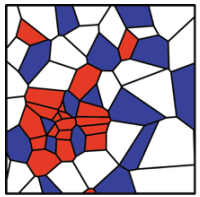
Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations

CCLS seminar, March 2022

Stefan Semrau

Universiteit Leiden
The Netherlands



Semrau lab

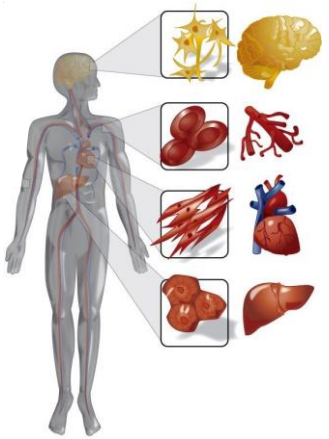
Quantitative single-cell biology

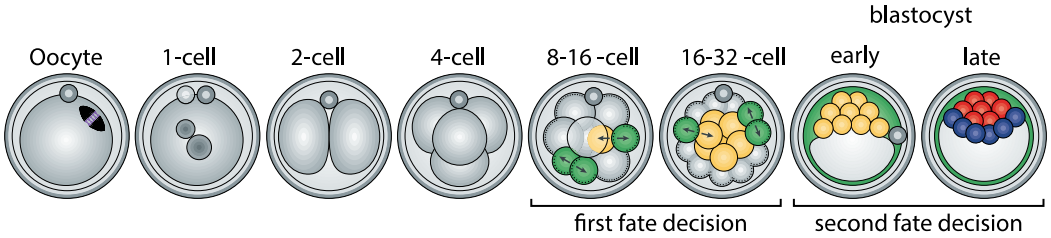
Our motivation: the central question of developmental biology

zygote
(1 cell type)

?

10^{13} cells with 1000s of different cell types

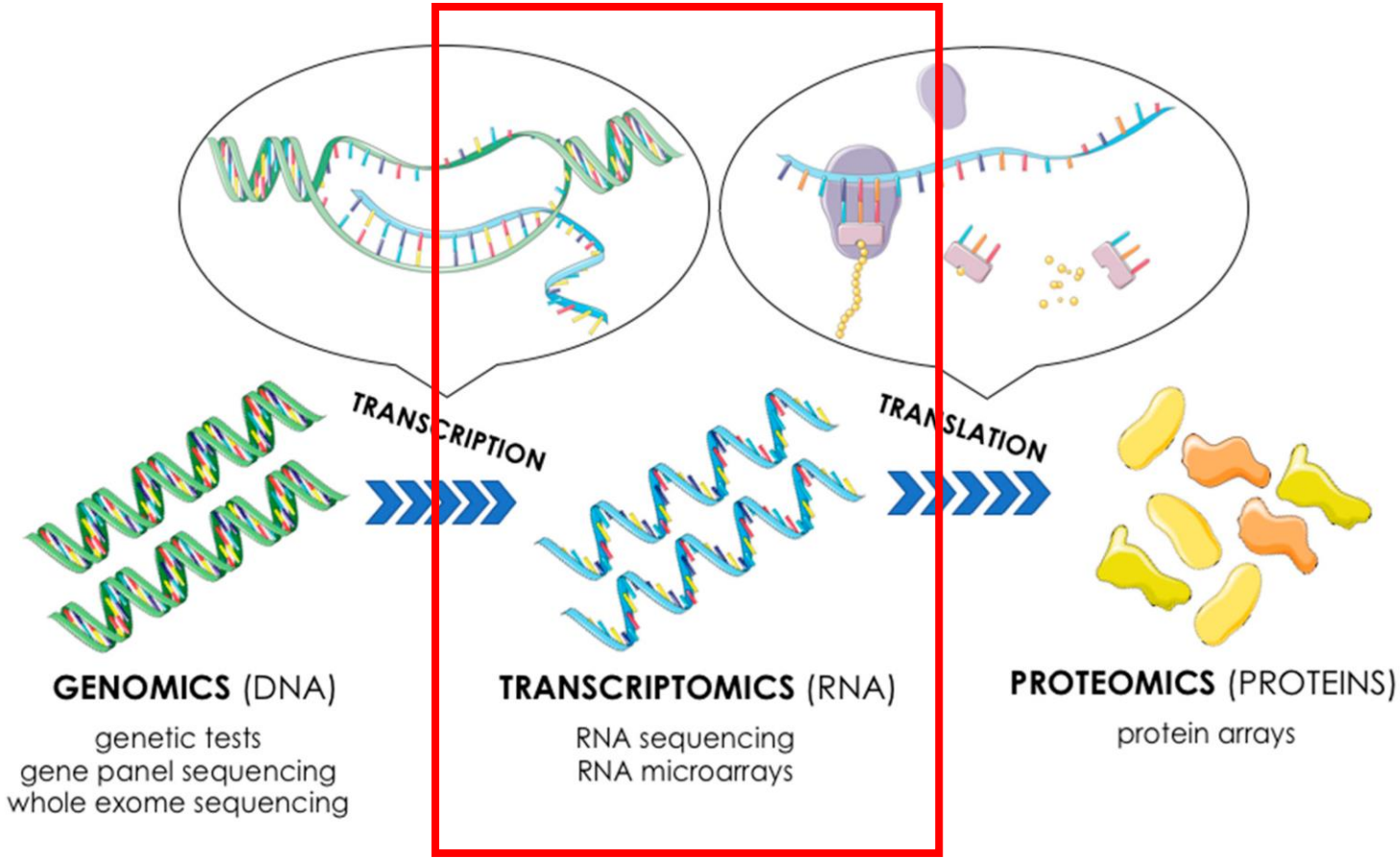




cell type diversity (= complexity) is created by a series of symmetry breaking events

P.W. Anderson, "More is different", Science, 1972

Molecular profiling (aka omics)



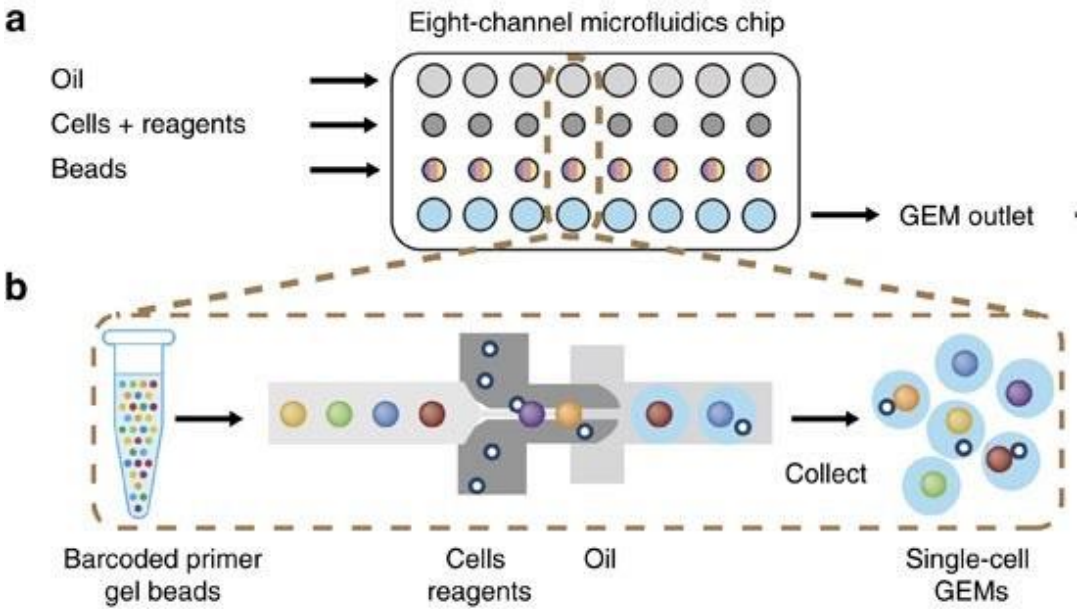
Bulk RNA-sequencing aka “fruit smoothie”

complex tissue

The single-cell smoothie

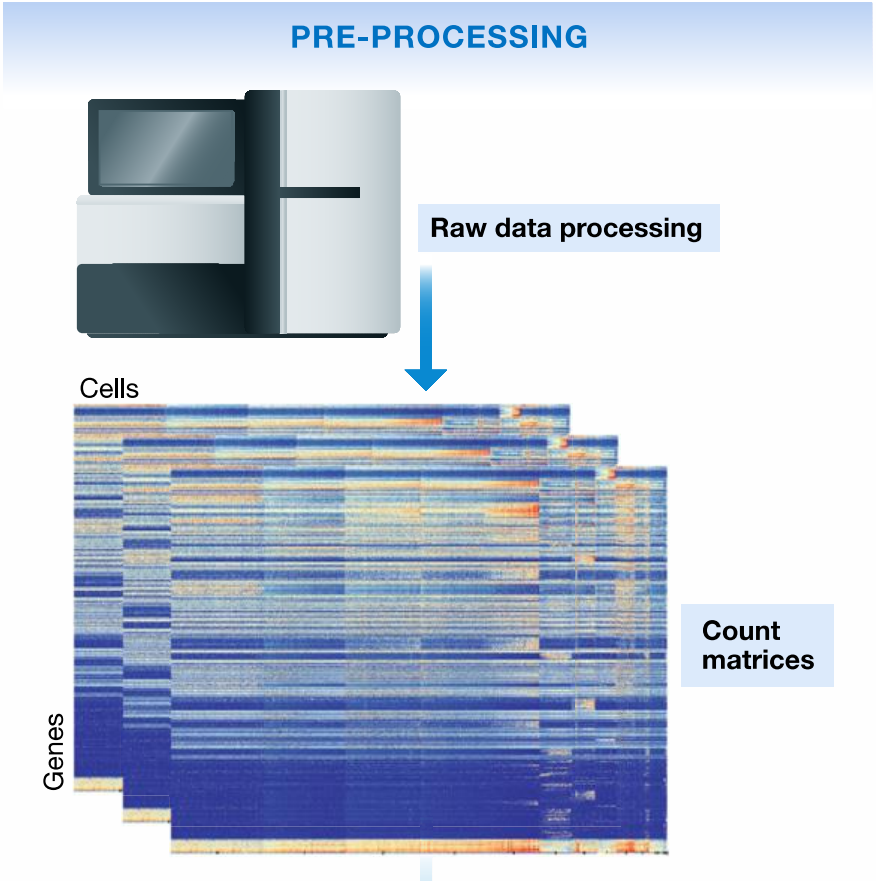
weight of a strawberry = 1 billlion x weight of a single cell

