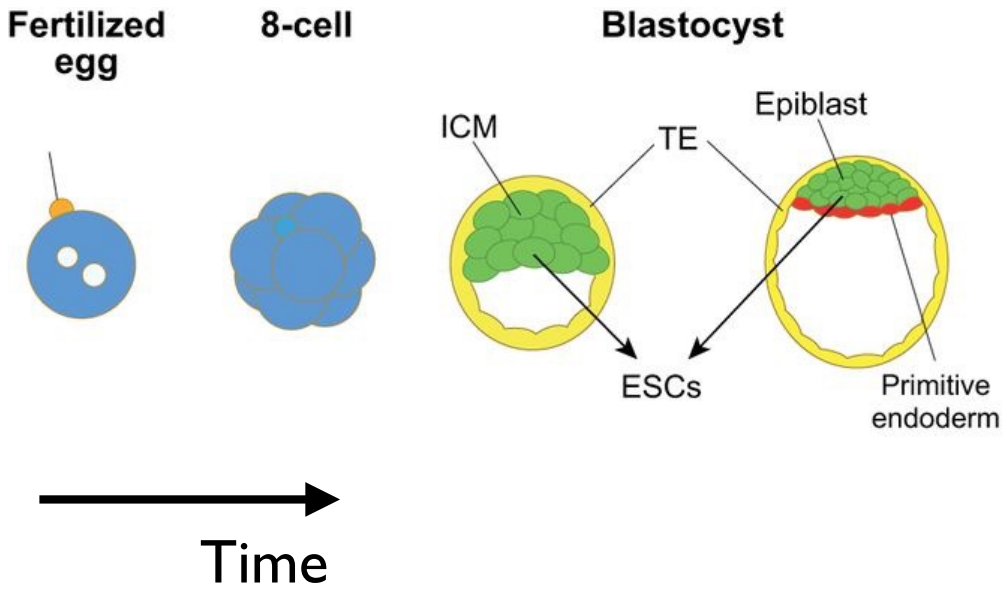
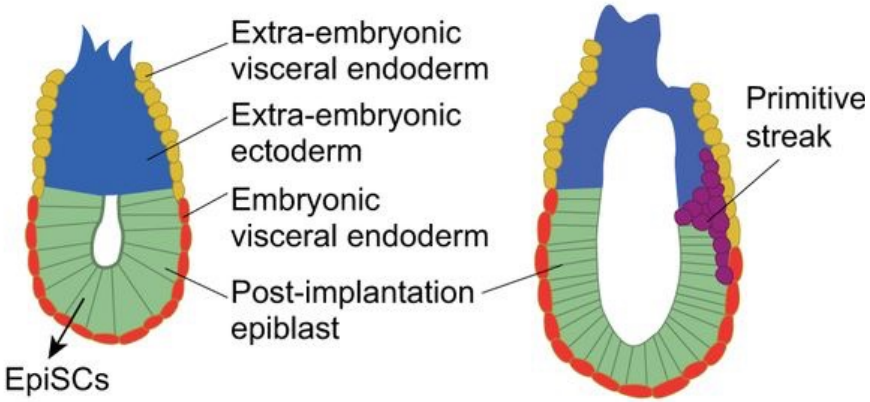
Pseudotime and lineage tree reconstruction from single cell transcriptomics data

