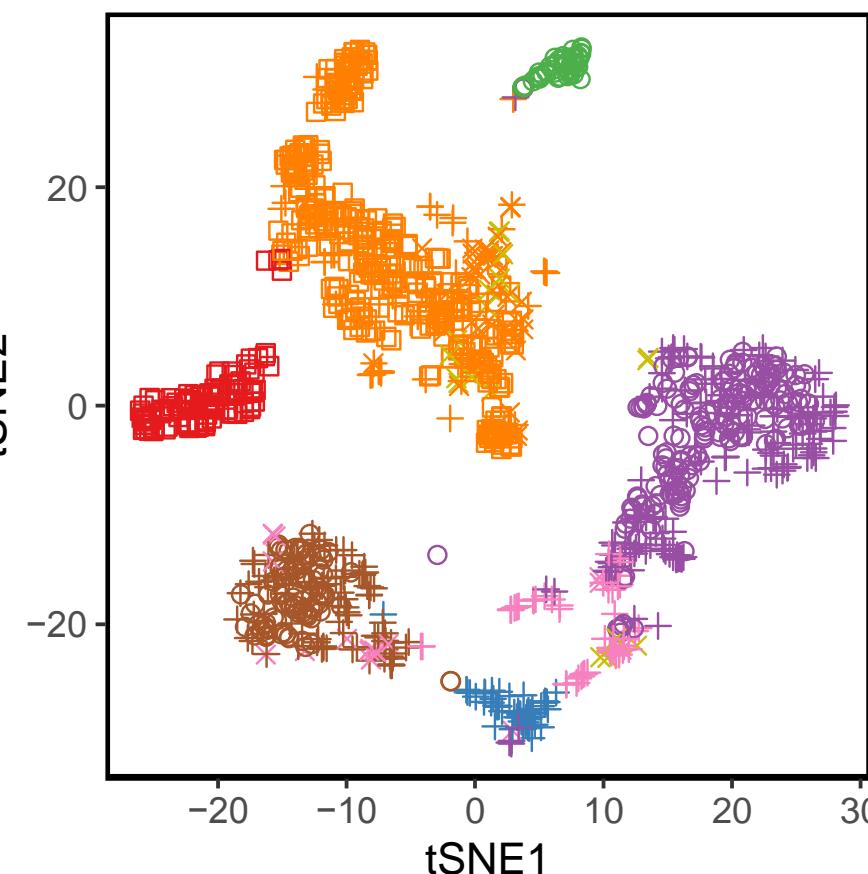
- GSE87038
- + GSE87795
- GSE90047
- × GSE96981