Single-cell RNA-seq

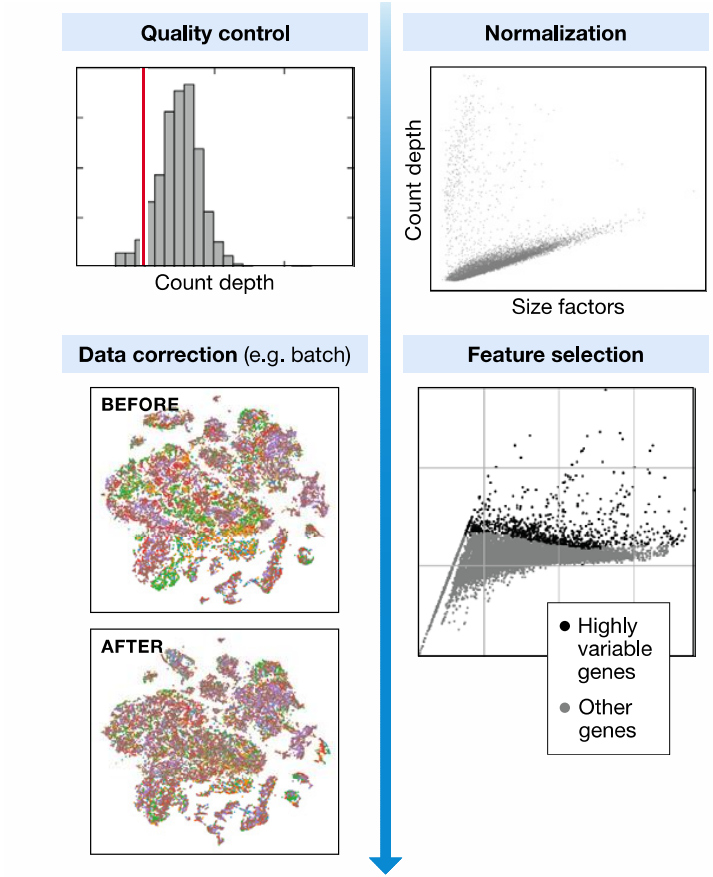


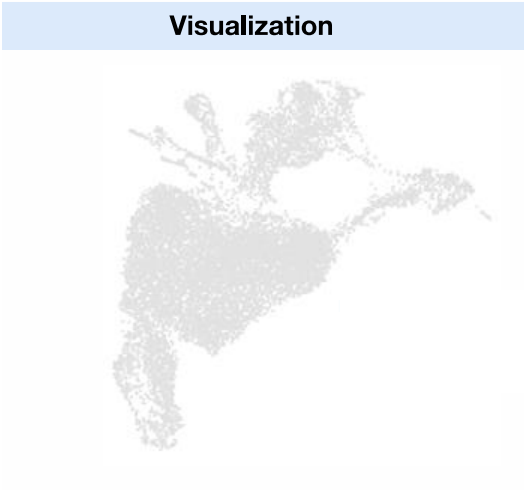
Zheng et al., Nature Comm., 2017

Generic preprocessing



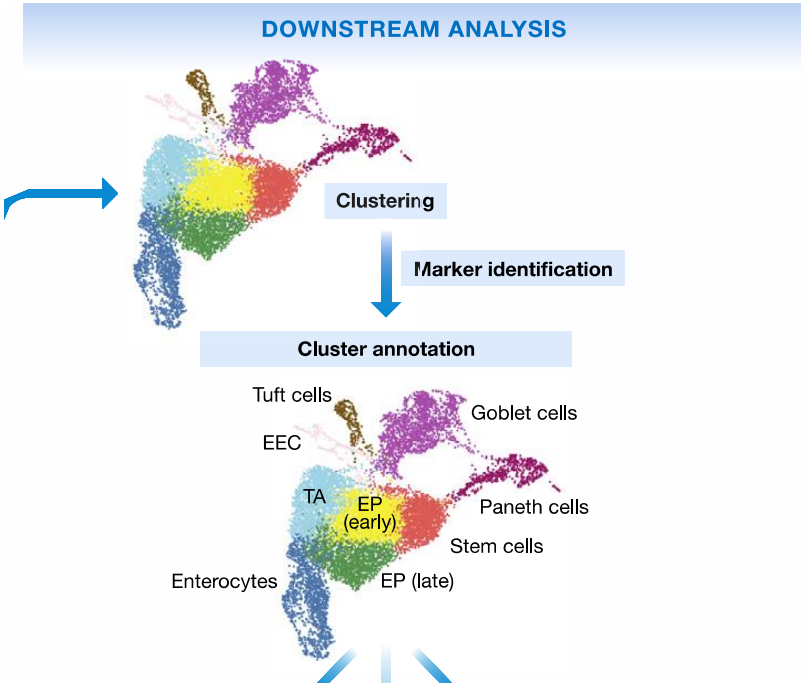
Further, sample-dependent preprocessing

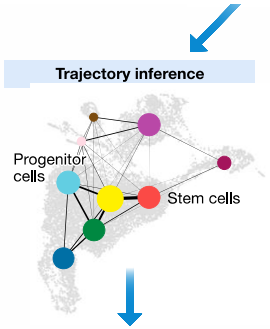


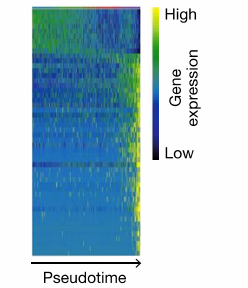


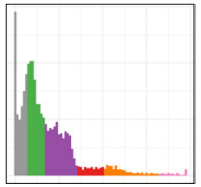
Luecken et al., Mol. Sys. Biol., 2019

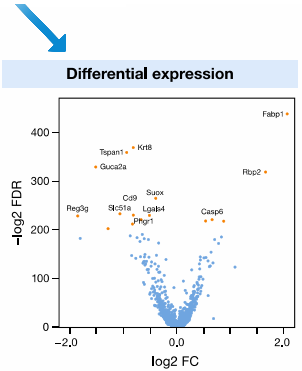
Downstream analysis

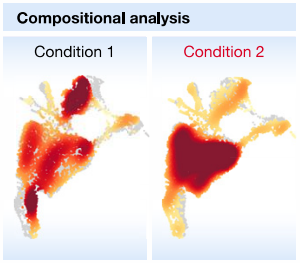






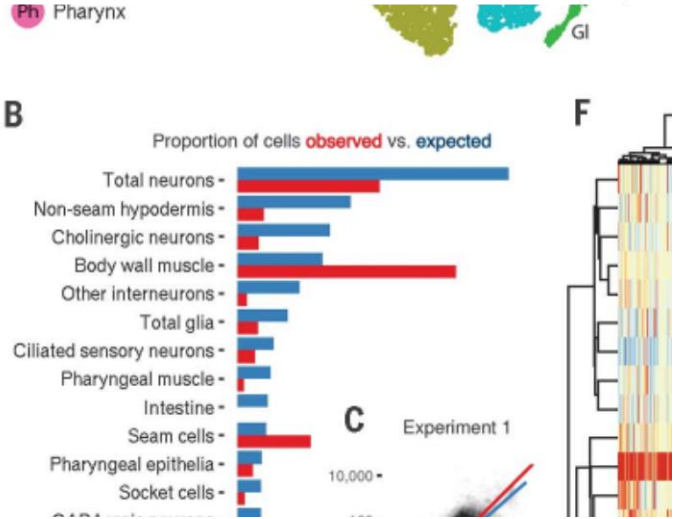






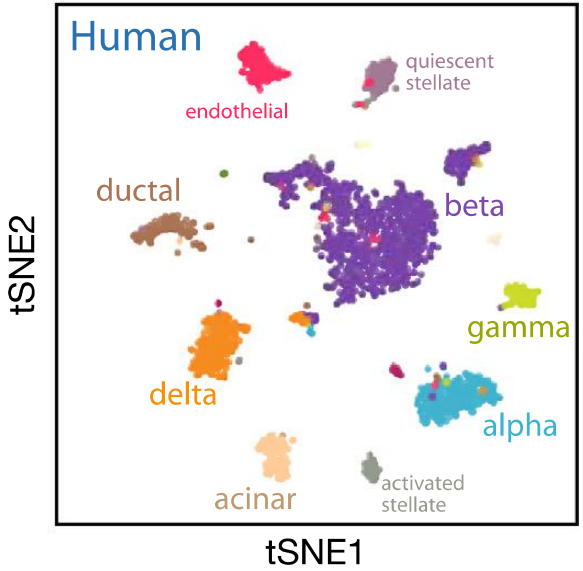
Single-cell RNA-seq examples

Complete C. Elegans worm



Cao et al., Science, 2017

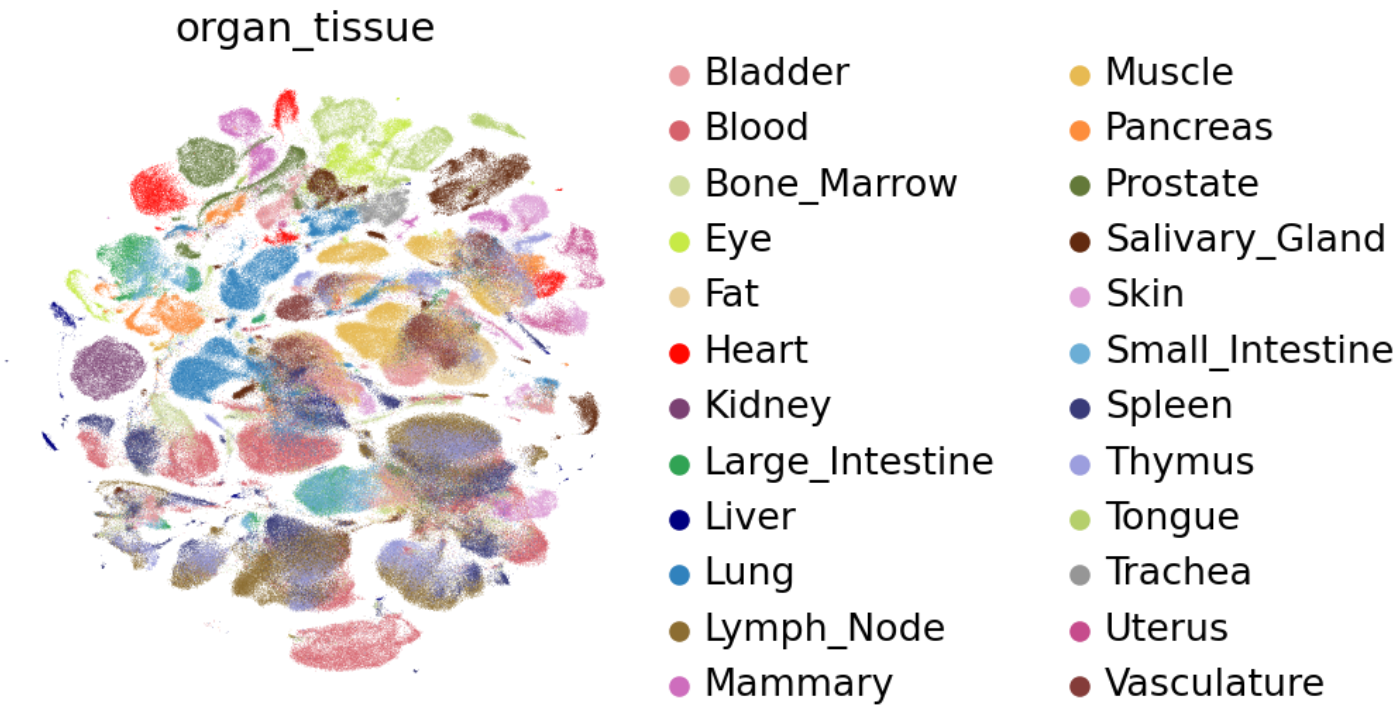
Human pancreas



Baron et al., Cell Systems, 2016

Single-cell RNA-seq examples

Tabula Sapiens



bioRxiv 2021.07.19.452956; doi: https://doi.org/10.1101/2021.07.19.452956

Single-cell transcriptomics of the human fetal kidney

PLOS BIOLOGY

METHODS AND RESOURCES

Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development

Mazène Hochane[1], Patrick R. van den Berg[1], Xueying Fan[2☯], Noémie Bérenger-Currias[1☯], Esmée Adegeest[1☯], Monika Bialecka[2], Maaike Nieveen[2], Maarten Menschaart[1], Susana M. Chuva de Sousa Lopes[2,3‡*], Stefan Semrau[1‡*]

1 Leiden Institute of Physics, Leiden University, Leiden, The Netherlands, **2** Department of Anatomy and Embryology, Leiden University Medical Center, Leiden, The Netherlands, **3** Department of Reproductive Medicine, Ghent University Hospital, Ghent, Belgium

☯ These authors contributed equally to this work.
‡ These authors share equal senior authorship on this work.
* Lopes@lumc.nl (SMCSL); semrau@physics.leidenuniv.nl (SS)