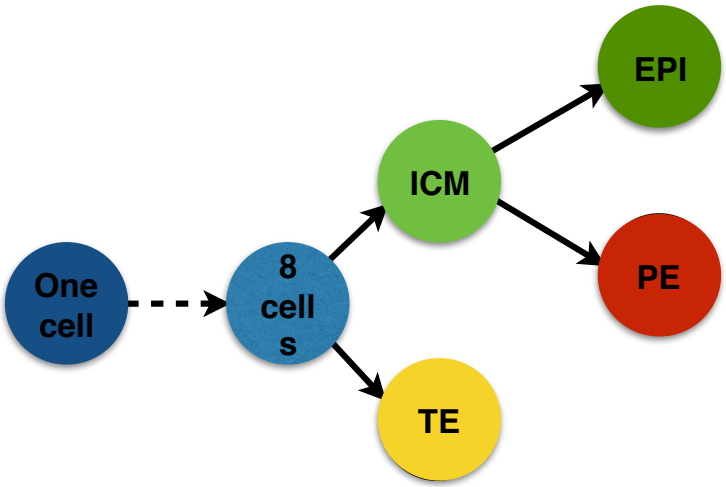
Laleh Haghverdi
July 2018

EMBL

Cell differentiation, lineage trees

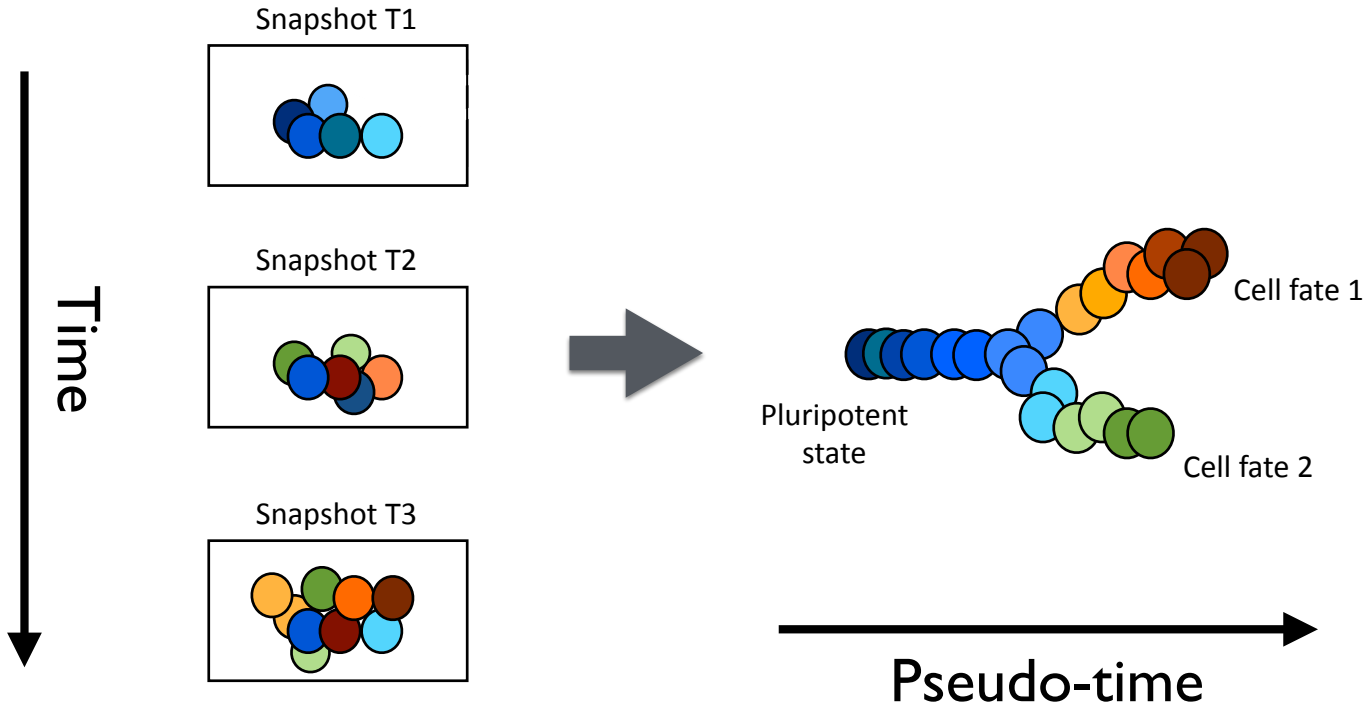






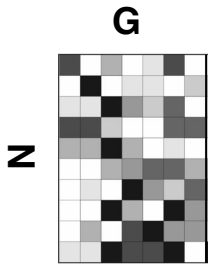
- Change of transcriptional profile

Lineage tree reconstruction



- Asynchronous development of cells —> bulk measurements fail

Single cell transciptomics data and challenges



- scRNA-seq, sc-qPCR, flow cytometry, mass cytometry.
- Data matrix of size n*G
- High-dimensional data, with too many irrelevant features (i.e. genes) noise dominates the signal —> selection of highly variable genes
- Big noise and dropouts

Single-cell data embedding and lineage tree analysis

- PCA (Wold et al. *Chemometrics and intelligent laboratory systems* 1987)
- t-SNE (Maaten, *Journal of machine learning research* 2008)
- Diffusion maps (Coifman *et al. PNAS 2005)*

(non-linear methods try to accommodate more information in lower dimensions)

- SPADE (Qiu, Peng, et al. *Nature biotechnology* 2011)
- *monocle2 (*Qiu, Xiaojie, et al. *Nature methods* 2017*)*
- *Wishbone (Setty et al. Nature biotechnology 2016)*
- *DPT (Haghverdi et al. Nature methods 2016)*

PCA

centred data matrix $X \quad [N, G]$

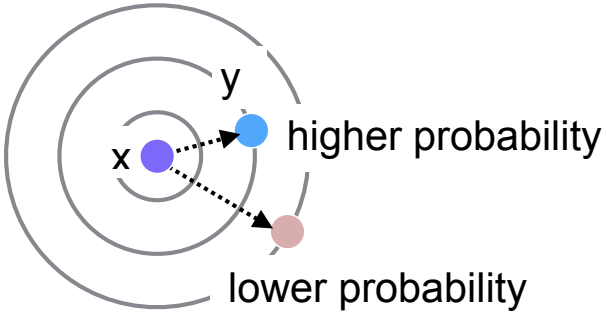
$$\text{argmax}_W \{ \underbrace{W^T X^T}_{Y^T} \underbrace{XW}_{Y} \}$$

$$Y^T * Y$$

Y is orthonormal, so what we are maximizing is just the variance of the embedded data Y

Diffusion maps

Gaussian kernel



$$T_{n\times n}$$

$$\psi_i \mathbf{T} = \lambda_i \psi_i \qquad 1 = \lambda_0 \geq \lambda_1 \geq ... \geq \lambda_{n-1}$$

$$D_t^2(x,y) = ||\mathbf{T}^t(x,.) - \mathbf{T}^t(y,.)||^2 \qquad \text{(for symmetric T)}$$

$$= \sum_{i=1}^{n-1} \lambda_i^{2t} (\psi_i(x) - \psi_i(y))^2$$

t-SNE

Original (high-dim) space X

$$p_{ij} = \frac{1}{z_X(i)} \cdot \exp(-\frac{||\mathbf{X}_i - \mathbf{X}_j||^2}{2\sigma^2})$$

$$z_X(i) = \sum_j \exp(-\frac{||\mathbf{X}_i - \mathbf{X}_j||^2}{2\sigma^2})$$

$$\mathbf{P} \leftarrow \frac{\mathbf{P} + \mathbf{P}'}{2n}$$

embedding (e.g. 2D) space Y

$$q_{ij} = \frac{1}{z_Y} \cdot \frac{1}{1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2}$$

$$z_Y = \sum_{k \neq l} \frac{1}{1 + ||\mathbf{Y}_k - \mathbf{Y}_l||^2}$$

- Minimize the cost function

$$C = KL(\mathbf{P}||\mathbf{Q}) = \sum_{i \neq j} p_{ij} \frac{\log(p_{ij})}{\log(q_{ij})}$$

- Initialise with $\mathbf{Y}^{(0)}$

$$\frac{\partial C}{\partial \mathbf{Y}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{Y}_i - \mathbf{Y}_j)(1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2)^{-1}$$