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

Before scMerge



After scMerge



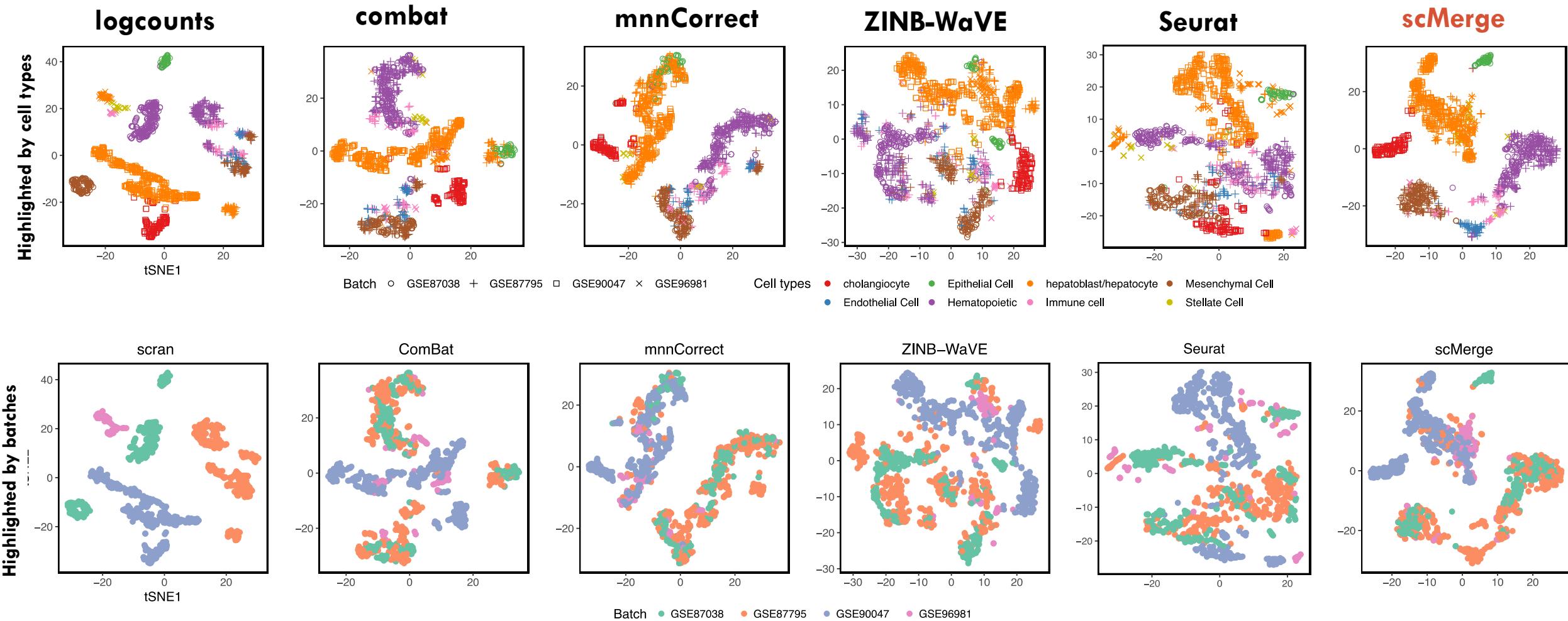
cell_types

- cholangiocyte
- Endothelial Cell
- Epithelial Cell
- Hematopoietic
- hepatoblast/hepatocyte
- Immune cell
- Mesenchymal Cell
- Stellate Cell