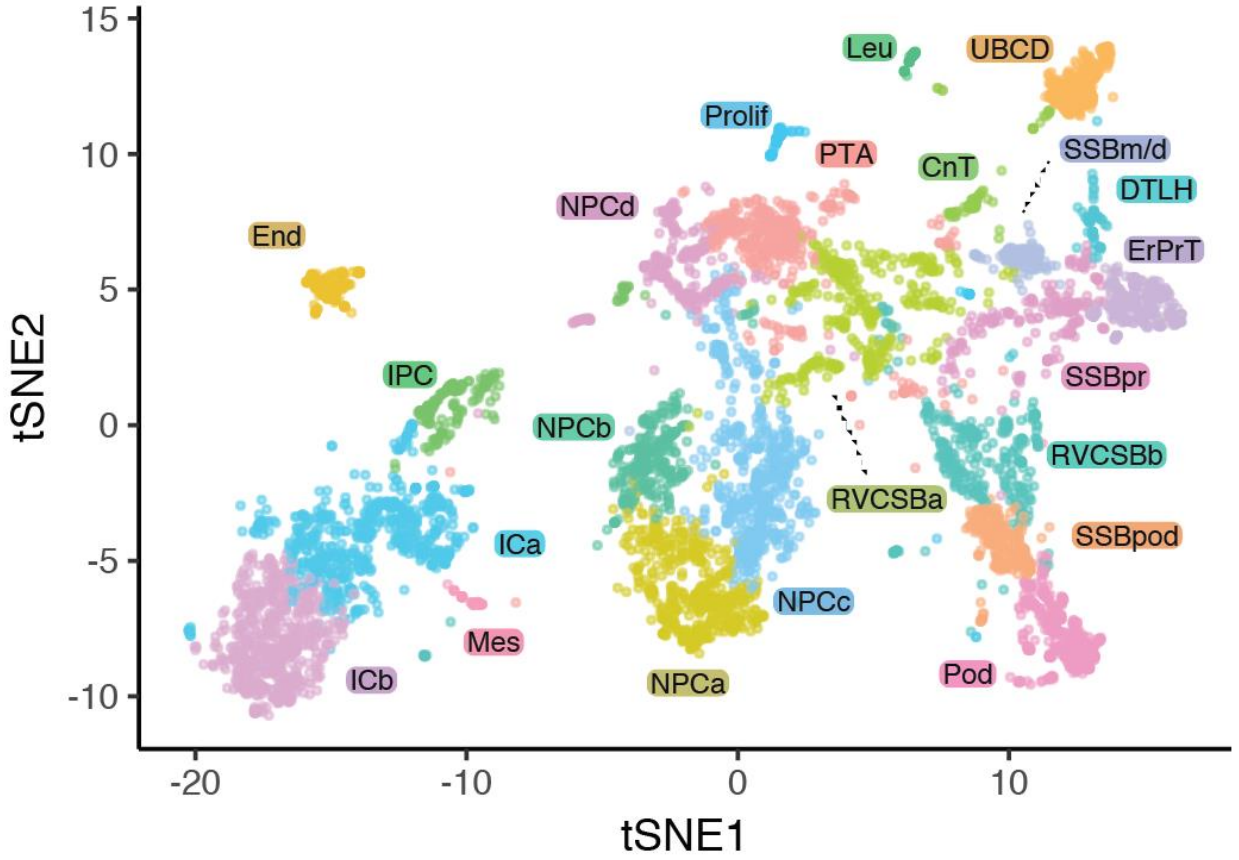
Mazène Hochane

Patrick van den Berg

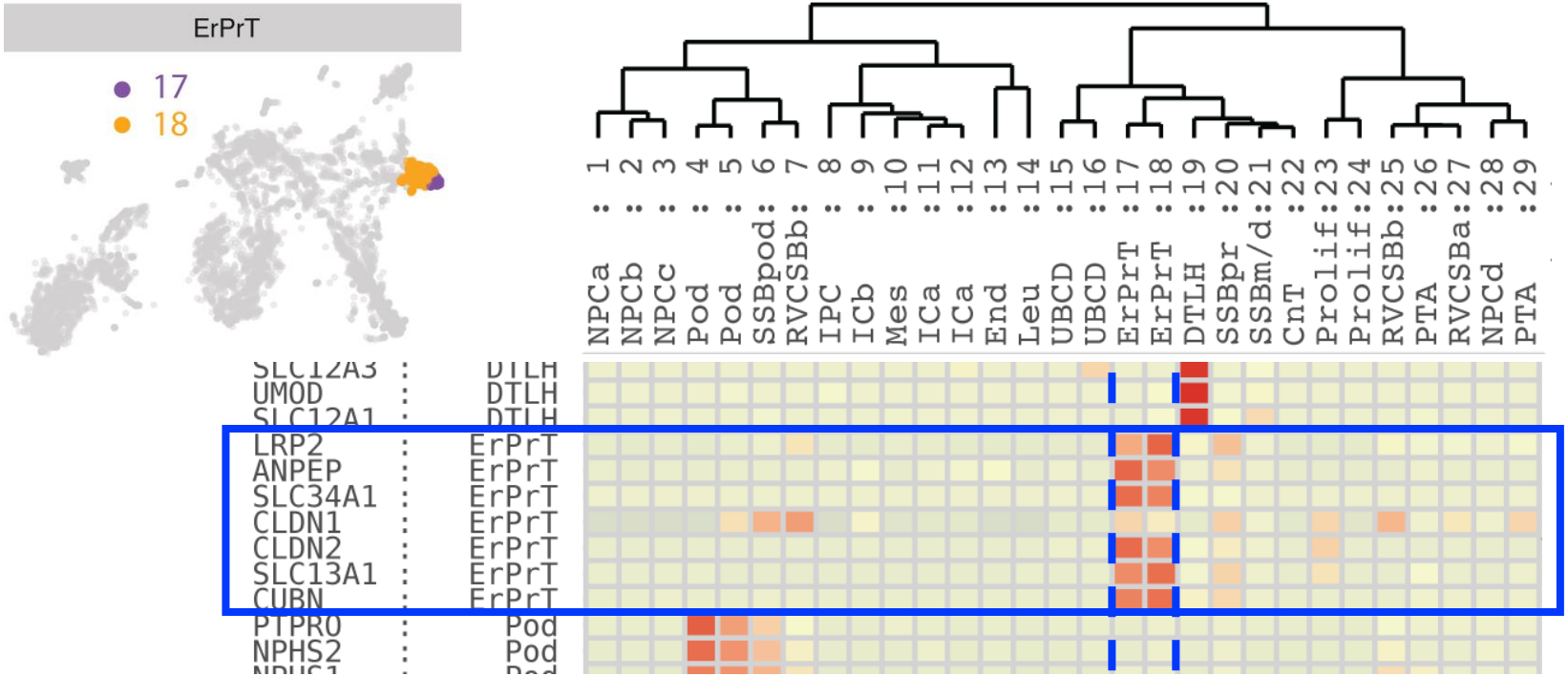
CHUVA SOUSA LOPES LAB

Dept Anatomy and Embryology, Leiden University Medical Center, The Netherlands

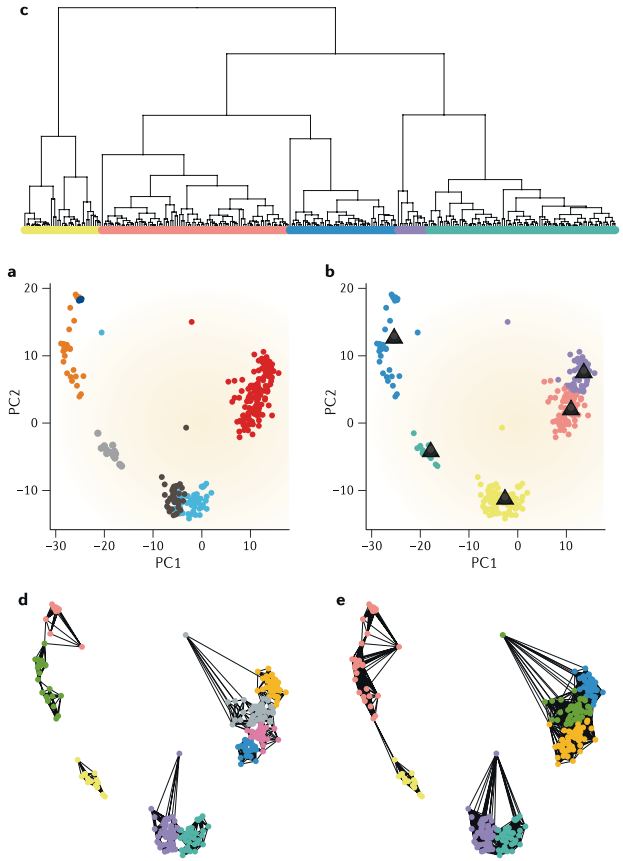
22 cell types could be distinguished



Clusters were merged based on interpretability



All clustering algorithms have tunable parameters



hierarchical clustering

k-means clustering

Louvain / Leiden community detection

Kiselev et al, Nat. Rev. Genetics, 2019

Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations

Mircea *et al. Genome Biology* (2022) 23:18
https://doi.org/10.1186/s13059-021-02590-x

Genome Biology

METHOD Open Access

Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations

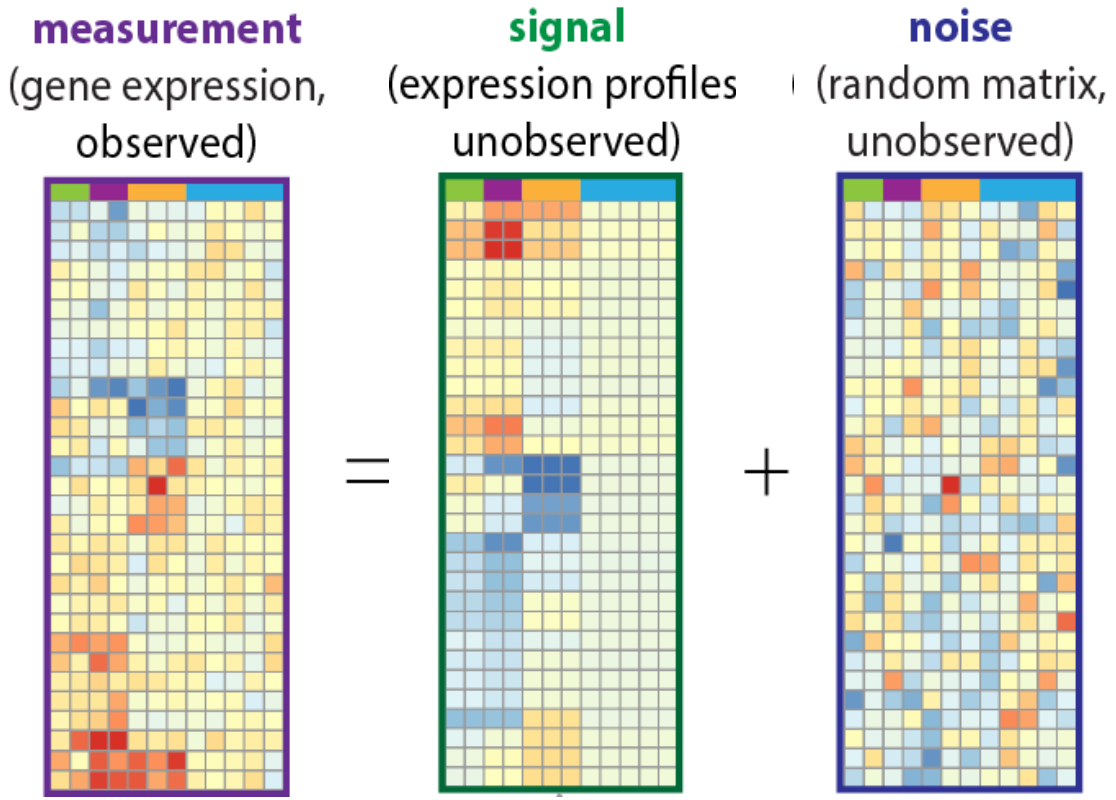
Maria Mircea[1], Mazène Hochane[2], Xueying Fan[3], Susana M. Chuva de Sousa Lopes[3], Diego Garlaschelli[1,4] and Stefan Semrau[1*]

Maria Mircea

Diego Garlaschelli

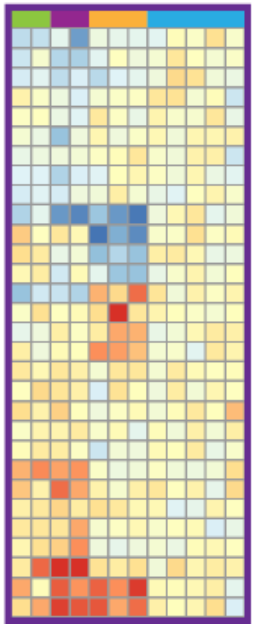
CHUVA SOUSA LOPES LAB
Dept Anatomy and Embryology, Leiden University Medical Center, The Netherlands

measurement = signal perturbed by noise



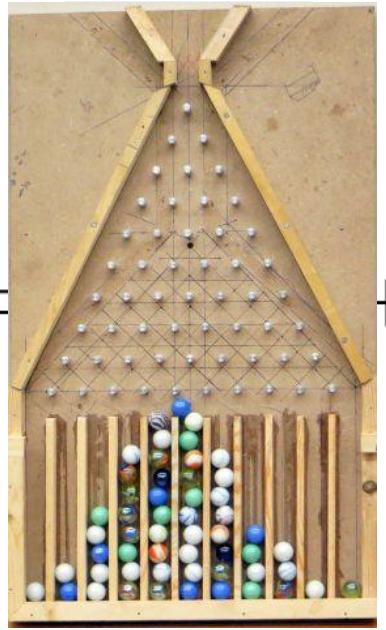
measurement = noise perturbed by signal

measurement
(gene expression, observed)



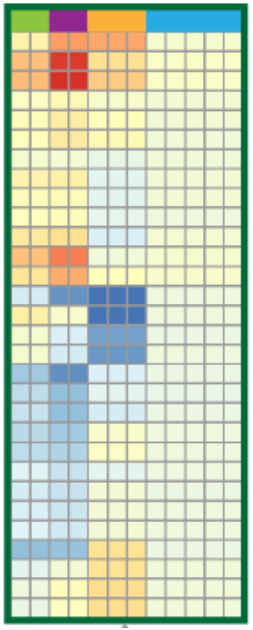
=

noise
(random matrix, unobserved)

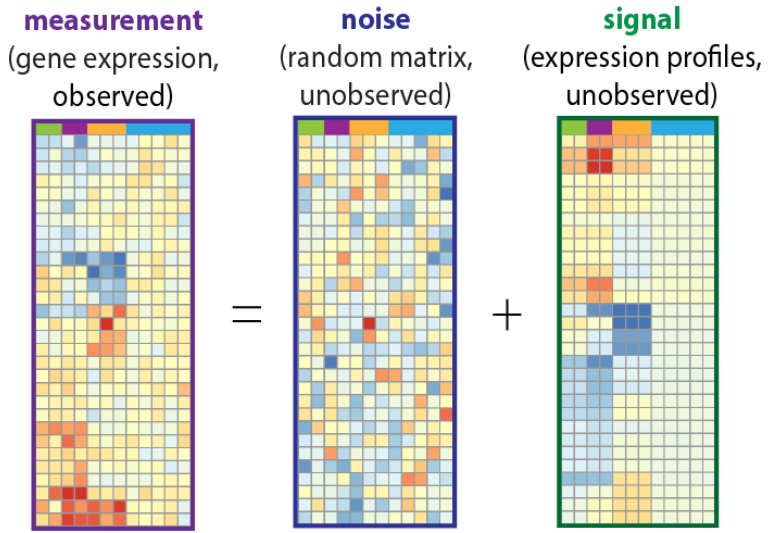


+

signal
(expression profiles, unobserved)



Random matrix theory predicts the singular value distribution



Marchenko-Pastur theorem predicts singular value distribution of covariance matrix for iid random processes with variance σ^2

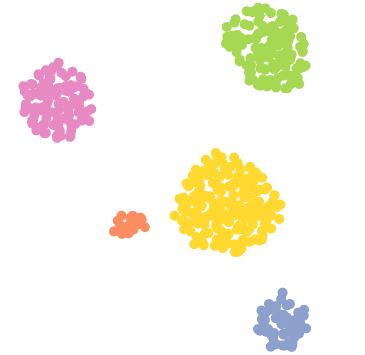
$$f[\lambda] = \begin{cases} \frac{T}{N}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda\sigma^2} & if\ \lambda \in [\lambda_-, \lambda_+] \\ 0 & if\ \lambda \notin [\lambda_-, \lambda_+] \end{cases}$$

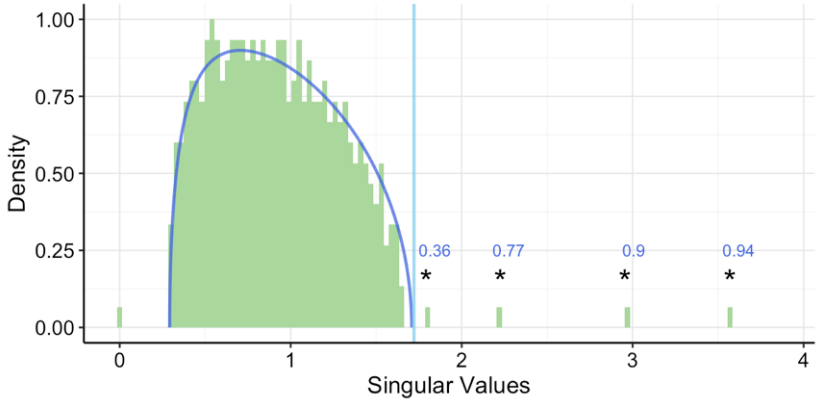
$$\lambda_+ = \sigma^2\left(1 + \sqrt{\frac{N}{T}}\right)^2 \ and\ \lambda_- = \sigma^2\left(1 - \sqrt{\frac{N}{T}}\right)^2$$