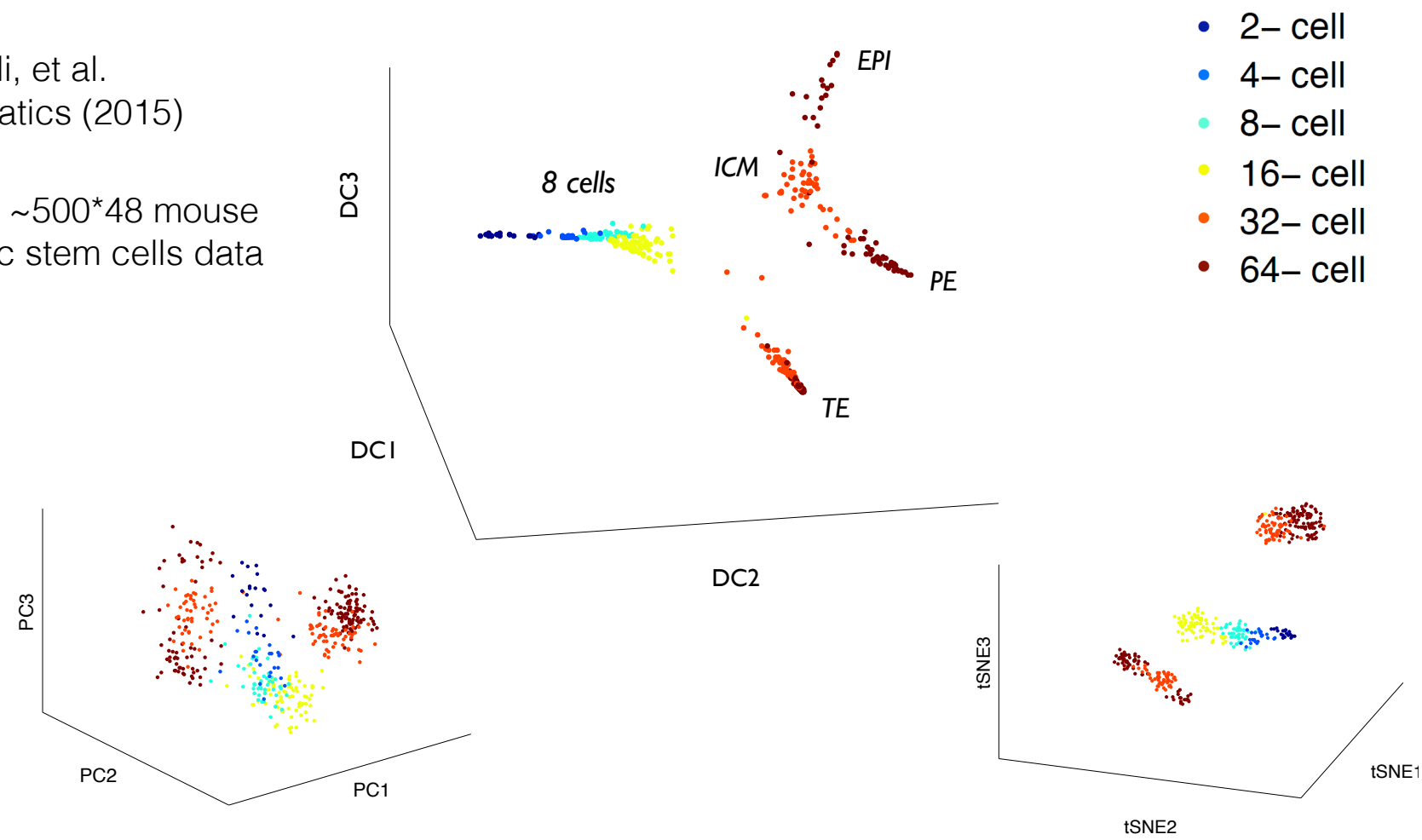
$$\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathbf{Y}_i} + \alpha(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)})$$

- The final solution for Y depends on the initialisation Y^0
- t-SNE does not respect the global continuity of data manifold

PCA, t-SNE, Diffusion maps

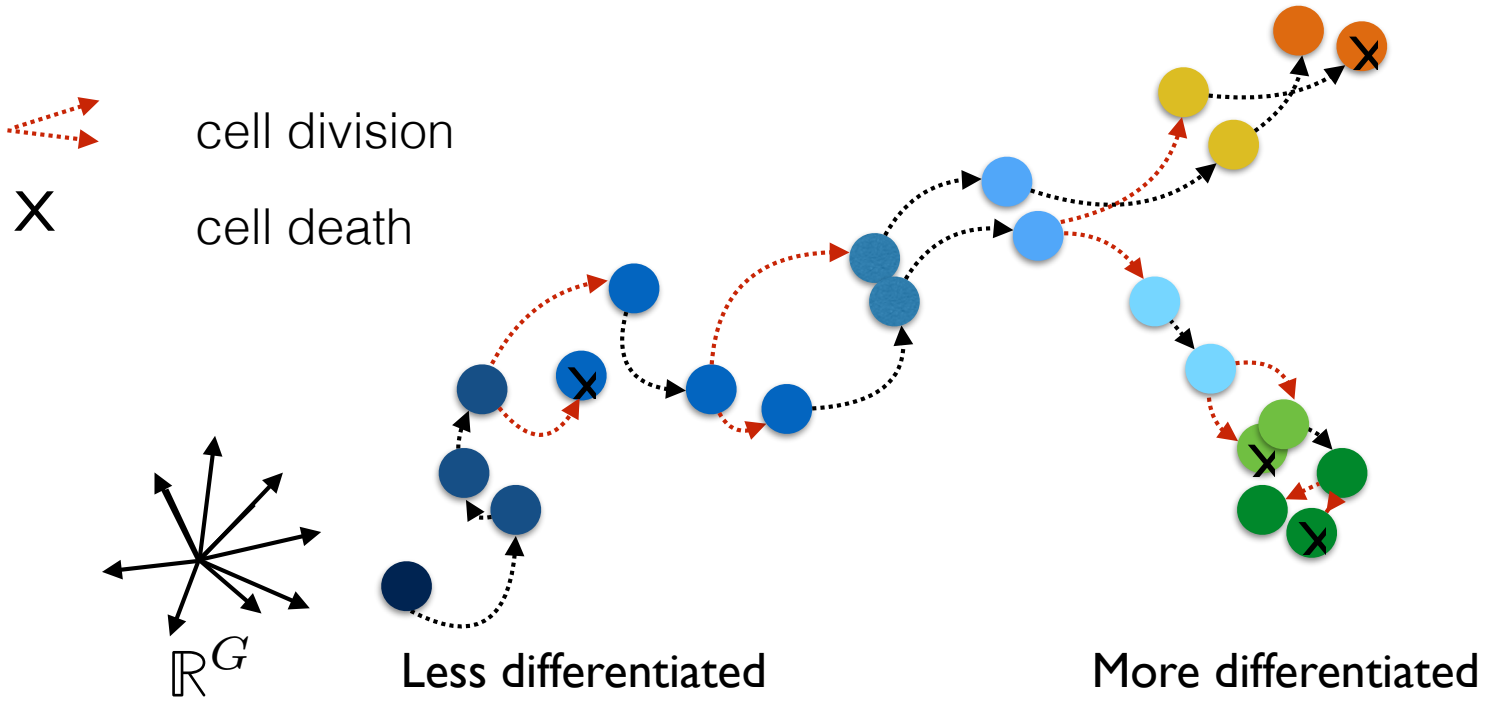
Haghverdi, et al.
Bioinformatics (2015)

sc-qPCR ~500*48 mouse embryonic stem cells data set



- Keeps the global continuity structure
- Robust to noise
- Intrinsically related to the biological process

Cell differentiation; a diffusion-like process



$$\frac{\partial p}{\partial t} = \frac{\partial^2}{\partial x^2}\left(Dp(t)\right) - \frac{\partial}{\partial x}\left(\mu p(t)\right) + \nabla \cdot p$$