batch

- GSE87038
- +
- GSE90047
- ×
- GSE96981

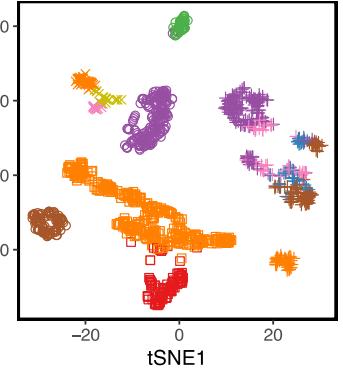
Results: liver datasets tSNE



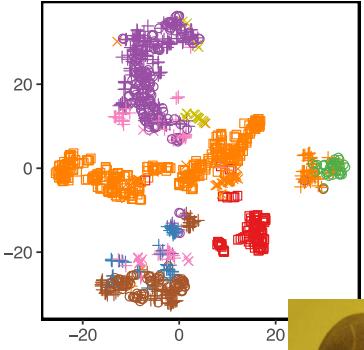
Results: liver datasets tSNE

Highlighted by cell types

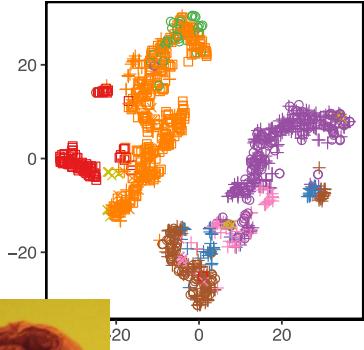
logcounts



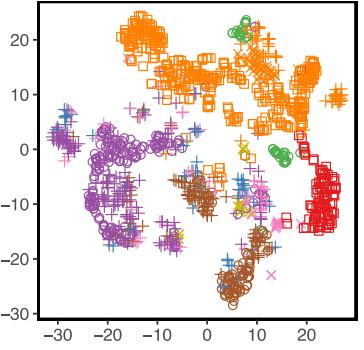
combat



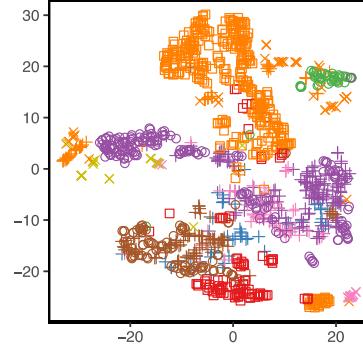
mnnCorrect



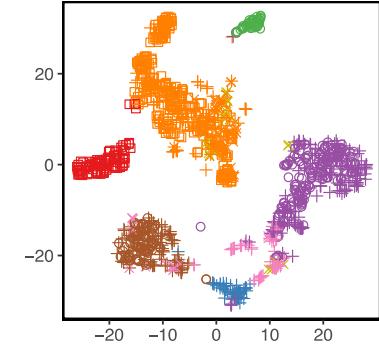
ZINB-WaVE



Seurat

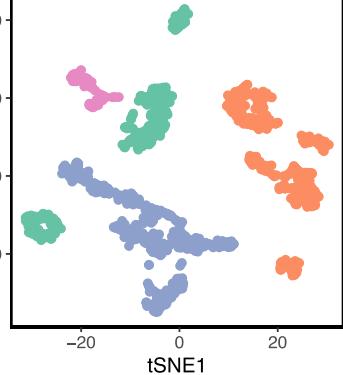


scMerge

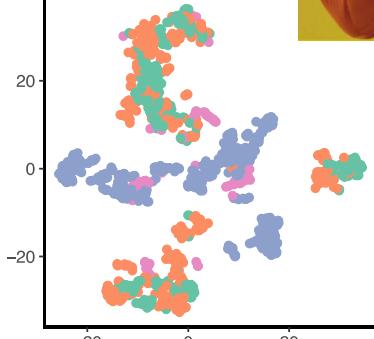


Highlighted by batches

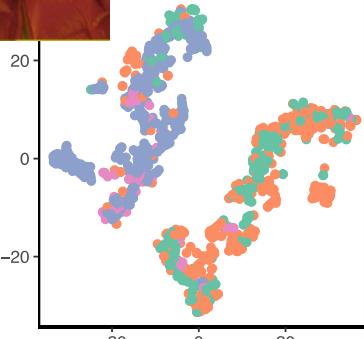
scran



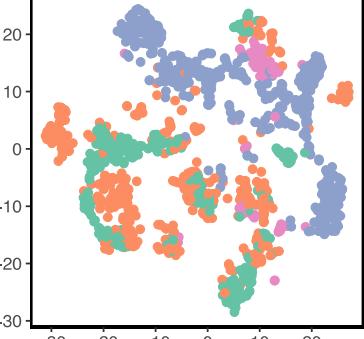
ComBat



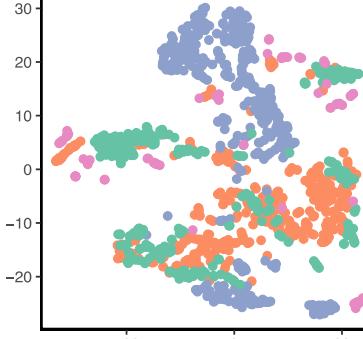
mnnCorrect