T: number of cells

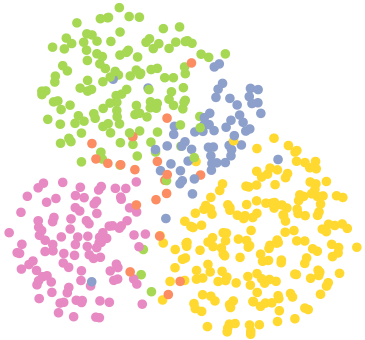
N: number of genes

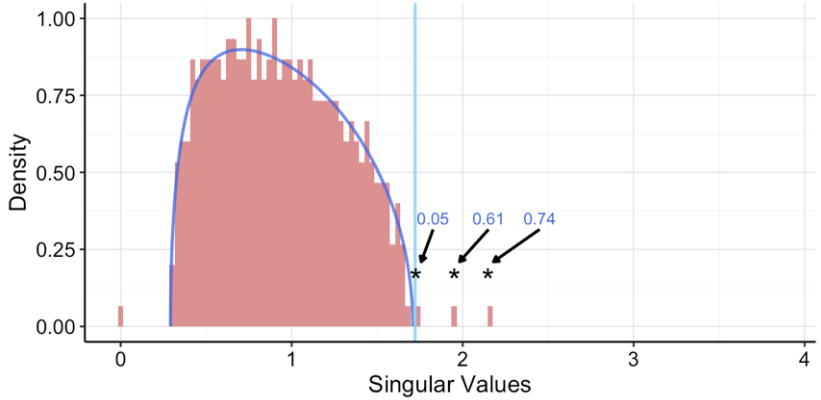
Distance of significant singular values from bulk distribution reflects signal-to-noise ratio





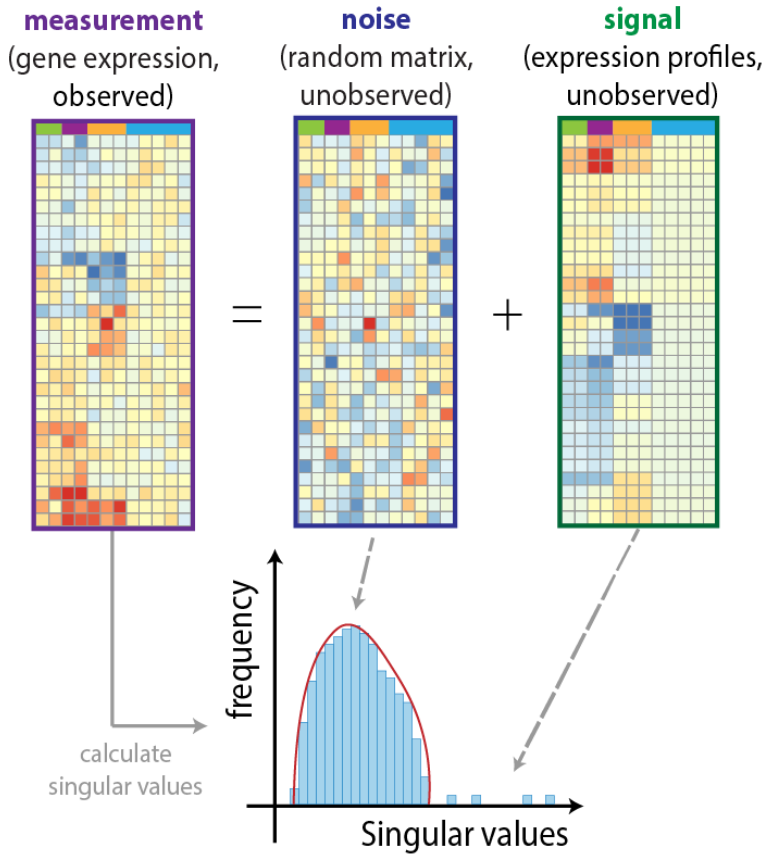
high signal-to-noise





low signal-to-noise

A useful measure can be defined based on the significant singular values



The distance between signal and measurement can be calculated from the singular values

The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices ⋆

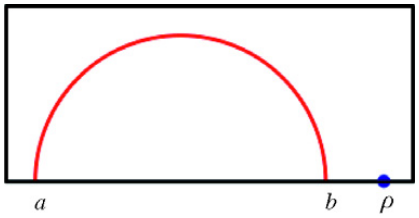
Florent Benaych-Georges [a,b], Raj Rao Nadakuditi [c,□]

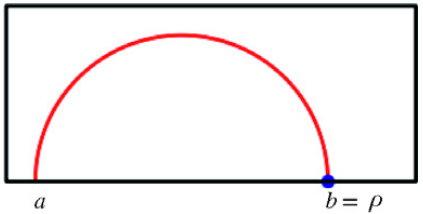
[a] LPMA, UPMC Univ Paris 6, Case courier 188, 4, Place Jussieu, 75252 Paris Cedex 05, France
[b] CMAP, École Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France
[c] Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109, USA

Received 12 October 2009; accepted 9 February 2011

Available online 23 February 2011

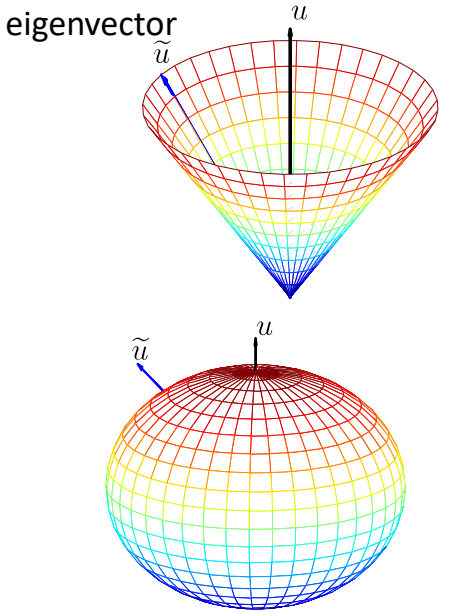
Communicated by Dan Voiculescu





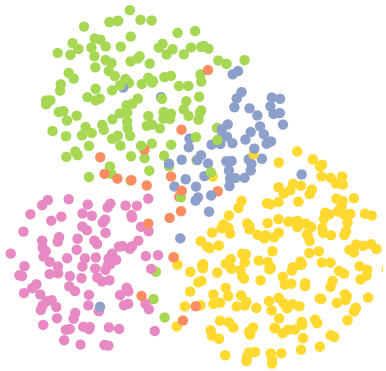
signal eigenvector

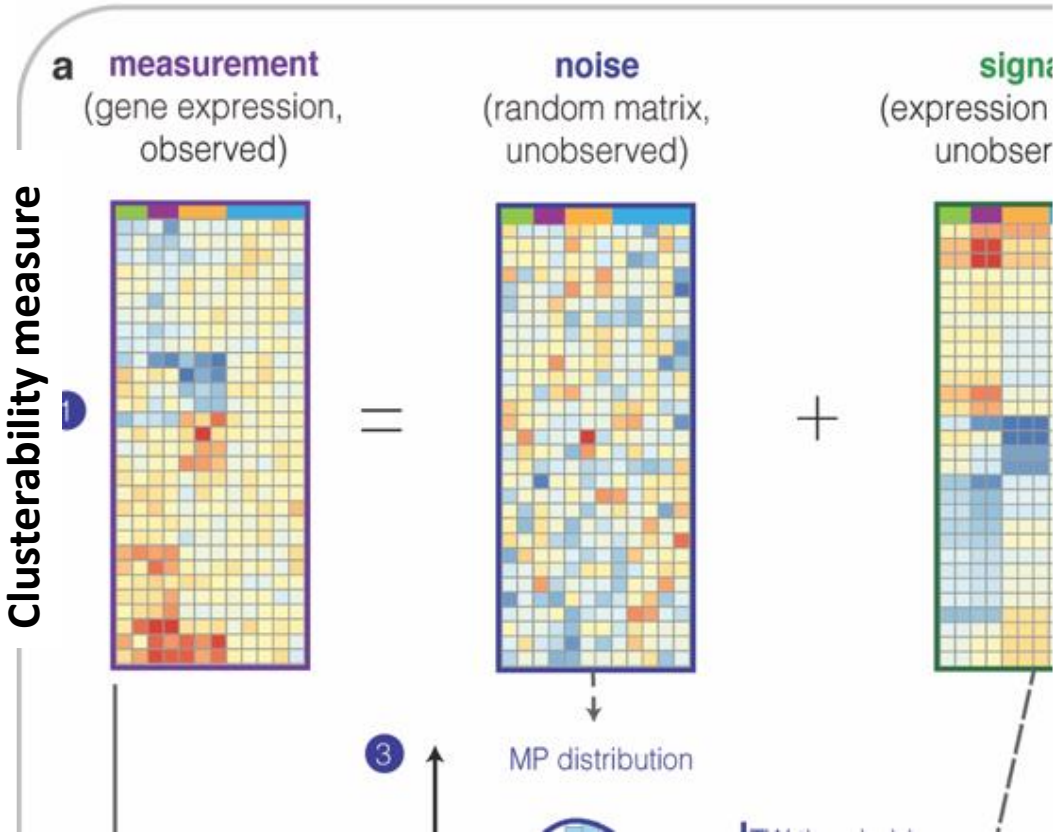
measured eigenvector

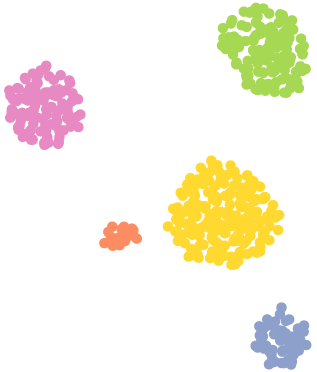


Clusterability measure = $\cos^2$(angle)

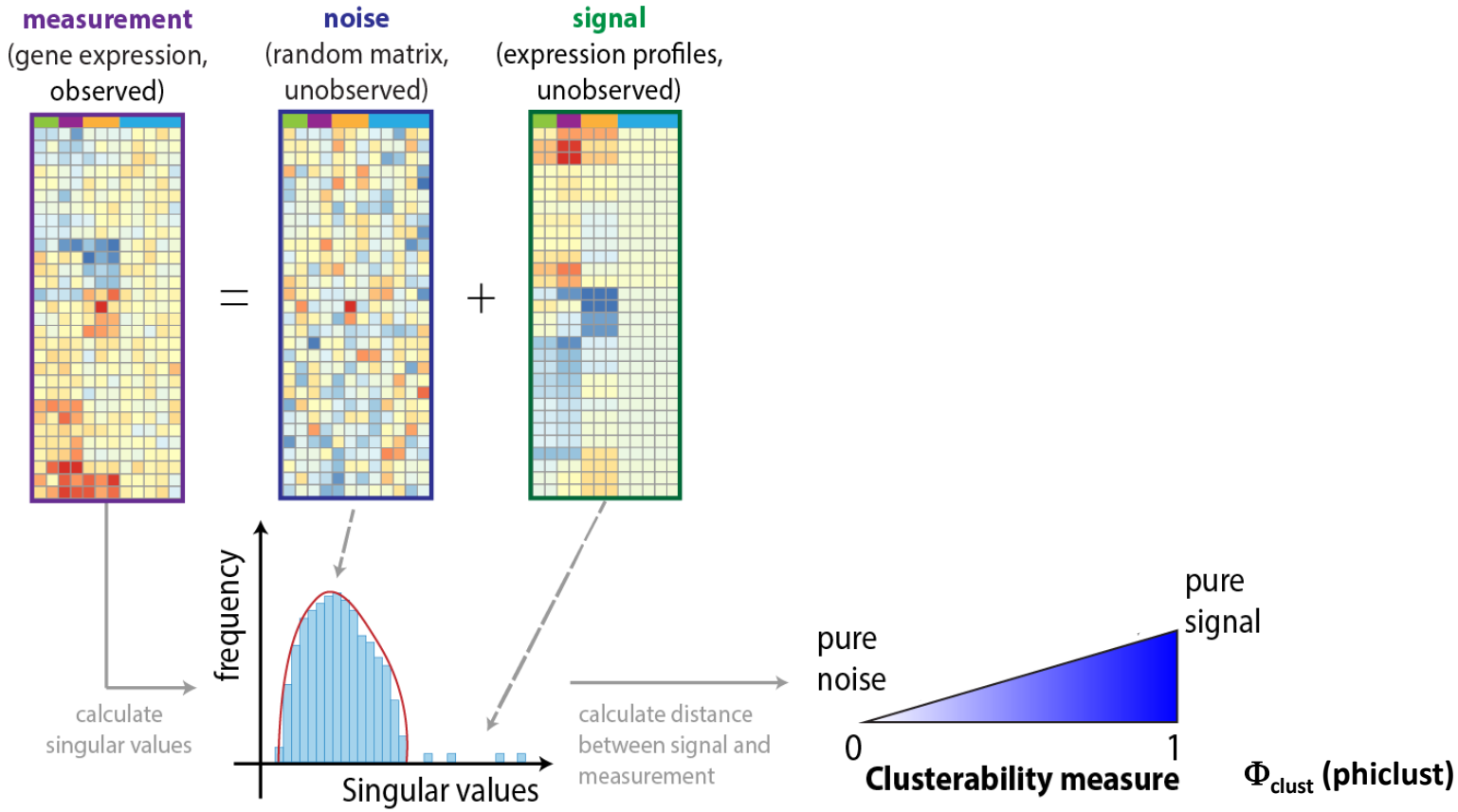
Distance between signal and measurement can be calculated from the singular values







