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

Presented by

Dr Shila Ghazanfar

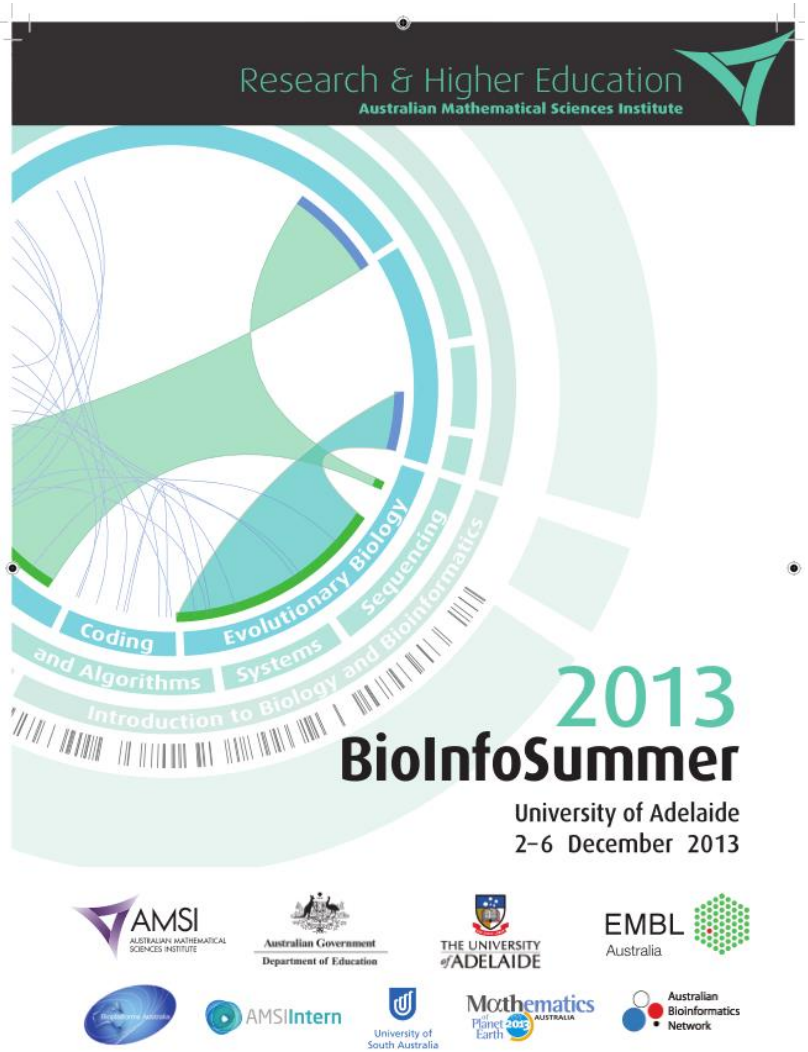
Judith and David Coffey LifeLab
Charles Perkins Centre

School of Mathematics and Statistics
The University of Sydney

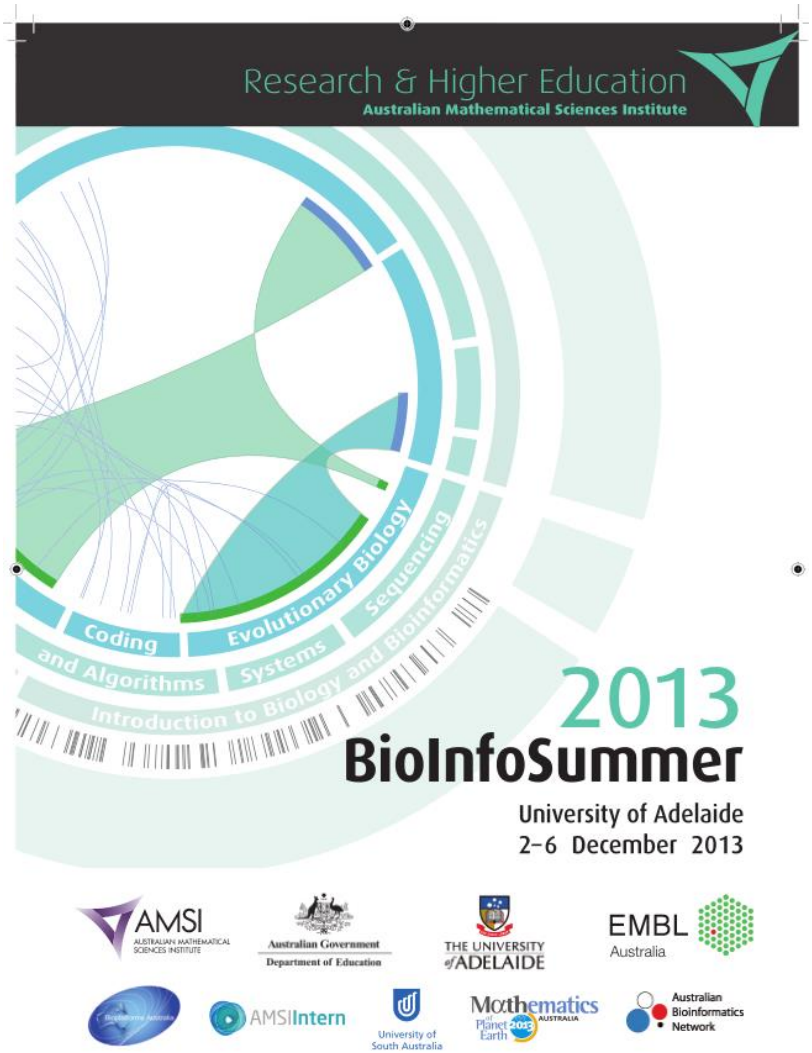
@shazanfar

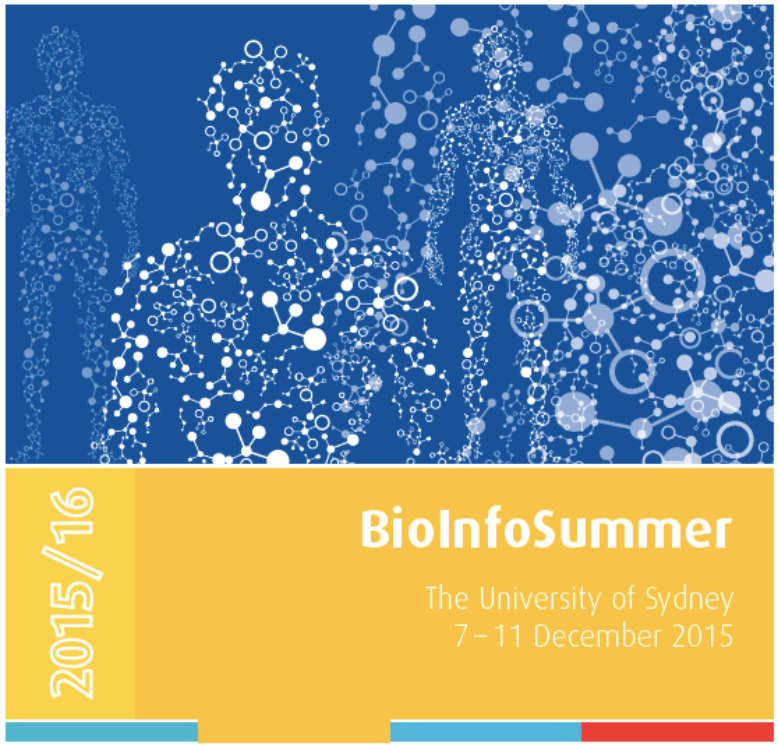
THE UNIVERSITY OF SYDNEY

A brief history

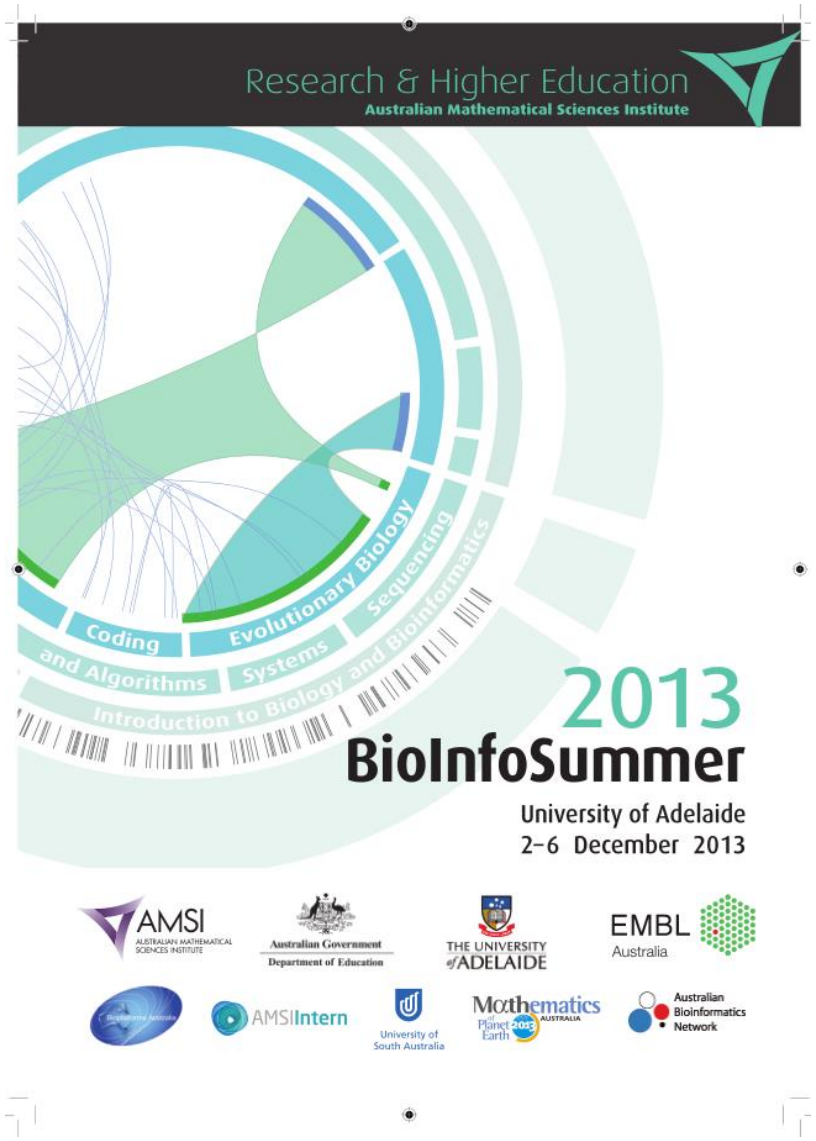


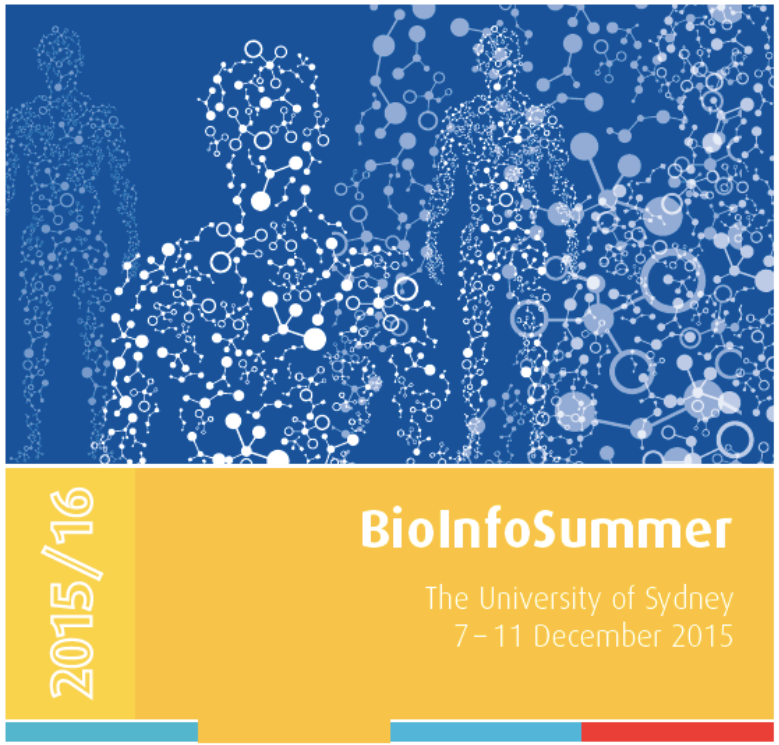
A brief history

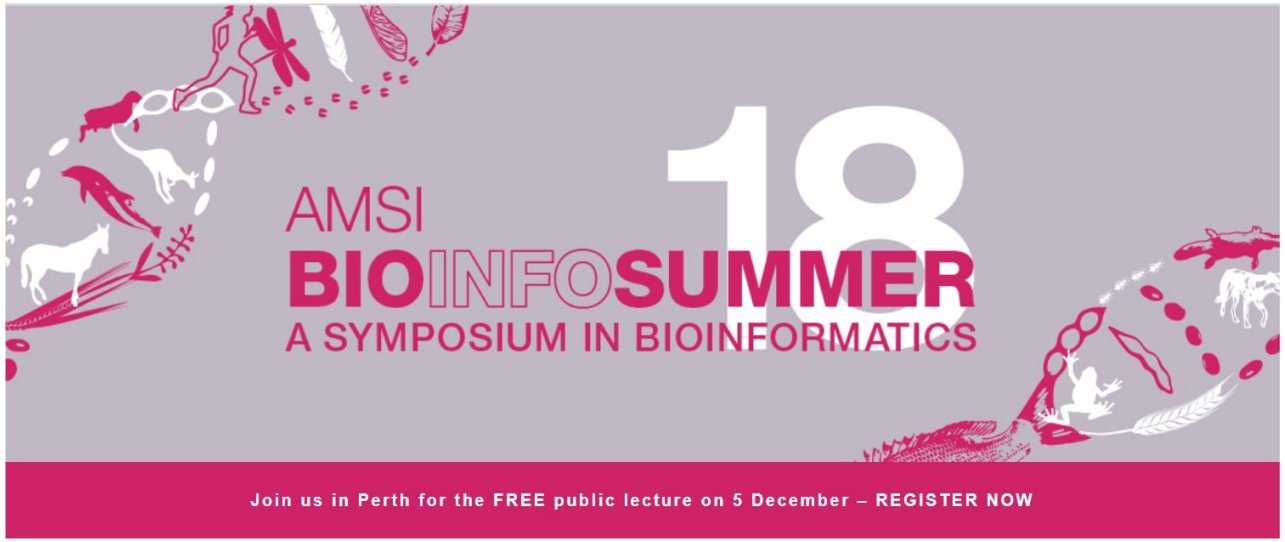




A brief history





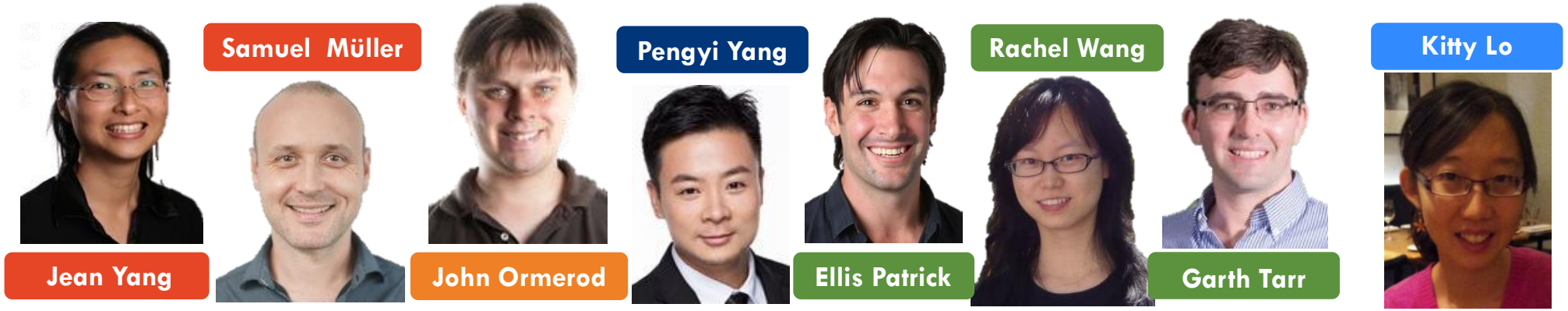


Sydney Precision Bioinformatics Research Group

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.

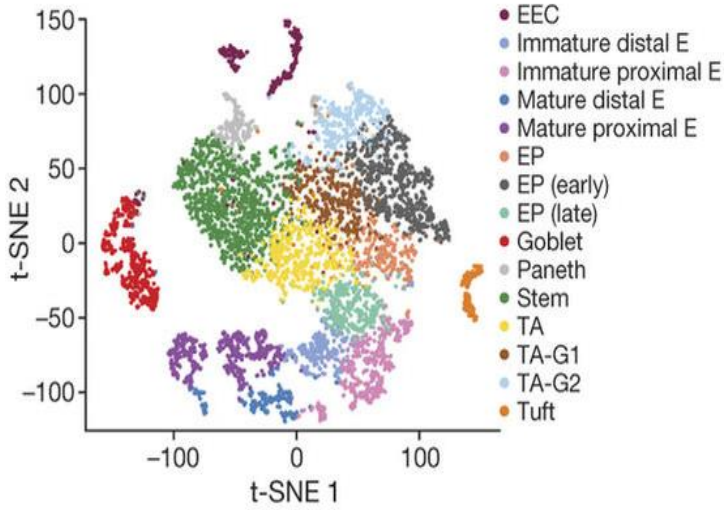
Meet our senior and junior research leaders …



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

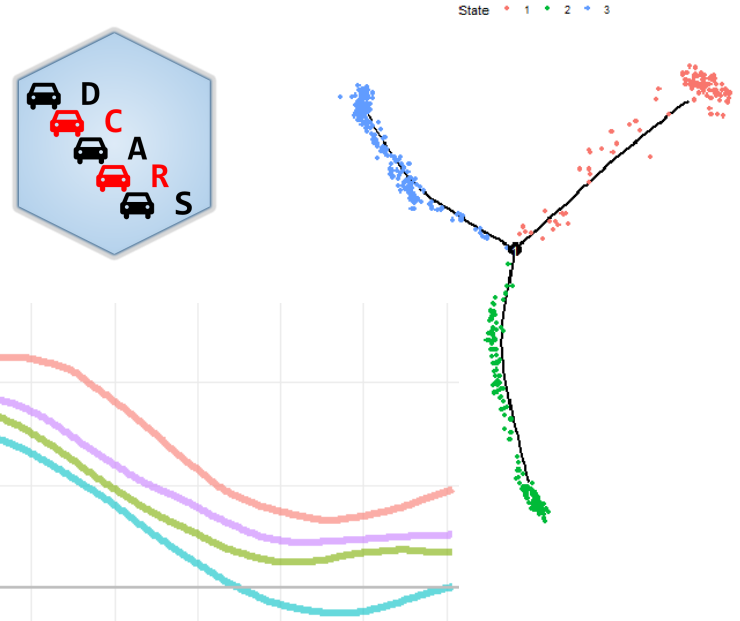
Find out more: http://www.maths.usyd.edu.au/bioinformatics/