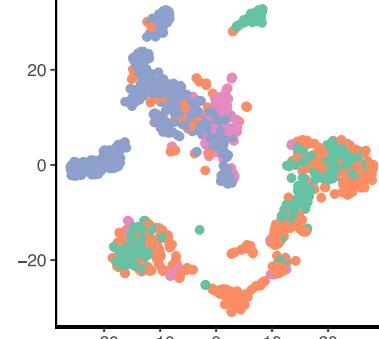
ZINB-WaVE



Seurat



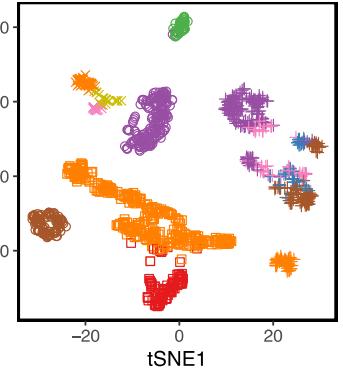
scMerge



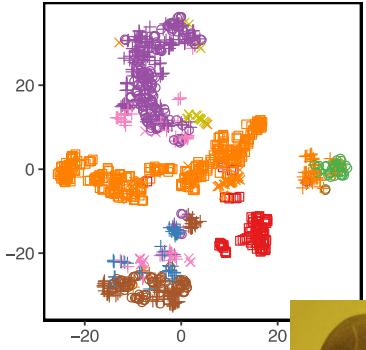
Results: liver datasets tSNE

Highlighted by cell types

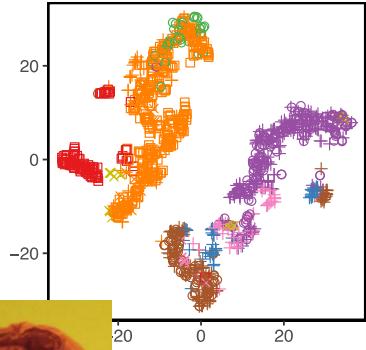
logcounts



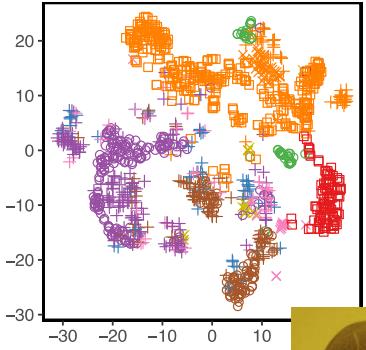
combat



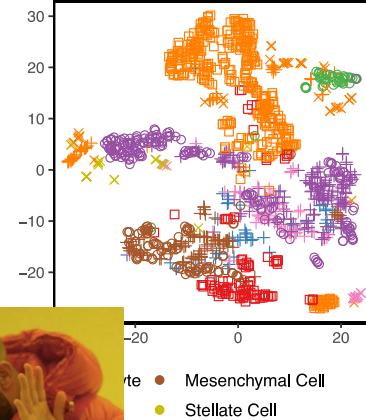
mnnCorrect



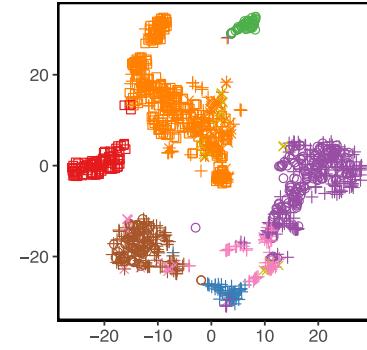
ZINB-WaVE



Seurat

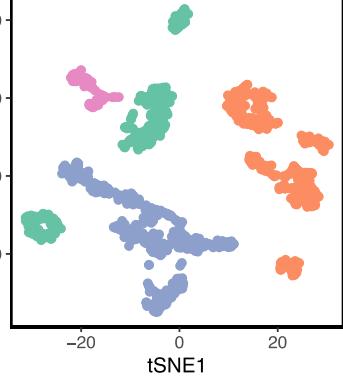


scMerge

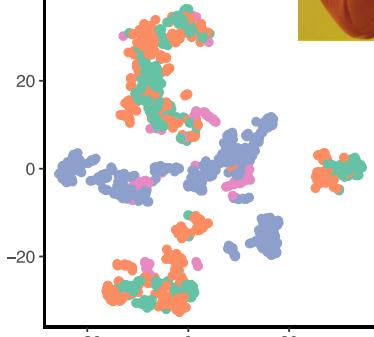


Highlighted by batches

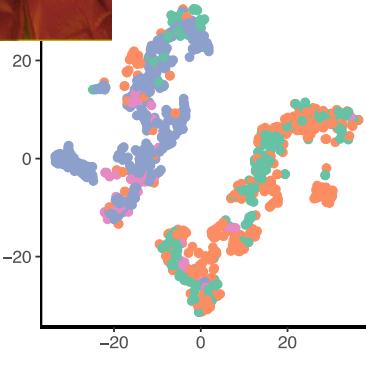
scran



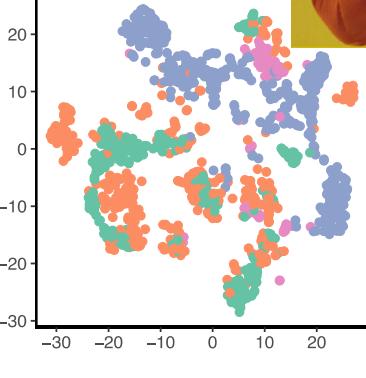
ComBat



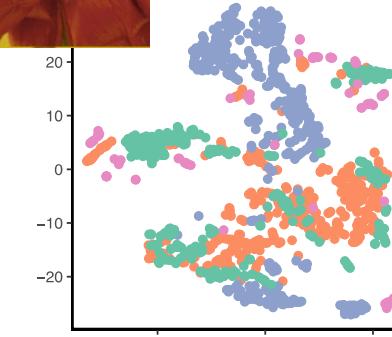
mnnCorrect



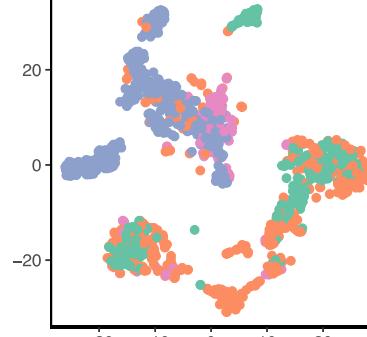