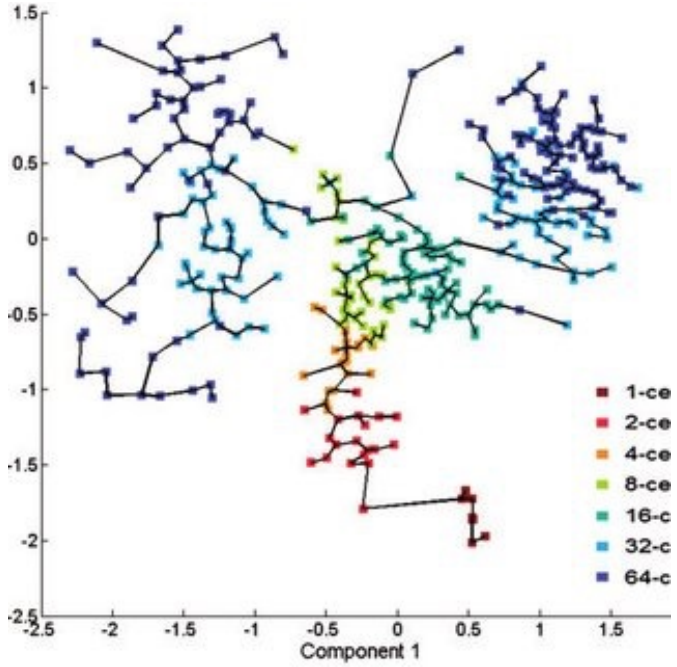
$$p(t) = p(t-1) + \Delta p \qquad p(t) = p(t-1)T$$

- Different types of time evolution dynamics can take place on the same manifold (Butler et al. Electronic journal of linear algebra 2007)

Going beyond mere embedding and visualisation; pseudo-time and branch identification

Minimum Spanning tree

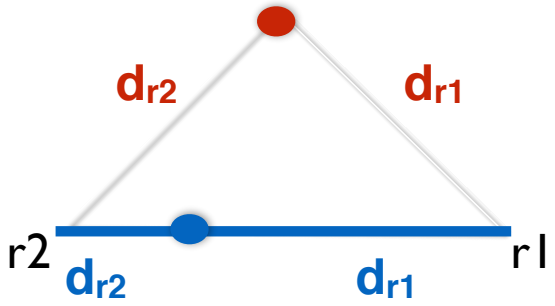
- Minimum total edges distance that goes through all nodes
- In general not robust; requires a neat low-dimensional embedding

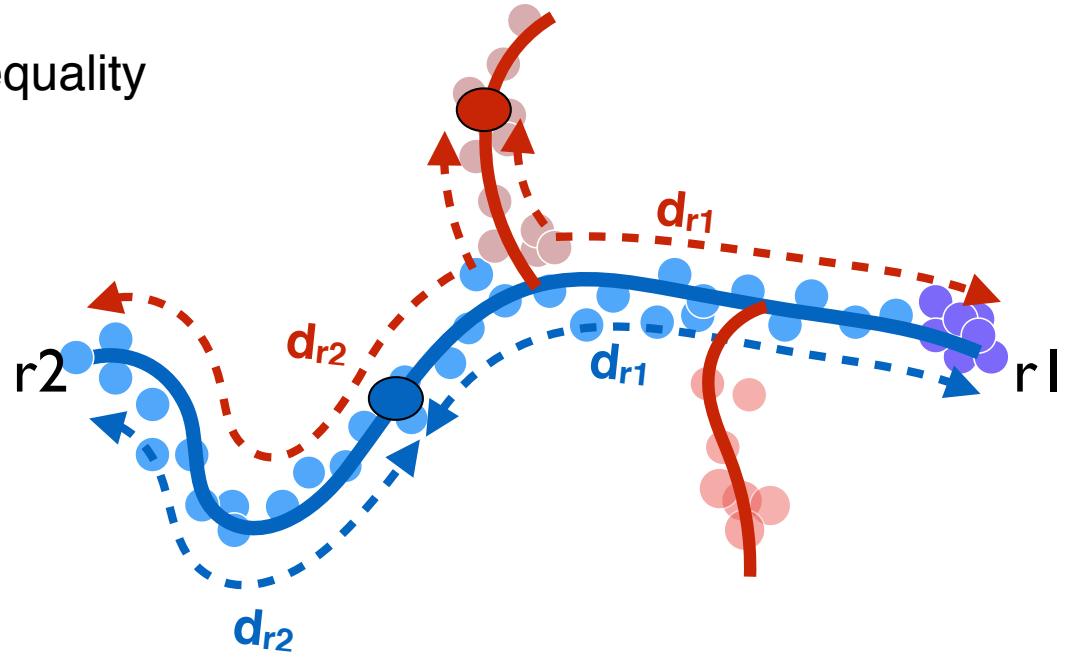


Going beyond mere embedding and visualisation; pseudo-time and branch identification

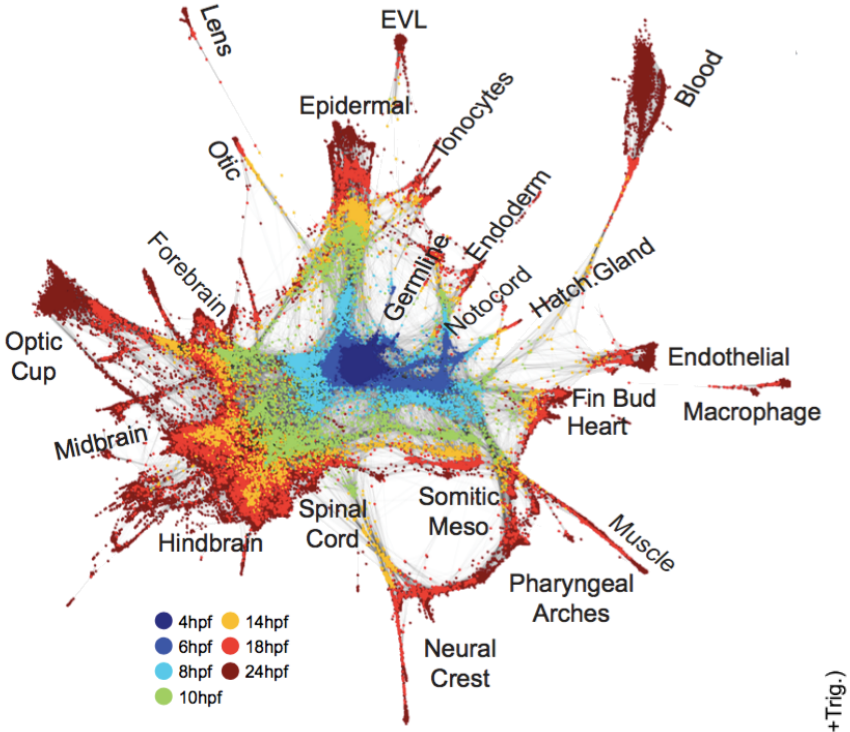
Using an on-manifold metric

- Pseudotime defined by on-manifold distance from the root cell
- Branch identification using triangle inequality