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



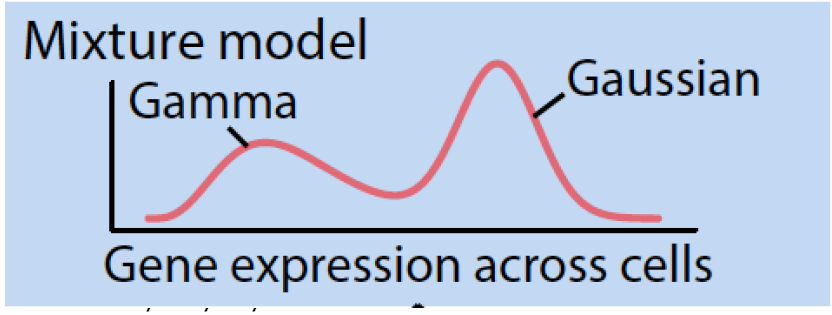
Clustering metrics

Sydney
Single cell

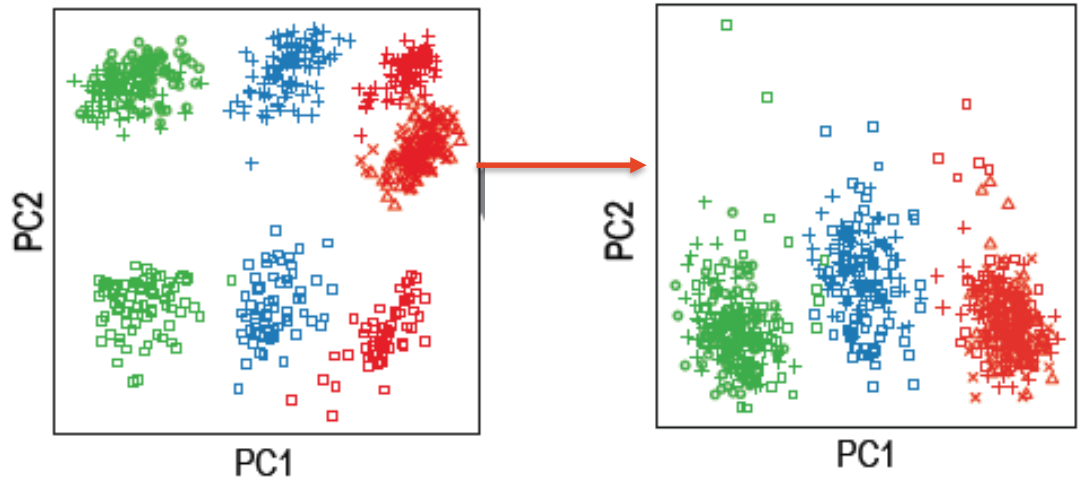


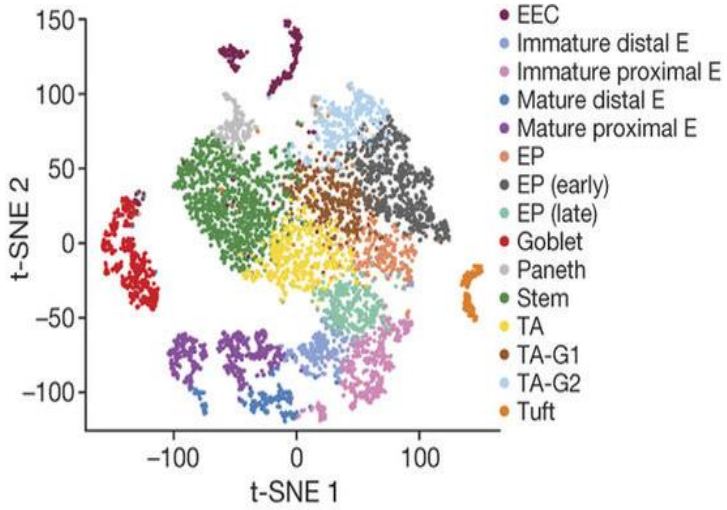
Differential correlation

Finding stably expressed genes



scMerge



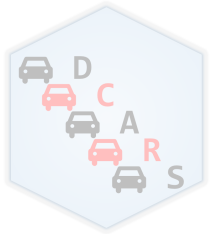


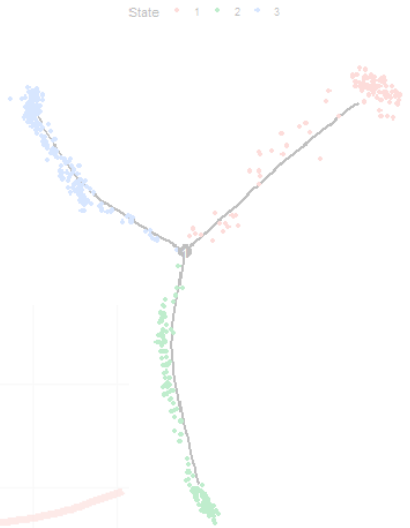
Clustering metrics

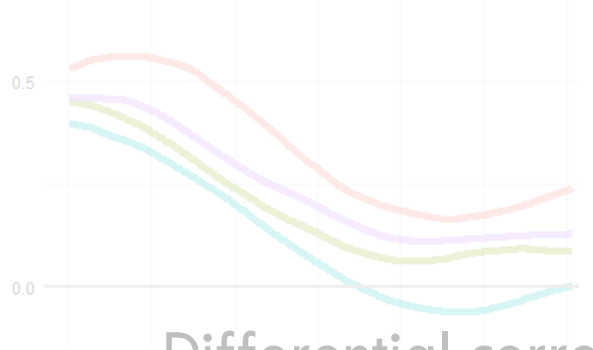
Pengyi Yang

Taiyun Kim

Sydney
Single cell

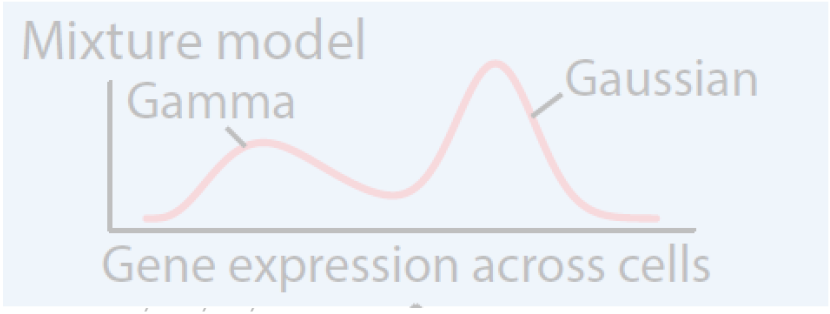




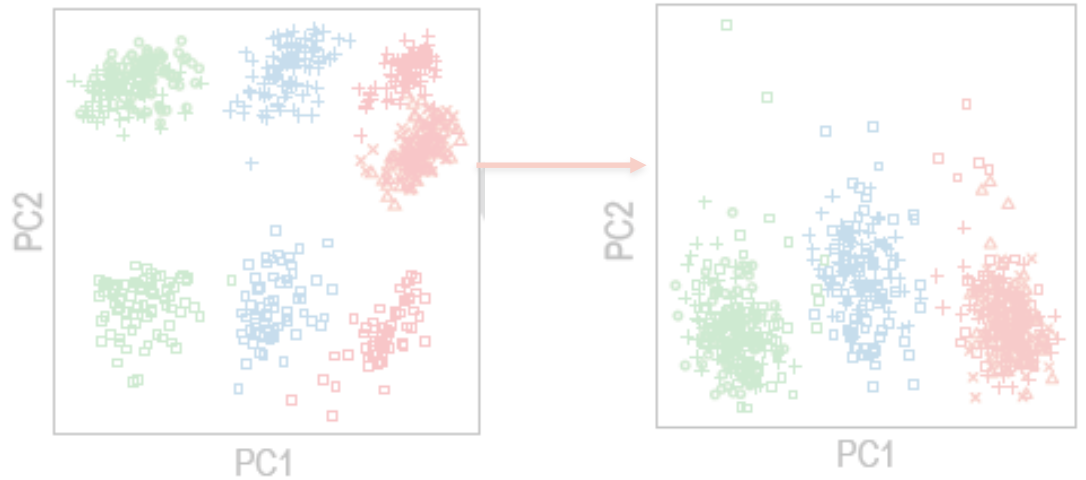


Differential correlation

Finding stably expressed genes



scMerge



Key underlying aspect of clustering? Similarity metrics

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

Key underlying aspect of clustering? Similarity metrics

- *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

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

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

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

Key underlying aspect of clustering? Similarity metrics

- *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

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

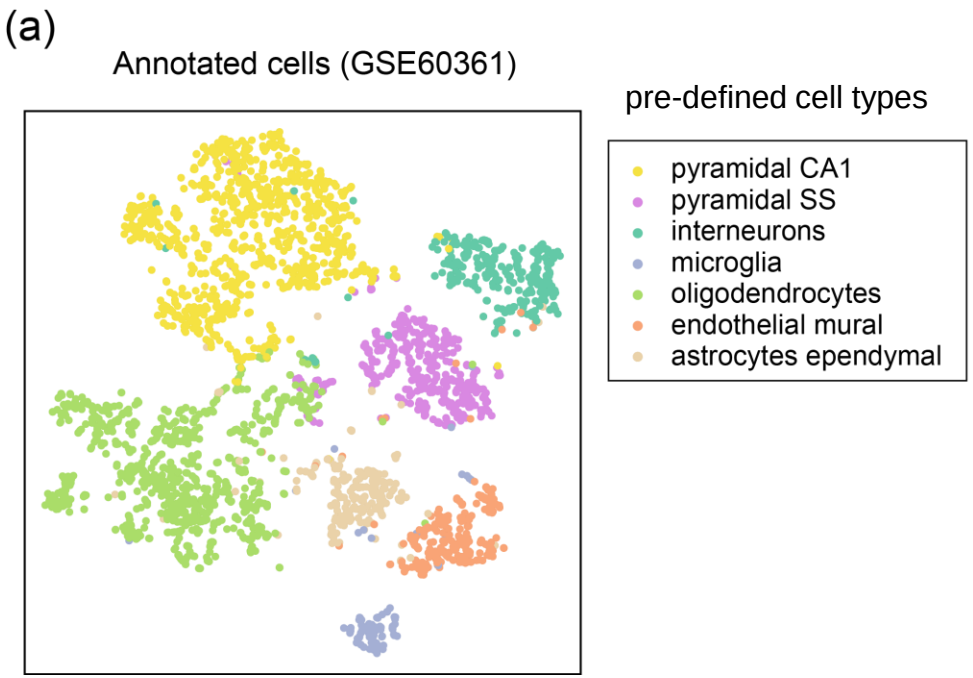
Spearman

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

Correlation-based

k-means Clustering on GSE60361

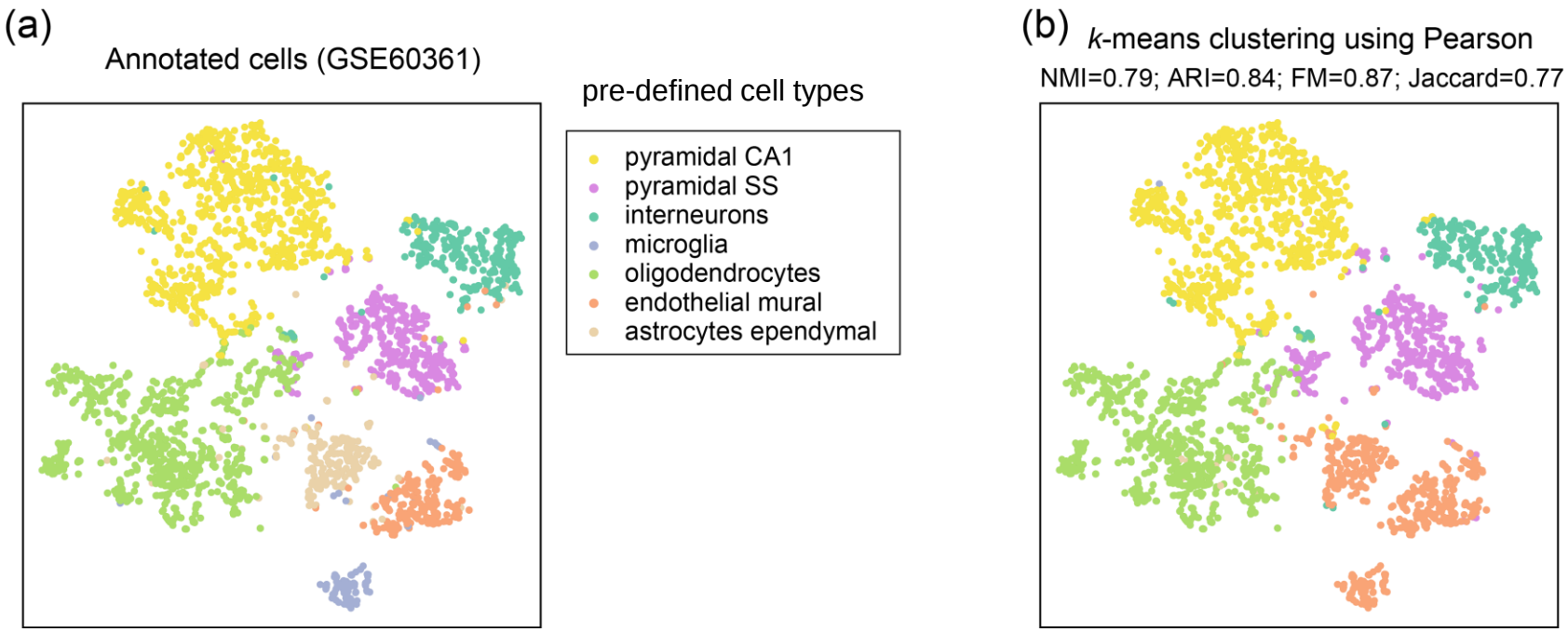
k-means



Zeisel A, et al. *Science* 2015

k-means Clustering on GSE60361

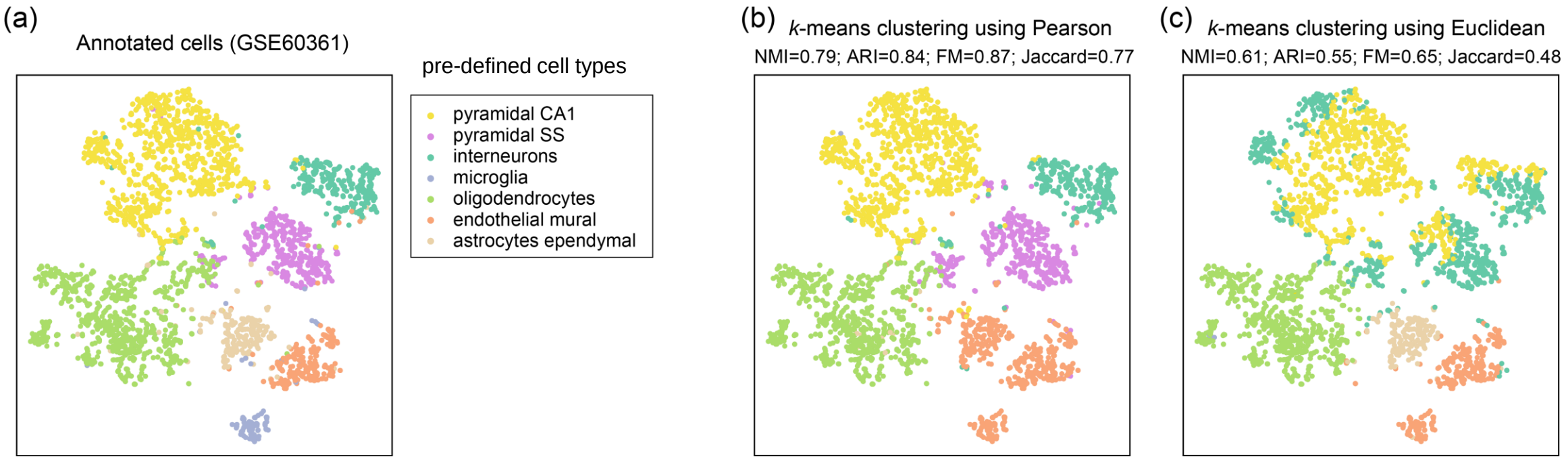
k-means



Zeisel A, et al. *Science* 2015

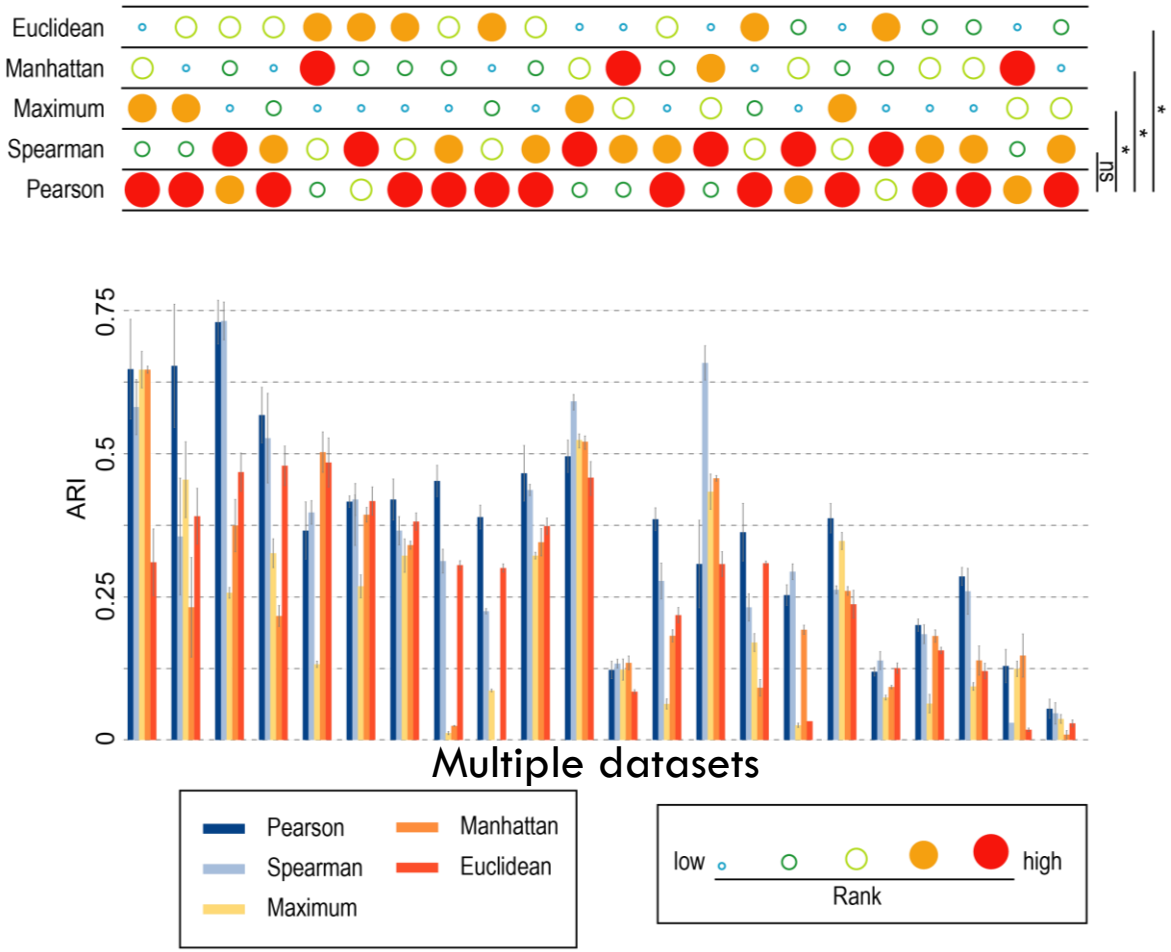
k-means Clustering on GSE60361

k-means

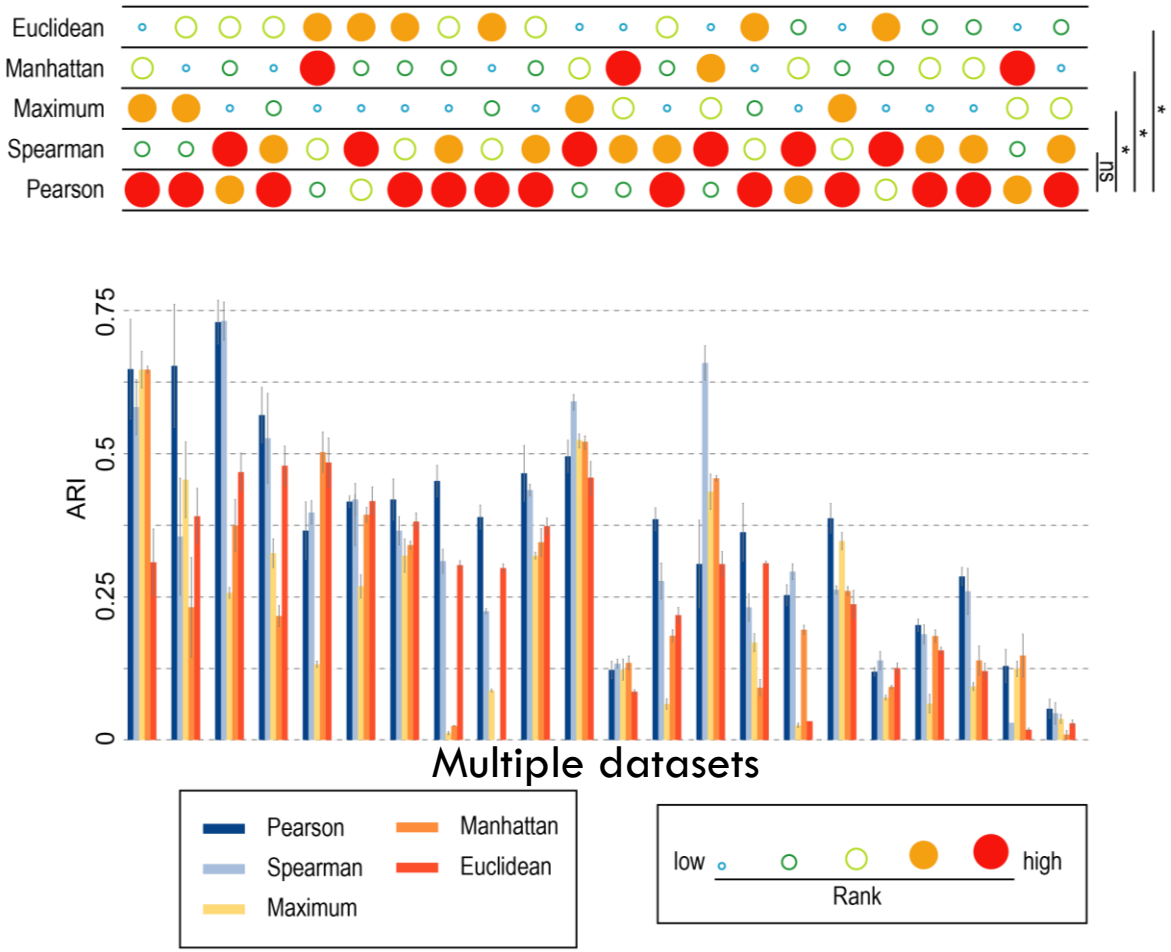


Zeisel A, et al. *Science* 2015

Evaluation results (against the pre-defined cell types)



Evaluation results (against the pre-defined cell types)



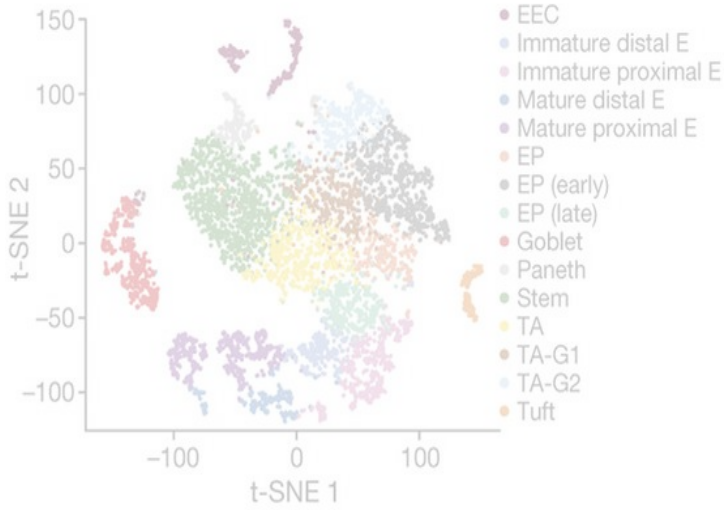
Impact of similarity metrics on single-cell RNA-seq data clustering

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

Briefings in Bioinformatics, bby076,

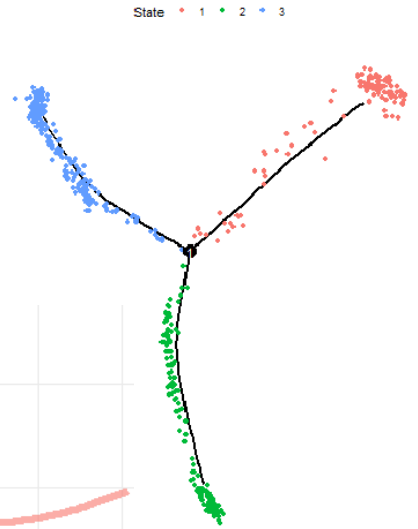
Pengyi Yang

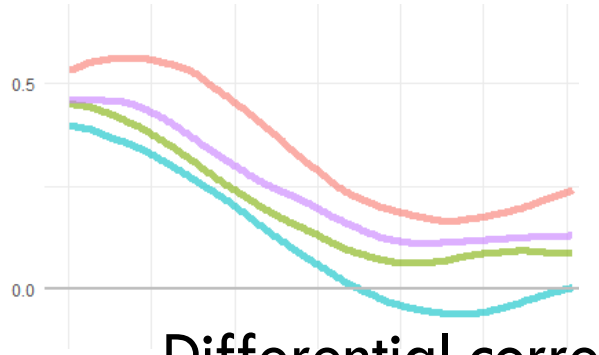
Taiyun Kim



Clustering metrics

Ellis Patrick

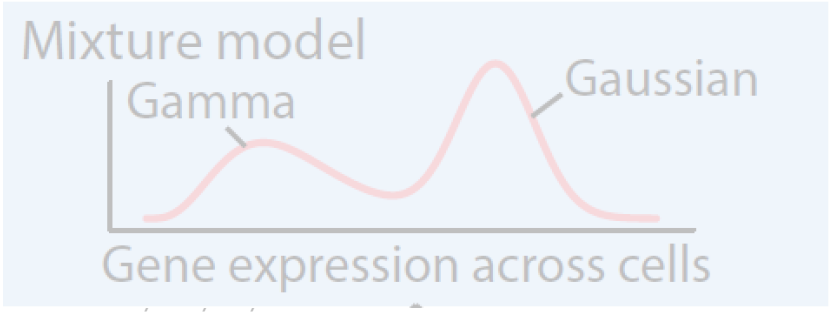




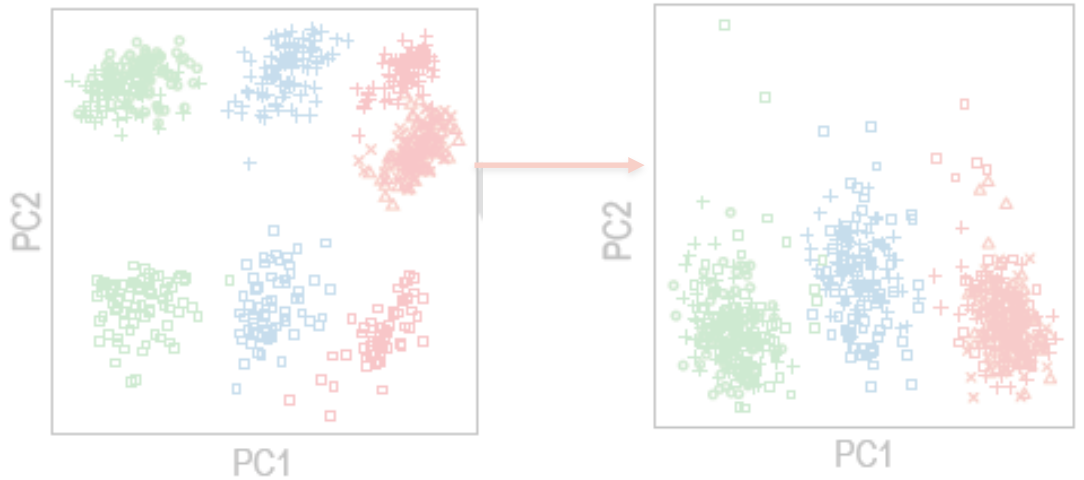
Differential correlation

Sydney
Single cell

Finding stably expressed genes

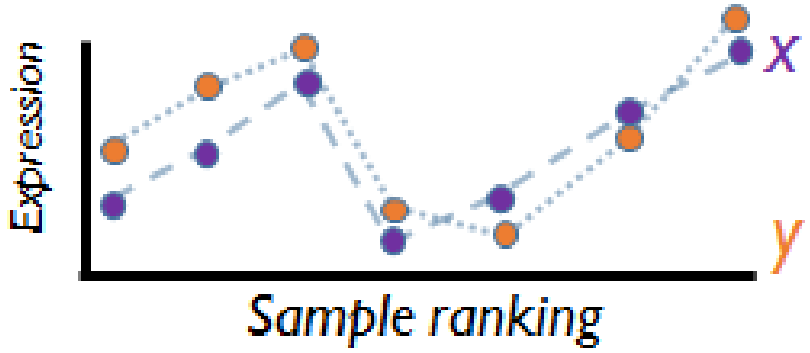


scMerge



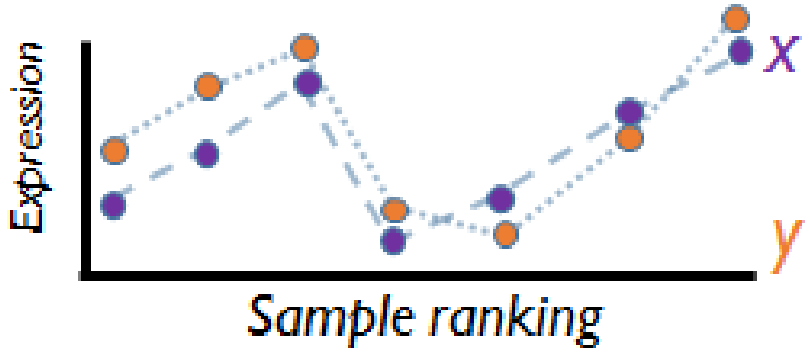
What is differential correlation?

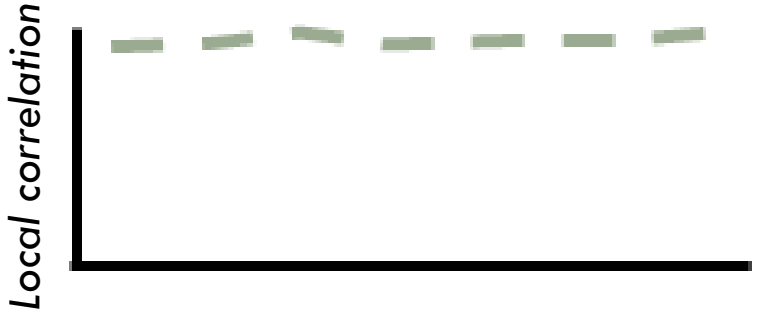
Consider the expression of gene x and gene y



What is differential correlation?

Consider the expression of gene x and gene y

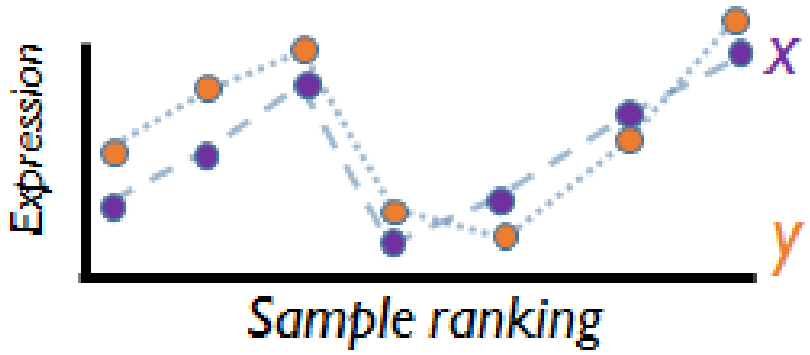




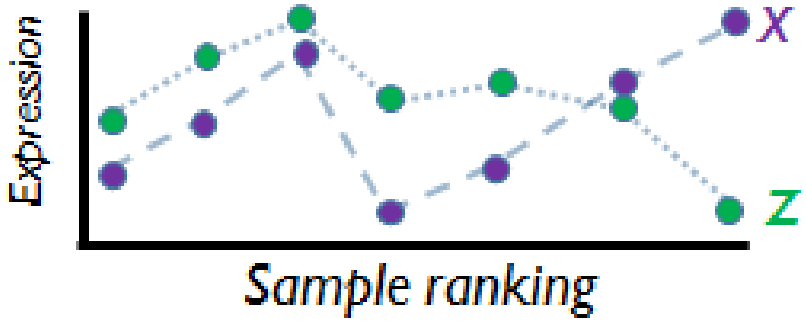
Constant correlation

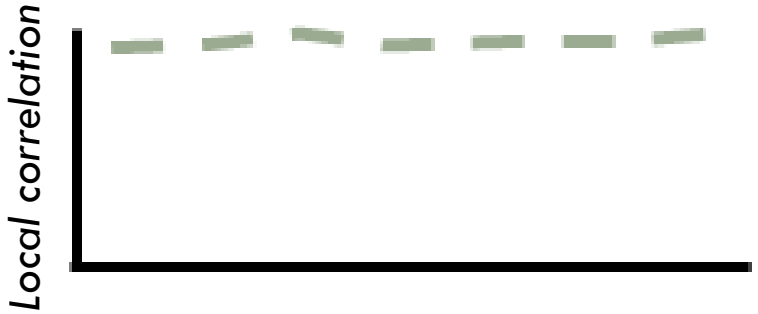
What is differential correlation?

Consider the expression of gene x and gene y



Now consider the expression of gene x and gene z

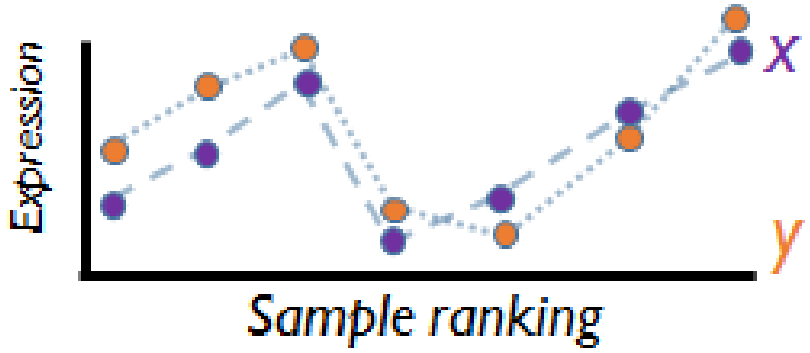


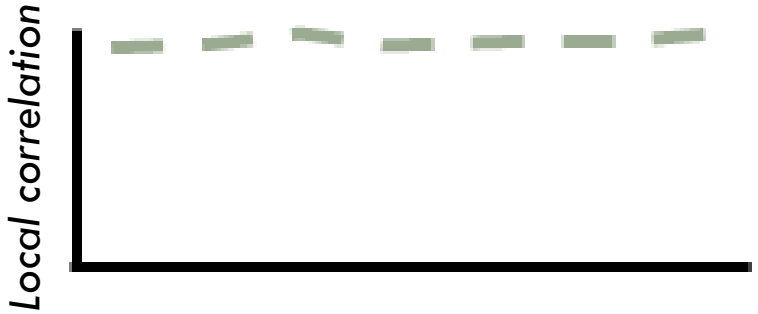


Constant correlation

What is differential correlation?