ZINB-WaVE



Seurat



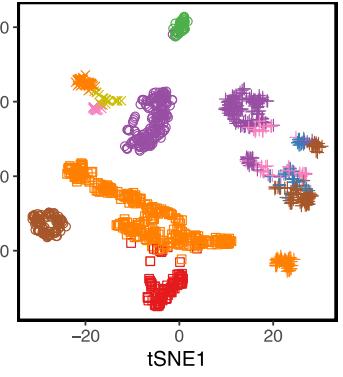
scMerge



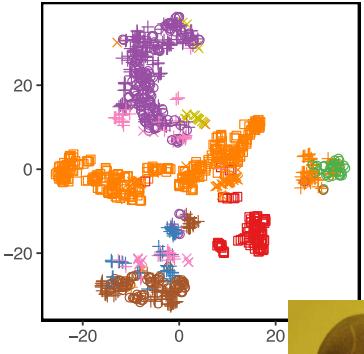
Results: liver datasets tSNE

Highlighted by cell types

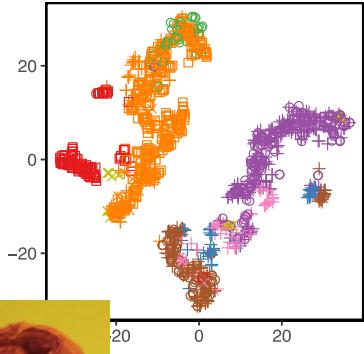
logcounts



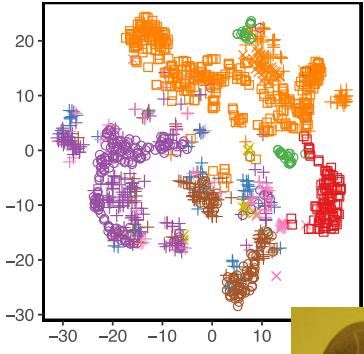
combat



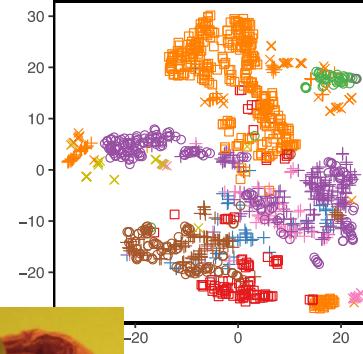
mnnCorrect



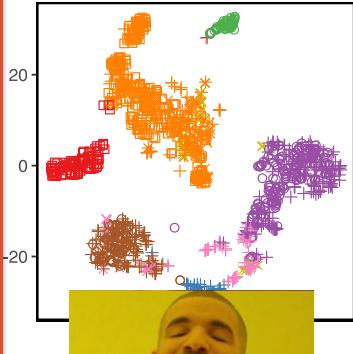
ZINB-WaVE



Seurat

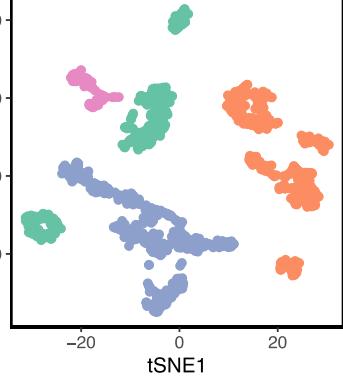


scMerge

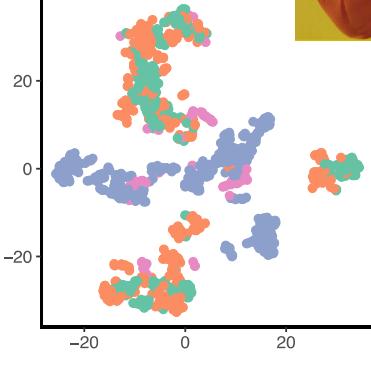


Highlighted by batches

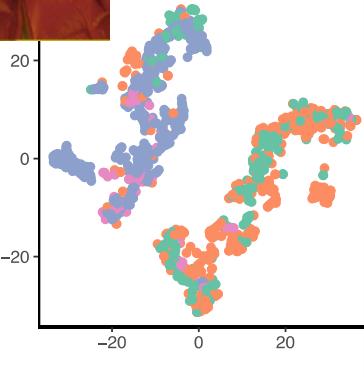
scran



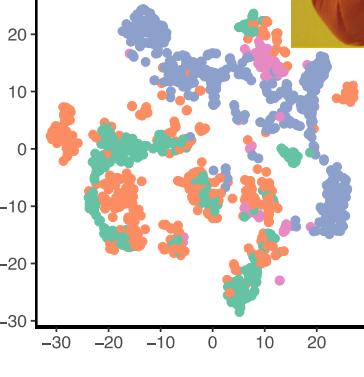
ComBat



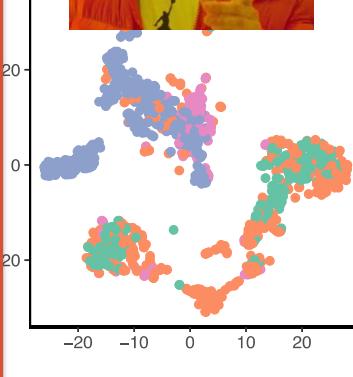
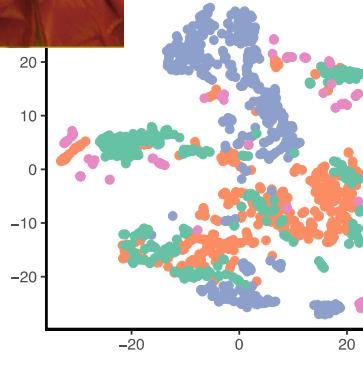
mnnCorrect