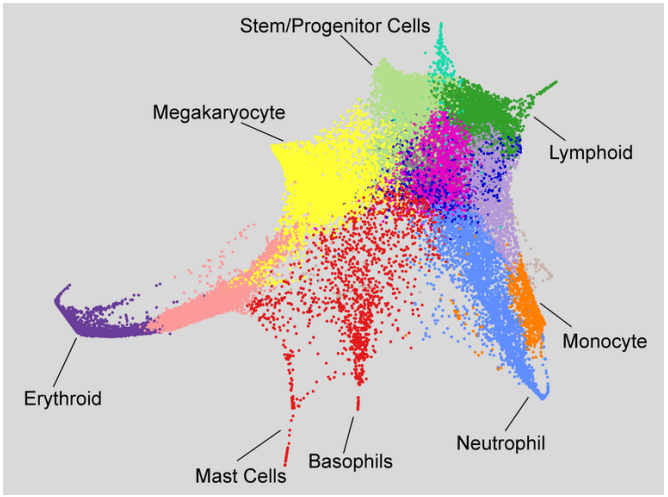




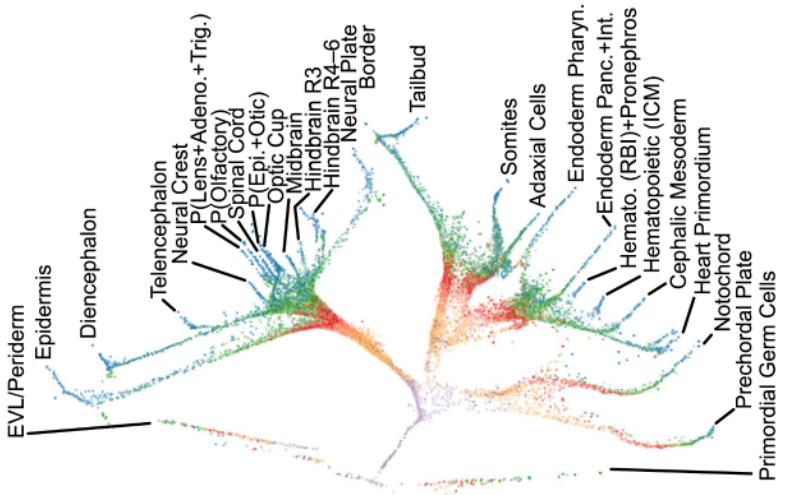
Complex manifolds with several branching events



Zebrafish (Wagner, Daniel E., et al. *Science* 2018)



Blood (Berthold Goetgens lab 2018)



Zebrafish (Farrell, Jeffrey A., et al. 2018)

Complex manifolds with several branching events

- Force-directed graphs
- Taking directed dynamics into account

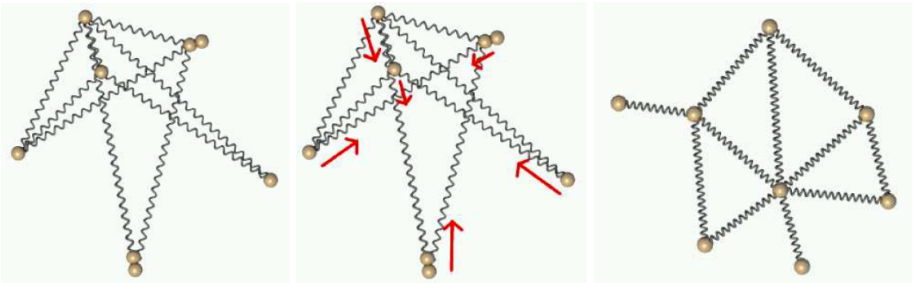
- SPRING (Weinreb et al. *Bioinformatics* 2017)
- URD (Farrell et al. *Science* 2018)
- Velocyto (La Manno, *bioRxiv* 2017)

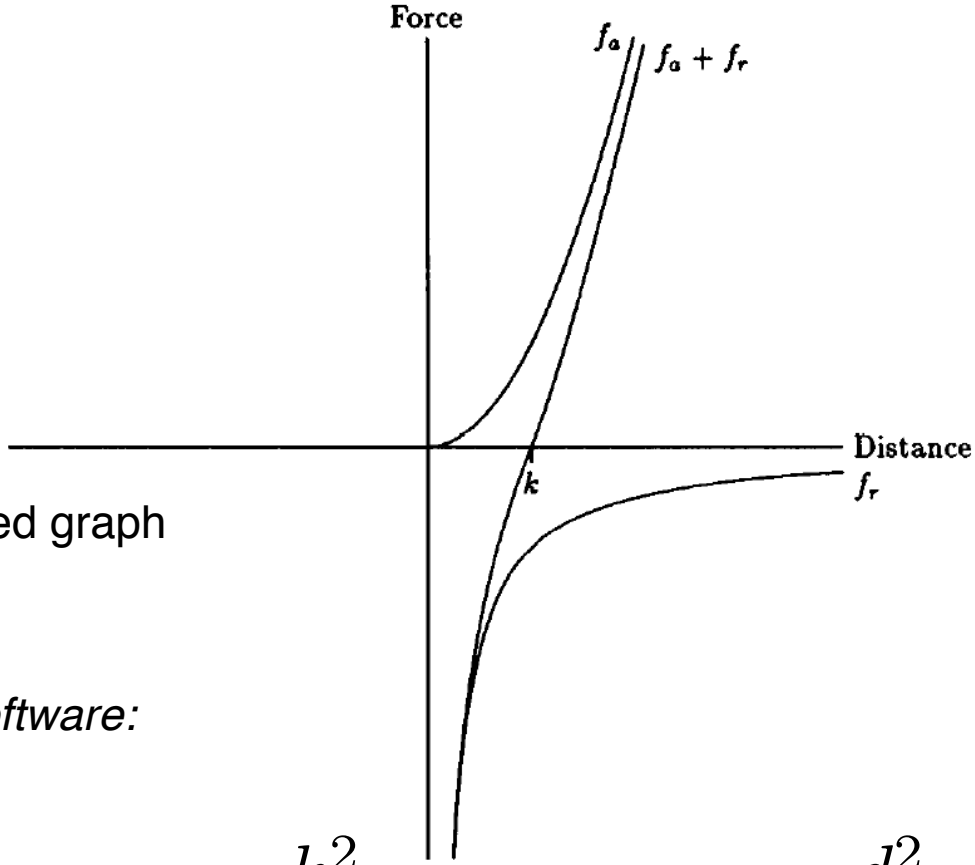
—> User-defined root cell and endpoints to infer directed diffusions

—> Estimate the velocity vector for each cell

Force-directed embedding

- Minimize the energy of a springs system





- Just a visualisation; requires a prior neatly weighted graph (e.g. careful knn graph construction)