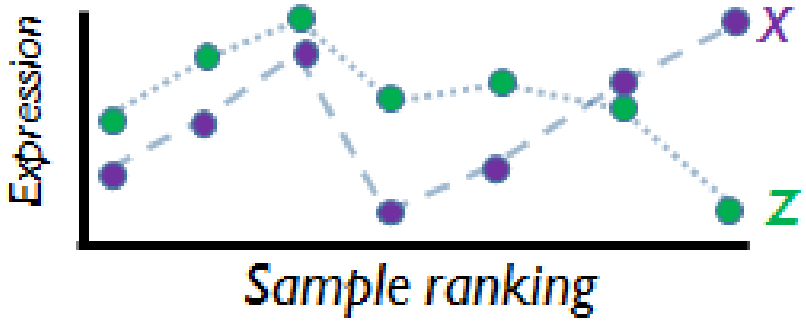
Consider the expression of gene x and gene y

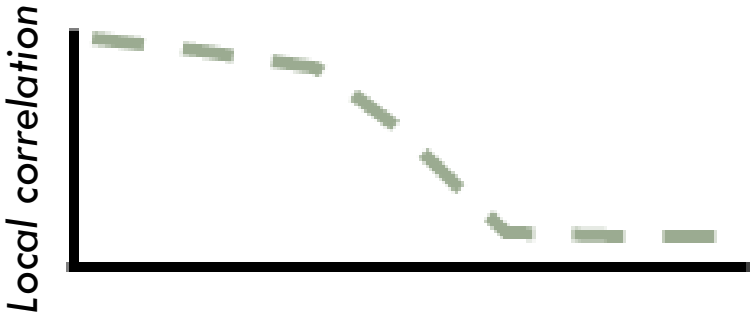




Constant correlation

Now consider the expression of gene x and gene z





Differential correlation

Bioinformatics, 2018, 1–7
doi: 10.1093/bioinformatics/bty698
Advance Access Publication Date: 9 August 2018
Original Paper

OXFORD

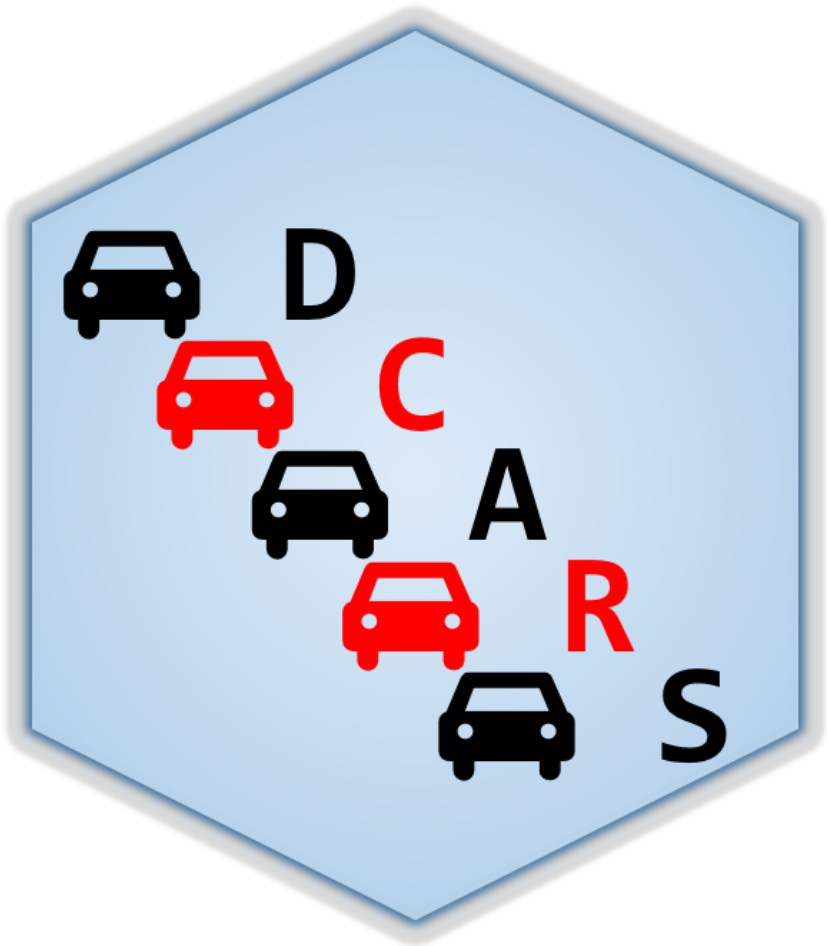
Systems biology

DCARS: differential correlation across ranked samples

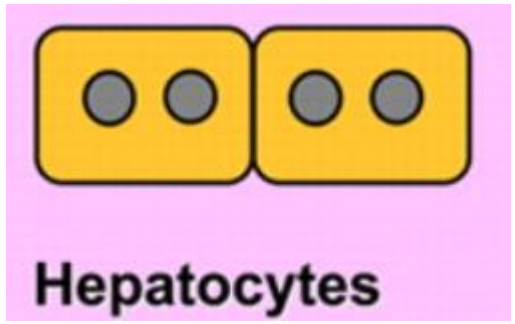
Shila Ghazanfar [1,2,*], Dario Strbenac[2], John T. Ormerod[2,3], Jean Y. H. Yang[2,1] and Ellis Patrick [2,4]

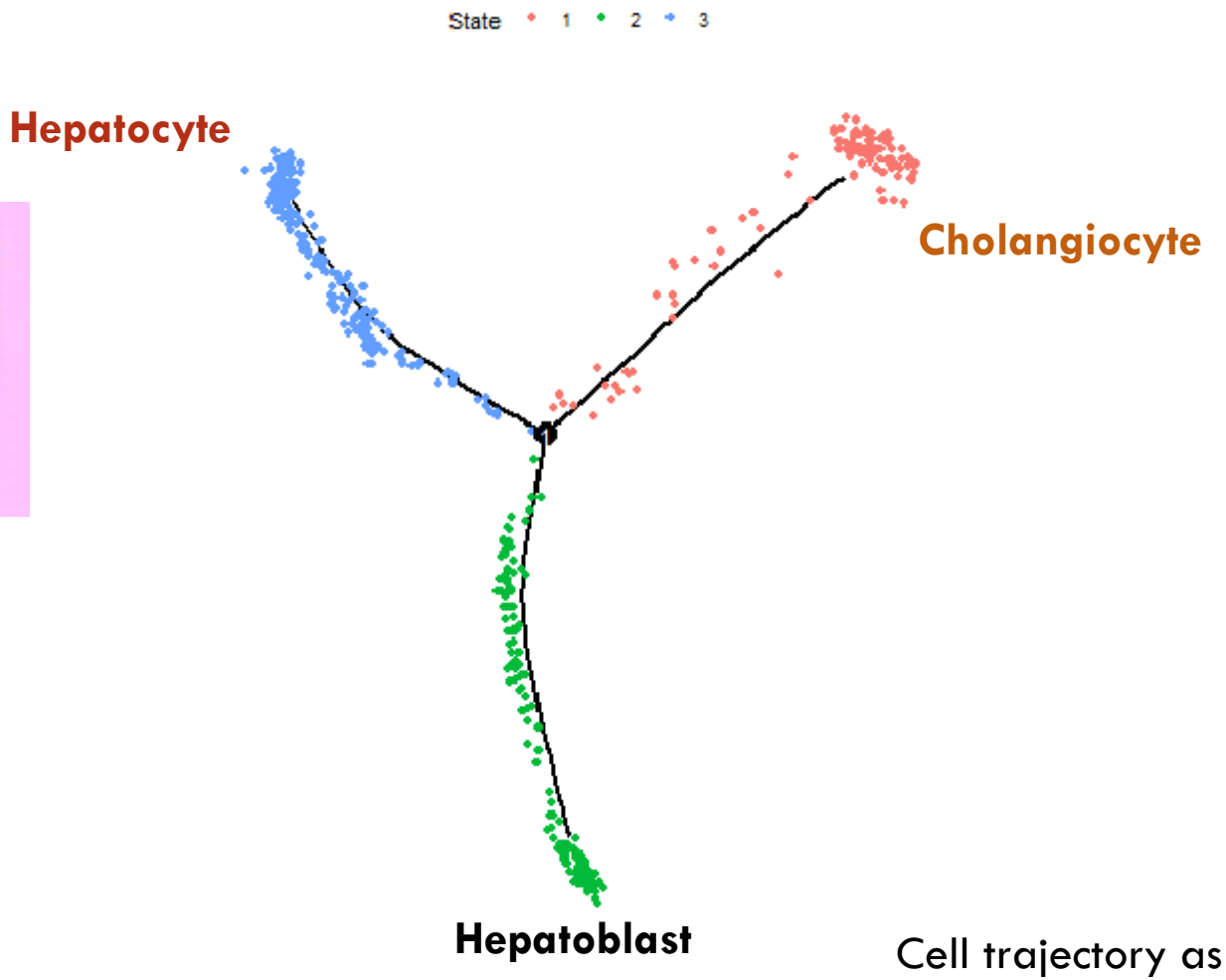
Installation

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



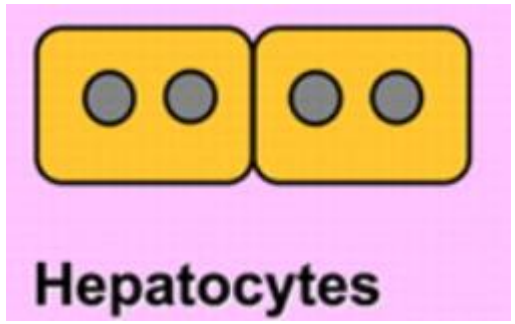
Differential correlation across pseudotime

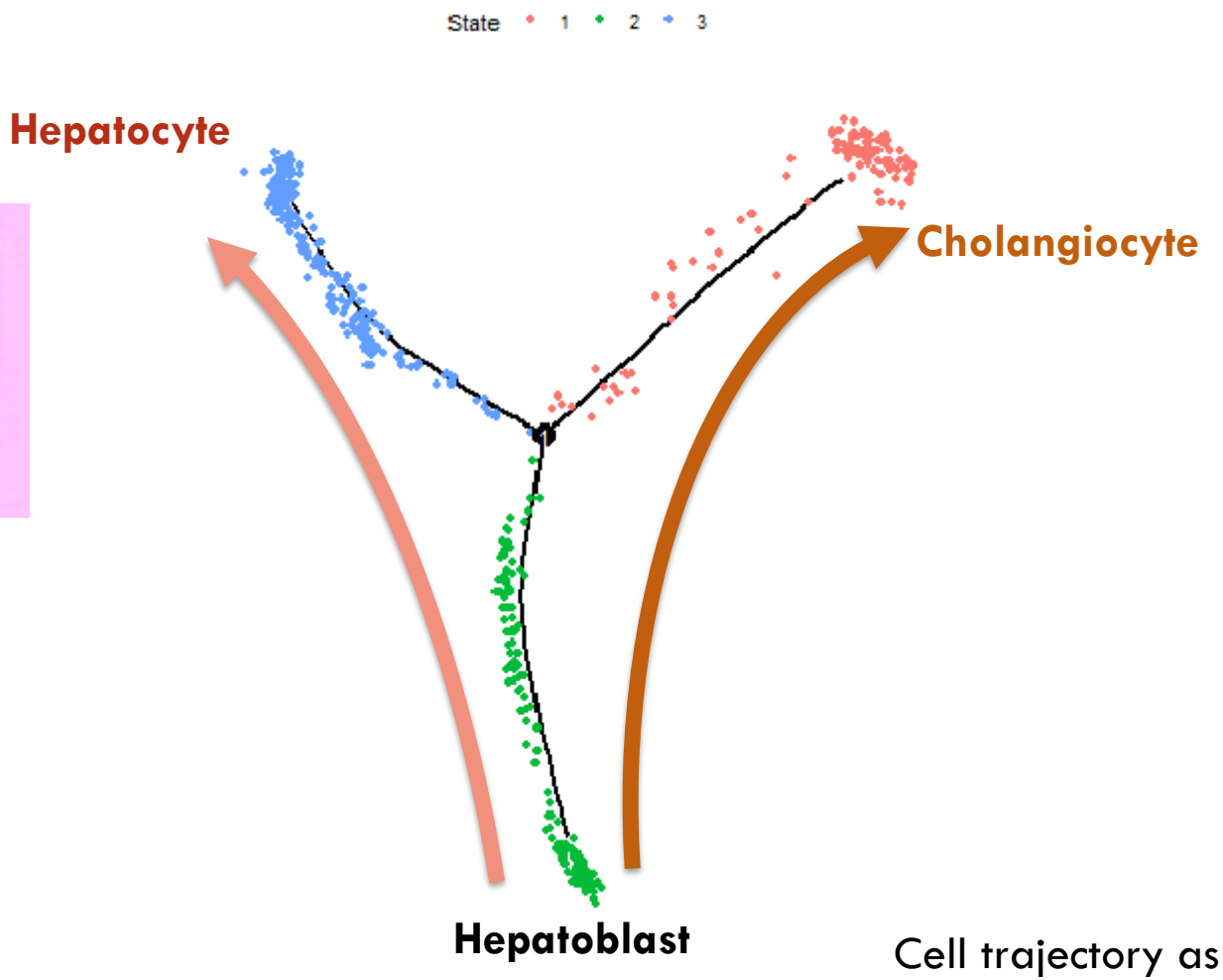




Cell trajectory as estimated using Monocle 2

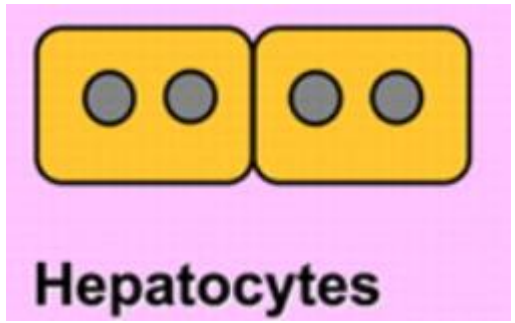
Differential correlation across pseudotime

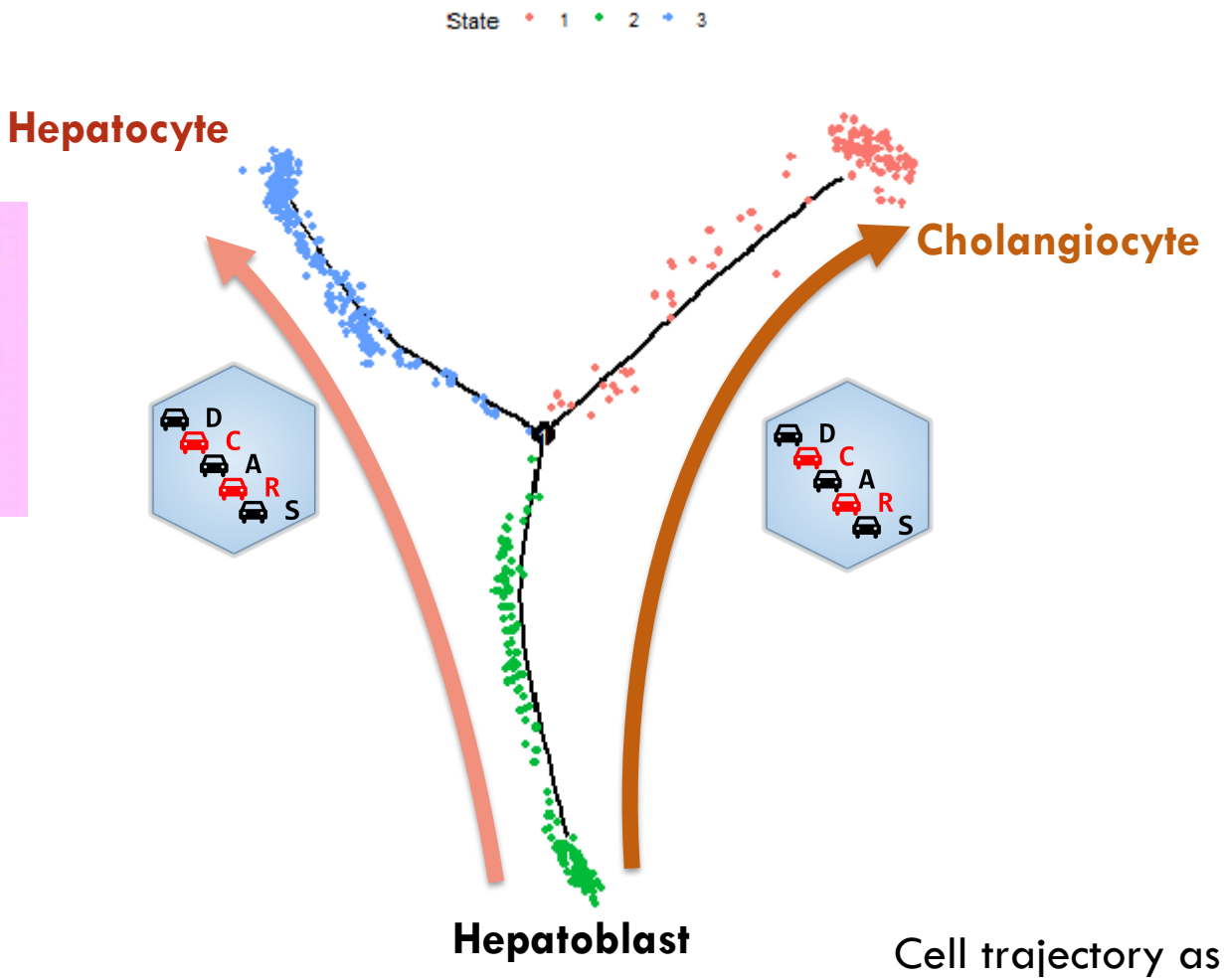




Cell trajectory as estimated using Monocle 2

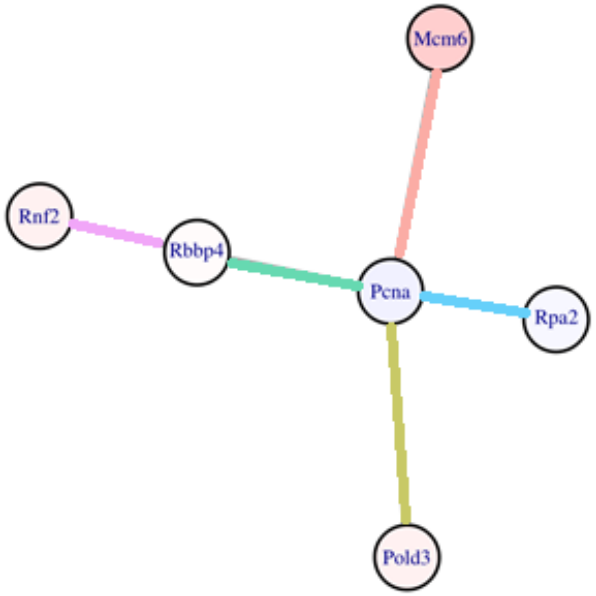
Differential correlation across pseudotime

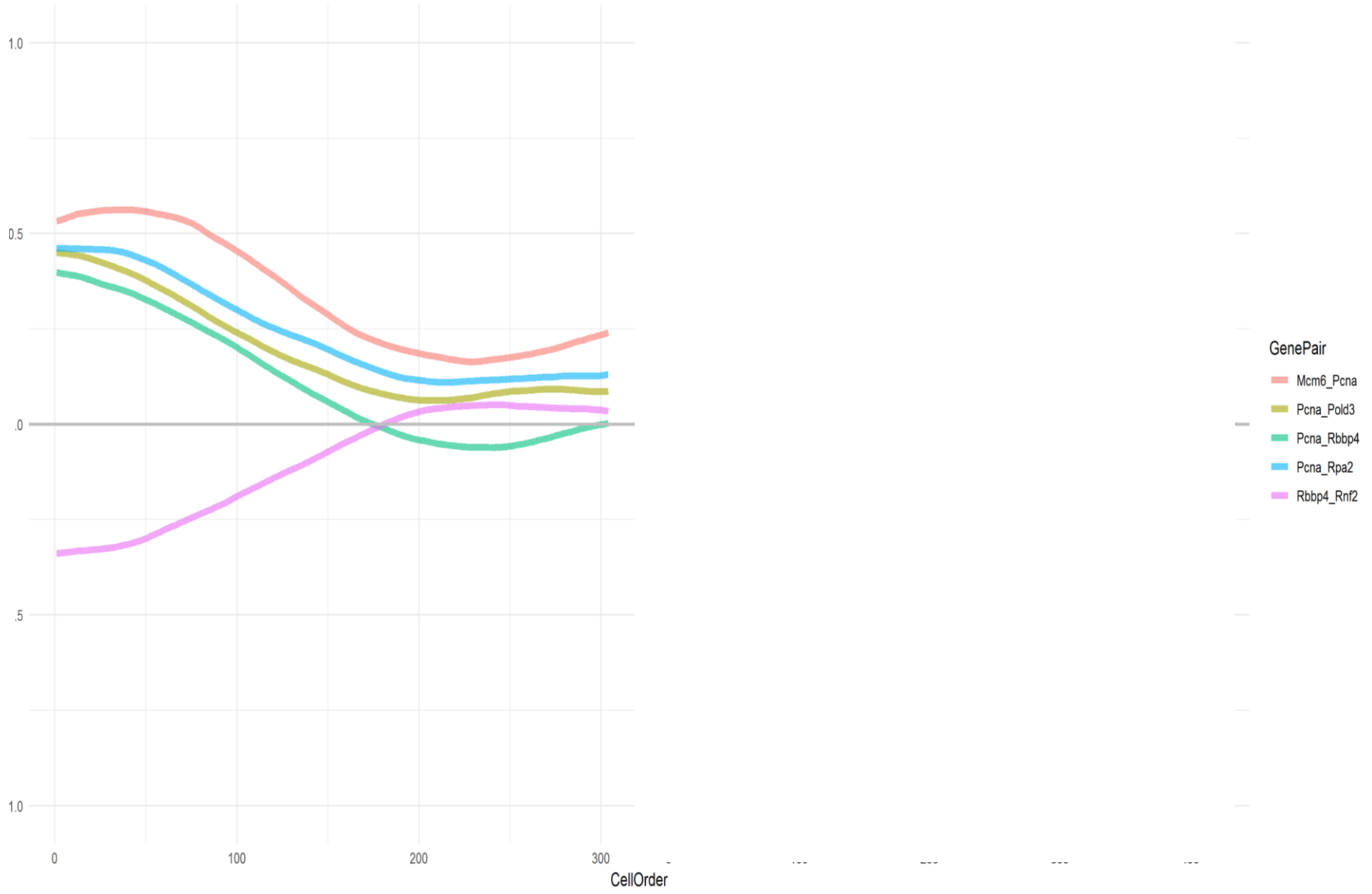




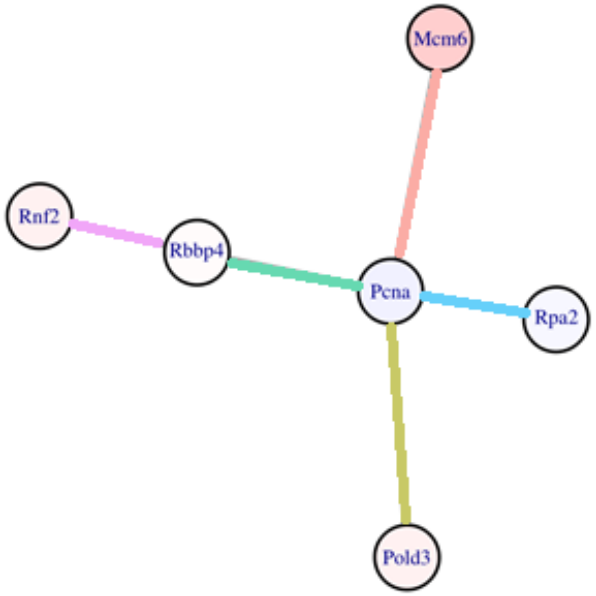
Cell trajectory as estimated using Monocle 2

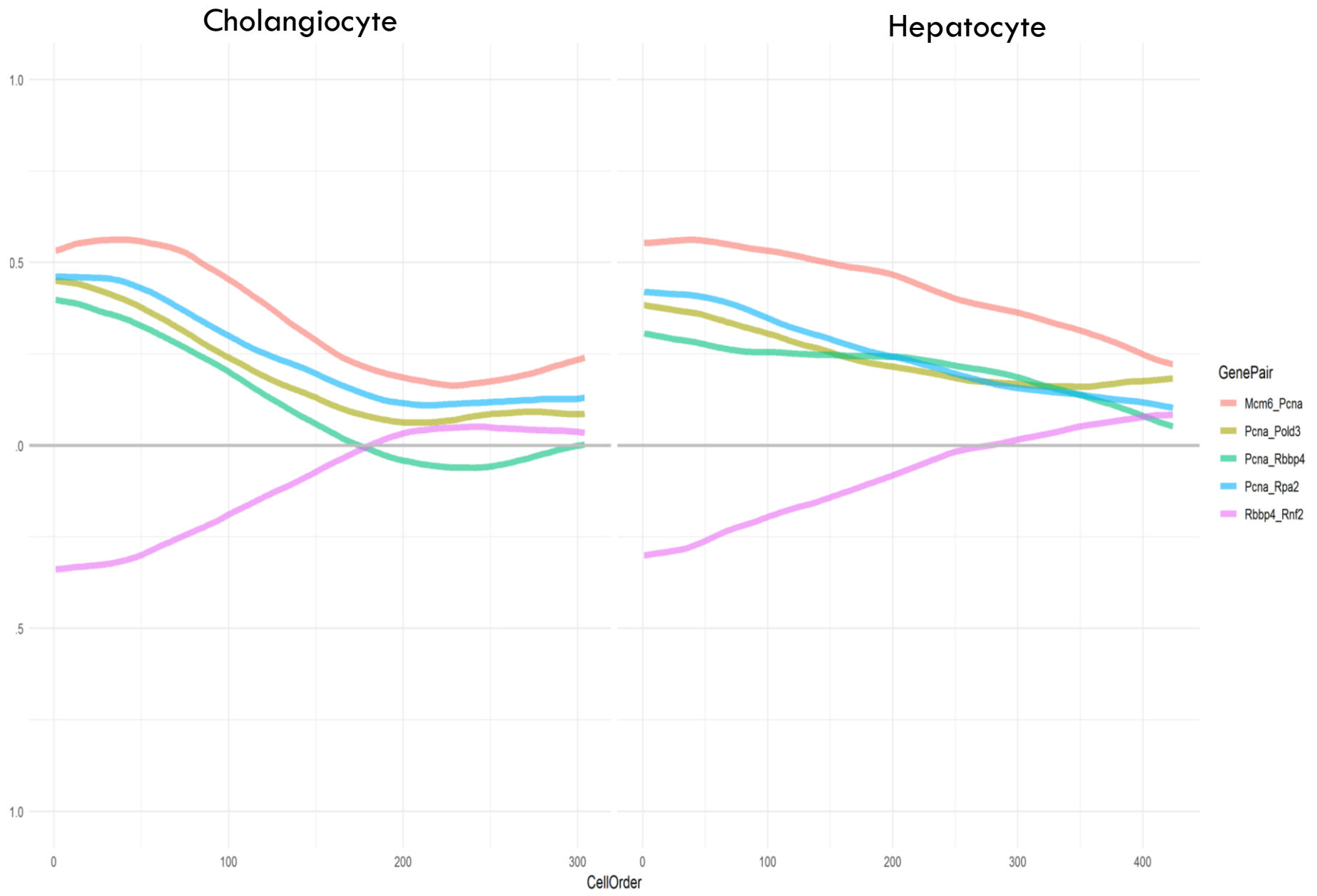
Pcna: DNA replication, cell cycle. Correlation lost during differentiation



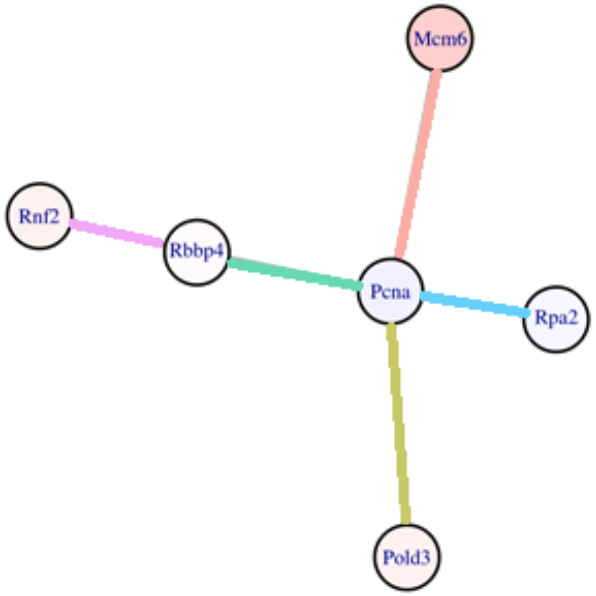


Pcna: DNA replication, cell cycle. Correlation lost during differentiation

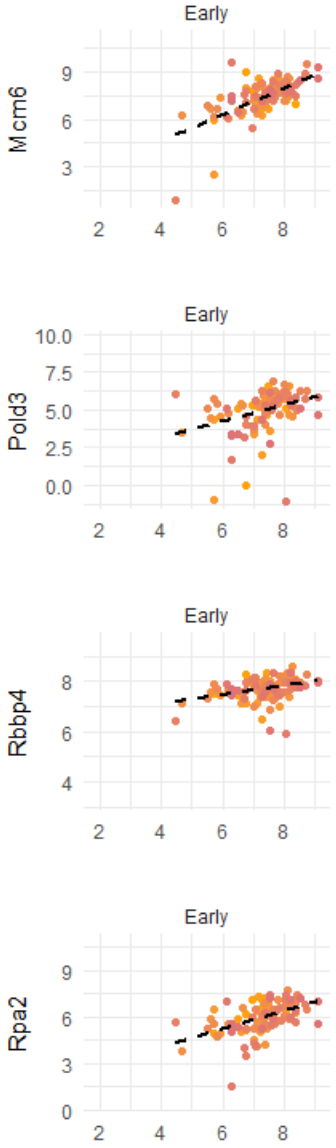




Pcna: DNA replication, cell cycle. Correlation lost during differentiation

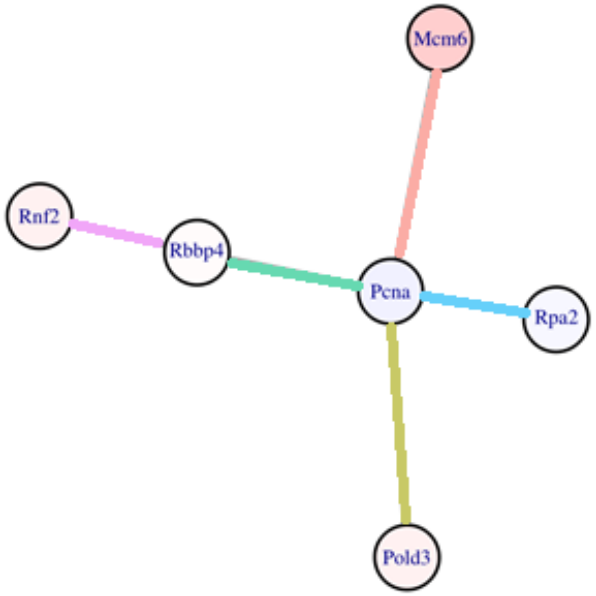


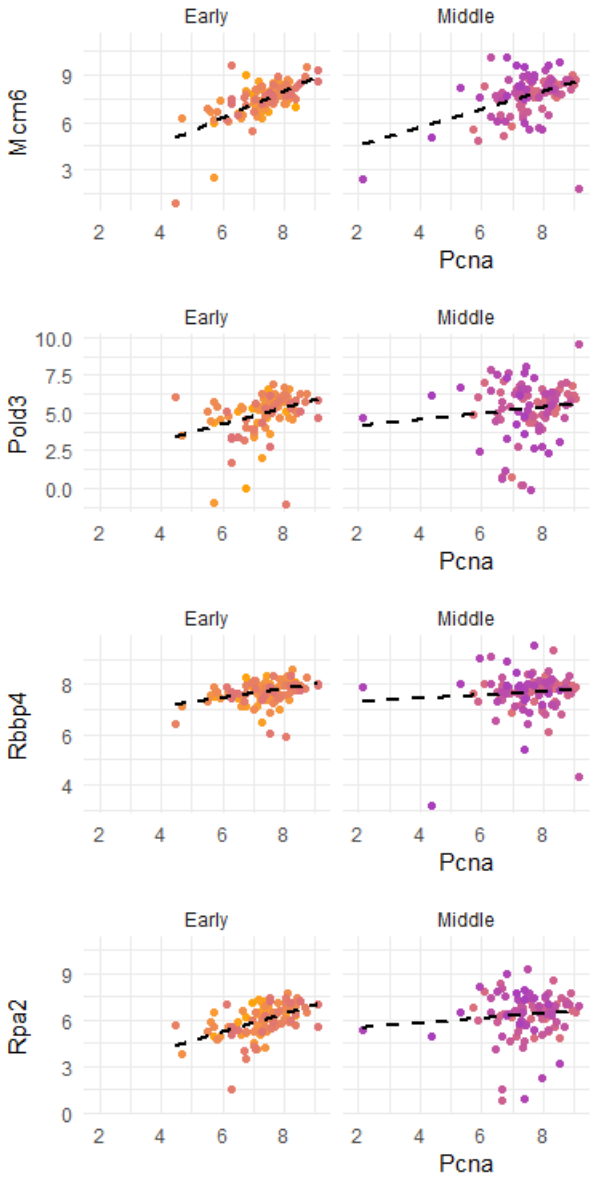
Cholangiocyte



Hepatocyte

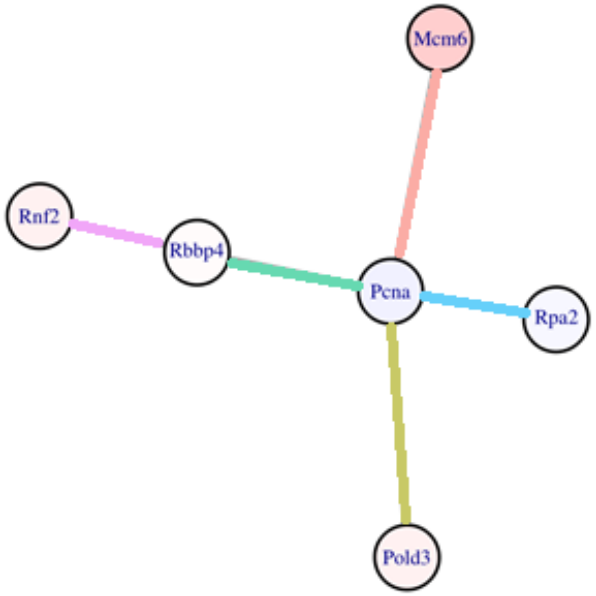
Pcna: DNA replication, cell cycle. Correlation lost during differentiation