ZINB-WaVE

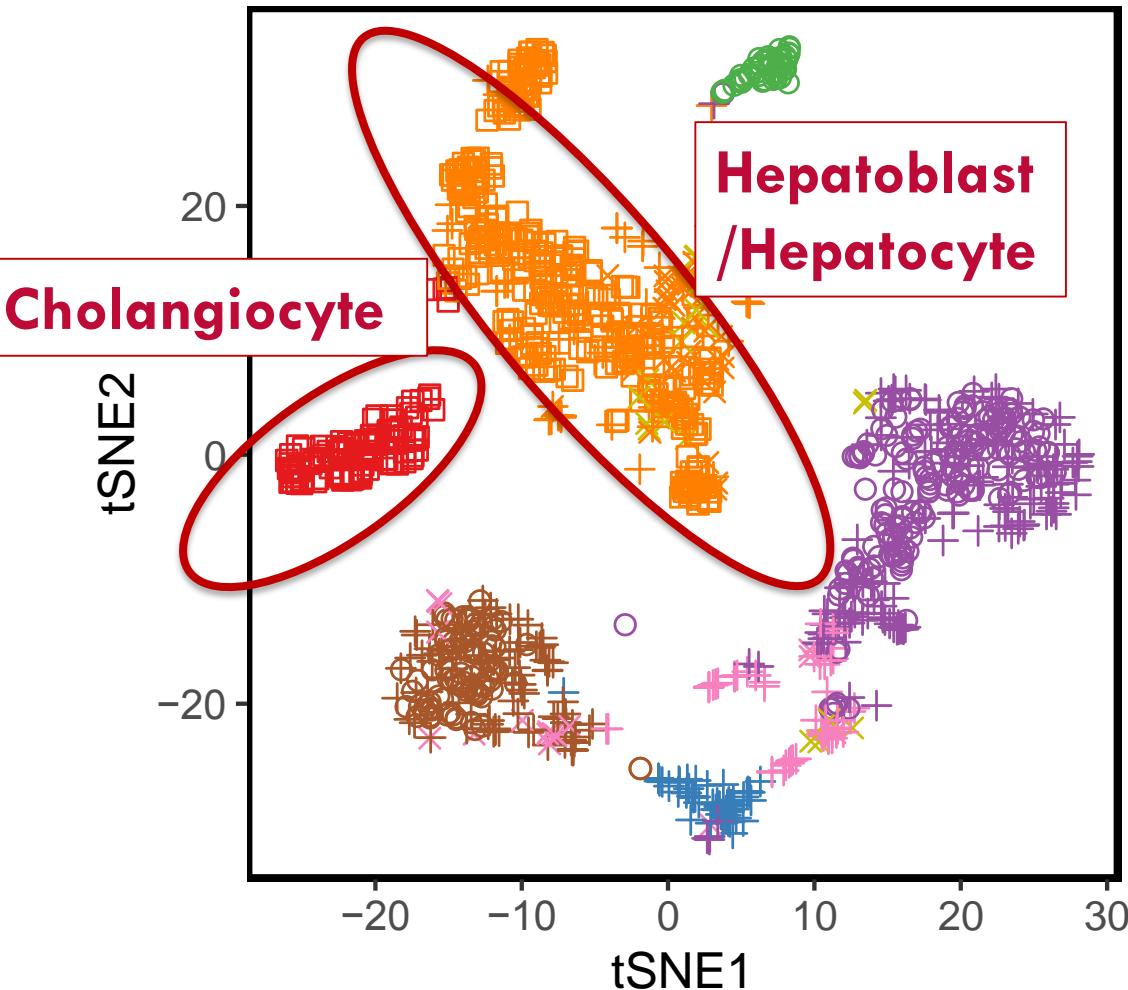


Seurat

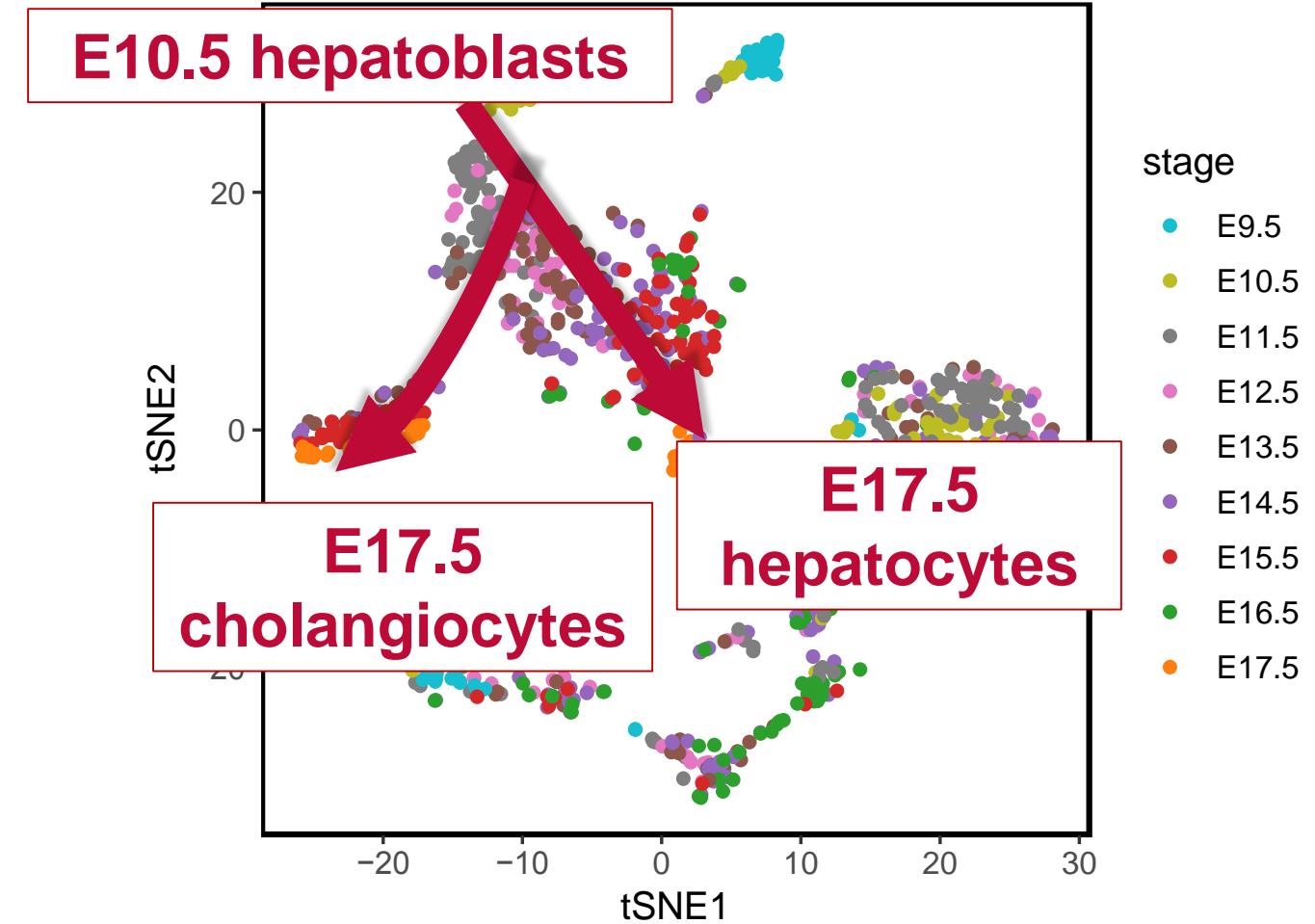


Results: liver datasets – tSNE retains rough trajectory

Highlighted by cell types



Highlighted by stage



More information

bioRxiv:

<https://www.biorxiv.org/content/early/2018/09/12/393280>



New Results

scMerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication

Yingxin Lin, Shila Ghazanfar, Kevin Wang, Johann A. Gagnon-Bartsch, Kitty K. Lo, Xianbin Su, Ze-Guang Han, John T. Ormerod, Terence P. Speed, Pengyi Yang, Jean Y. H. Yang

doi: <https://doi.org/10.1101/393280>

scMerge R package and website:

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

scMerge 0.1.14 Vignette Reference Case Study ▾

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.

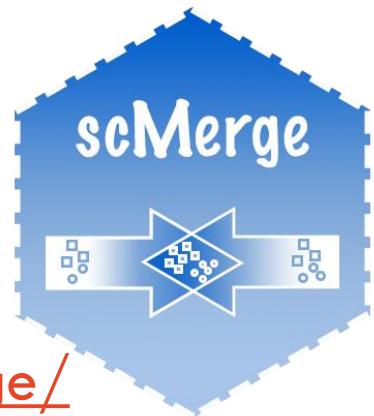
```
# Some CRAN packages required by scMerge
install.packages(c("ruv", "rsvd", "igraph", "pdist", "proxy", "foreach", "doSNOW", "distr", "Rcpp", "RcppEigen"))