- Fruchterman and Reinhold (Fruchterman et al. *Software: Practice and experience,* 1991)

skips local minima

$$f_r(d) = \frac{k^2}{d^2} \qquad f_a(d) = -\frac{d^2}{k}$$

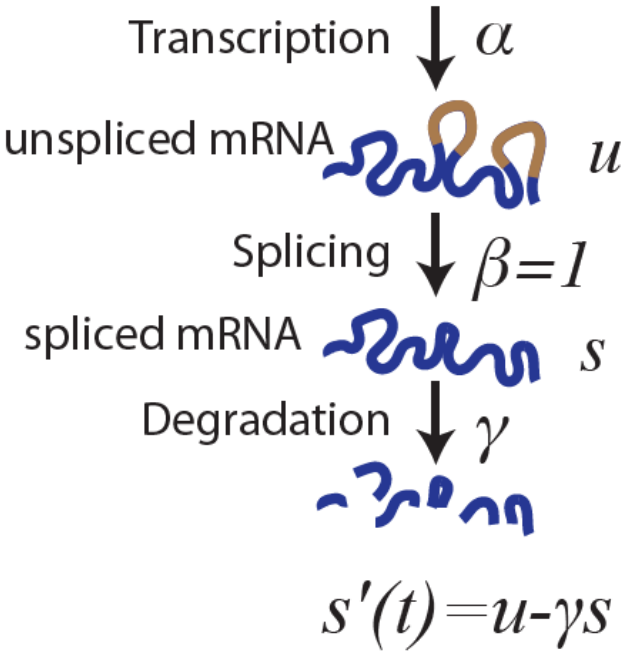
t-SNE (as a force-directed method)

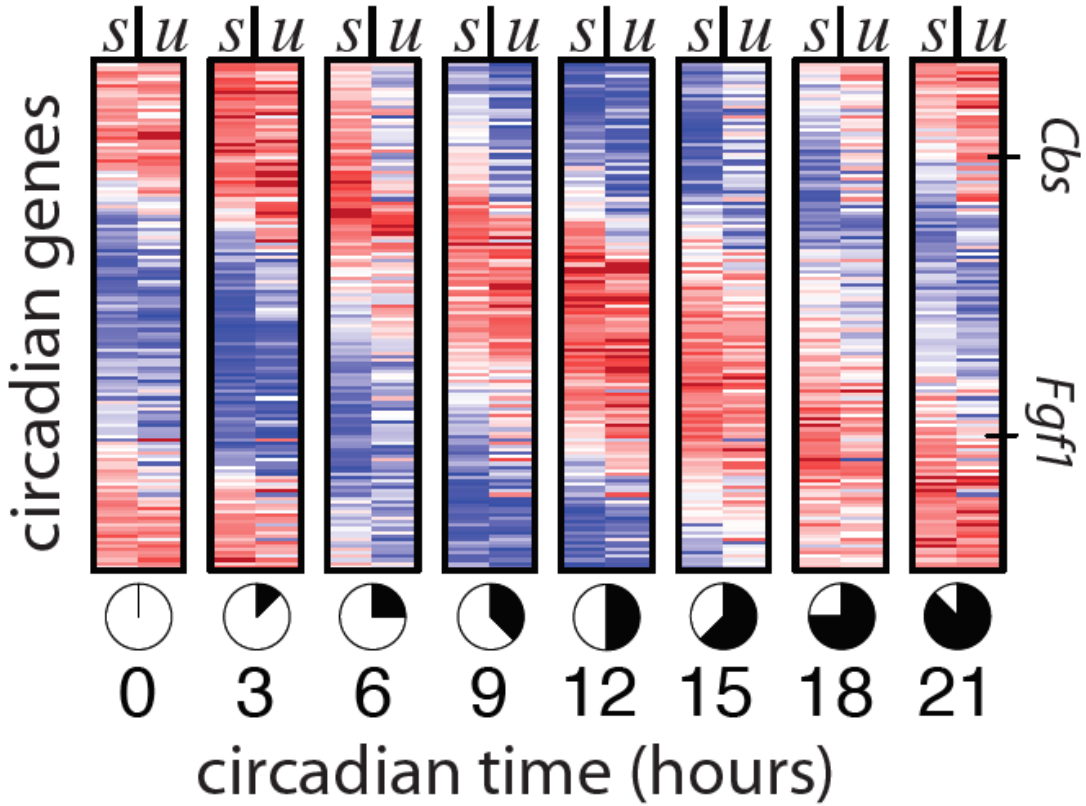
$$\frac{\partial C}{\partial \mathbf{Y}_i} = 4 \left(\sum_j p_{ij} q_{ij} z_Y (\mathbf{Y}_i - \mathbf{Y}_j) - \sum_j q_{ij}^2 z_Y (\mathbf{Y}_i - \mathbf{Y}_j) \right)$$

$$= 4 \left(\underbrace{\sum_j p_{ij} \frac{\mathbf{Y}_i - \mathbf{Y}_j}{1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2}}_{\text{Repulsive}} - \underbrace{\sum_j \frac{1}{z_Y} \cdot \frac{\mathbf{Y}_i - \mathbf{Y}_j}{(1 + ||\mathbf{Y}_i - \mathbf{Y}_j||^2)^2}}_{\text{Attractive}} \right)$$

RNA velocity

(La Manno, *bioRxiv* 2017)