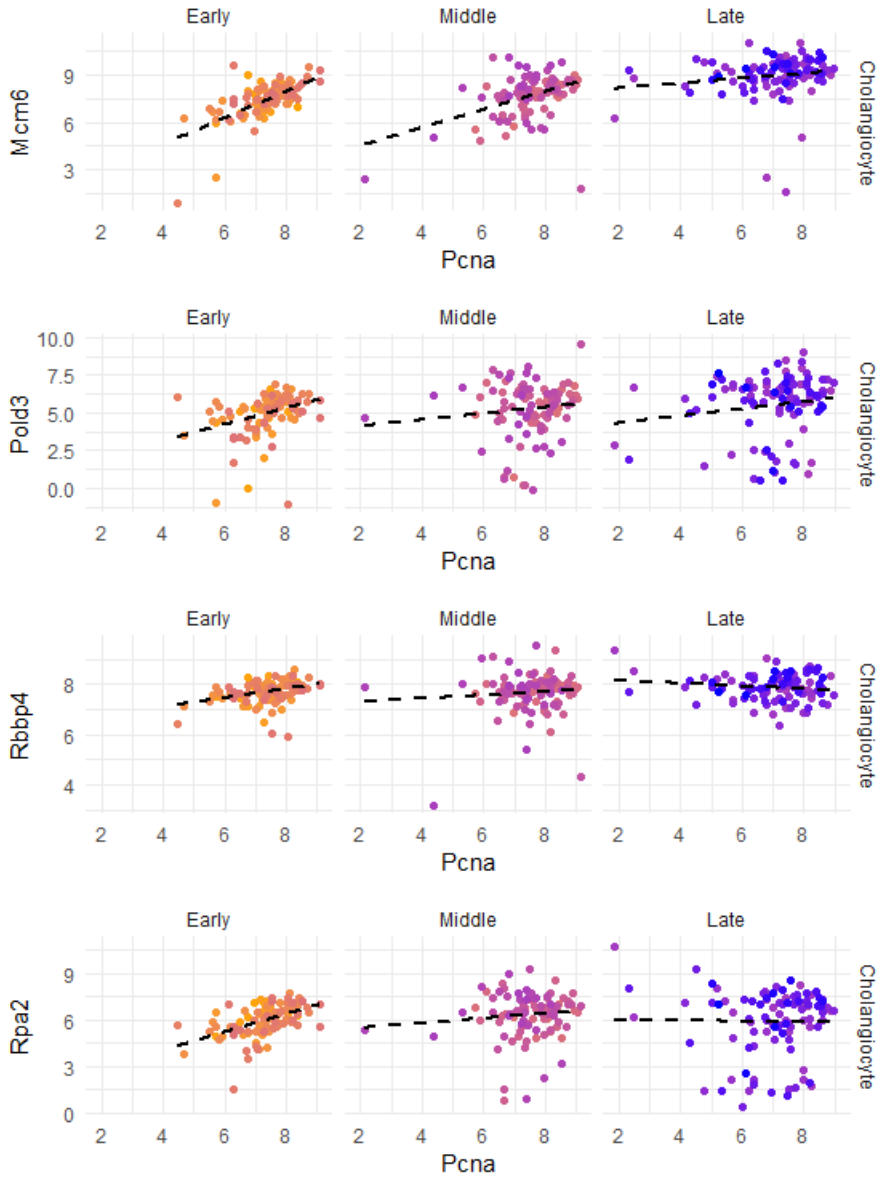
Hepatocyte

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

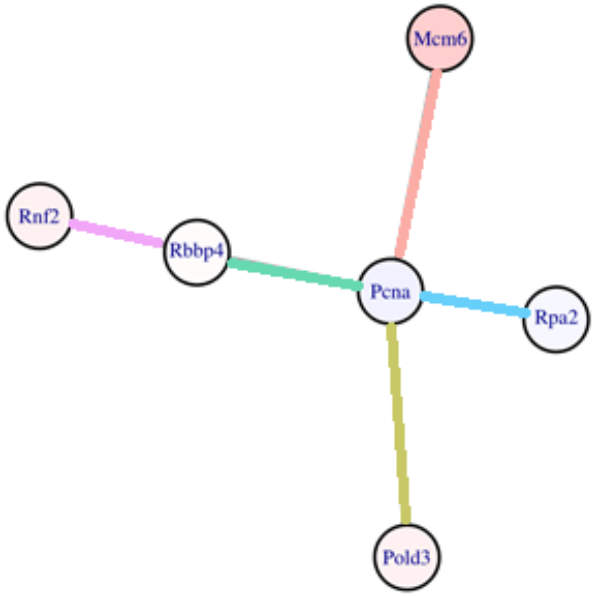


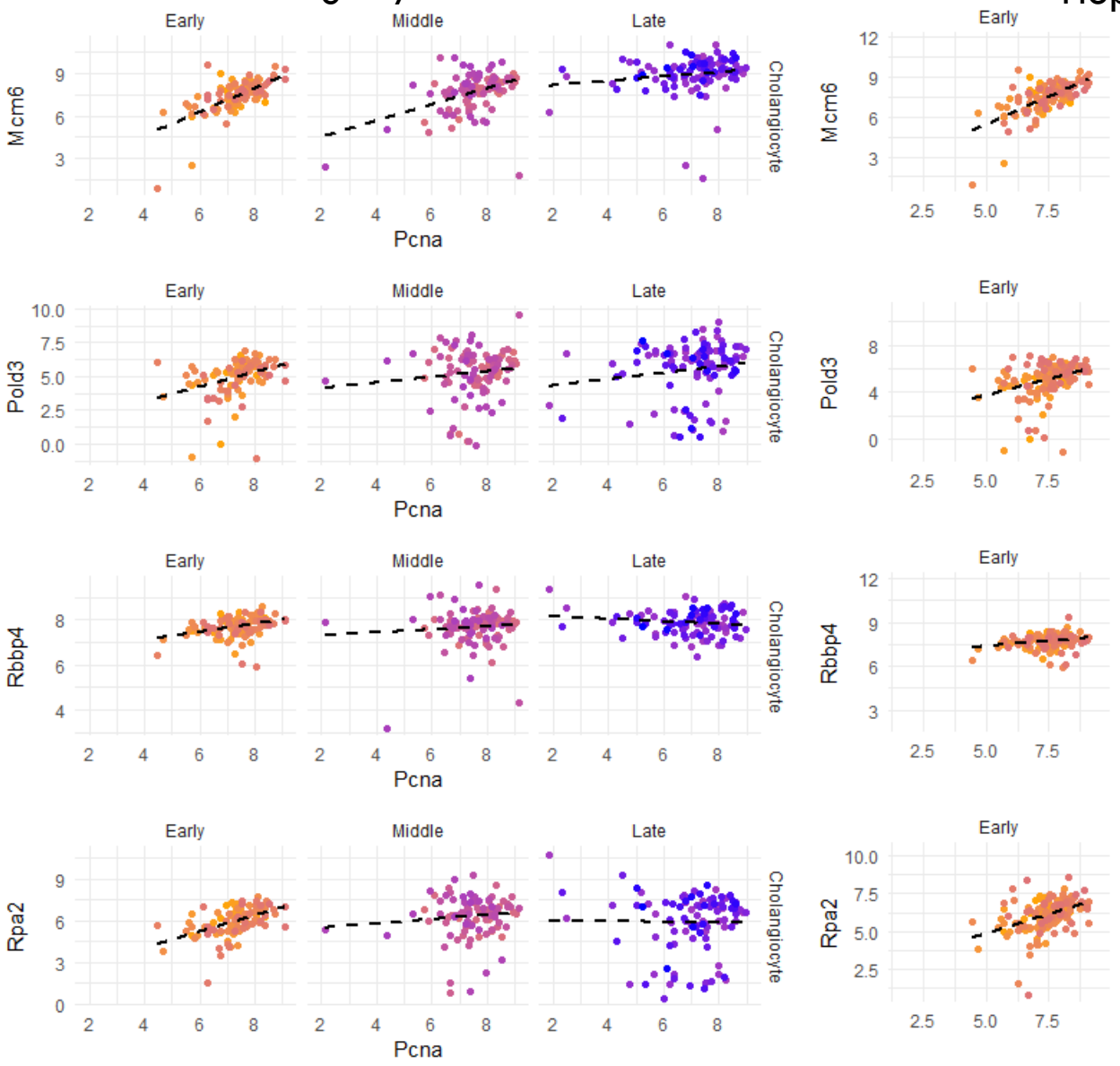
Cholangiocyte

Hepatocyte

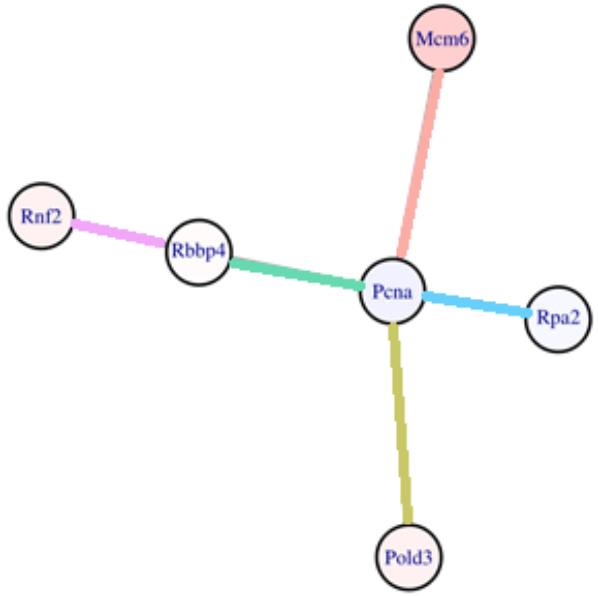


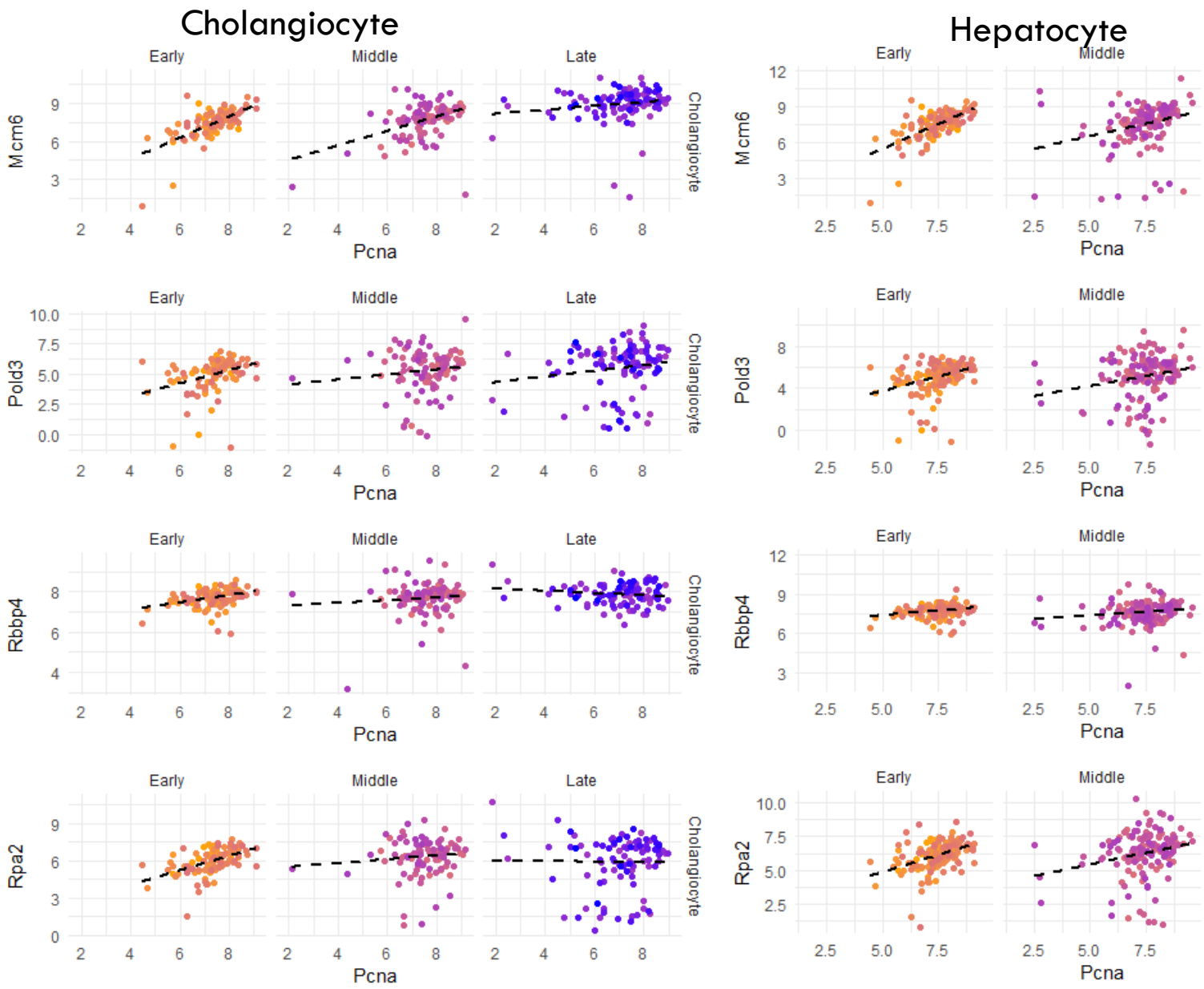
Pcna: DNA replication, cell cycle. Correlation lost during differentiation



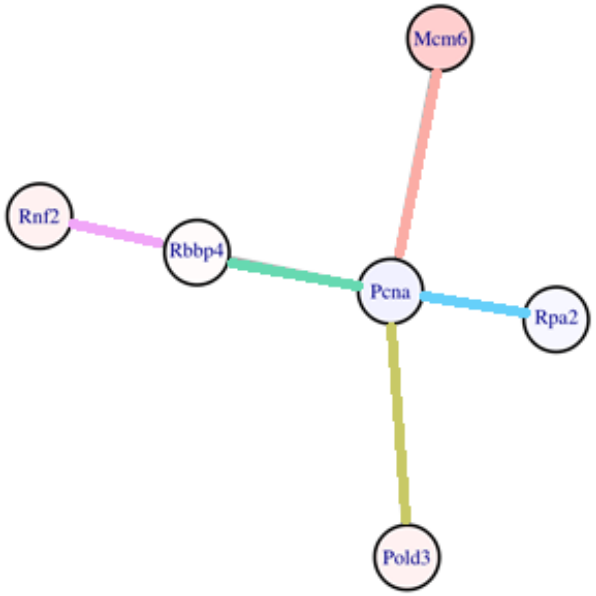


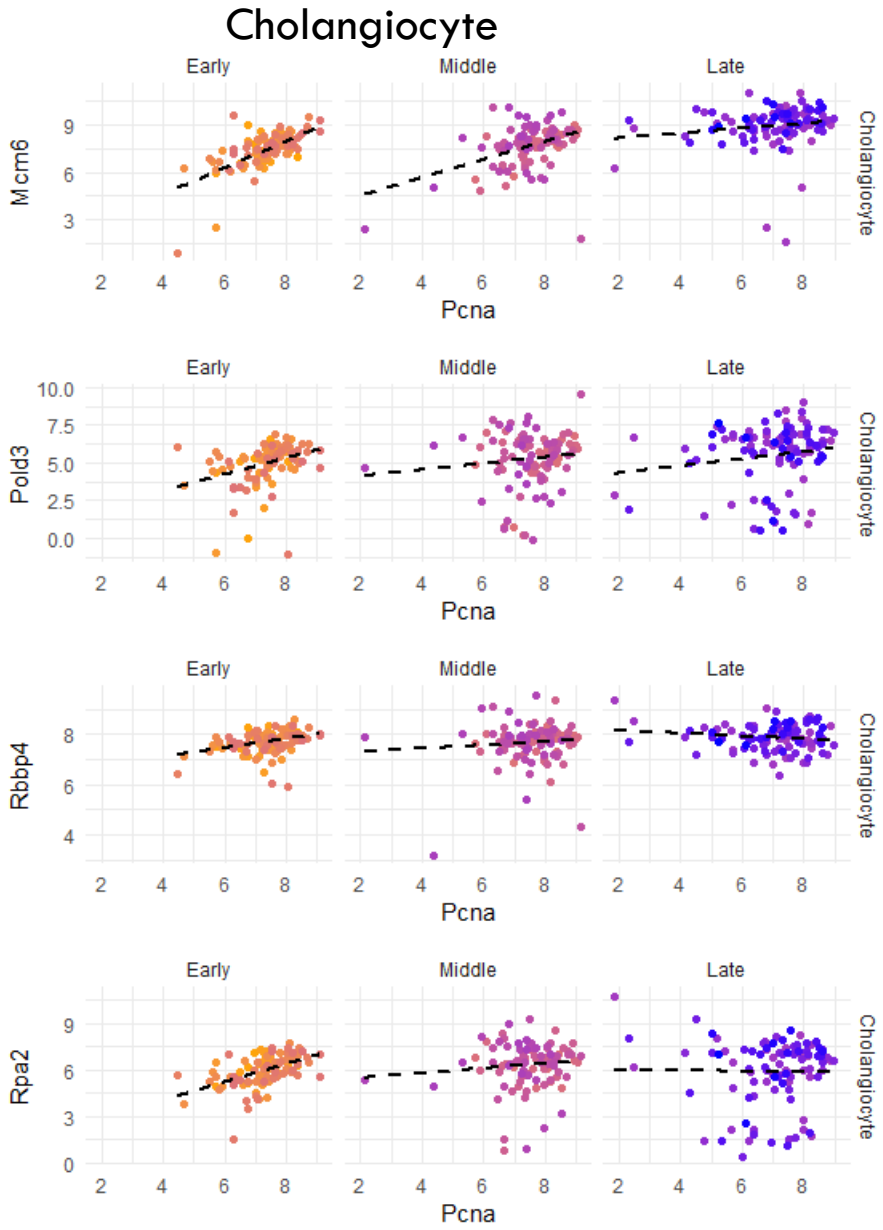
Pcna: DNA replication, cell cycle. Correlation lost during differentiation

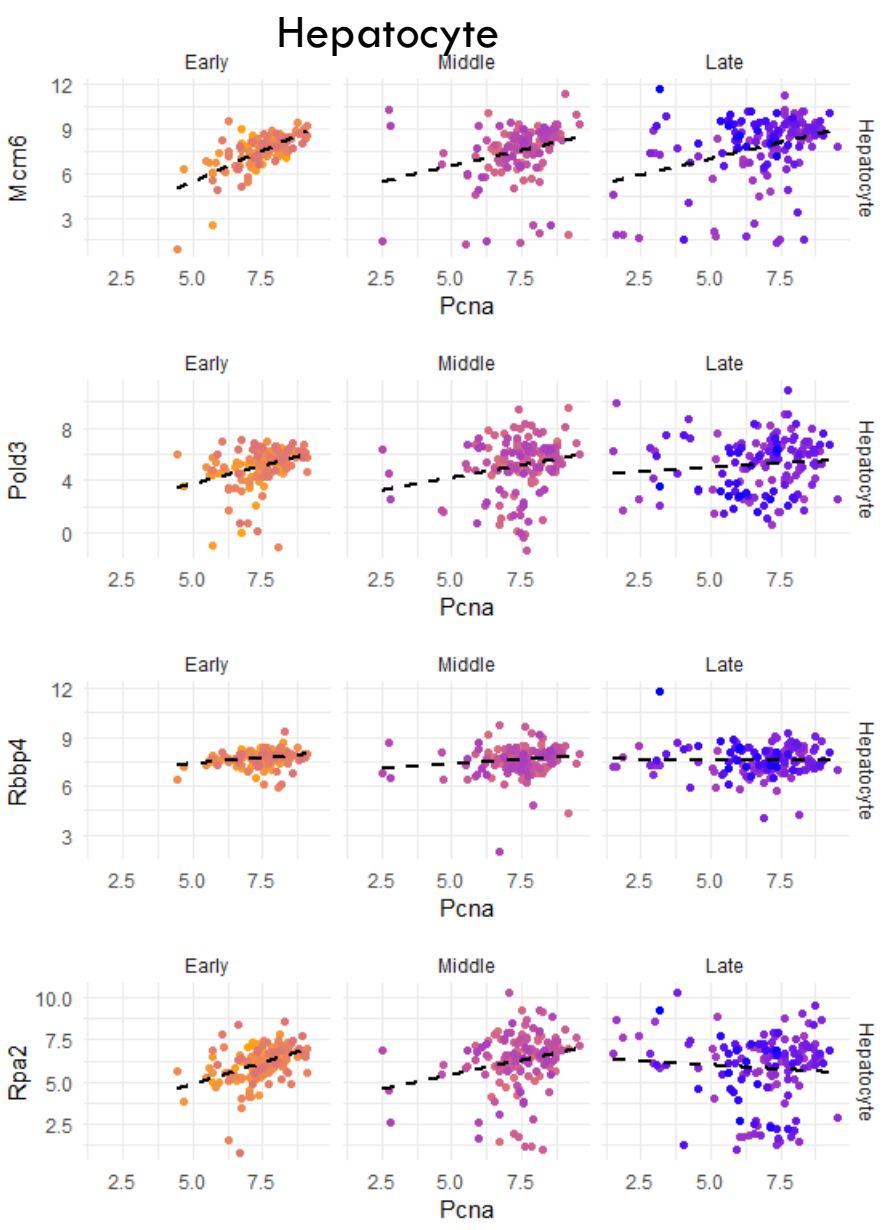


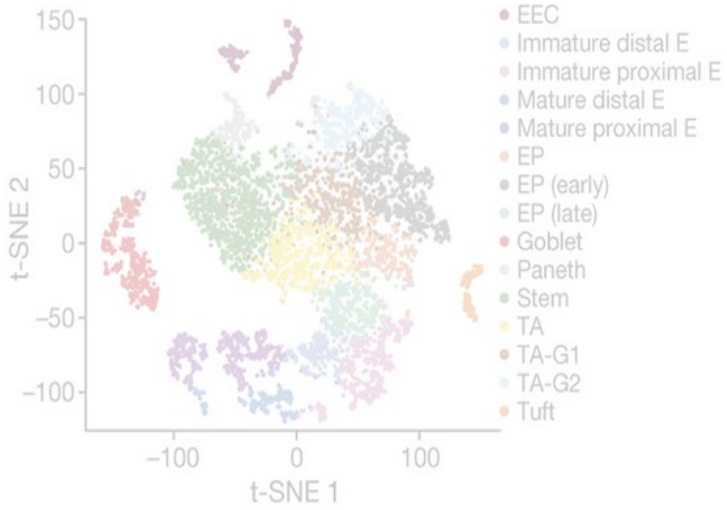


Pcna: DNA replication, cell cycle. Correlation lost during differentiation







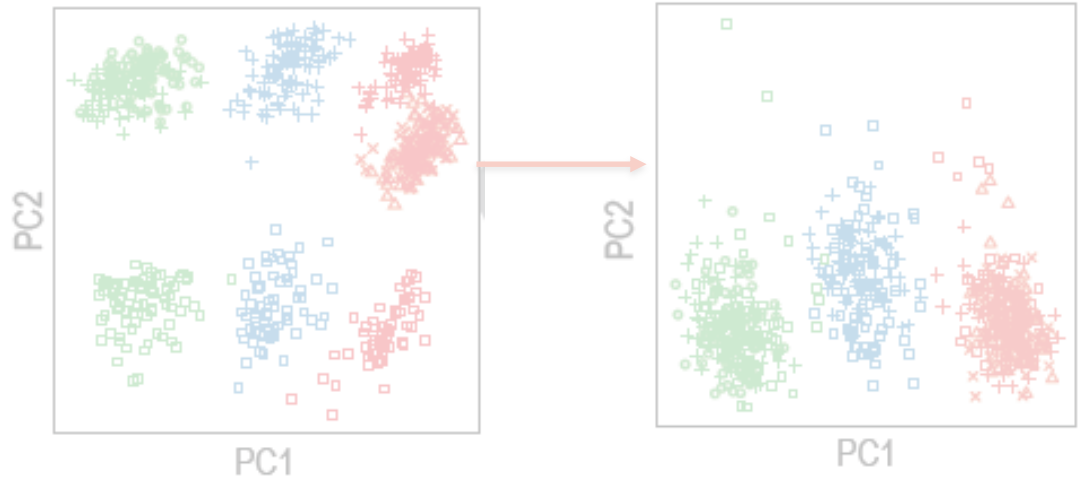


Clustering metrics

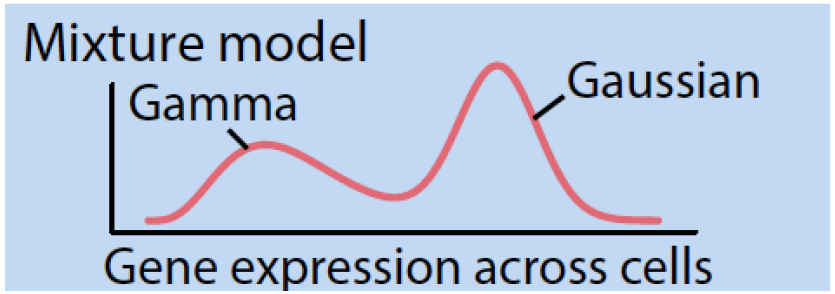
Differential correlation

Sydney
Single cell

scMerge

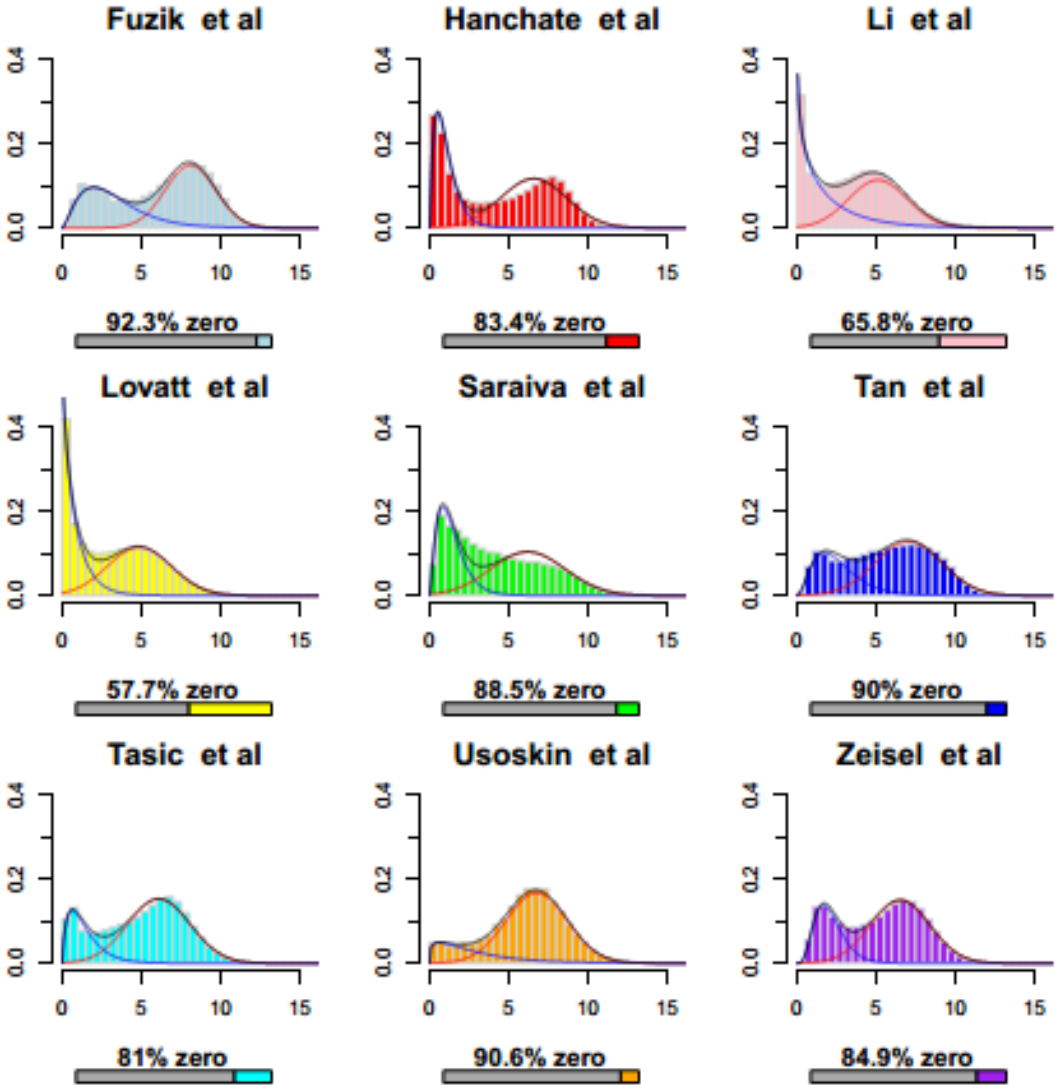


Finding stably expressed genes



Yingxin Lin

Previous work on mixture modelling

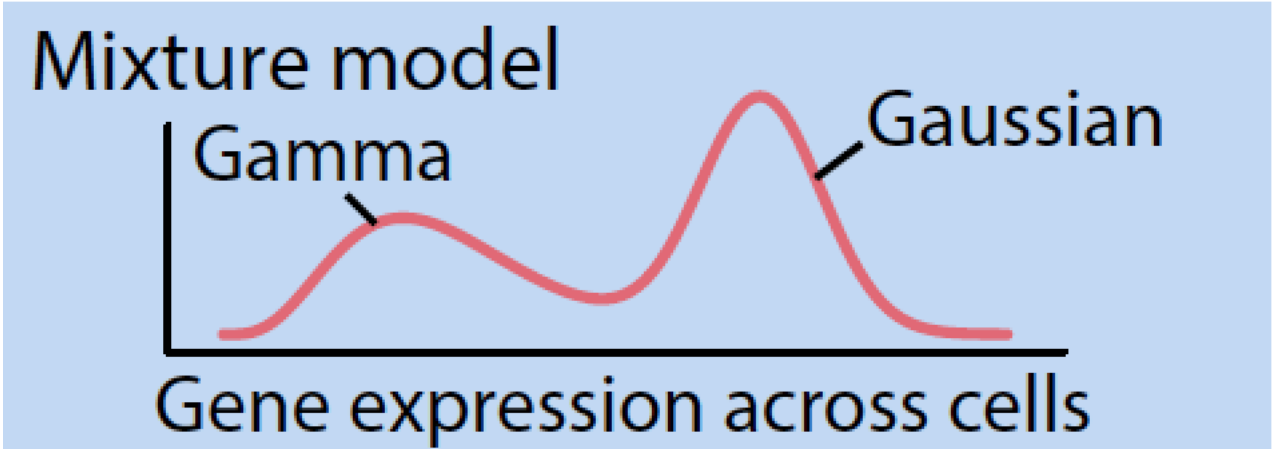


BioMed Central The Open Access Publisher | BMC Systems Biology | this article | search | submit a manuscript | register