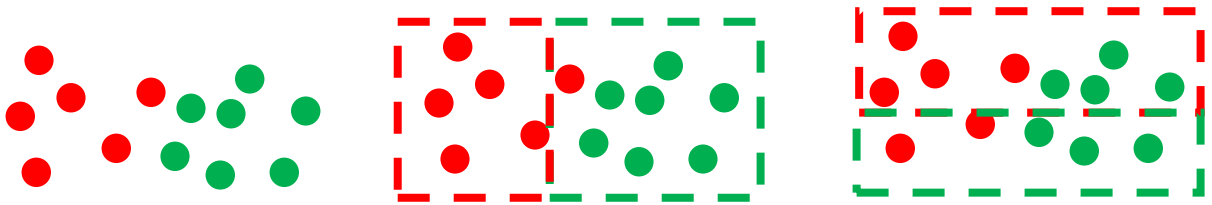
The measure can be shown to relate to clusterability



The adjusted Rand index (ARI) quantifies clustering quality

Rand index RI
measure to assess the quality of a clustering;
ground truth is required; between 0 and 1

$$RI = \frac{\text{number of pairs of cells correctly put in the same cluster} + \text{number of pairs of cells correctly put in different clusters}}{\text{number of all possible pairs of cells}}$$



RI = 66/78 =0.85

good clustering

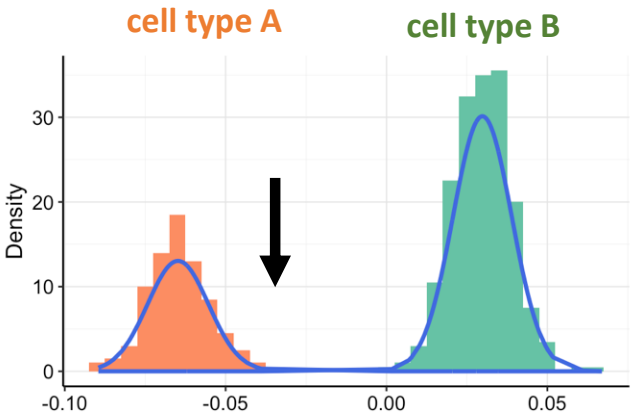
RI = 36/78 = 0.46

bad clustering

Adjusted Rand index ARI
Rand index relative to random clustering

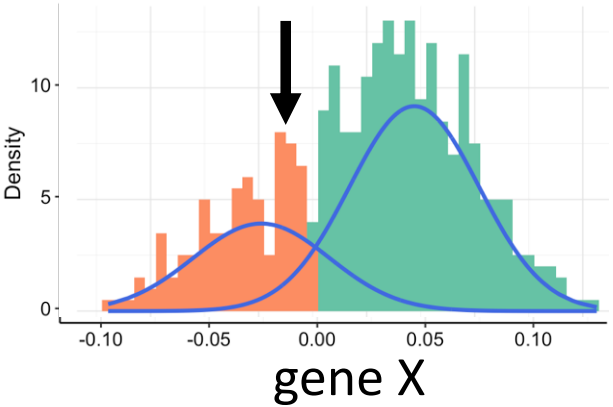
The theoretically achievable ARI (tARI) is limited by the Bayesian error rate

low noise



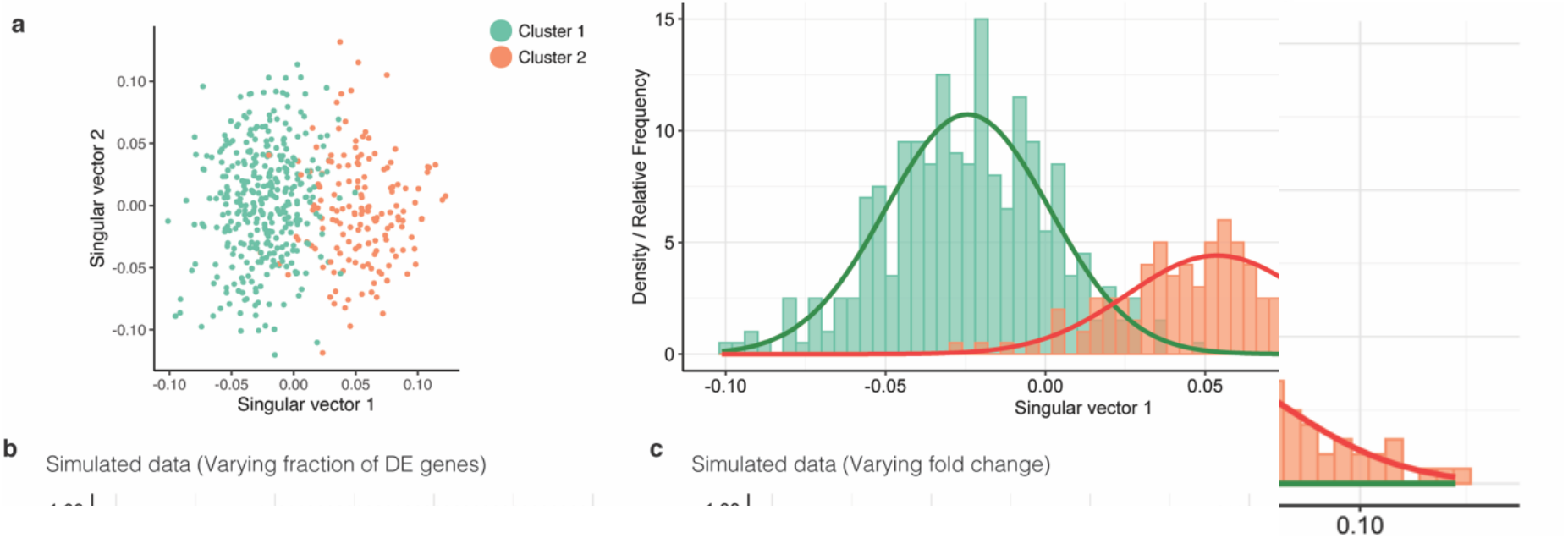
cell can is assigned
to A with low error rate

high noise

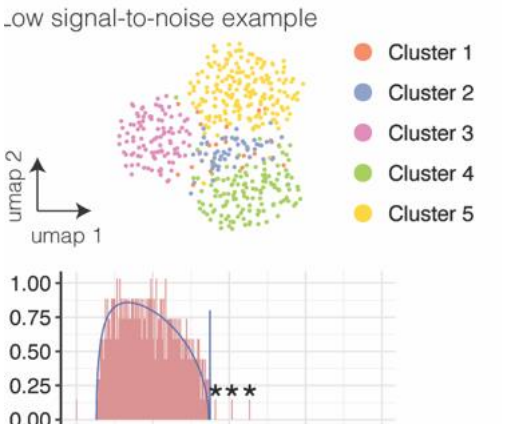


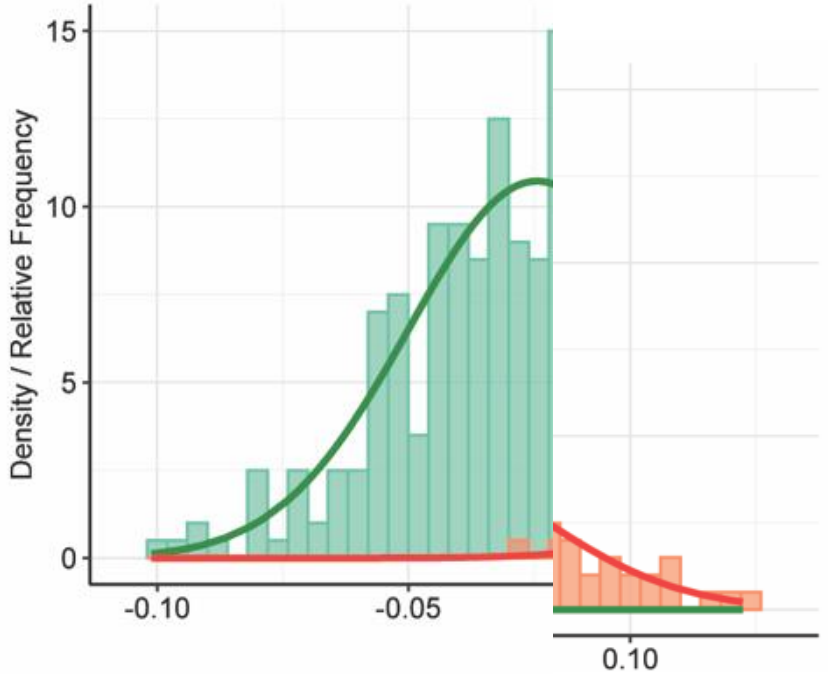
cell is assigned to A
with higher error rate

ARI for simulated data



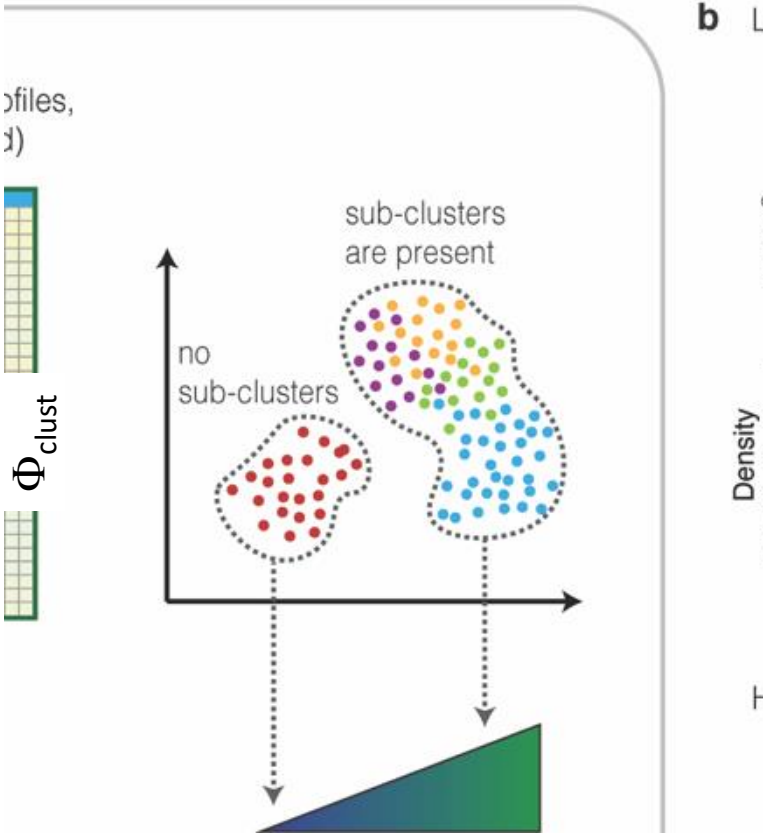
ARI for synthetic data



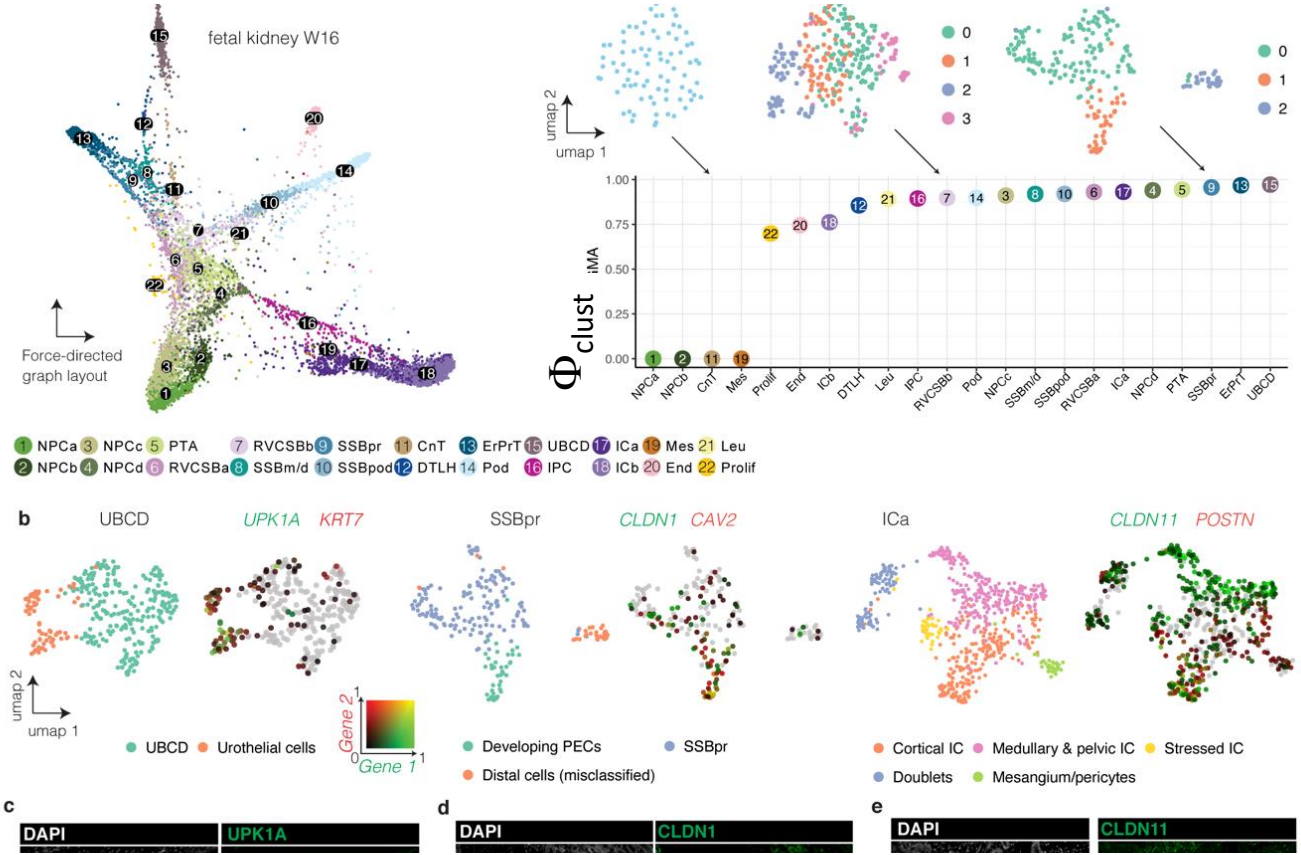


c

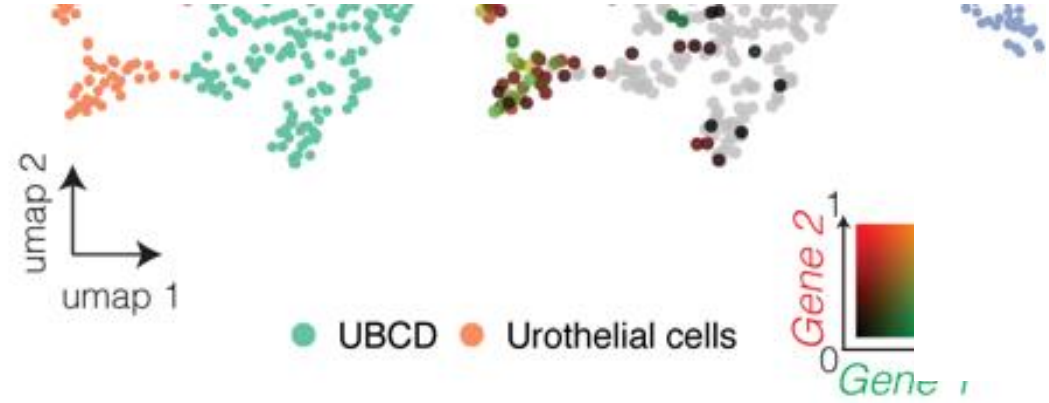
Φ_{clust} is a proxy of the achievable ARI