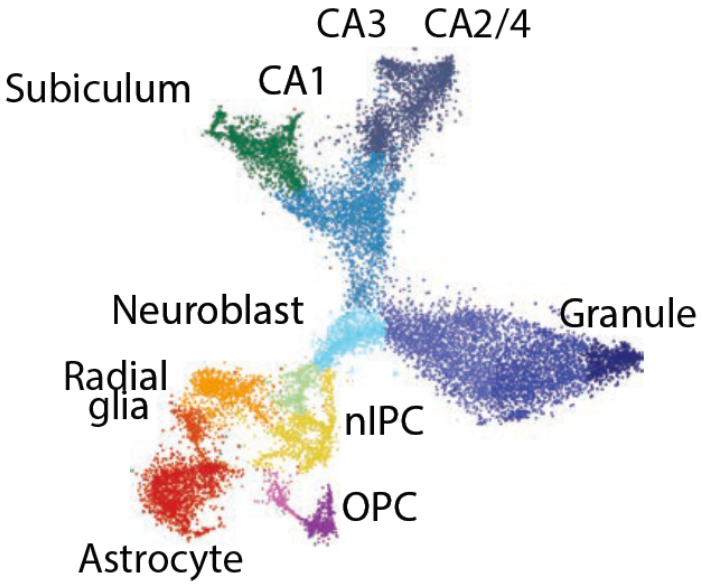


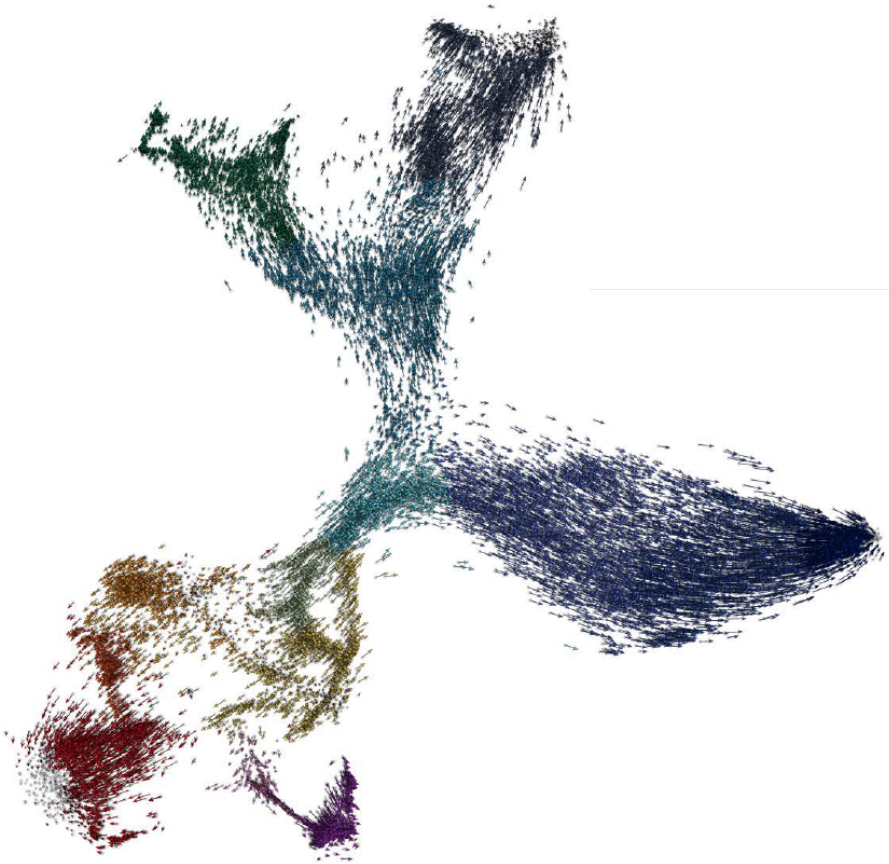


RNA velocity

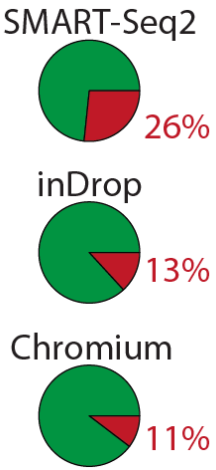
(La Manno, *bioRxiv* 2017)

Developing mouse
hippocampus dataset

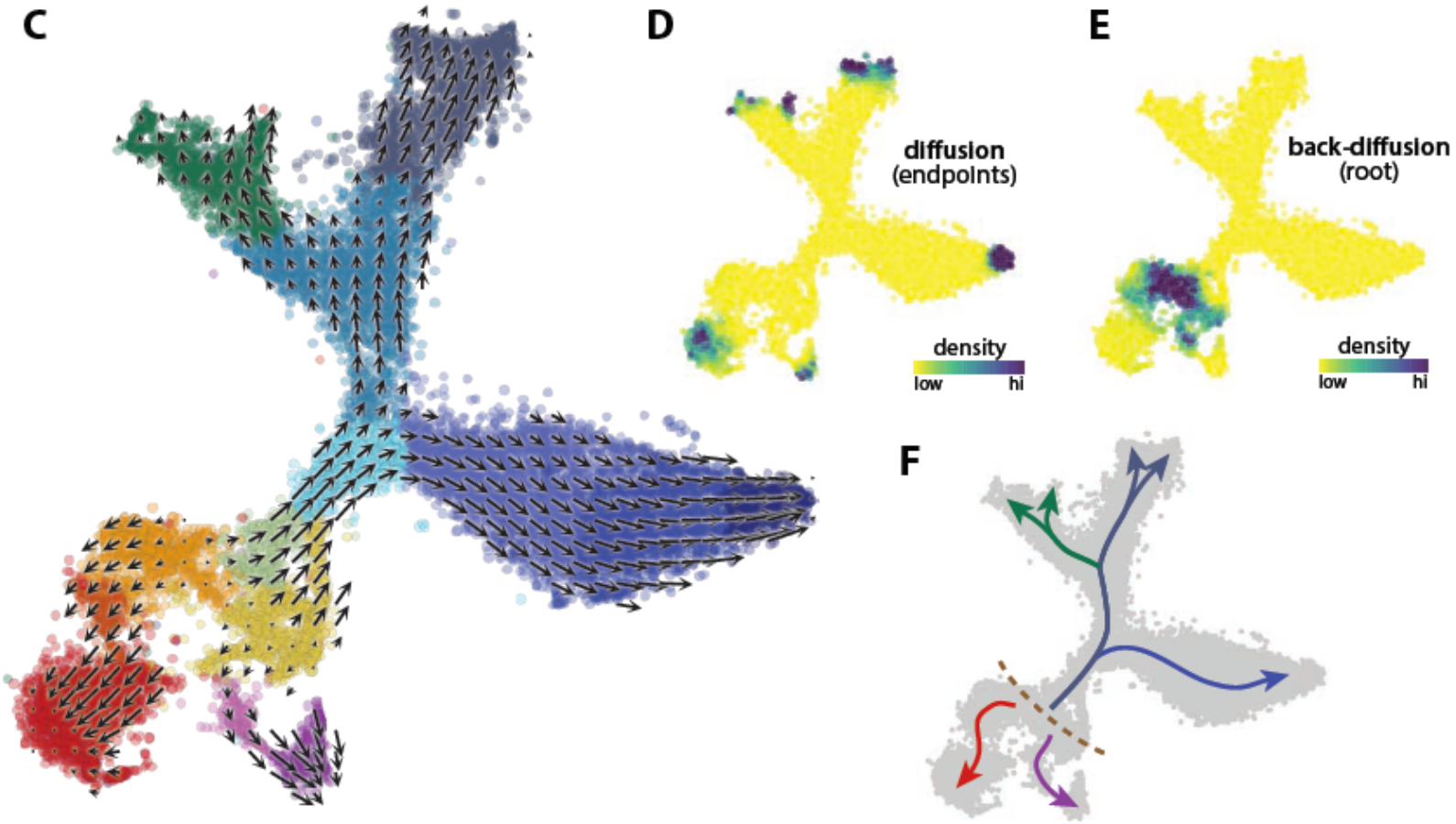




Unspliced mRNA
percentage



$$T_{ij} = \underbrace{pW_{ij}}_{\text{Geometrical diffusion}} + \underbrace{(1-p)D_{ij}}_{\text{Velocity-driven (drift)}}$$