BMC Syst Biol. 2016; 10(Suppl 5): 127.
Published online 2016 Dec 5. doi: 10.1186/s12918-016-0370-4
PMCID: PMC5249008
PMID: 28105940

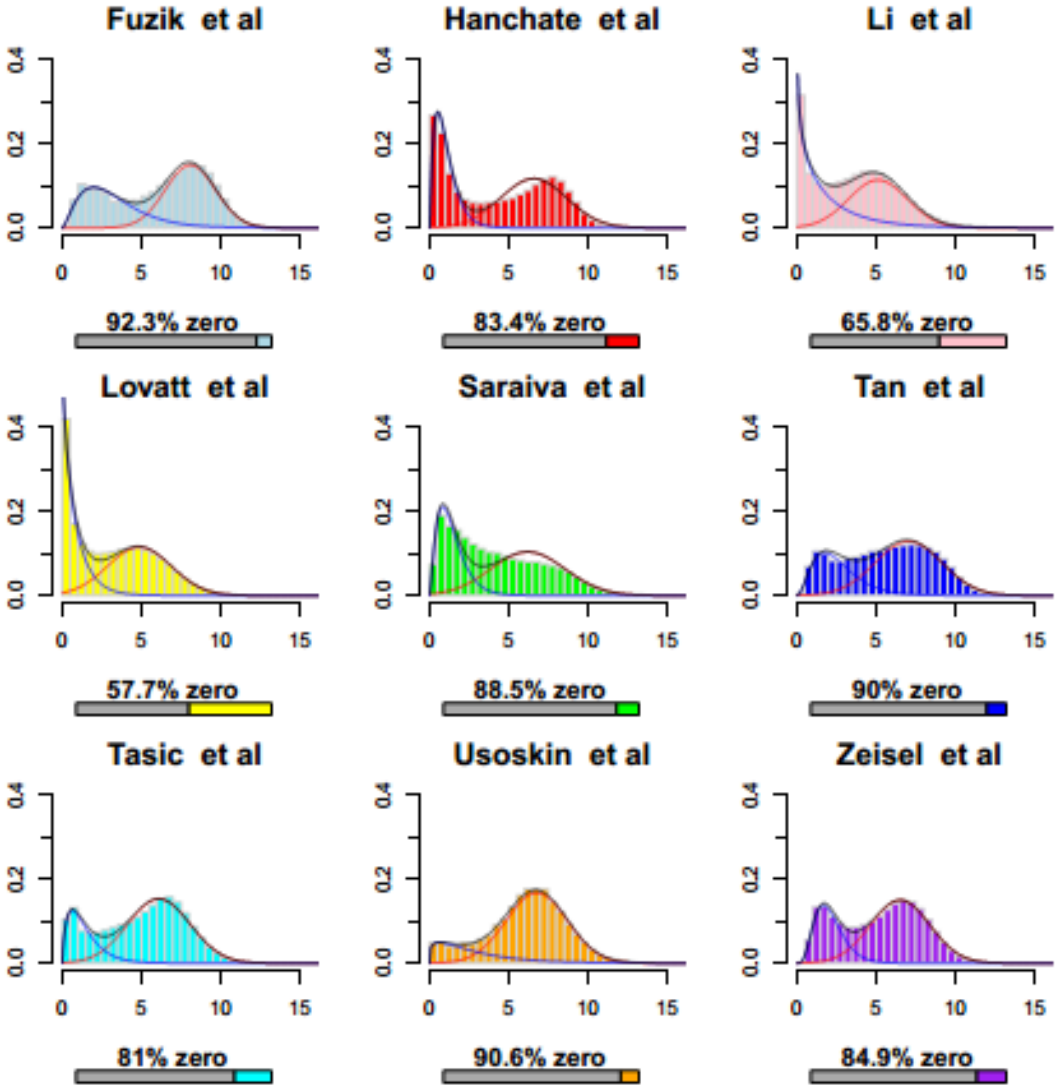
Integrated single cell data analysis reveals cell specific networks and novel coactivation markers

Shila Ghazanfar,[✉1] Adam J. Bisogni,[2] John T. Ormerod,[1,3] David M. Lin,[2] and Jean Y. H. Yang[1]



Ghazanfar et al BMC Systems Biology 2016

Previous work on mixture modelling

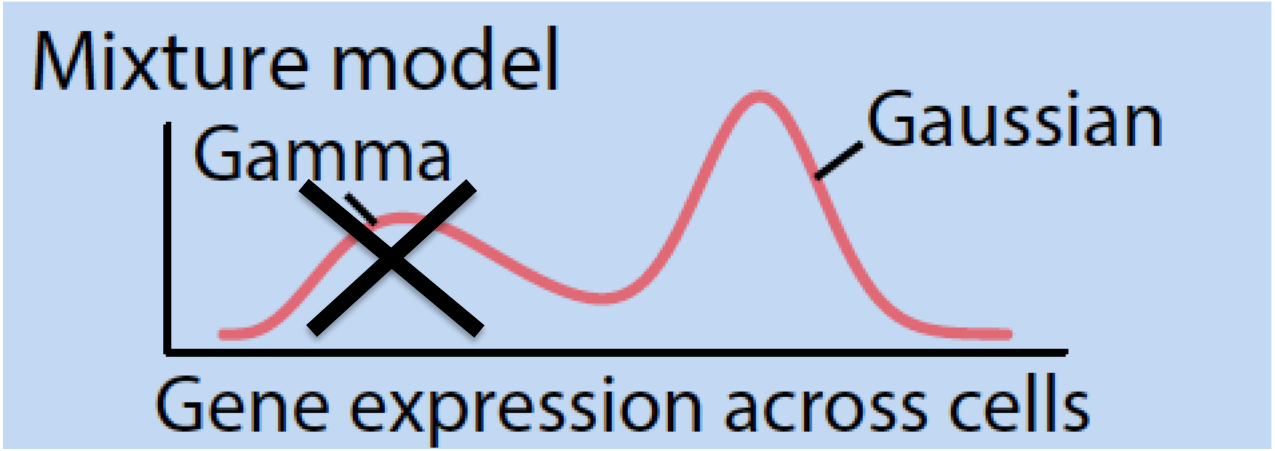


BioMed Central The Open Access Publisher | BMC Systems Biology | this article | search | submit a manuscript | register

BMC Syst Biol. 2016; 10(Suppl 5): 127.
Published online 2016 Dec 5. doi: 10.1186/s12918-016-0370-4
PMCID: PMC5249008
PMID: 28105940

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. Yang[1]



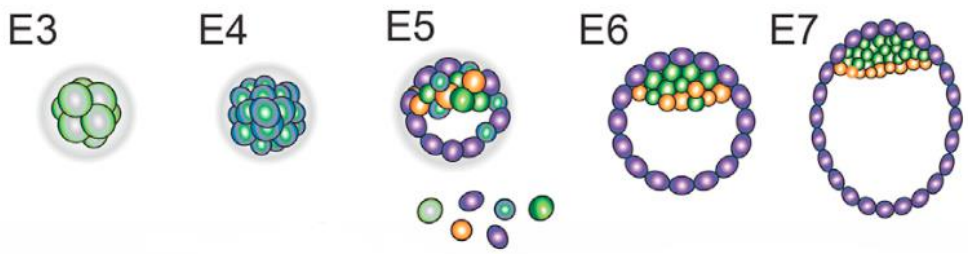
Ghazanfar et al BMC Systems Biology 2016

Establishing a good set of stably expressed genes

Datasets with wide ranges of cell types

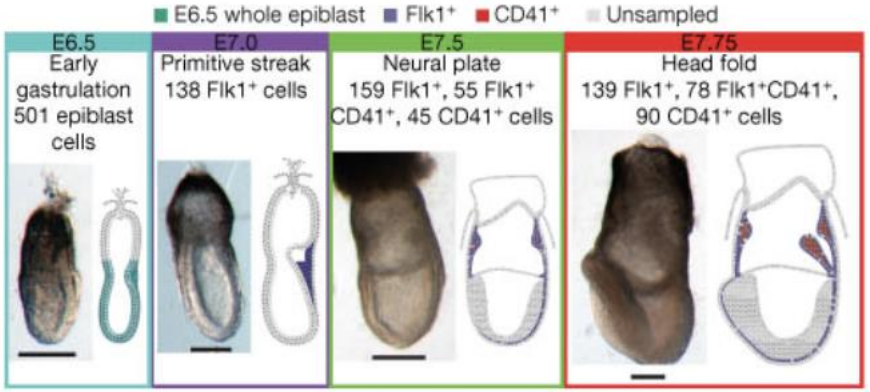
Human early developmental dataset (Petropoulos et al.):

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



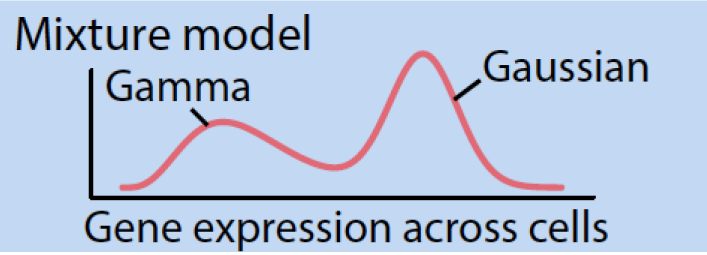
Building features for stably expressed genes

- 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):

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



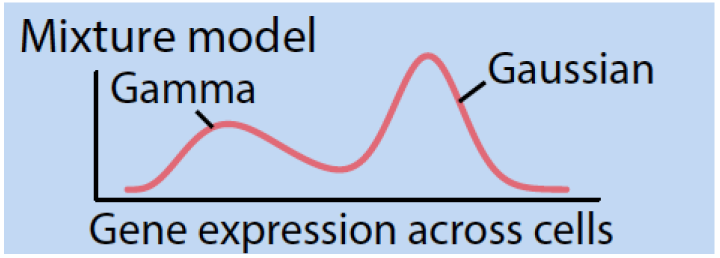
Building features for stably expressed genes

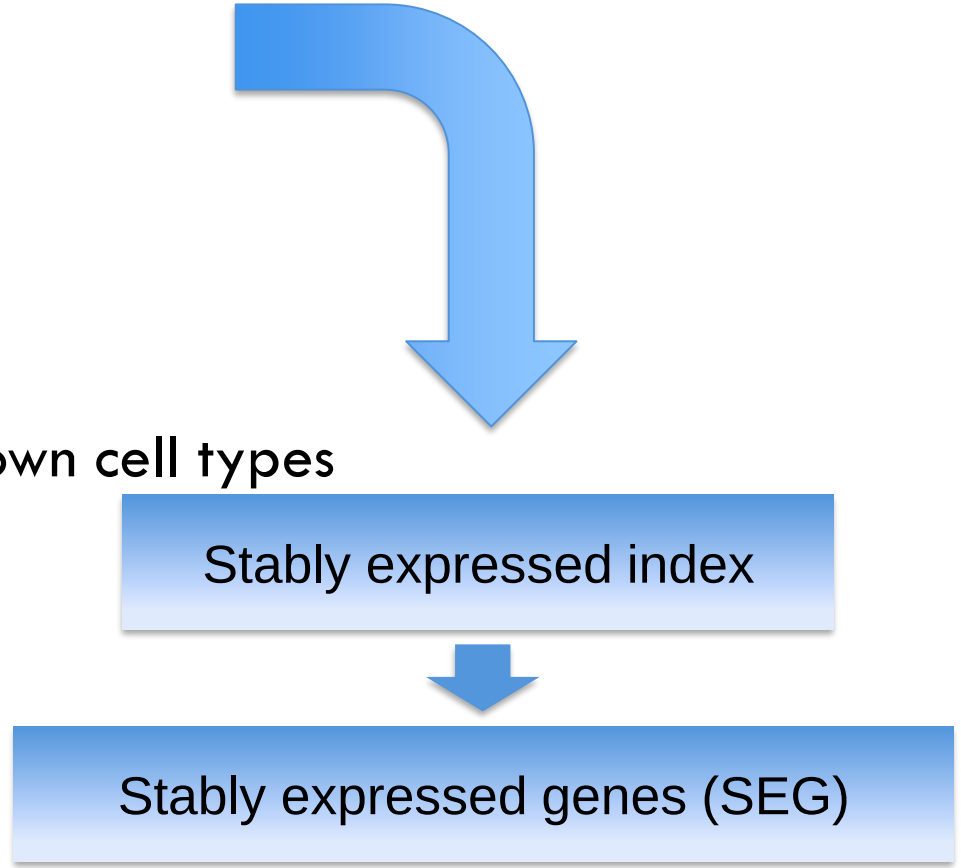
- Mixture Model:
  - **Mixing Proportion** of the second component (λ_i)
  - **Standard deviation** of the second component (σ_i)

- **Zero Proportion** of each gene across all the cells (ω_i)

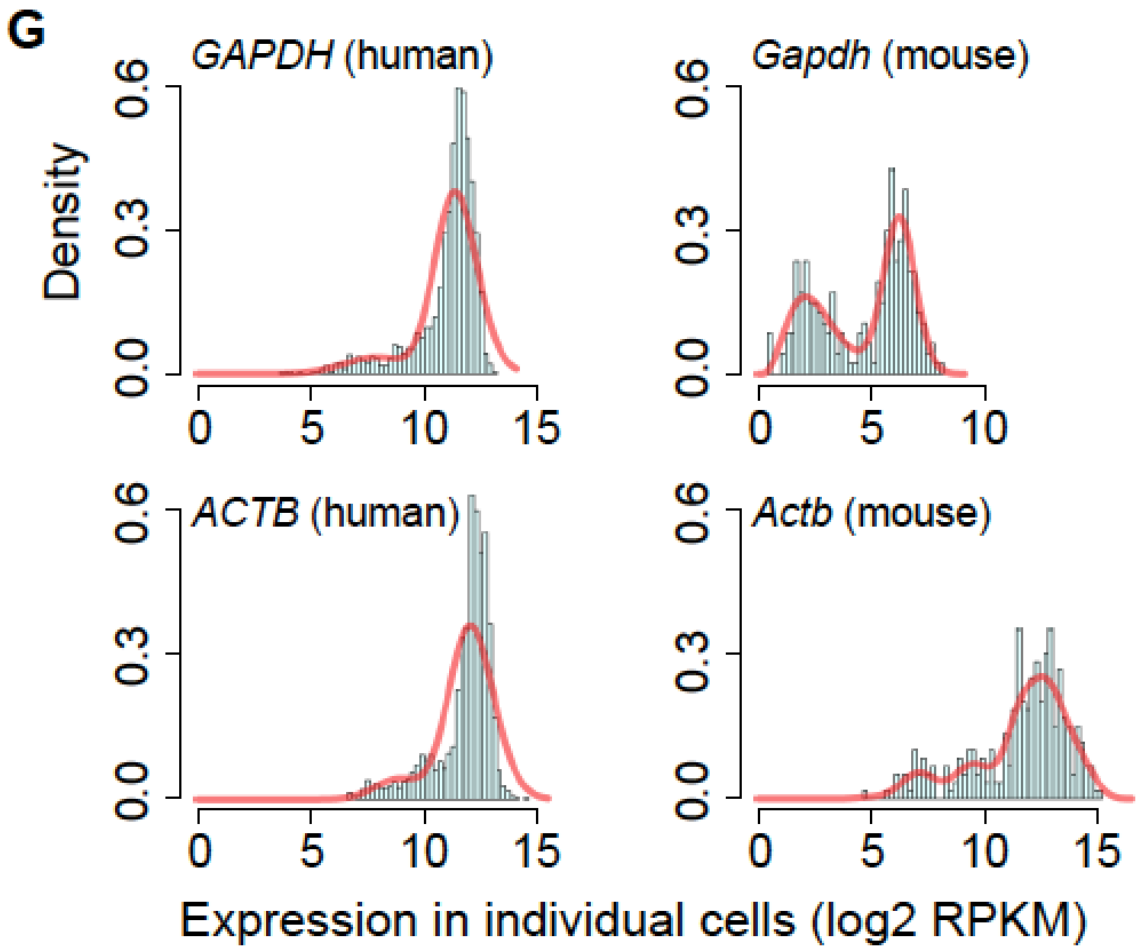
- **F-statistics** from one-way ANOVA (F_i):

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



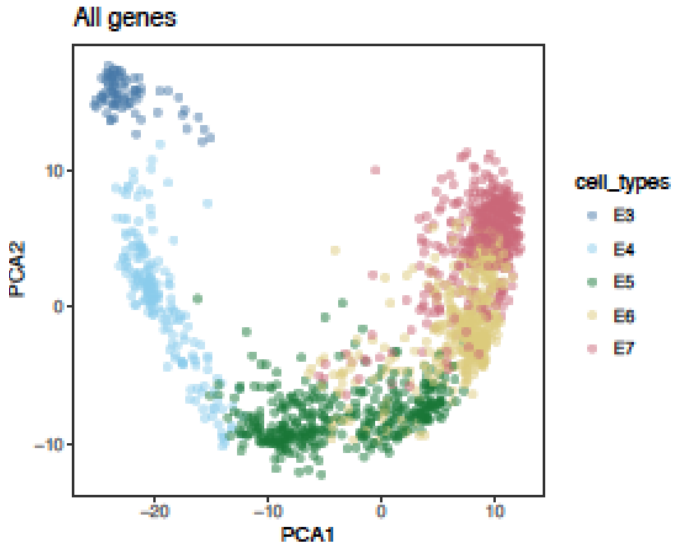


What about GAPDH?



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

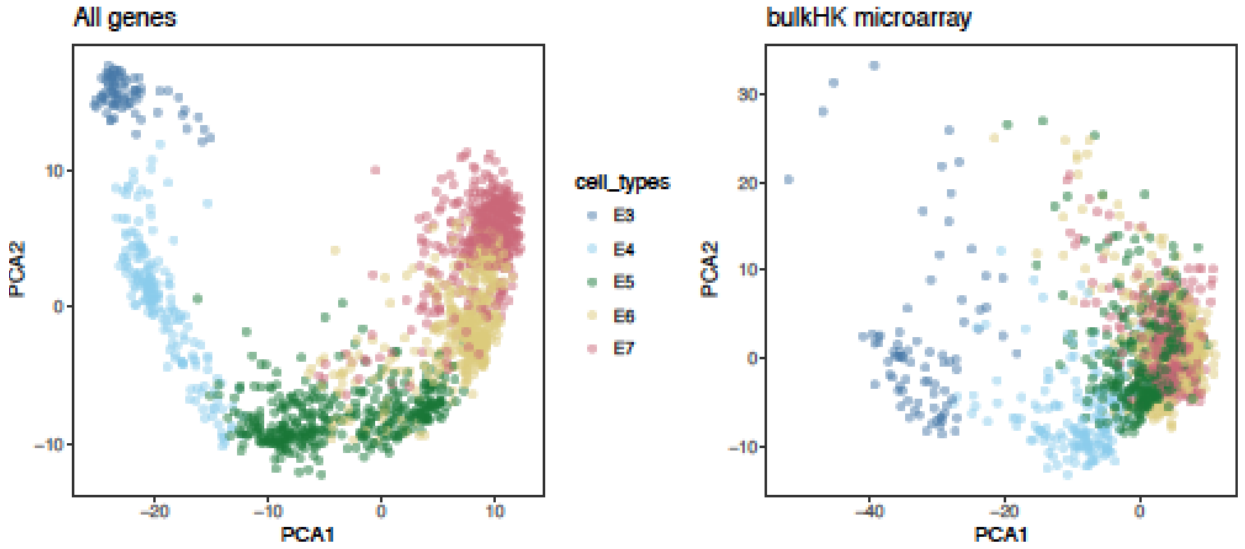
Proposed scSEG

Eisenberg and Levanon 2013. Human housekeeping genes, revisited. Trends Genet 29 :

Eisenberg and Levanon. 2003. Human housekeeping genes are compact. Trends Genet 19

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

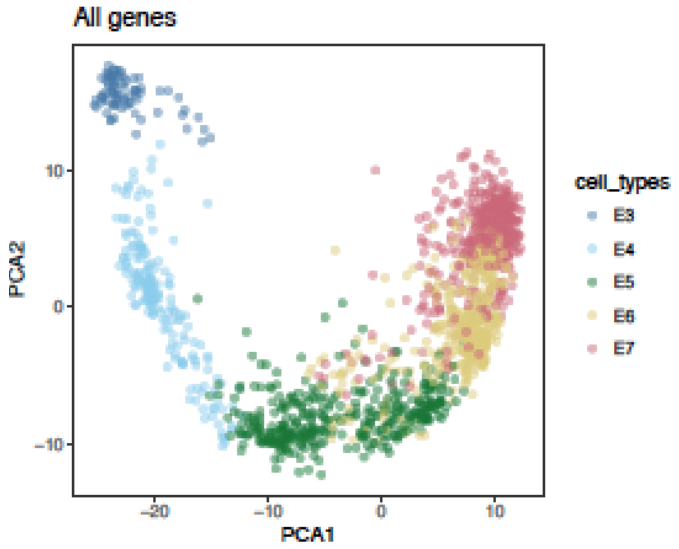
Proposed scSEG

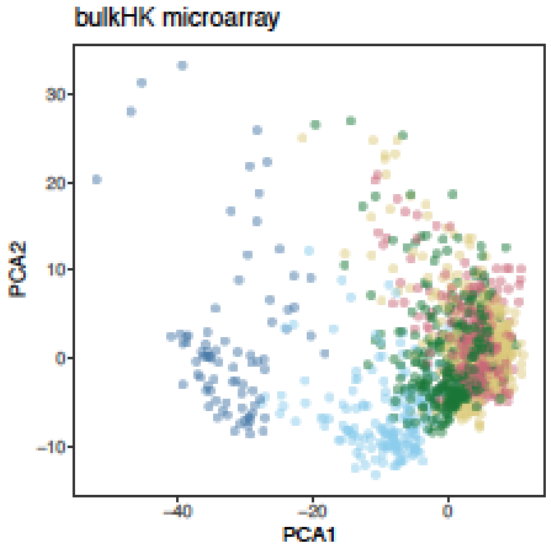
Eisenberg and Levanon 2013. Human housekeeping genes, revisited. Trends Genet 29 :

Eisenberg and Levanon. 2003. Human housekeeping genes are compact. Trends Genet 19

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

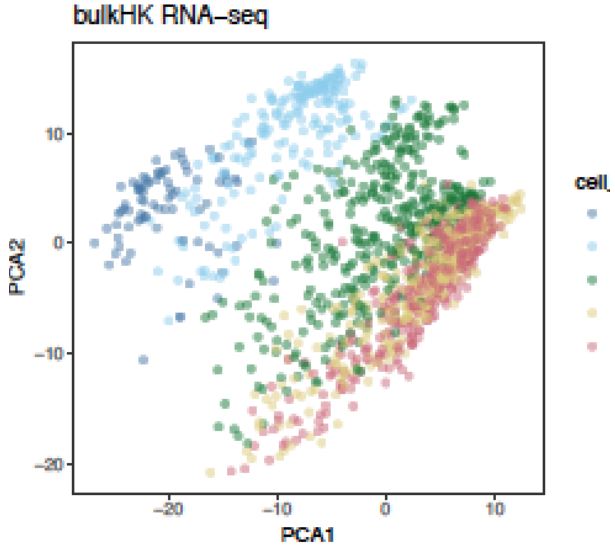
ALL genes





House-keeping genes based on microarray data

House keeping genes based on bulk RNA-seq



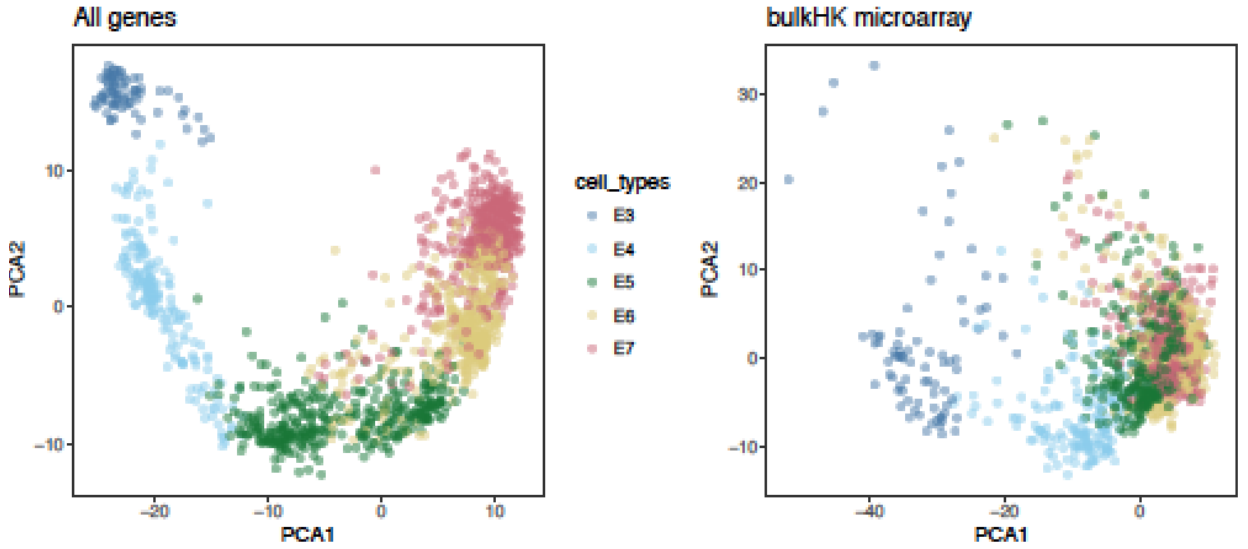
Proposed scSEG

Eisenberg and Levanon 2013. Human housekeeping genes, revisited. Trends Genet 29 :

Eisenberg and Levanon. 2003. Human housekeeping genes are compact. Trends Genet 19