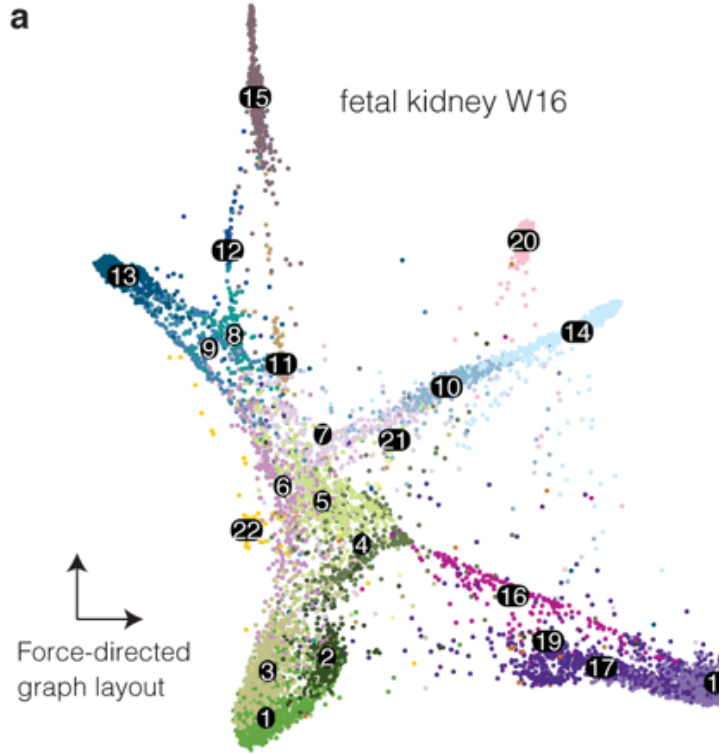


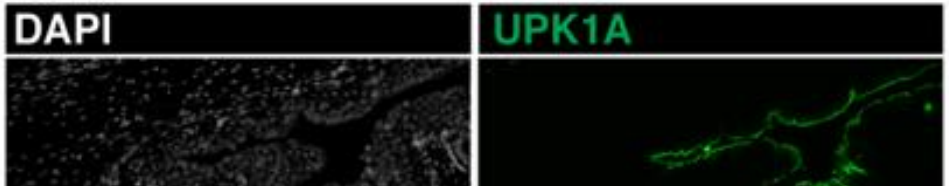
Application to fetal human kidney data



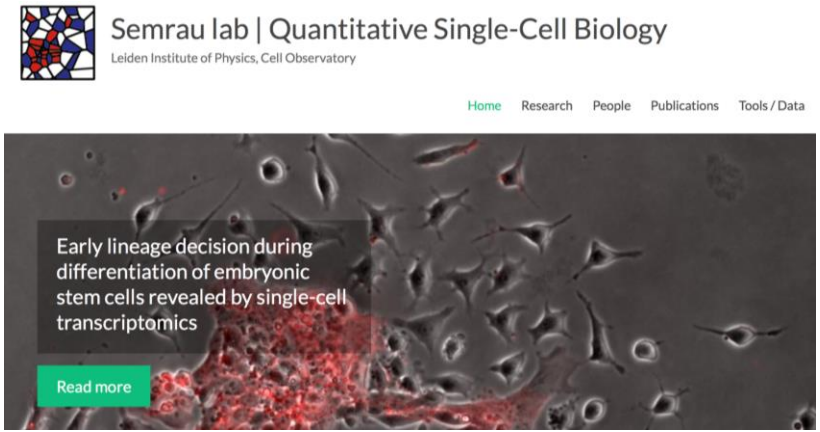
Application to fetal human kidney data







Universiteit Leiden
The Netherlands



www.semraulab.com

Twitter: **@SemrauLab**

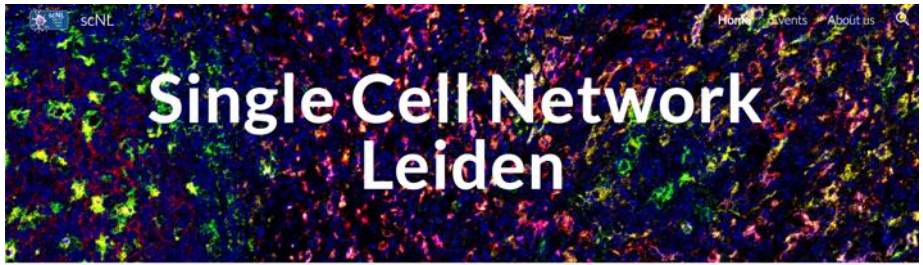
semrau@physics.leidenuniv.n
l

Single-cell Netherlands

scNL
single cell
Network
Leiden

Single Cell Network Leiden

A platform to ***exchange*** experiences, to ***connect*** researchers with complementary expertise, and to ***strengthen*** the single cell community in Leiden



www.singlecell.nl

@scNL4

singlecell.nl@gmail.com

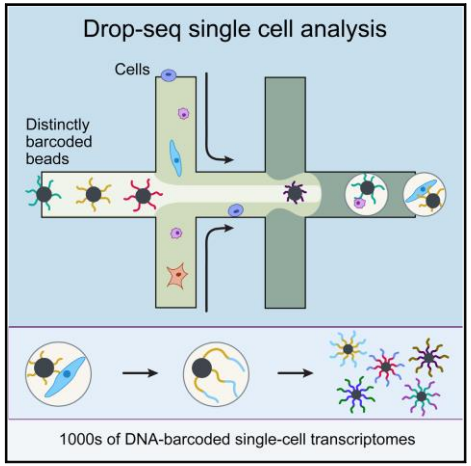
Thank you!

Powered by SURFsara

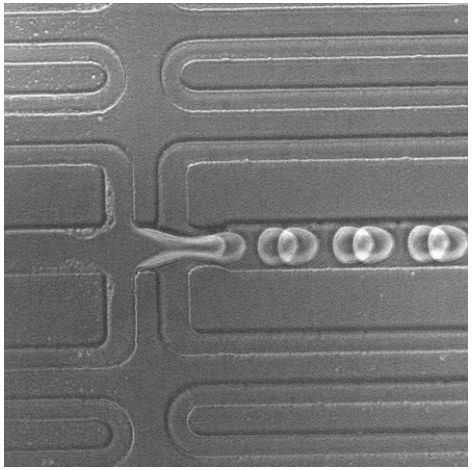
Backup

Experimental challenges of single-cell RNA-seq

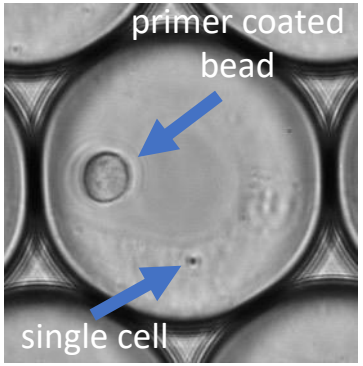
Drop-seq microfluidics

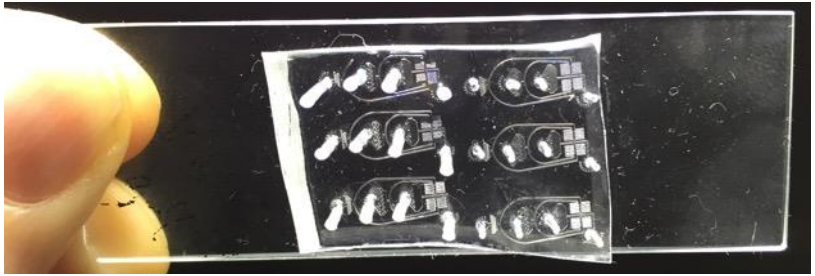


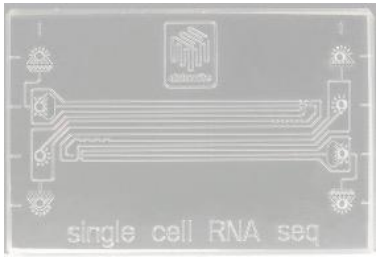
Macosko et al., Cell, 2015



our home made PDMS device
(1000 libraries / 5 min)

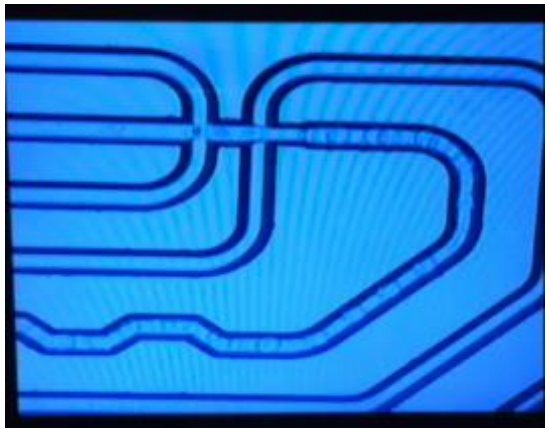


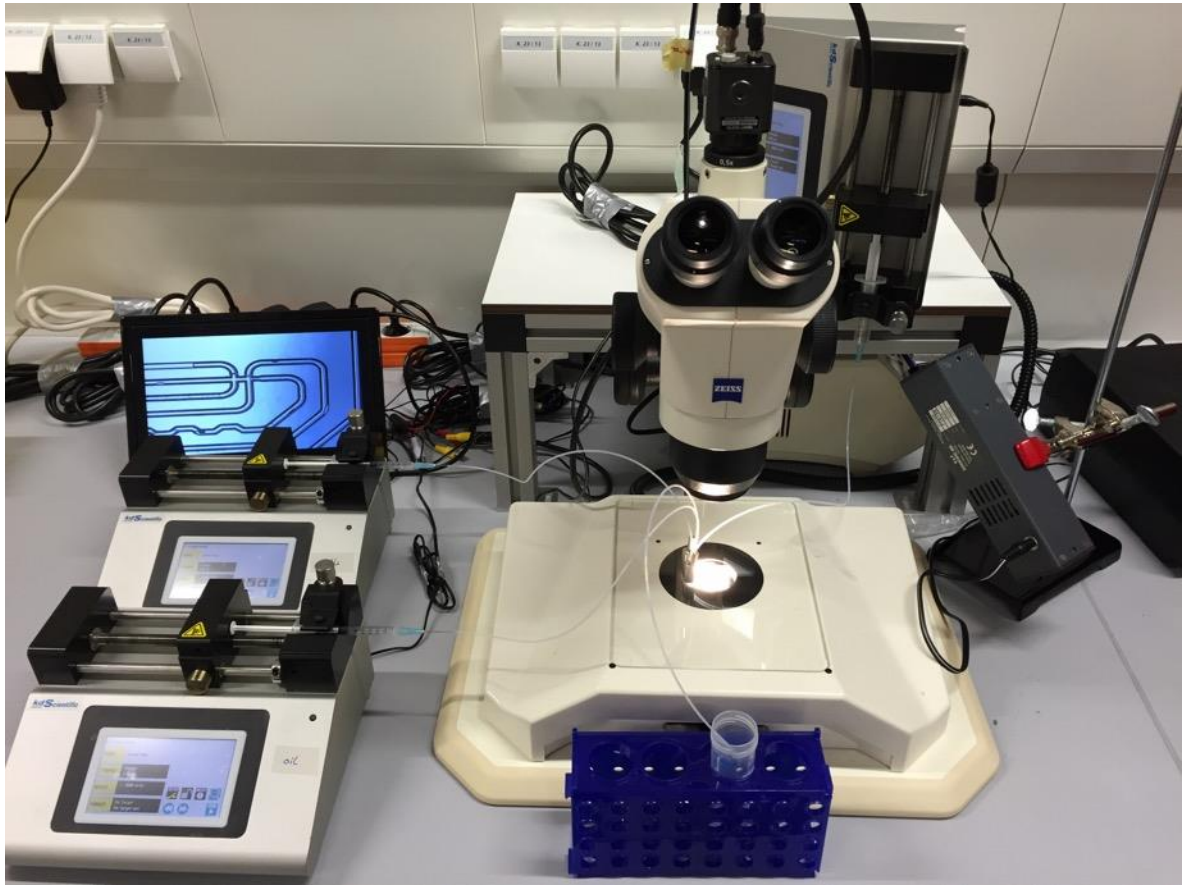




dolomite microfluidics

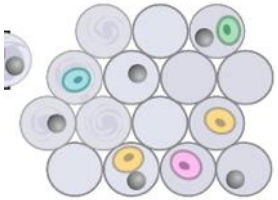
Drop-seq setup



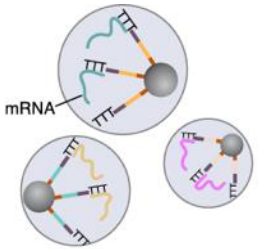


Single-cell RNA-seq principle (drop-seq)