devtools::install_github("theislab/kBET")

# 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>.



More information

bioRxiv:

<https://www.biorxiv.org/content/early/2018/09/12/393280>



THE PREPRINT SERVER FOR BIOLOGY

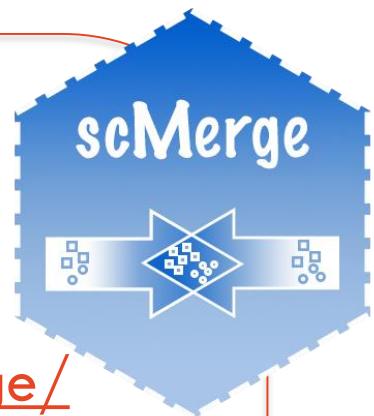
New Results

scMerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication

Yingxin Lin, Shila Ghazanfar, Kevin Wang, Johann A. Gagnon-Bartsch, Kitty K. Lo, Xianbin Su, Ze-Guang Han, John T. Ormerod, Terence P. Speed, Pengyi Yang, Jean Y. H. Yang

doi: <https://doi.org/10.1101/393280>

Thursday workshop!



scMerge R package and website:

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

scMerge 0.1.14 Vignette Reference Case Study ▾

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.

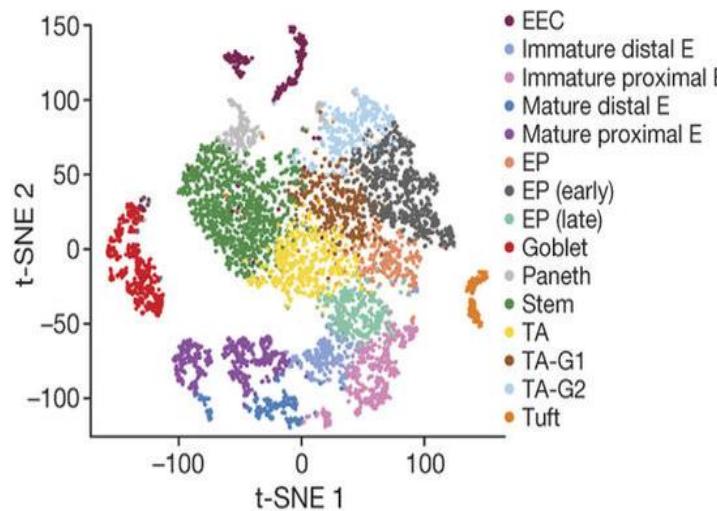
```
# Some CRAN packages required by scMerge
install.packages(c("ruv", "rsvd", "igraph", "pdist", "proxy", "foreach", "doSNOW", "distr", "Rcpp", "RcppEigen"))
devtools::install_github("theislab/kBET")