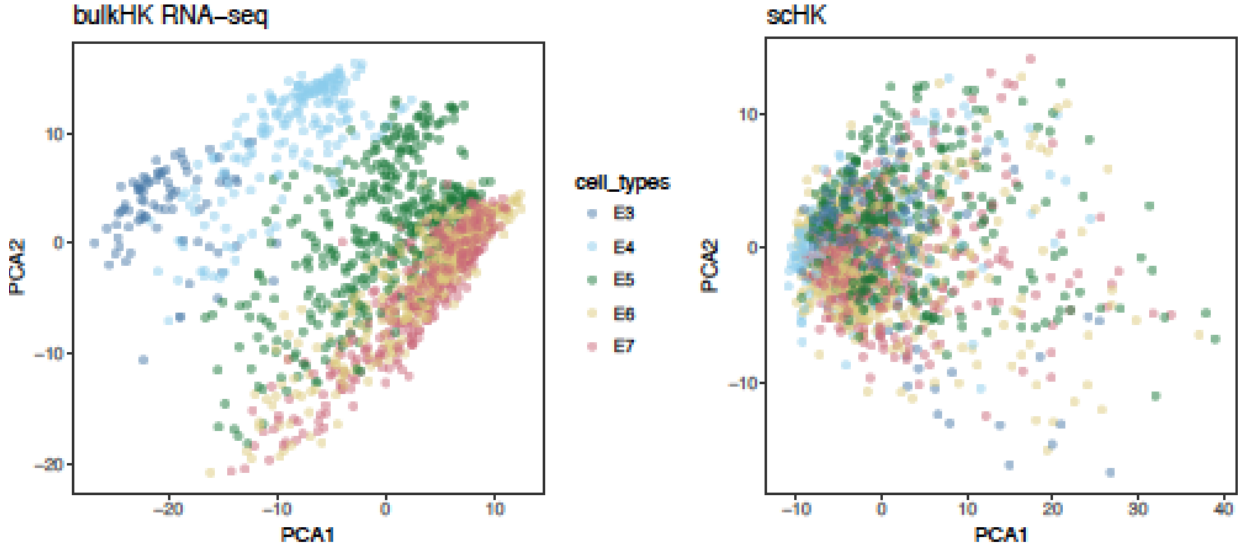
Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



House-keeping genes based on microarray data

House keeping genes based on bulk RNA-seq



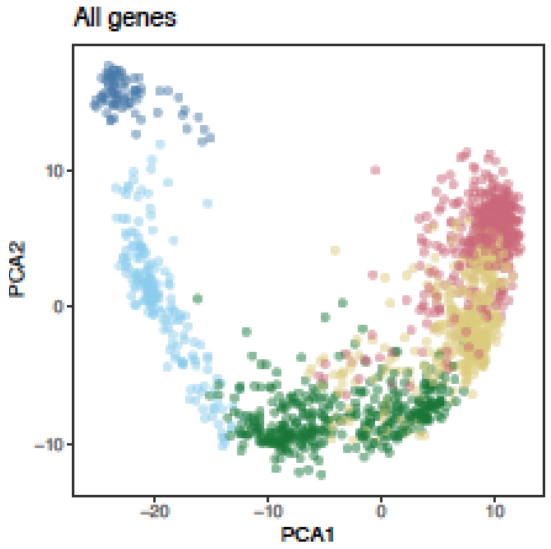
Proposed scSEG

Eisenberg and Levanon 2013. Human housekeeping genes, revisited. Trends Genet 29 :

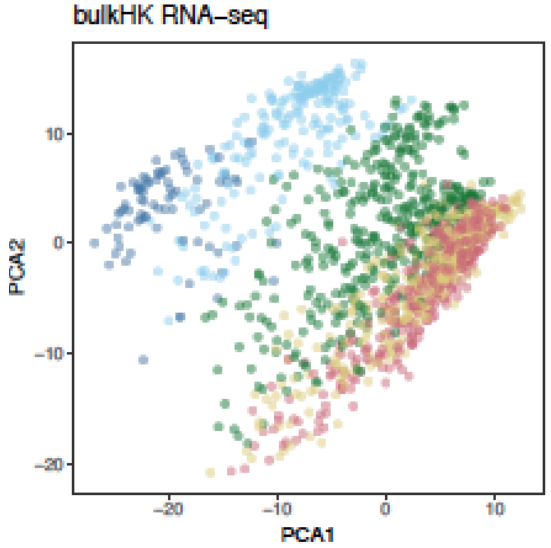
Eisenberg and Levanon. 2003. Human housekeeping genes are compact. Trends Genet 19

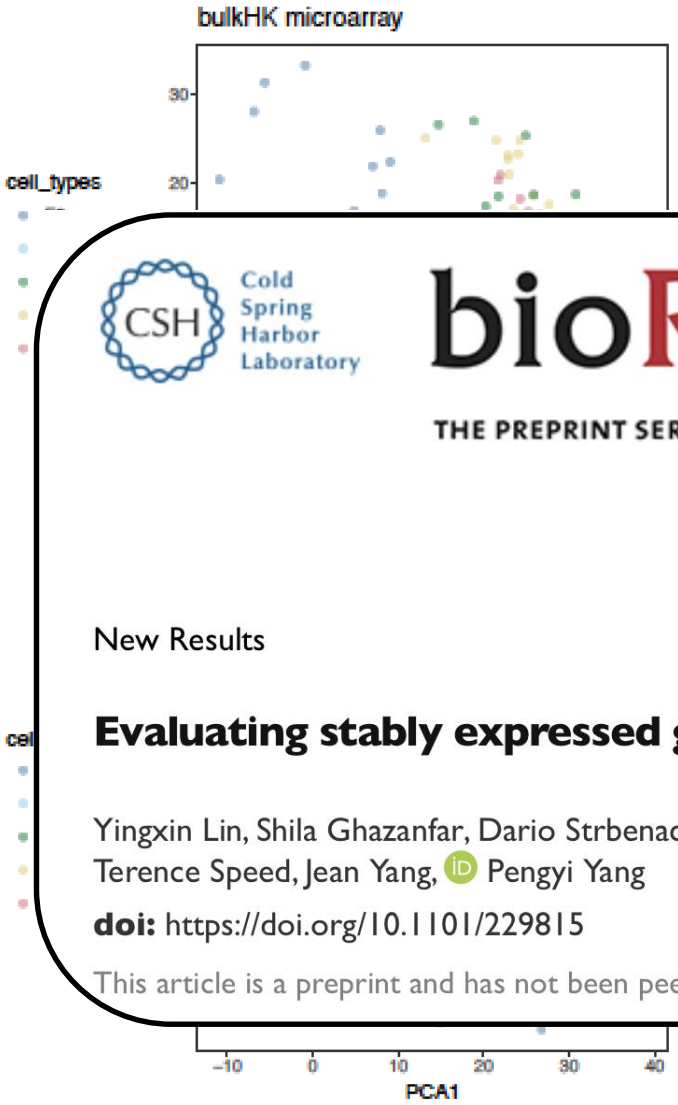
Evaluation of scSEG (PCA plot) - Human early developmental dataset (Petropoulos et al)

ALL genes



House keeping genes based on bulk RNA-seq





House-keeping genes based on microarray data

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Evaluating stably expressed genes in single cells

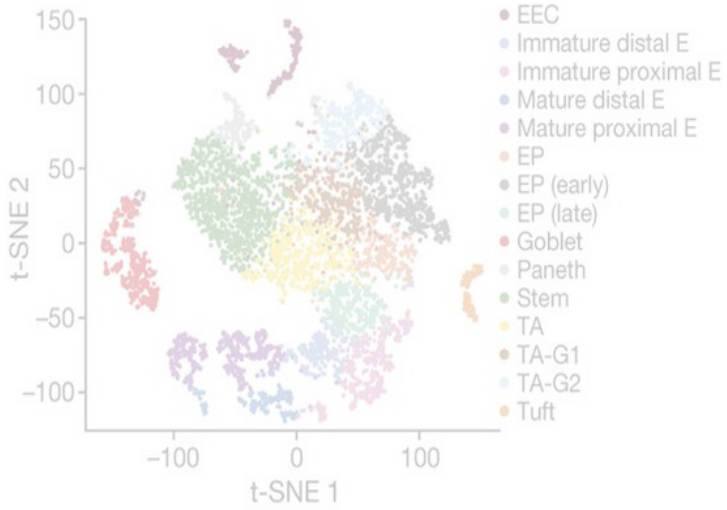
Yingxin Lin, Shila Ghazanfar, Dario Strbenac, Andy Wang, Ellis Patrick, Dave Lin, Terence Speed, Jean Yang, Pengyi Yang

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

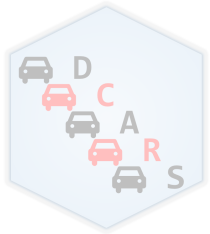
This article is a preprint and has not been peer-reviewed [what does this mean?].

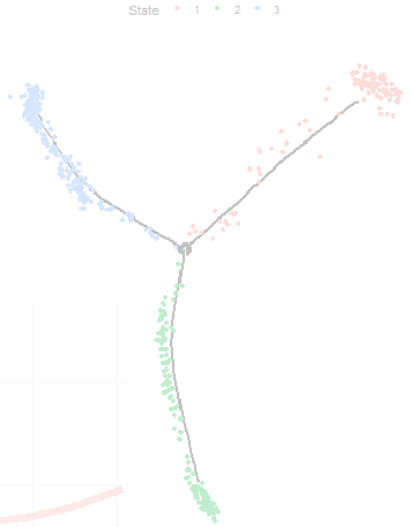
Eisenberg and Levanon 2013. Human housekeeping genes, revisited. Trends Genet 29 :

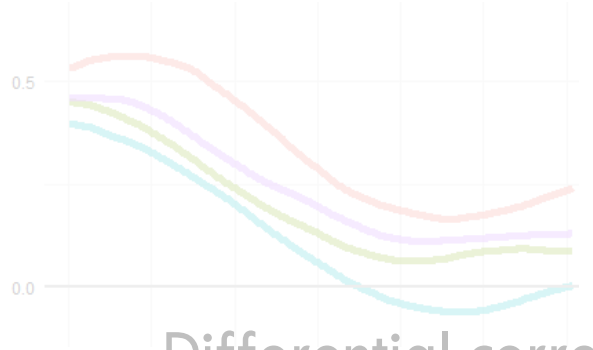
Eisenberg and Levanon. 2003. Human housekeeping genes are compact. Trends Genet 19



Clustering metrics



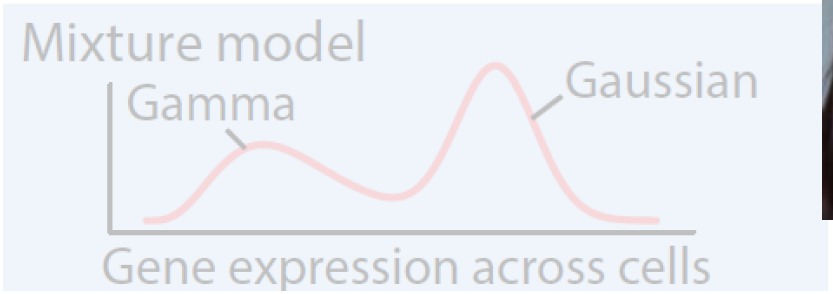




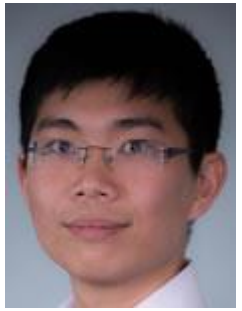
Differential correlation

Sydney Single cell

Finding stably expressed ge

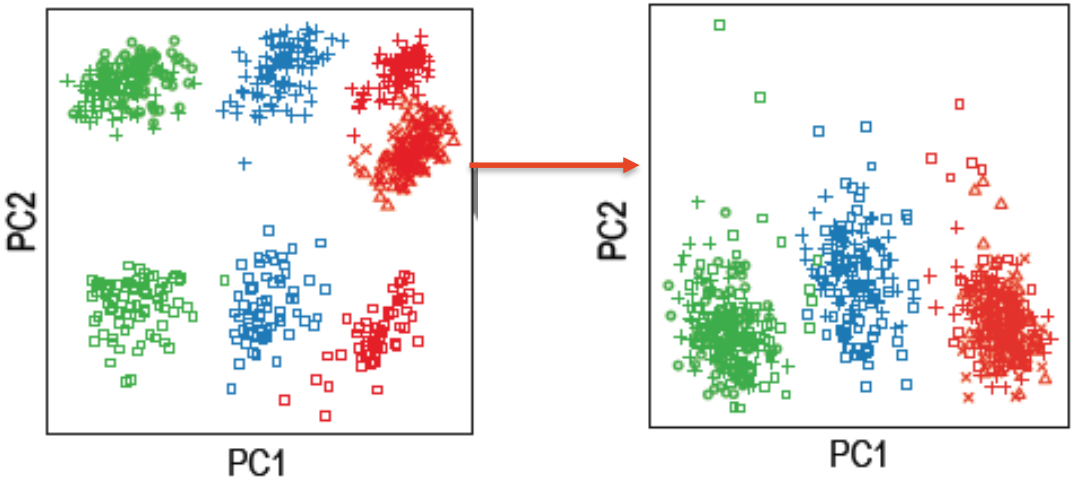


Yingxin Lin



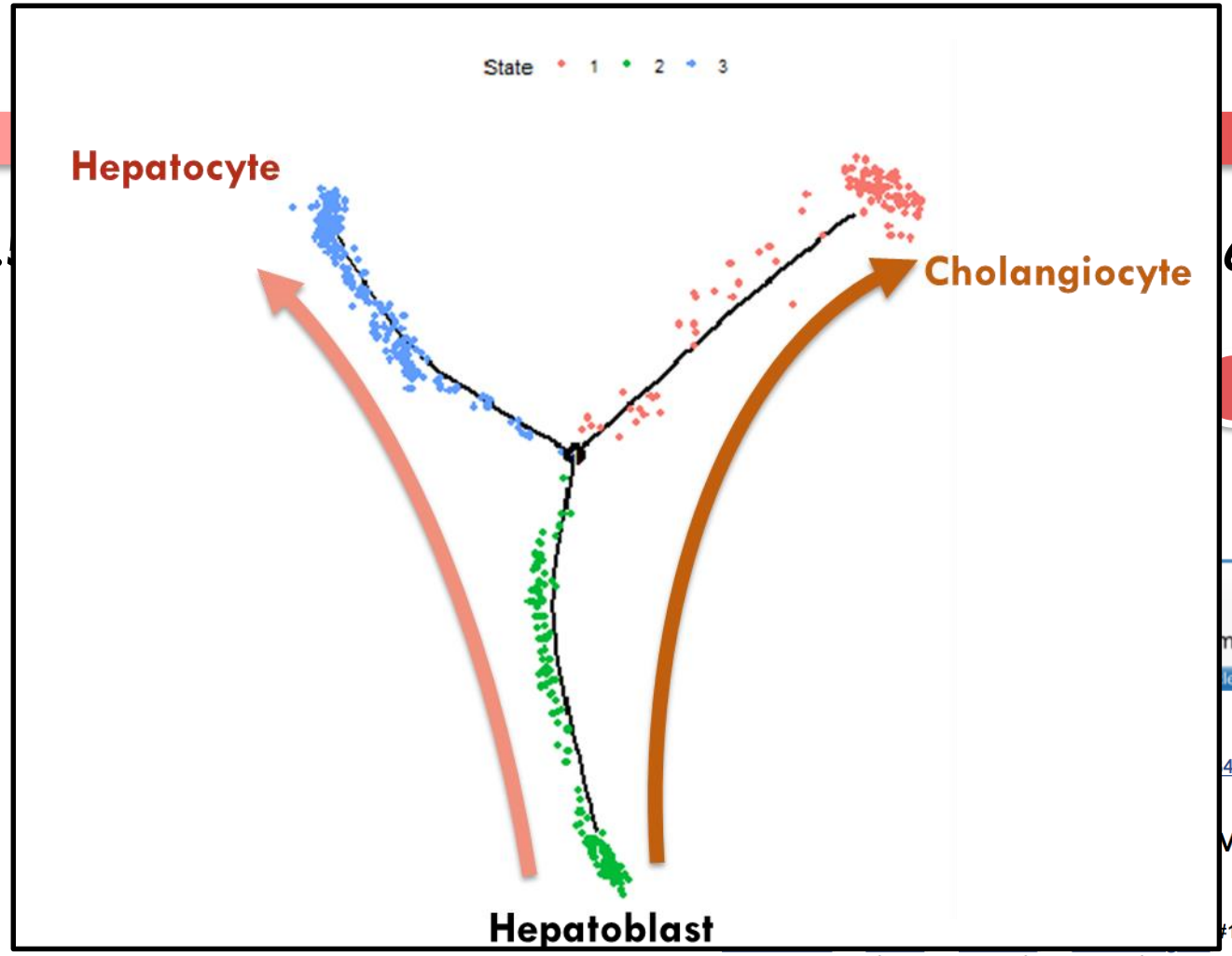
Kevin Wang

scMerge



Liver fetal development time course datasets – integrating multiple datasets

GSE87795
Su et al.



E9.

6.5 E17.5

nics
le search submit a manuscript register

PMCID: PMC5715535
PMID: 29202695

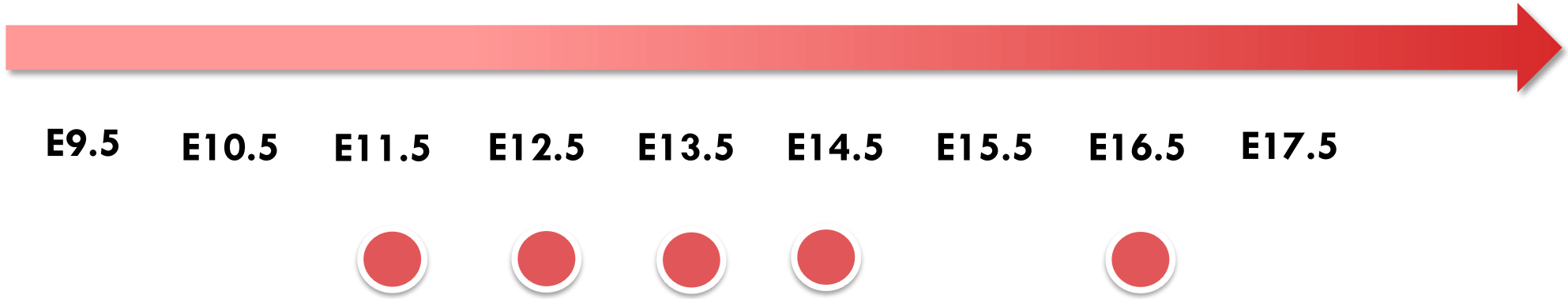
42-x

veals dynamic trajectories during

#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



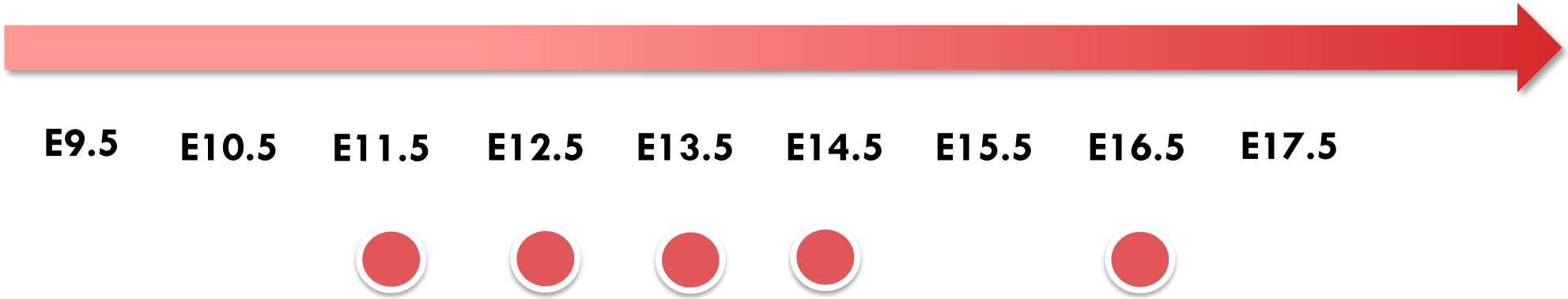
GSE87795
Su et al.

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]

Liver fetal development time course datasets – integrating multiple datasets



GSE87795
Su et al.

LETTER

doi:10.1038/nature22796

Multilineage communication regulates human liver bud development from pluripotency

J. Gray Camp[1]*, Keisuke Sekine[2]*, Tobias Gerber[1], Henry Loeffler-Wirth[3], Hans Binder[3], Malgorzata Gac[1], Sabina Kanton[1], Jorge Kageyama[1], Georg Damm[4,5], Daniel Seehofer[4,5], Lenka Belicova[6], Marc Bickle[6], Rico Barsacchi[6], Ryo Okuda[2], Emi Yoshizawa[2], Masaki Kimura[2], Hiroaki Ayabe[2], Hideki Taniguchi[2], Takanori Takebe[2,7] & Barbara Treutlein[1,6]

BioMed Central
The Open Access Publisher

BMC Genomics

this article | search | submit a manuscript | register

s. 2017; 18: 946.
e 2017 Dec 4. doi: 10.1186/s12864-017-4342-x

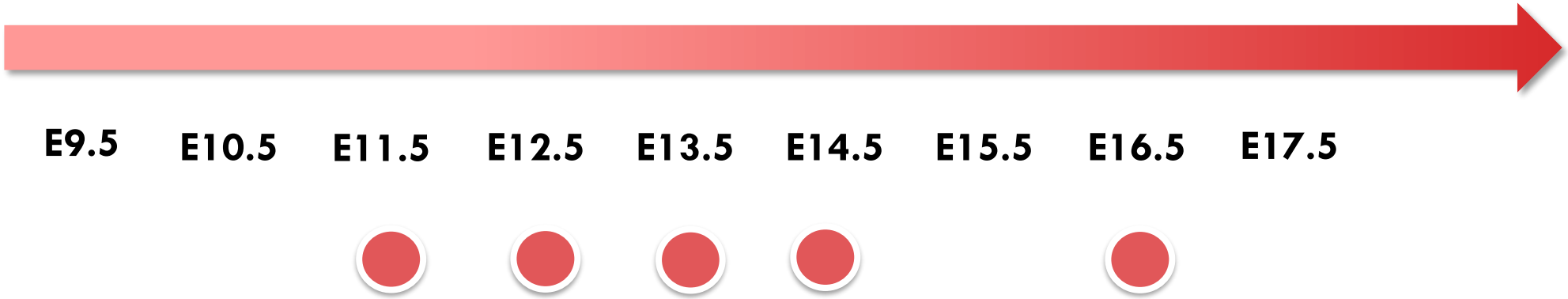
PMCID: PMC5715535
PMID: 29202695

ell RNA-Seq analysis reveals dynamic trajectories during
ver development

[1] 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]

tion ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Liver fetal development time course datasets – integrating multiple datasets



GSE87795
Su et al.

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

LET

A Single-Cell Transcriptomic Analysis Reveals Precise Pathways and Regulatory Mechanisms Underlying Hepatoblast Differentiation

Li Yang,[1,2*] Wei-Hua Wang,[1,2*] Wei-Lin Qiu,[1,3*] Zhen Guo,[1] Erfei Bi,[4] and Cheng-Ran Xu[1]

Multilin
bud dev

J. Gray Camp[1*], Kei
Jorge Kageyama[1], C
Emi Yoshizawa[2], M

Med Central
en Access Publisher
BMC Genomics
this article search submit a manuscript register

946.
4. doi: 10.1186/s12864-017-4342-x
PMCID: PMC5715535
PMID: 29202695

A-Seq analysis reveals dynamic trajectories during
velopment

Xin Zou,[#1] Zhao-Ning Lu,[#1] Gangcai Xie,[2] Jean Y. H. Yang,[3] Chong-Chao Wu,[1]
an 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

LET

Multilin
bud dev

J. Gray Camp[1]*, Kei
Jorge Kageyama[1], C
Emi Yoshizawa[2], M

A Single-Cell Transcriptom
Analysis Reveals Precise Pat
and Regulatory Mechanisms
Hepatoblast Differentiation

Li Yang,[1,2]* Wei-Hua Wang,[1,2]* Wei-Lin Qiu,[1,3]* Zhen Guo,[1] Erfei Bi,[4] and Cheng-Ran Xu[1]

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

Genome Biology

RESEARCH ARTICLE Open Access

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*]

cript register

PMCID: PMC5715535
PMID: 29202695

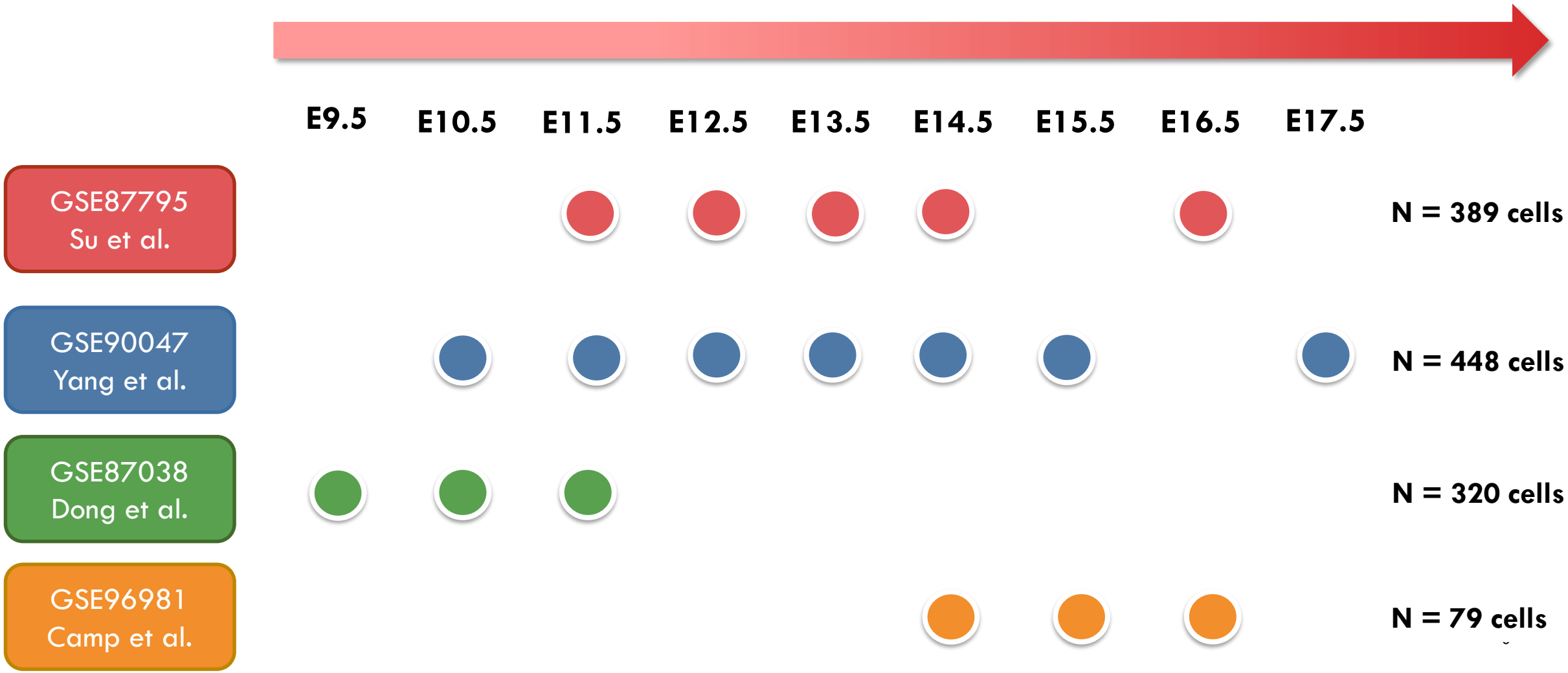
jectories during

velopment

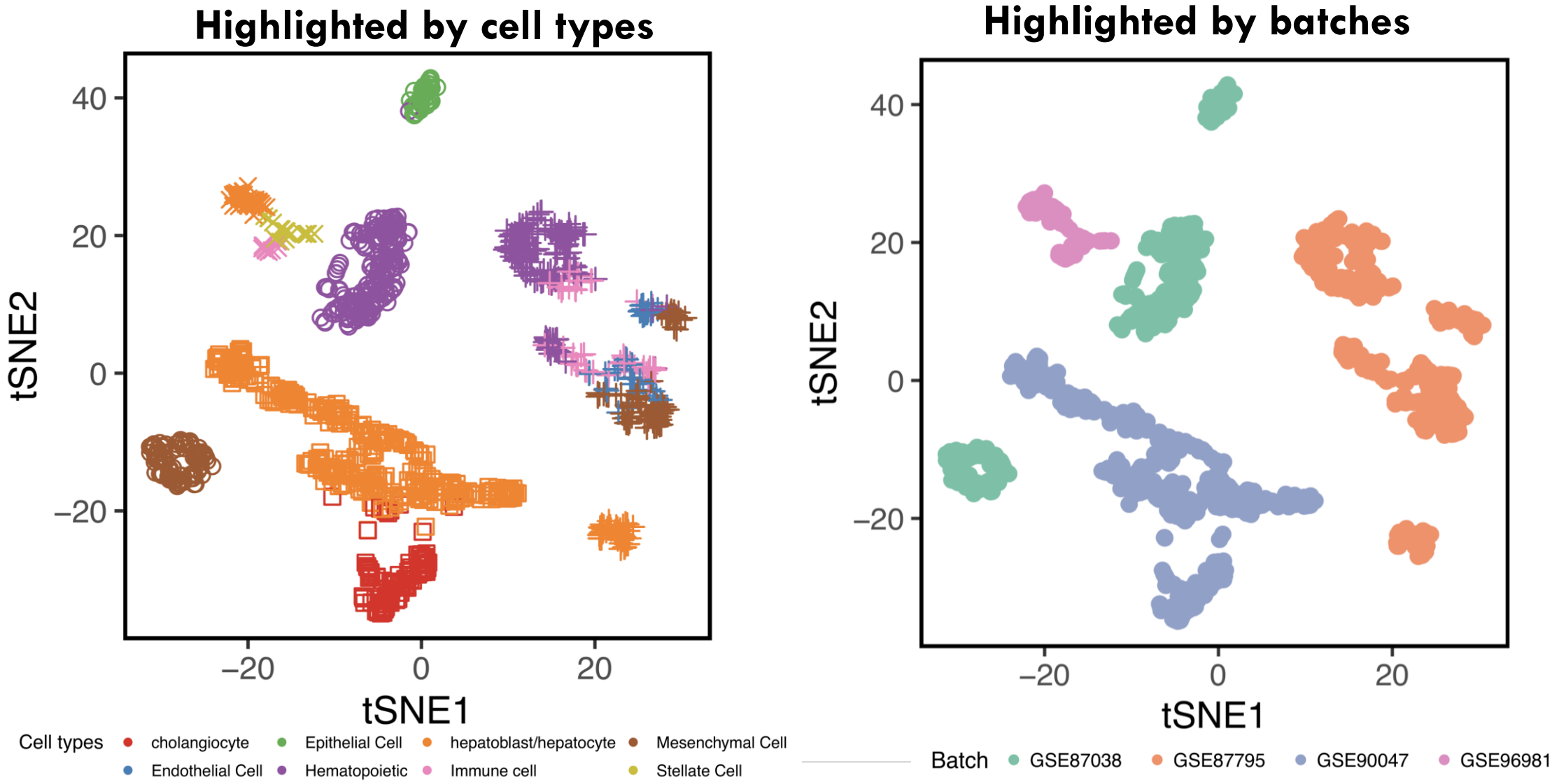
Xin Zou,[#1] Zhao-Ning Lu,[#1] Gangcai Xie,[2] Jean Y. H. Yang,[3] Chong-Chao Wu,[1]
Yan He,[1] Qing Luo,[1] Yu-Lan Qu,[1] Na Wang,[1] Lan Wang,[1] and Ze-Guang Han[1,4]