1. cell co-encapsulation and lysis

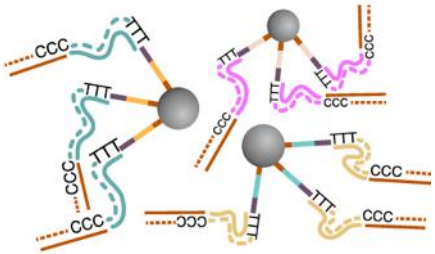


2 . capture of transcripts on primer coated bead

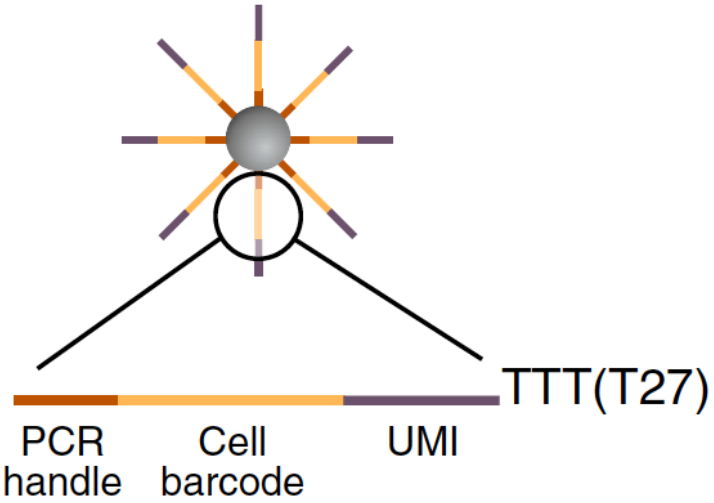


3. droplet breakage (= pooling)

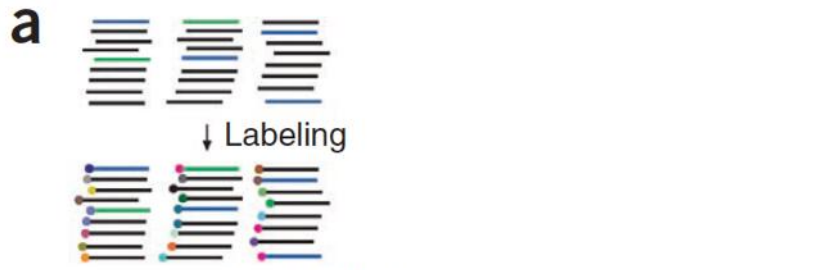
4. template switch RT, single-primer PCR



5. tagmentation (NEXTERA) & library amplification

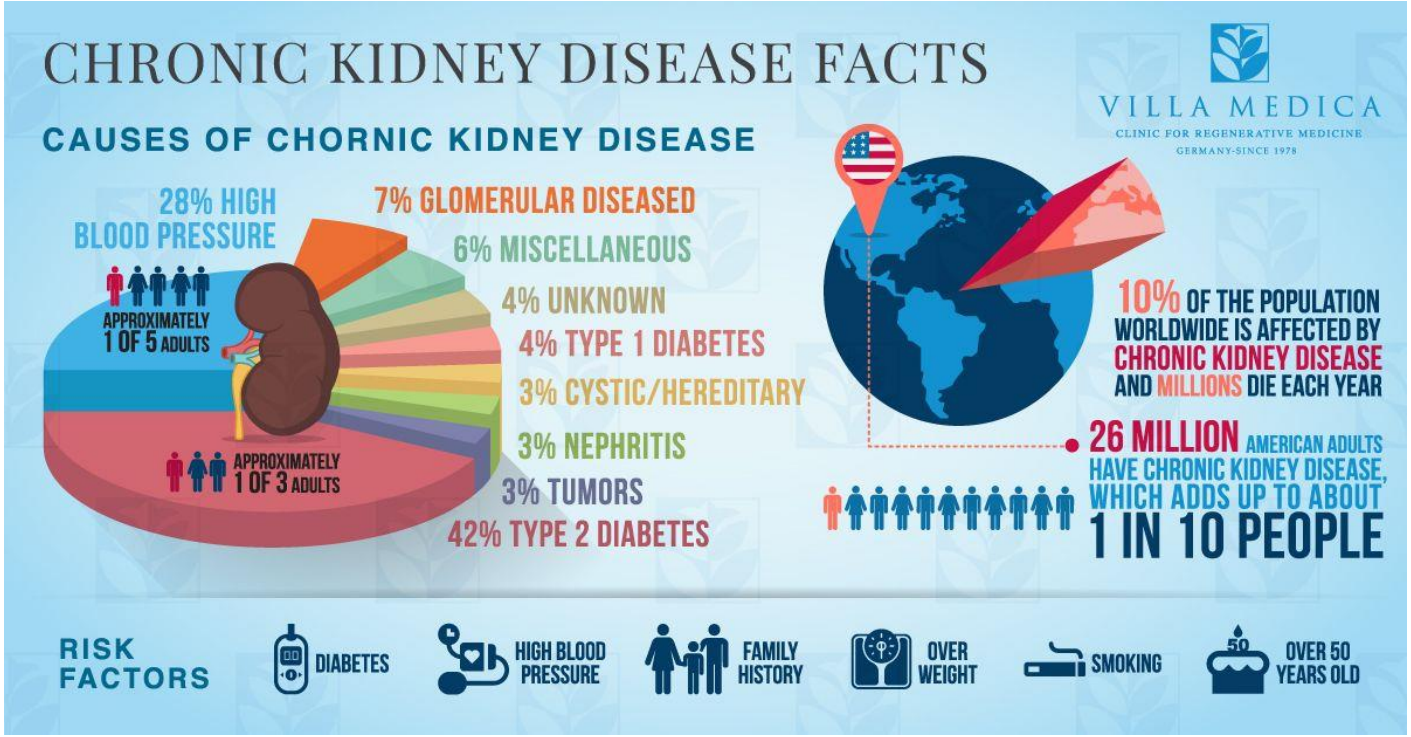


Unique molecular identifiers (UMIs)

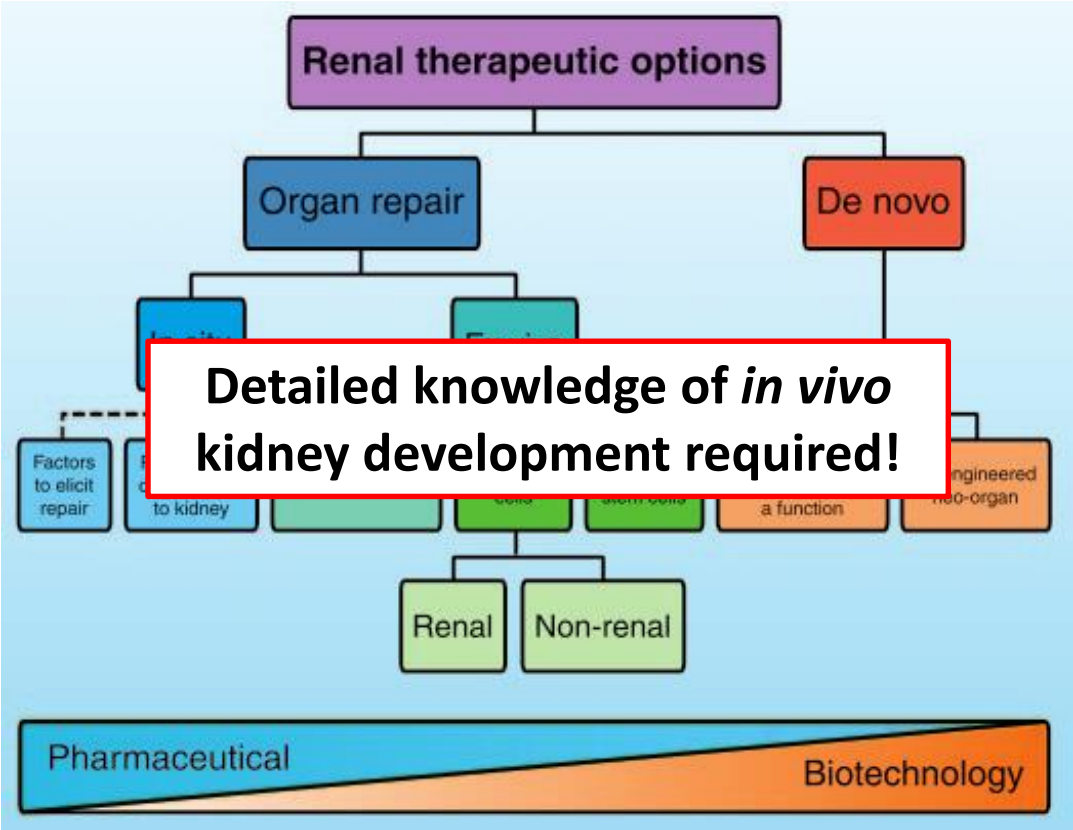


Kivioja et al., Nature Meth., 2011

Chronic kidney disease is a prevalent disease worldwide



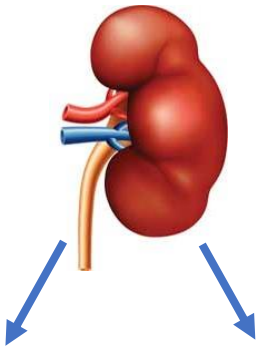
Regenerative medicine approaches for treating kidney disease



Detailed knowledge of *in vivo* kidney development required!

Little, JASN, 2006

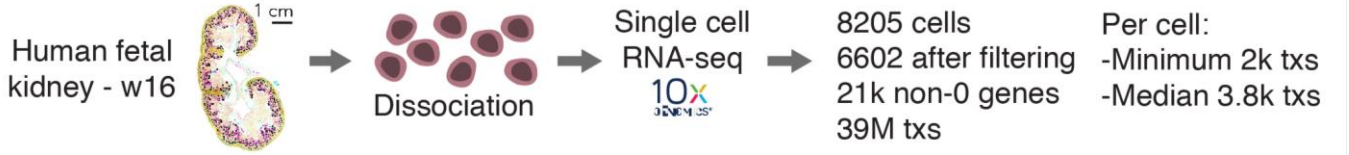
Transcriptomics of individual cell in the kidney (TRICK)



human fetal kidneys

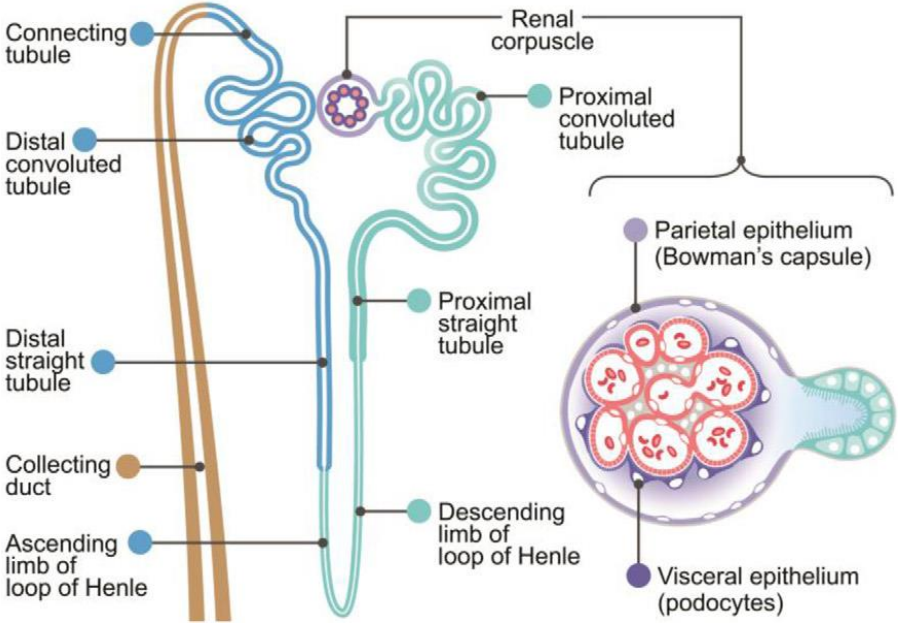
single-cell RNA-seq (10x genomics)