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

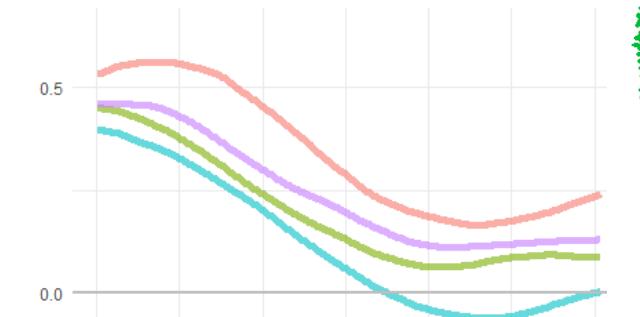
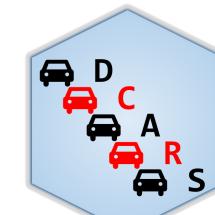
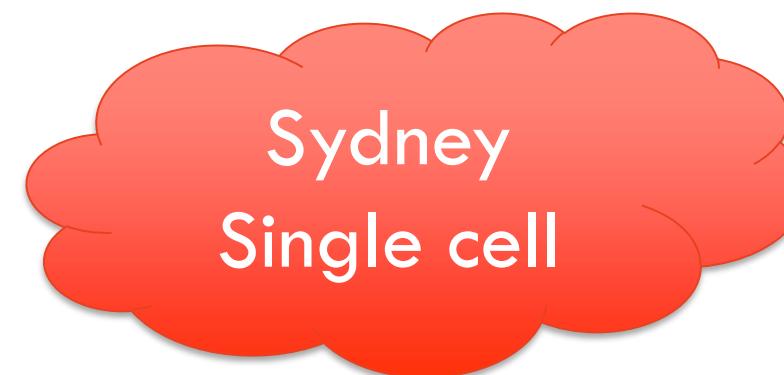
# Installing scMerge and the data files using
devtools::install_github("SydneyBioX/scMerge.data")
devtools::install_github("SydneyBioX/scMerge")
```

Vignette

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

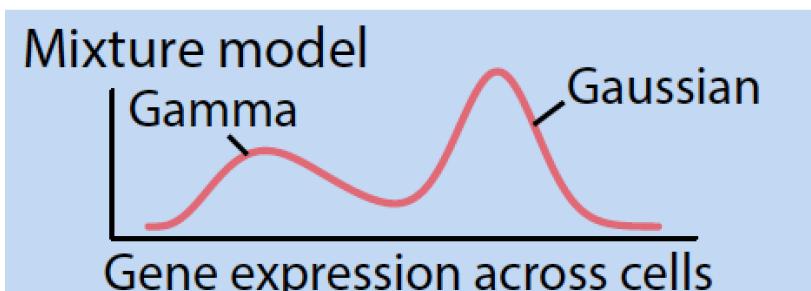


Clustering metrics

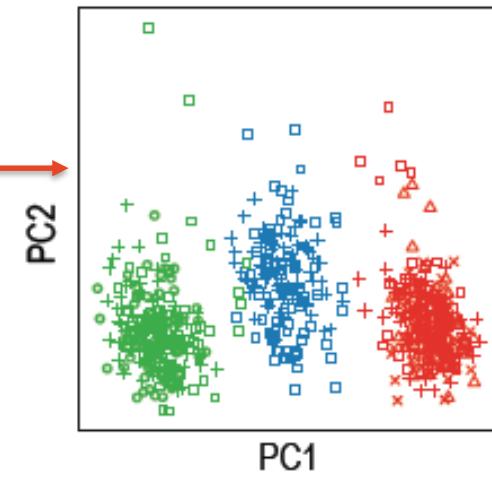
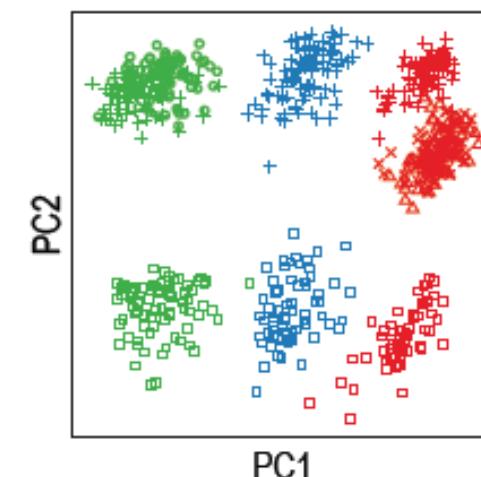


Differential correlation

Finding stably expressed genes



The University of Sydney



scMerge

Acknowledgements

Utsyd School of Mathematics and Statistics

- Jean Yang
- Pengyi Yang
- John Ormerod
- Yingxin Lin
- Kevin Wang
- Taiyun Kim
- Irene Chen
- Andy Wang

Utsyd Faculty of Science

- Kitty Lo

WEHI

- Terry Speed

University of Michigan

- Johann Gagnon-Bartsch

Shanghai Jiao Tong University

- Zeguang Han
- Xianbin Su



THE UNIVERSITY OF
SYDNEY