tion ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

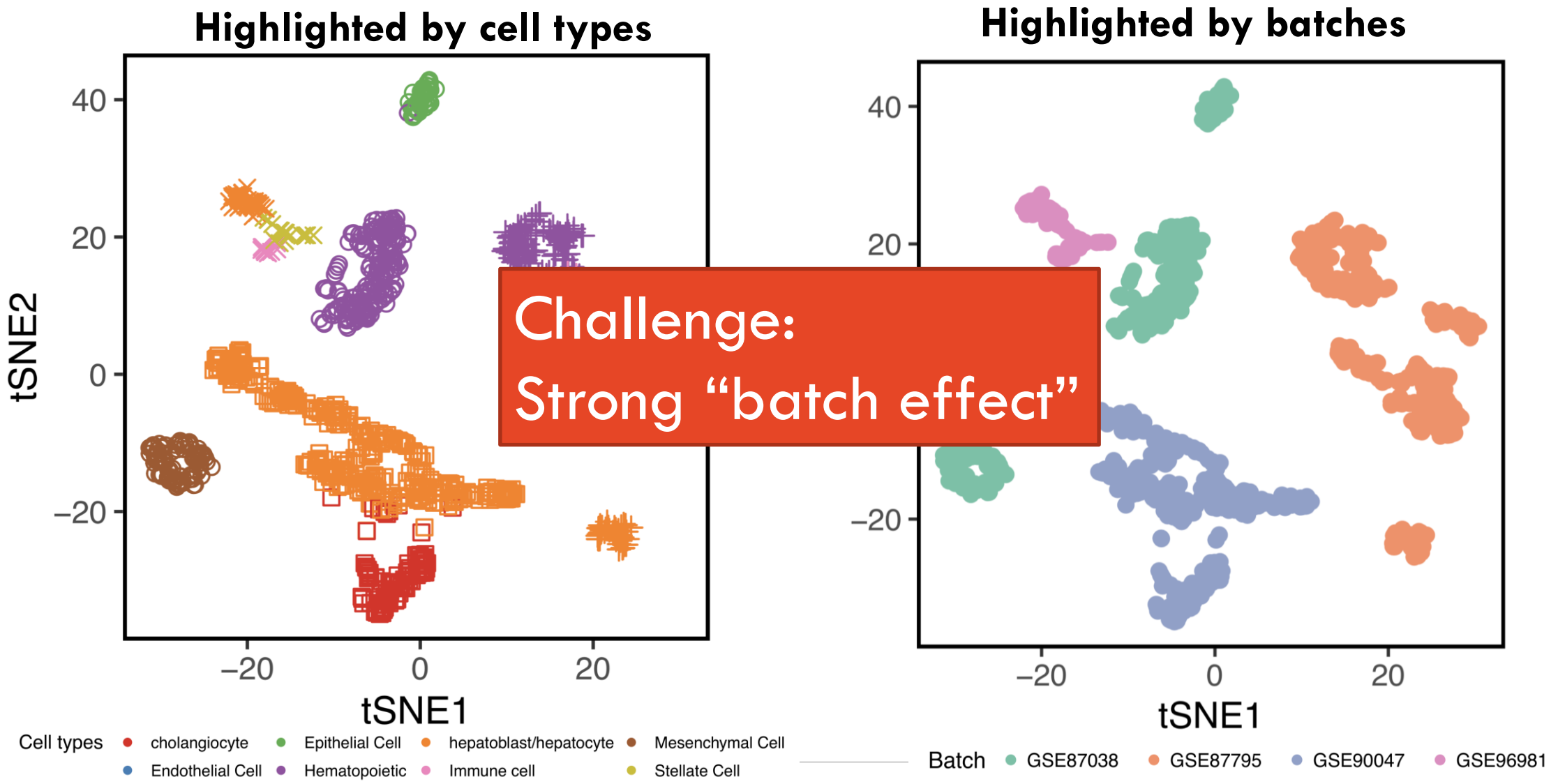
Liver fetal development time course datasets



tSNE of liver fetal development time course datasets

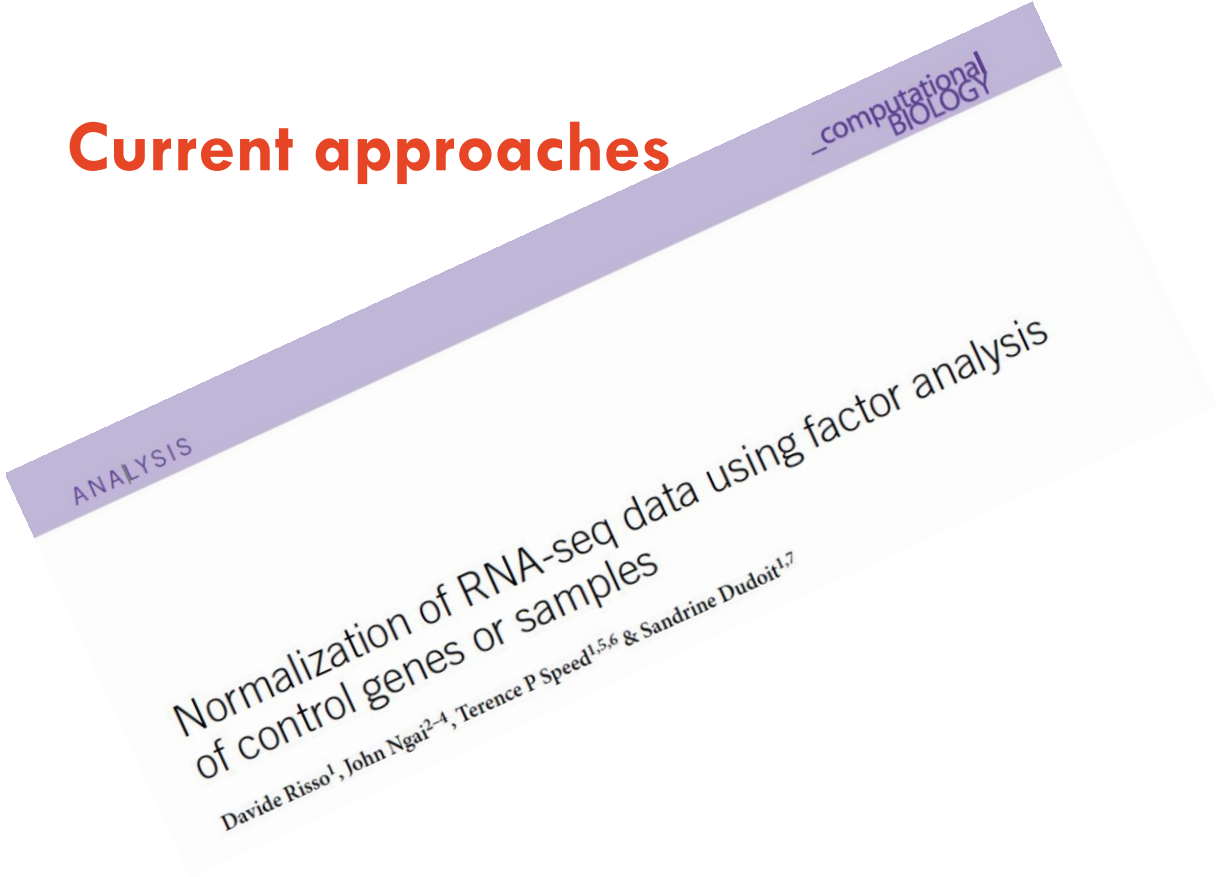


tSNE of liver fetal development time course datasets



Current approaches

Current approaches



Current approaches

ANALYSIS

_computational BIOLOGY

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

Davide Risso[1], John Ngai[2–4], Terence P Speed[1,5,6] & Sandrine Dudoit[1,7]

Biostatistics

Issues Advance articles Submit Purchase Alerts About

Article Navigation

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

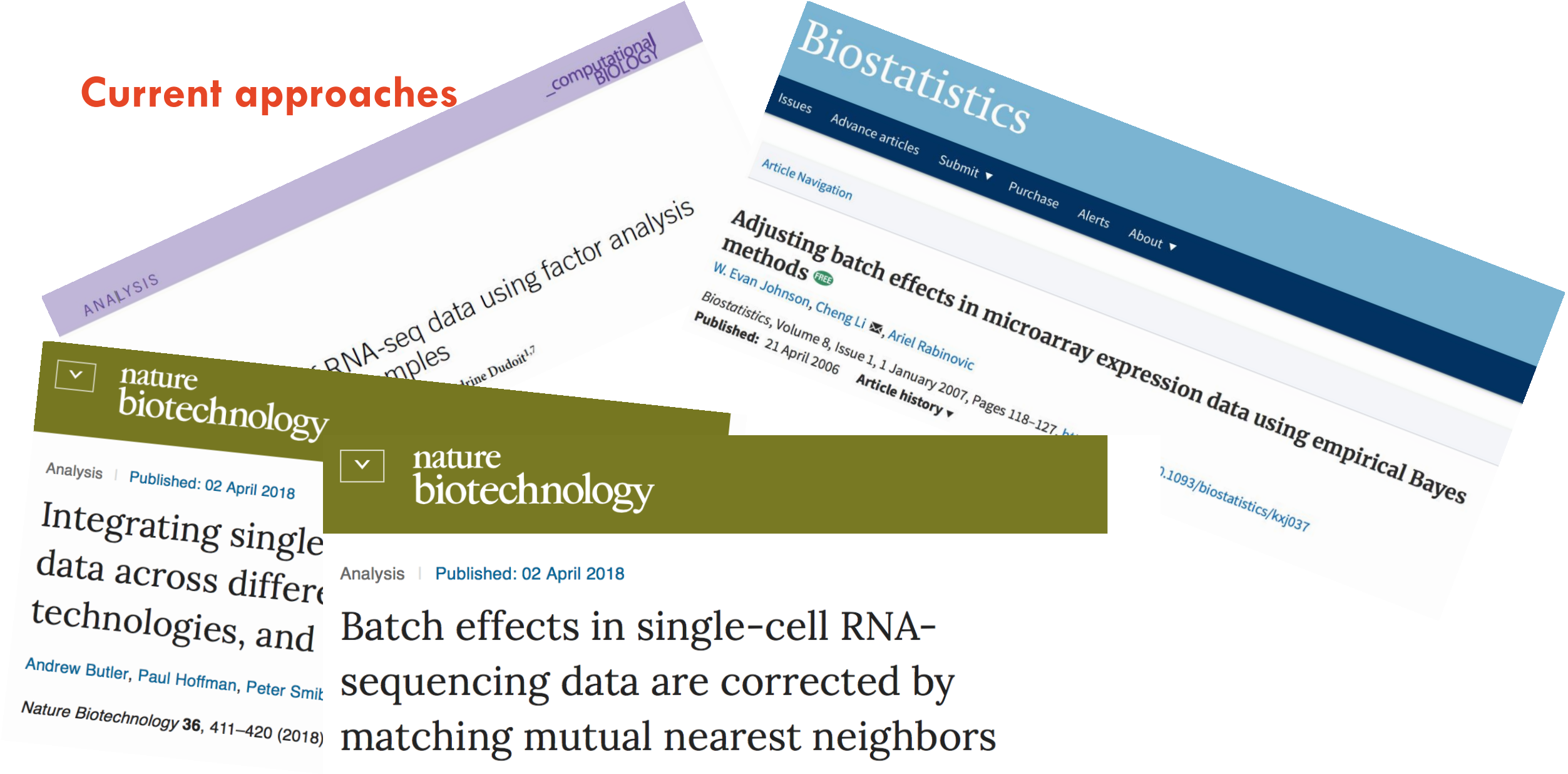
W. Evan Johnson, Cheng Li, Ariel Rabinovic

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

Published: 21 April 2006 **Article history**

Current approaches

Current approaches



Current approaches

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

scMerge: algorithm

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

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

Negative controls

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

RUVIII

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

Estimated by negative controls

scMerge: algorithm

Negative controls

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

RUVIII

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

Estimated by negative controls

scMerge: algorithm

Single-cell stably expressed genes (SEG)

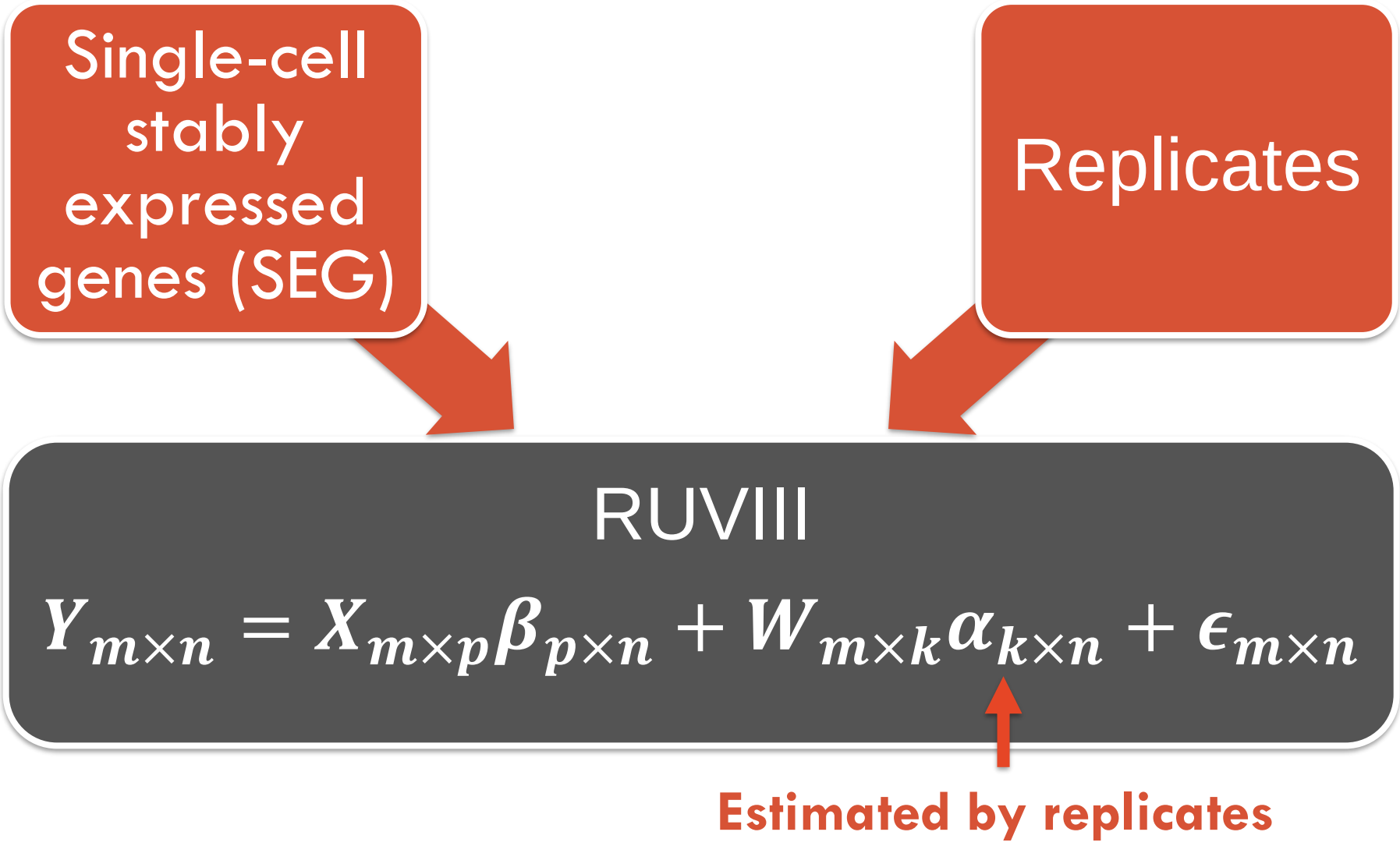
Genes with stable expression level are not affected by factors of interest i.e.,
Stably expressed across different cell types or developmental stages.

RUVIII

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

Estimated by negative controls

scMerge: algorithm



scMerge: algorithm

Sin
s
exp
gen

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

Sin
s
exp
gen

But cells don't have replicates!

$Y_{m\times$

Estimate pseudo-replicates

Estimated by replicates

scMerge: algorithm

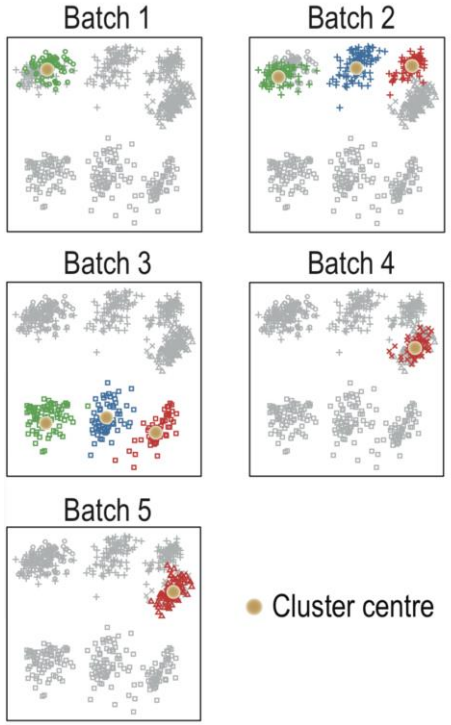
Idea:

- 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)

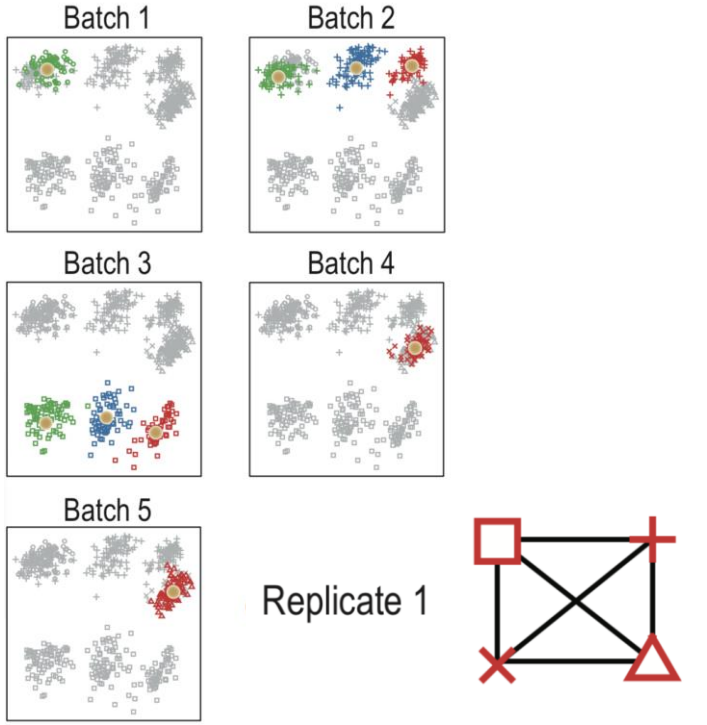


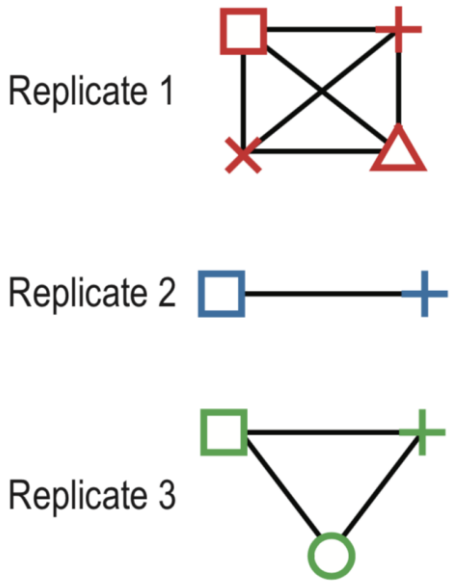
Pseudo-replicates

scMerge: algorithm

Clustering for each batch
(k-means by default)

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

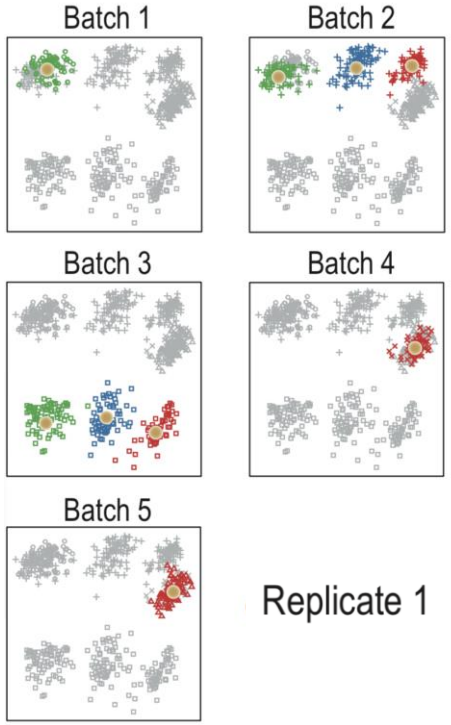




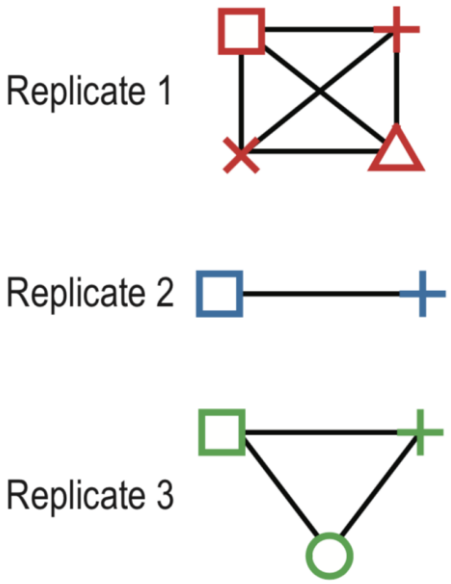
Pseudo-replicates

scMerge: algorithm

Clustering for each batch
(k-means by default)

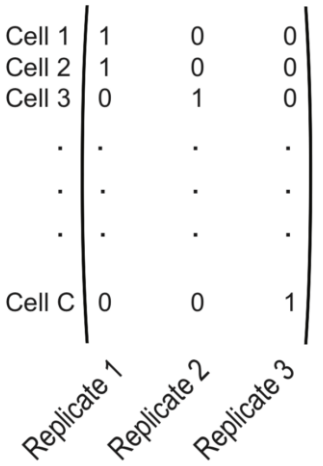


Find **Mutual Nearest Clusters** as pseudo-replicates

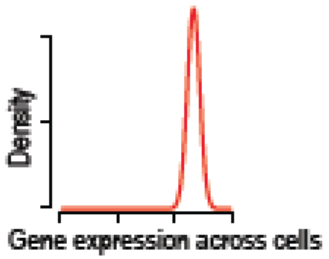


Pseudo-replicates

Frame as pseudo-replicate information

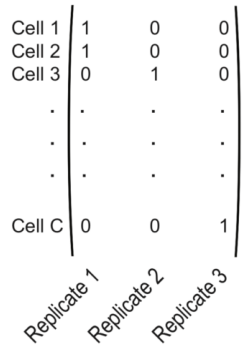


scMerge: algorithm



Single-cell stably expressed genes (SEG)

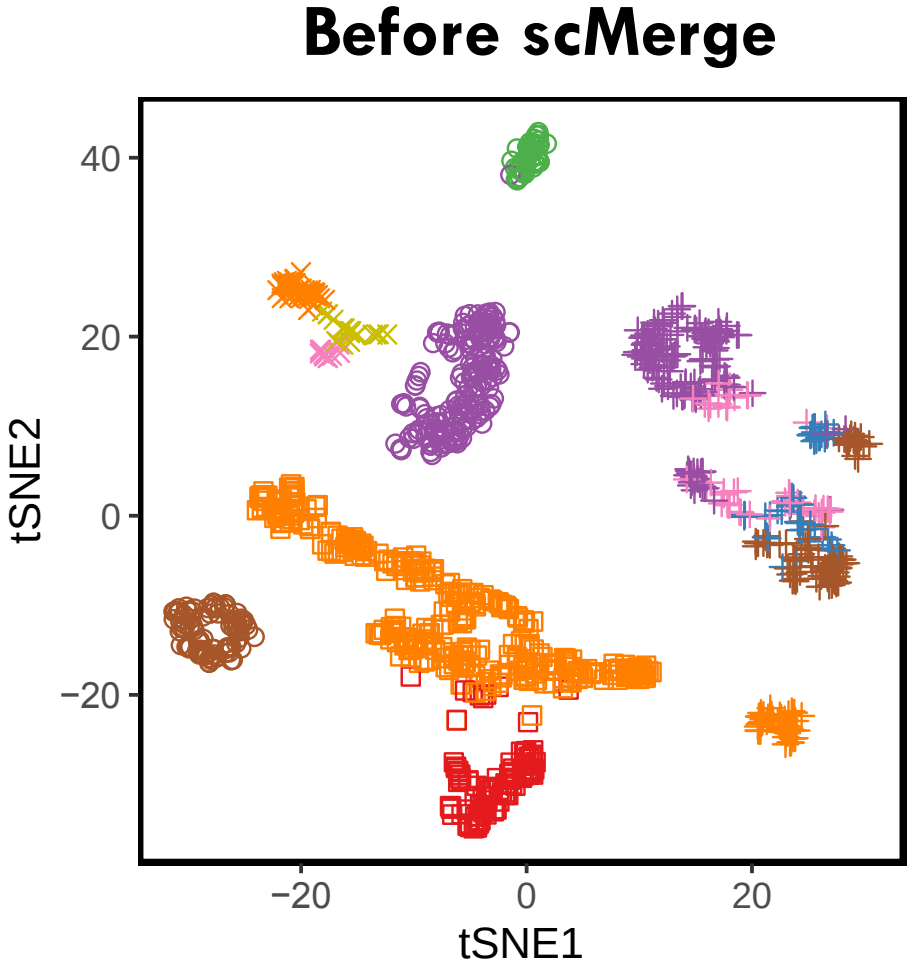
Pseudo-replicates

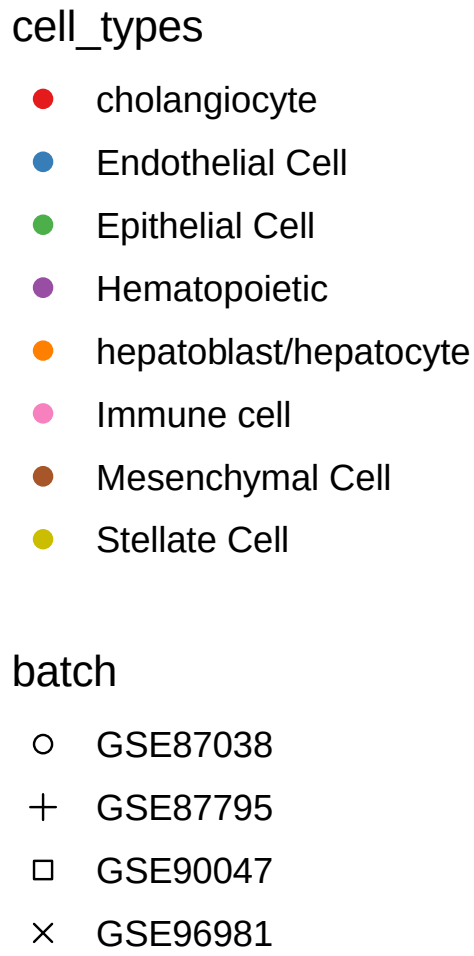


RUVIII

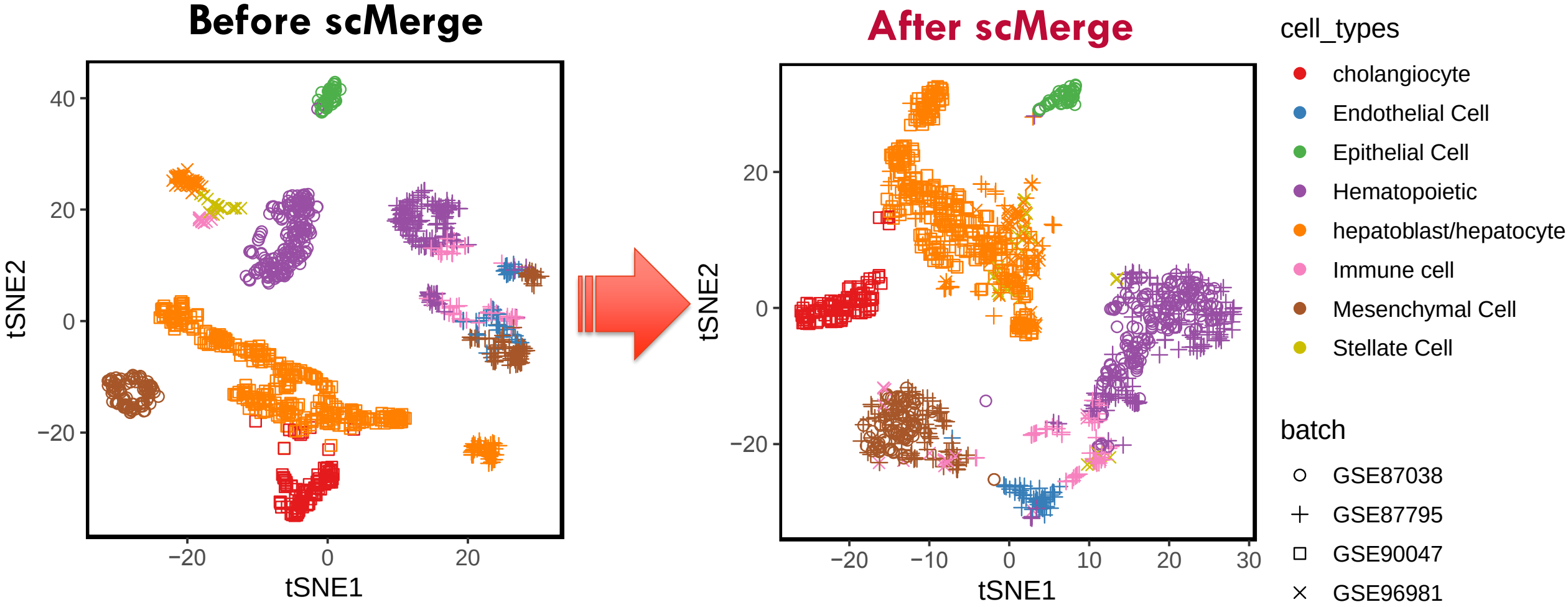
$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + W_{m\times k}\alpha_{k\times n} + \epsilon_{m\times n}$$