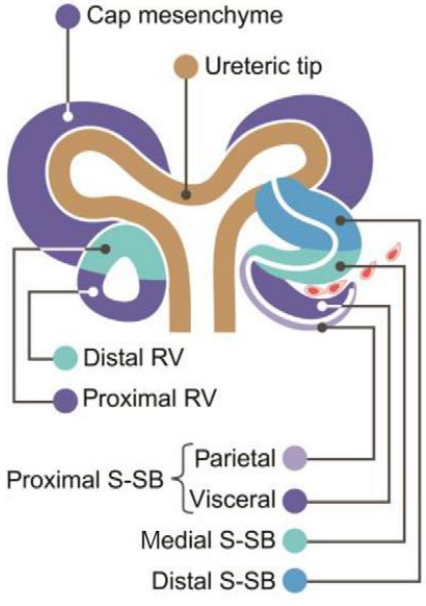
validation by imaging



Embryonic kidney development

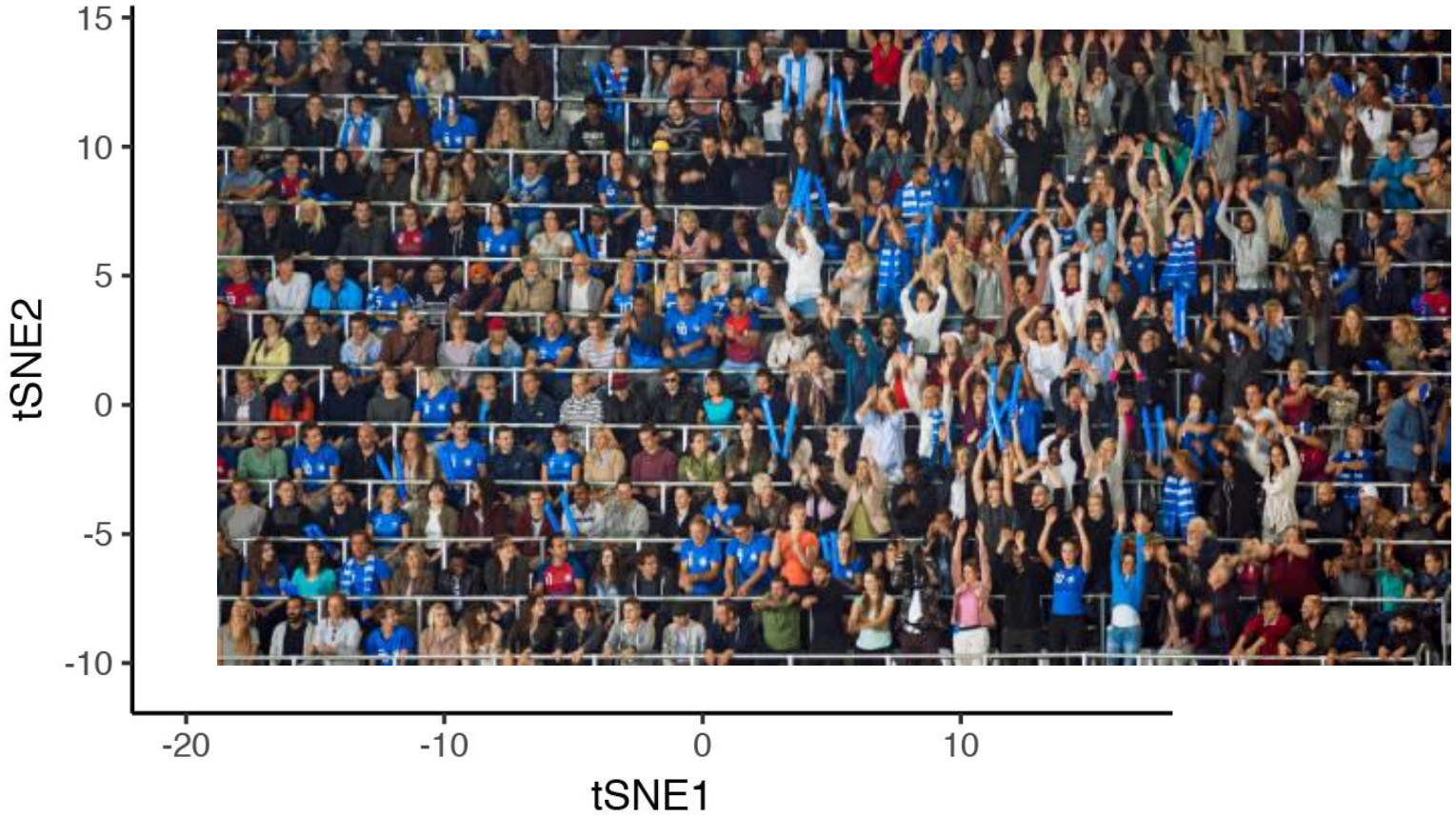
nephron – functional unit of the kidney



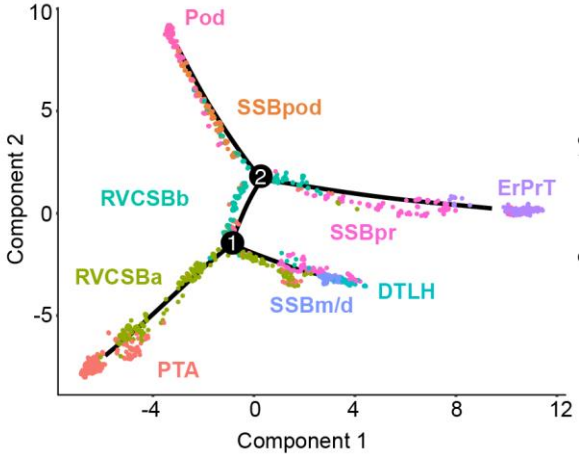


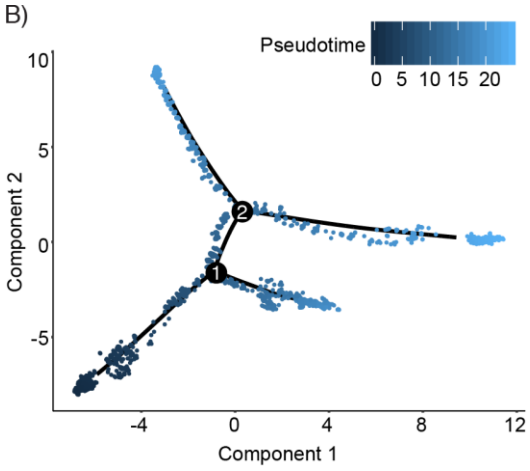
McMahon, Essays on Developmental Biology, 2016

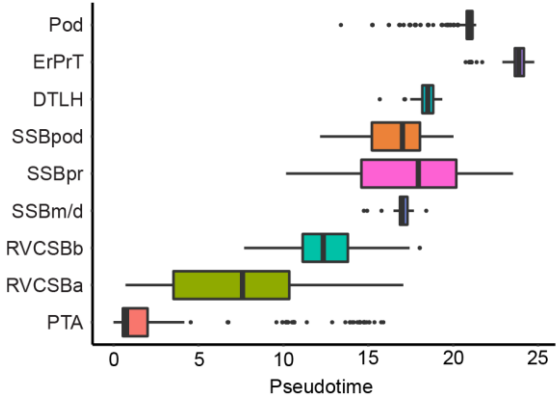
22 cell types could be distinguished



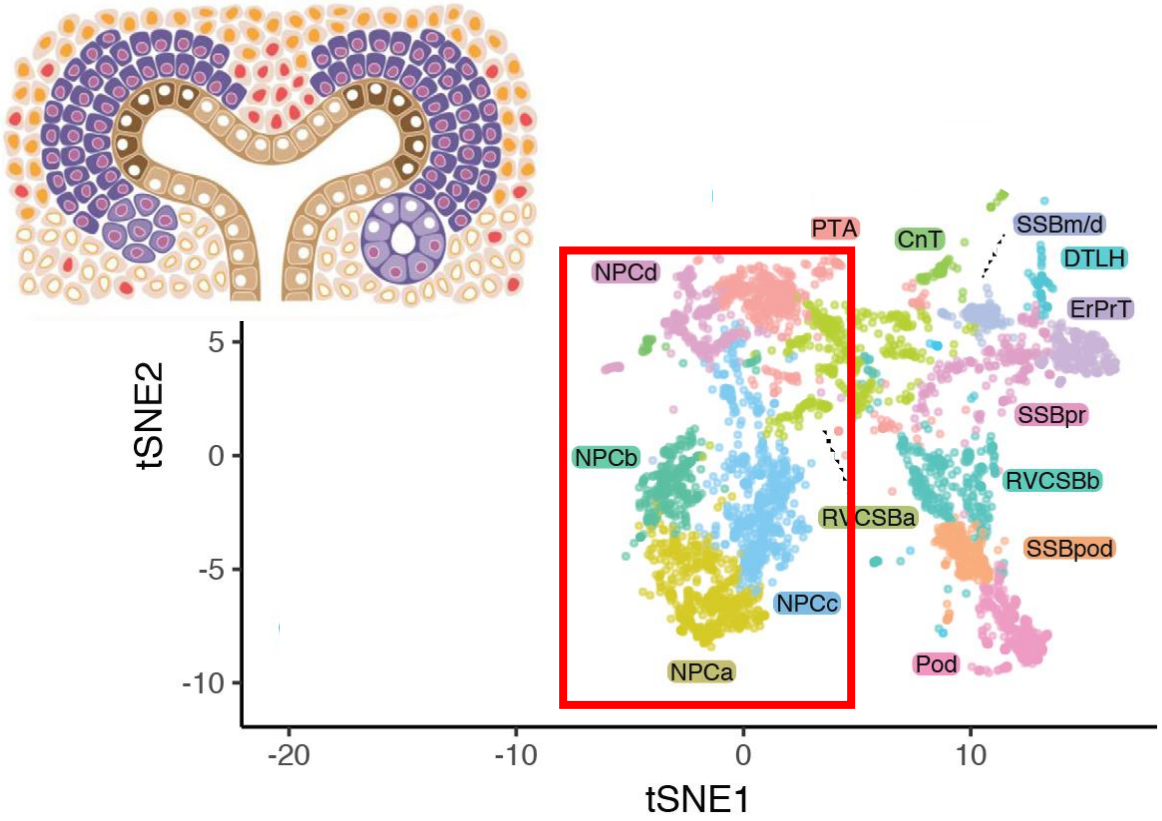
Trajectory inference with monocle 2 confirms developmental flow



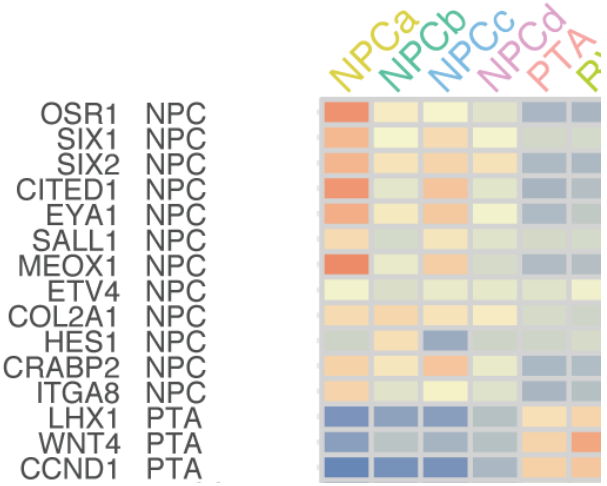


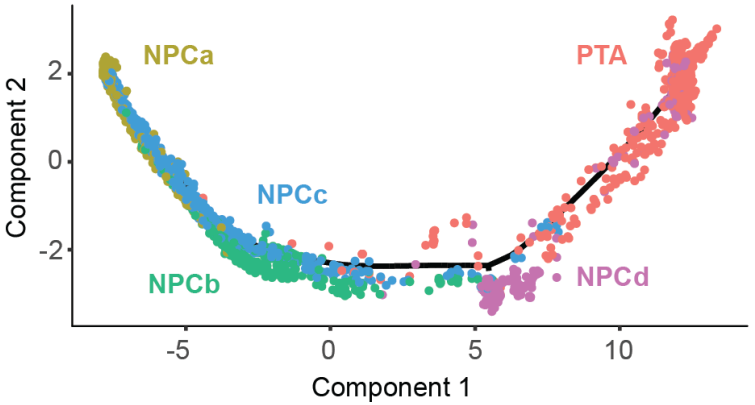


Heterogeneity in the nephrogenic niche

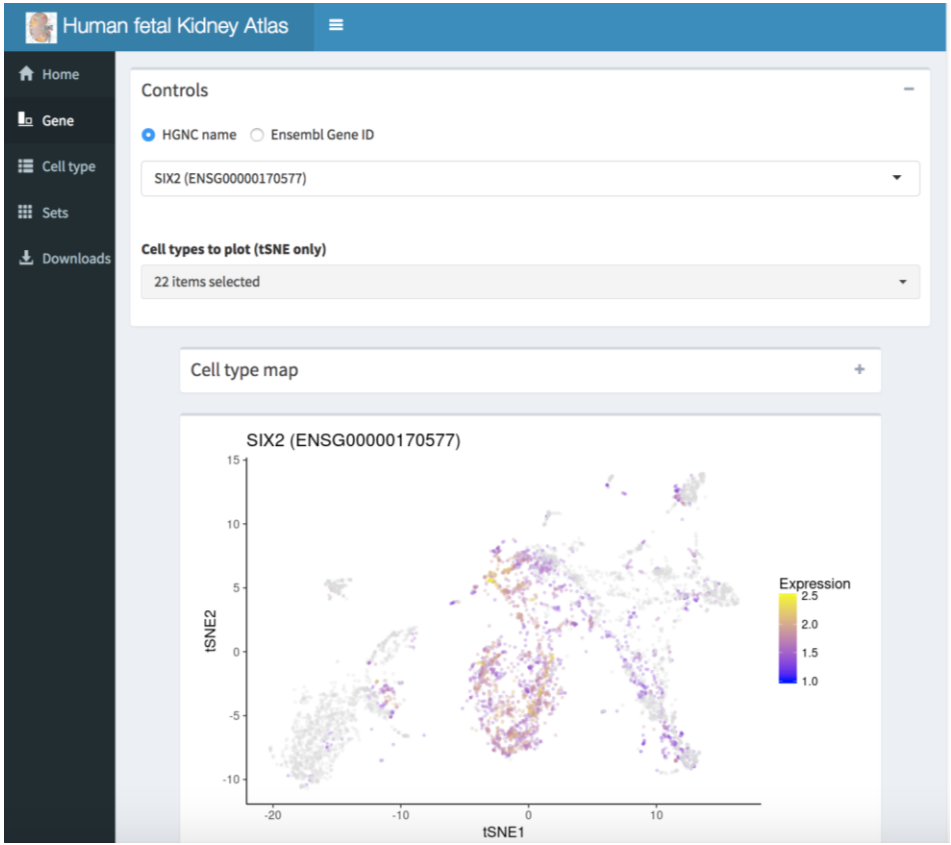


Gene expression and Monocle 2 suggest temporal order of NPCs





Data can be explored with an interactive web app



www.semraulab.com/kidney