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

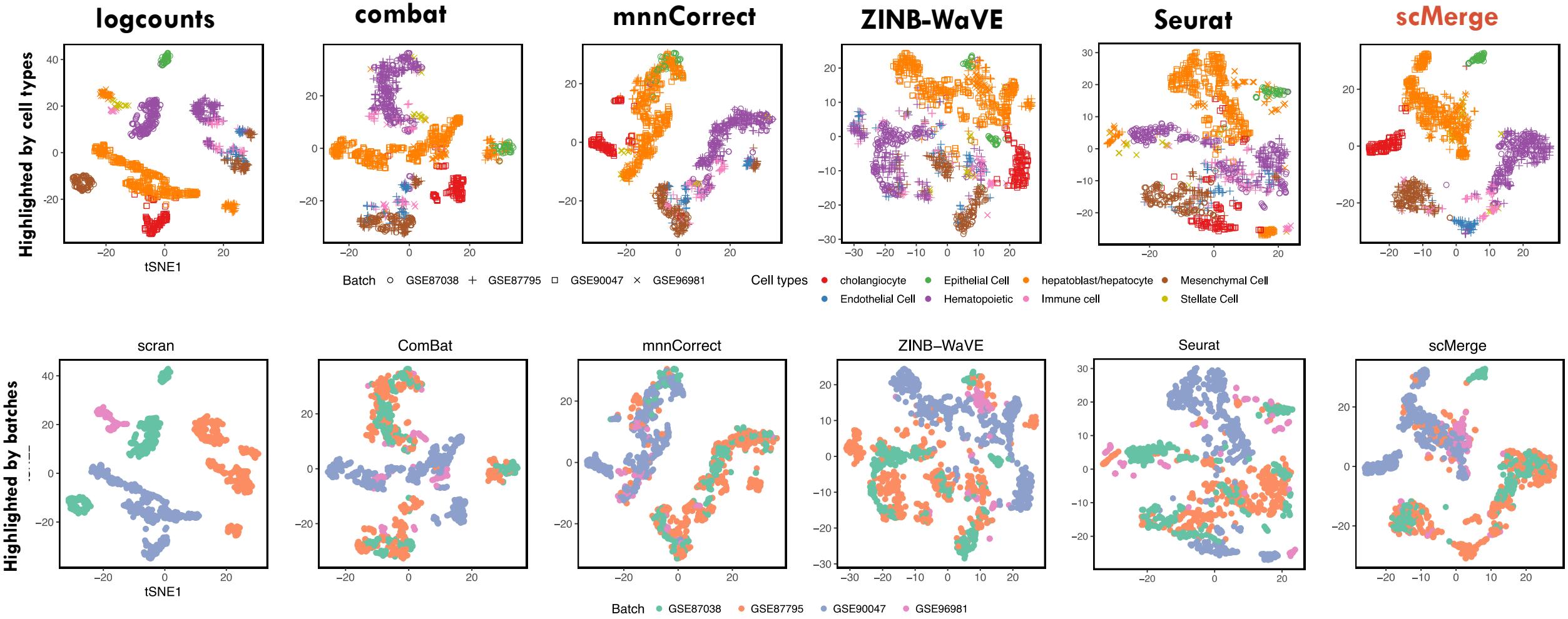




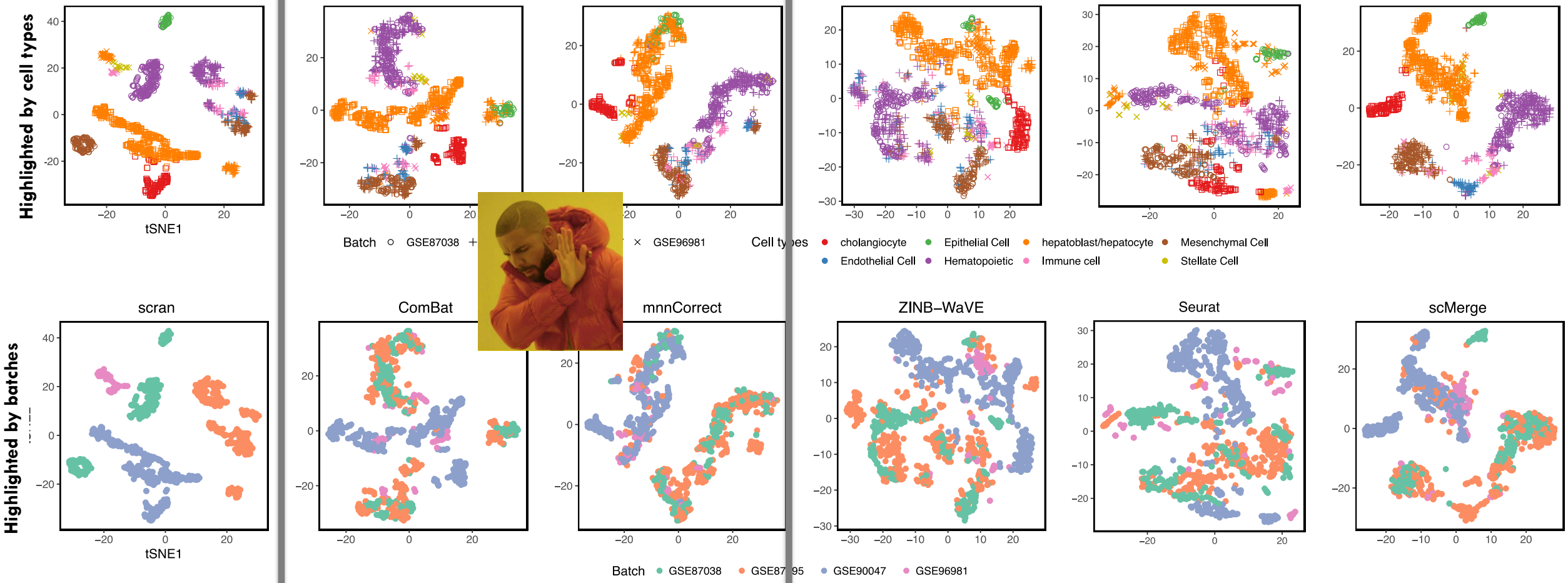
Coming back to our motivational data –
Liver fetal development time course datasets



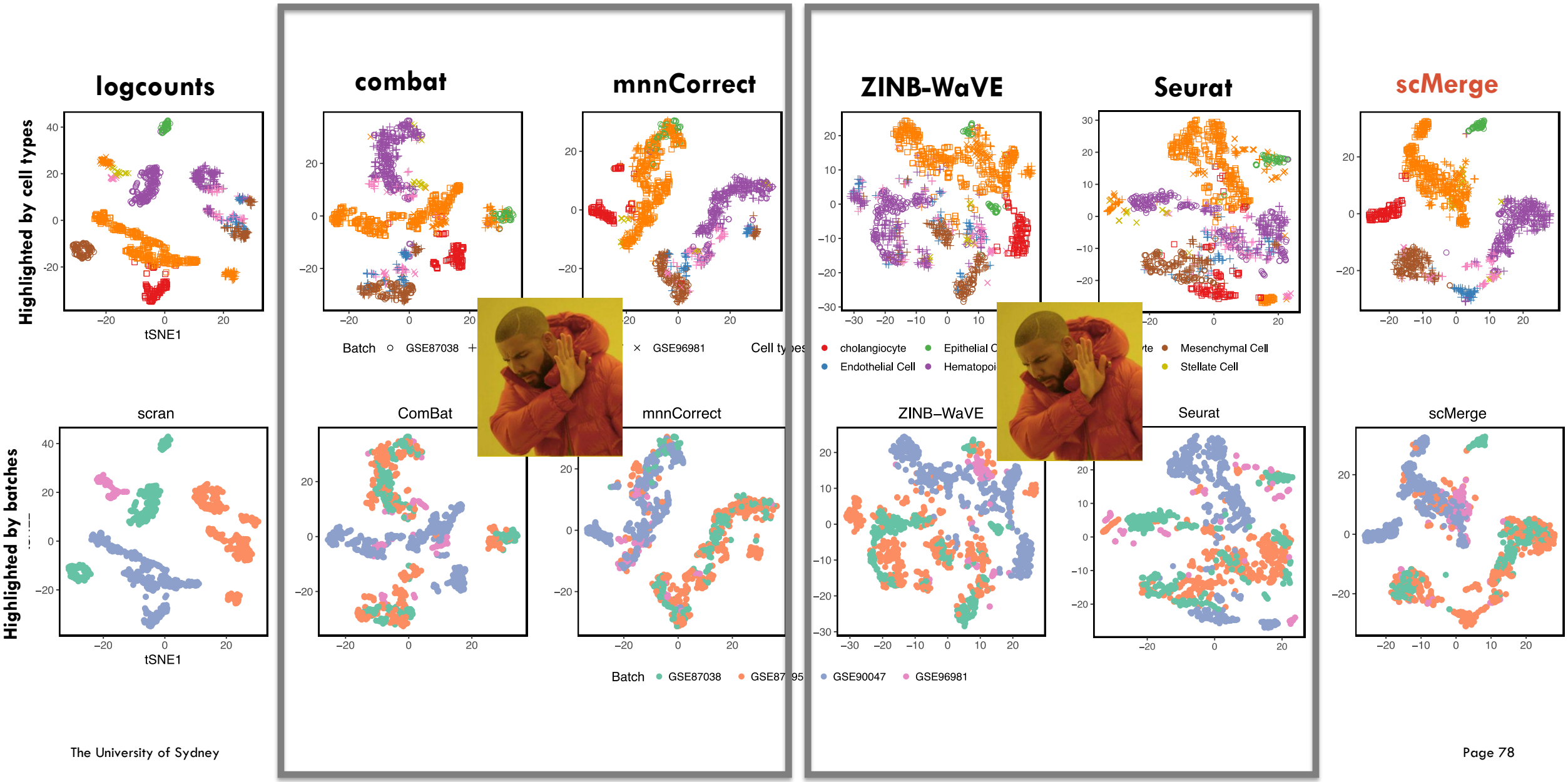
Results: liver datasets tSNE



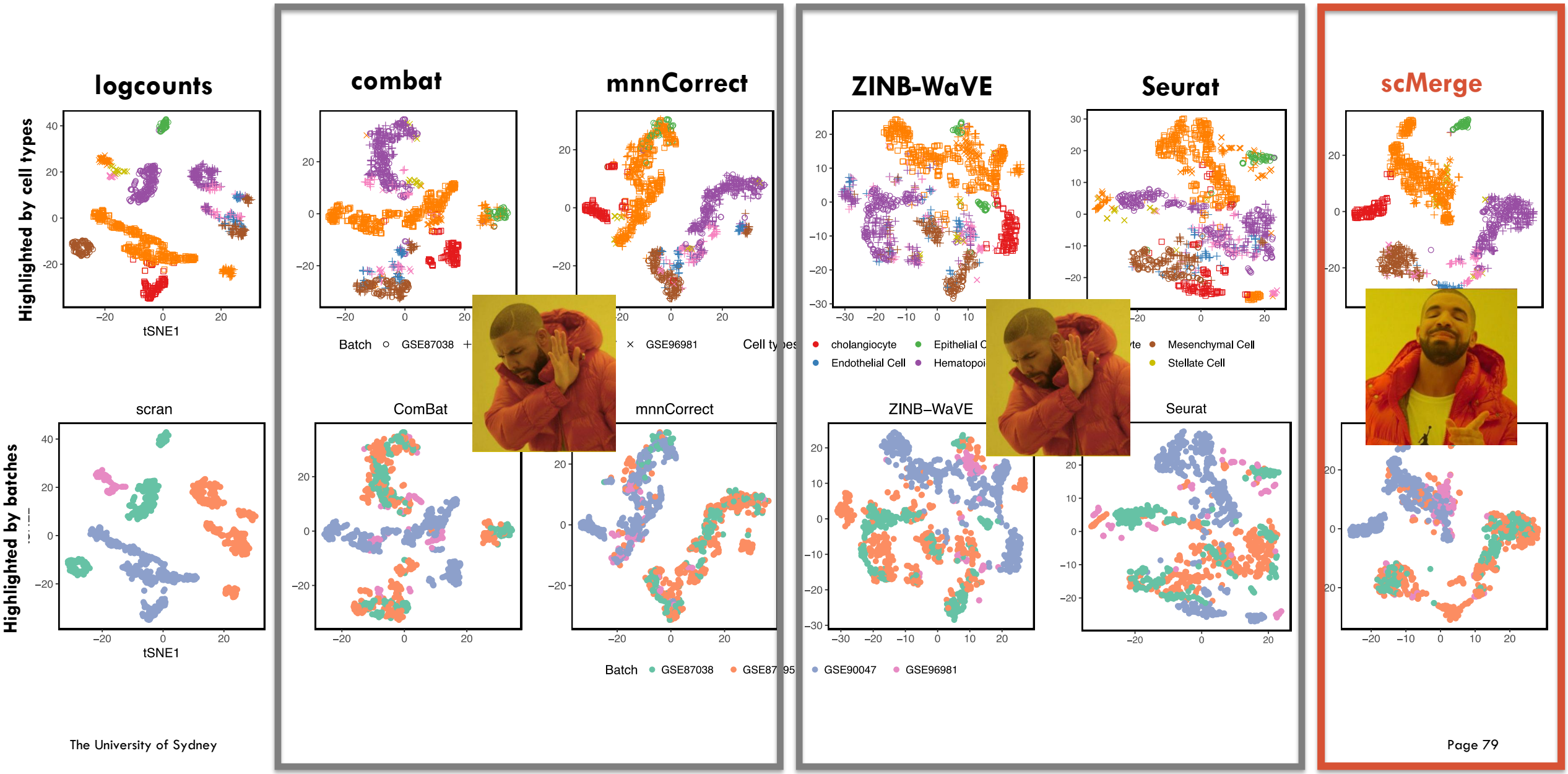
Results: liver datasets tSNE



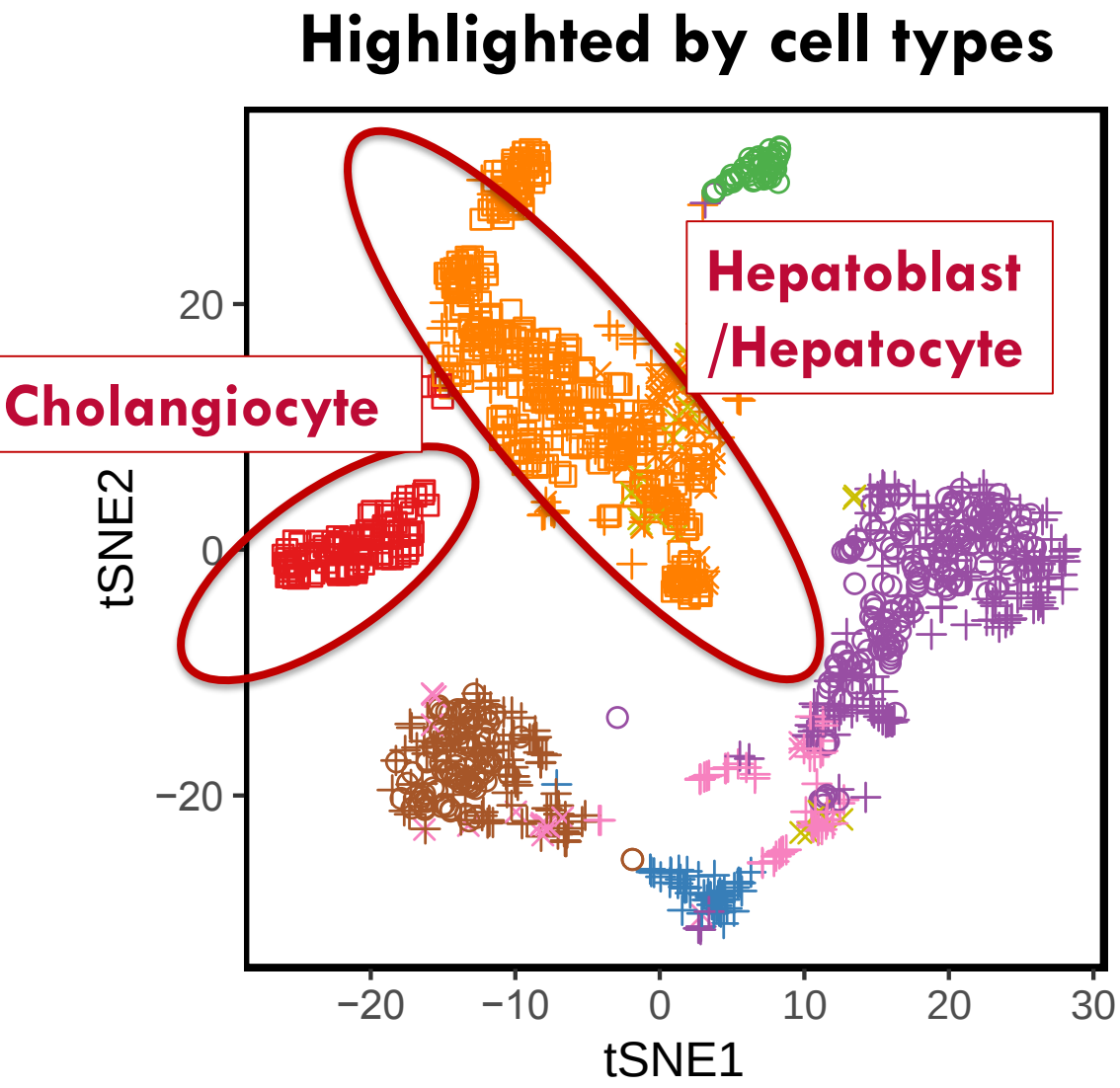
Results: liver datasets tSNE

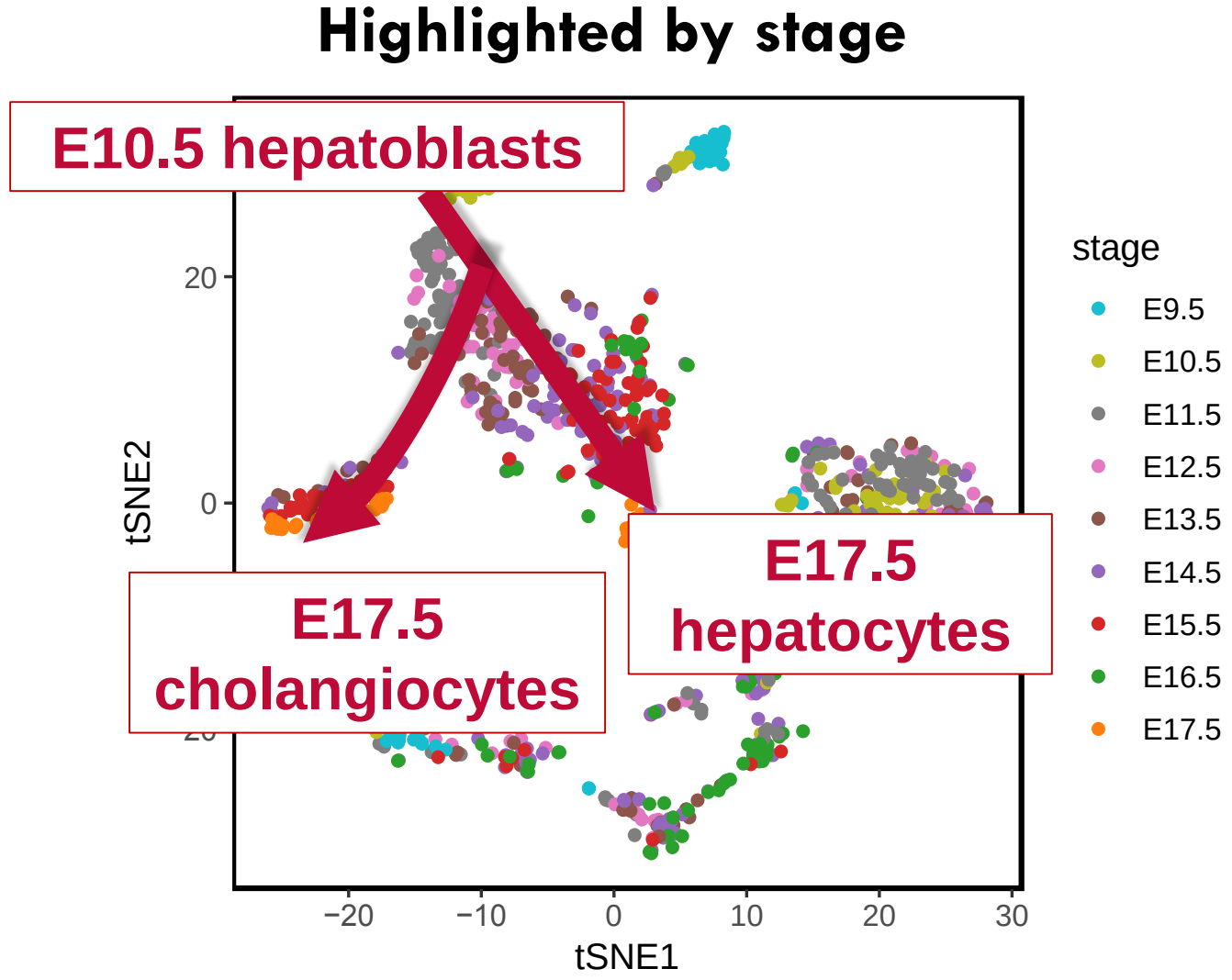


Results: liver datasets tSNE



Results: liver datasets – tSNE retains rough trajectory





More information

bioRxiv:

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

Cold Spring Harbor Laboratory

bioRxiv

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

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", "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.

More information

Thursday workshop!

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

Cold Spring Harbor Laboratory

bioRxiv

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

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.

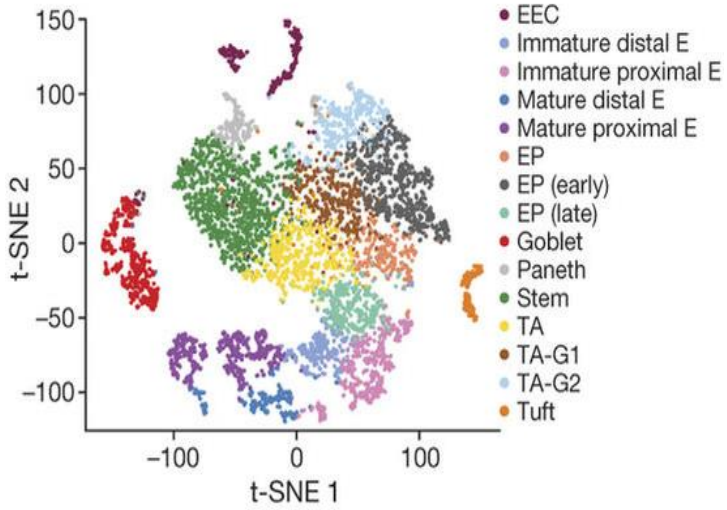
```
# Some CRAN packages required by scMerge
install.packages(c("ruv", "rsvd", "igraph", "pdist", "proxy", "foreach", "doSNOW", "distr", "Rcpp", "RcppEi
devtools::install_github("theislab/kBET")

# 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")
```

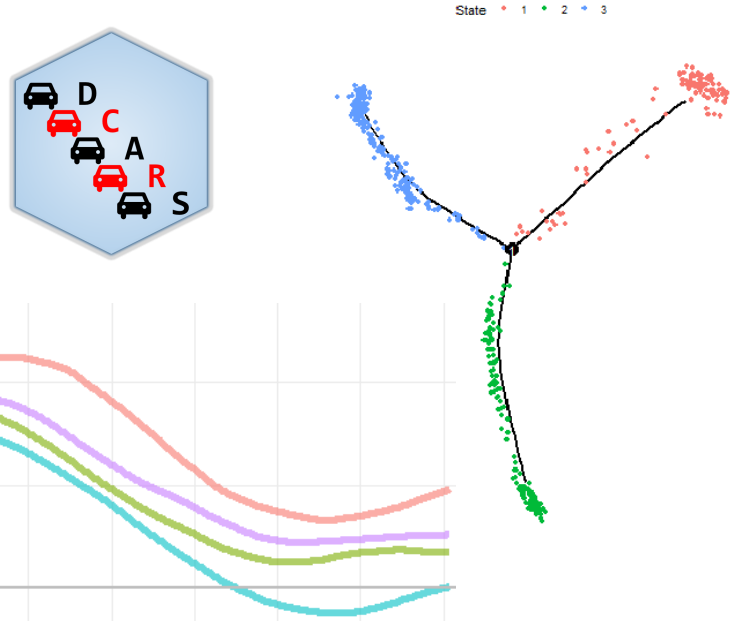
Vignette

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



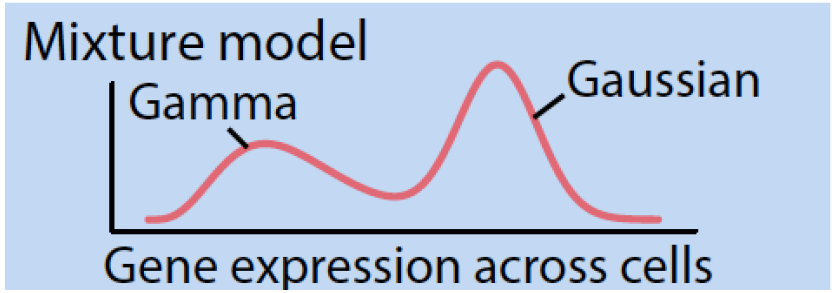
Clustering metrics

Sydney
Single cell

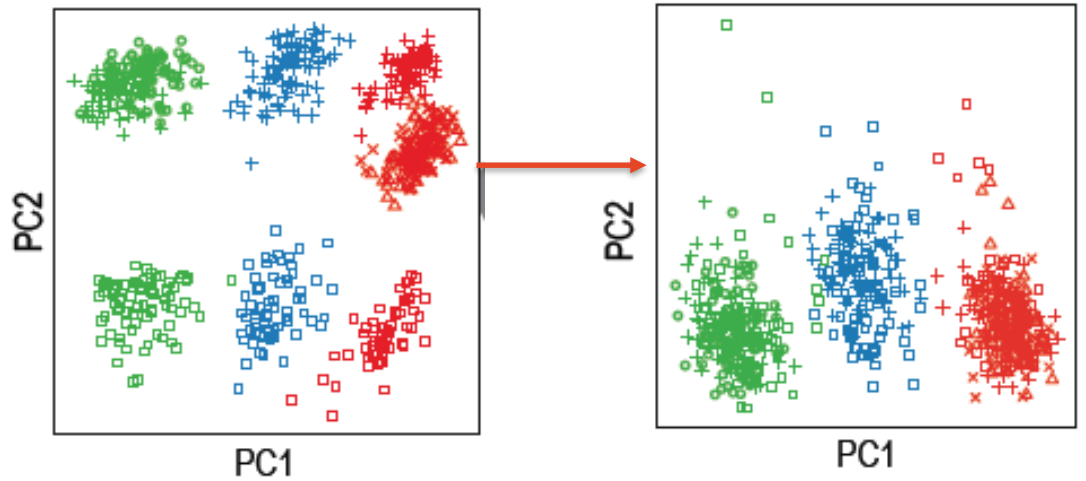


Differential correlation

Finding stably expressed genes



scMerge



Acknowledgements

Usyd School of Mathematics and Statistics

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

Usyd 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