Summary

- Development and lineage trees
- Single- cell transcriptomics
- Dimension reduction methods
- Lineage reconstruction methods with few branchings
- Lineage reconstruction methods for complex topologies
- Integrating actual time information (RNA velocity)

Batch effects in single cell transcriptomics data

Laleh Haghverdi
July 2018

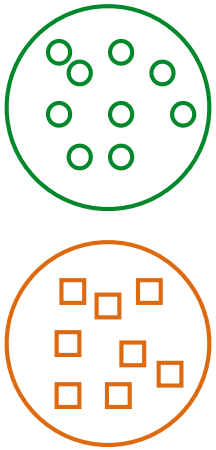
EMBL

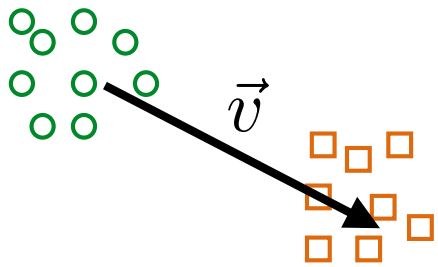
Batch correction in single-cell data

- Single-cell data growing in scales
- Human Cell Atlas project
- Split data collection between different labs possibly using different facilities and different protocols

Confounded vs. balanced experimental design

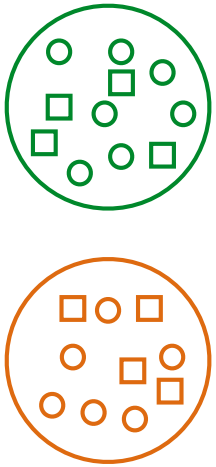
- Totally confounded

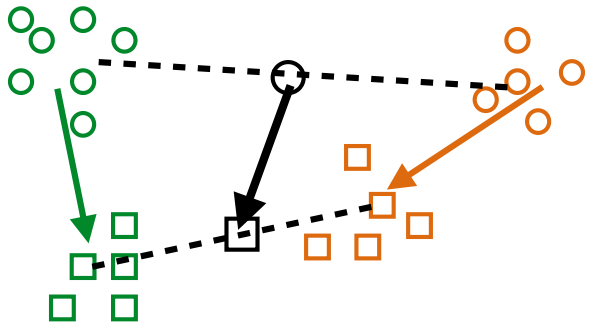




$$\overrightarrow{v} = \overrightarrow{biological} + \overrightarrow{batch}$$

- Balanced





average biological vector
e.g. limma and Combat

Batch correction for single-cell RNA-seq data

- Different laboratories, different protocols and different cell type compositions
- Scalable

- limma (Ritchie et al. Nucleic Acids Res. 2015)
- Combat (Johnson et al. Bioinformatics 2007)
- RUVseq (Risso et al. Nature biotechnology 2014)

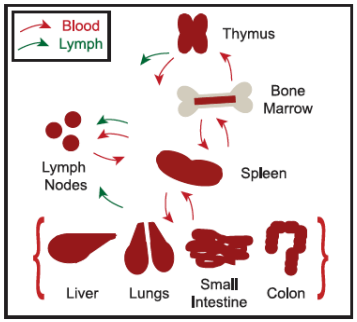
- Projection on a reference embedding
- Scaffold (Spitzer, et al. *Science* 2015)
- Mutual Nearest Neighbours matching (Haghverdi, et al. *Nature biotechnology* 2018)
- CCA approach + alignment (Butler, et al. *Nature biotechnology* 2018)
- Autoencoders approach (Lopez et al. arxiv 2017, Eraslan et al. biorxiv 2018)

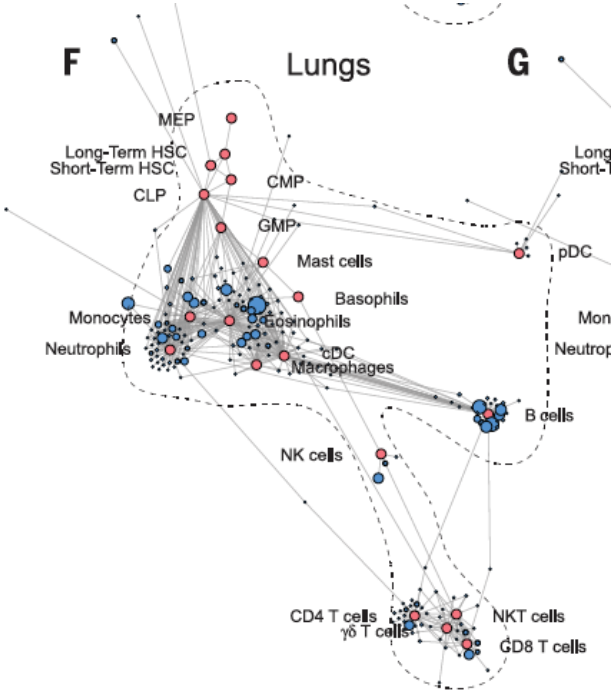
Projection methods

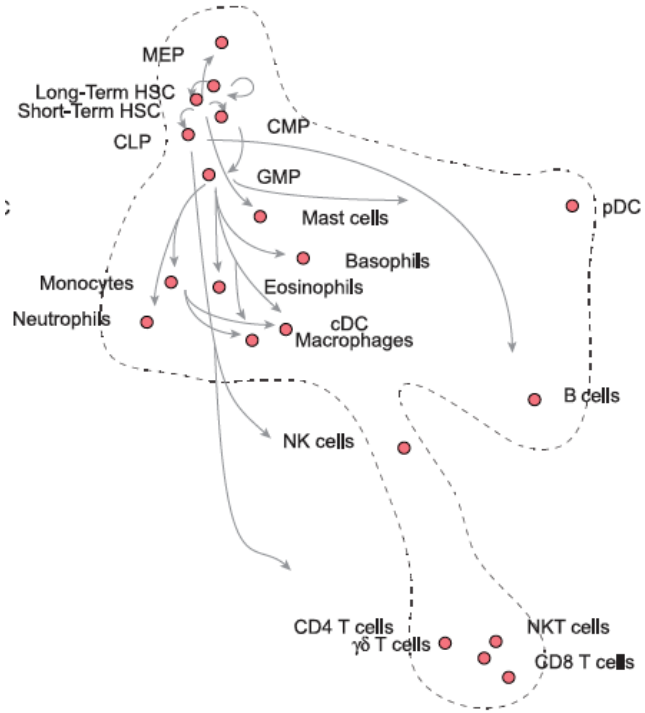
- Inter-batch and intra-batch distances are not comparable and can not be mixed
- Embedding of the reference batch (inter-batch only) + projection of the other batch on that embedding (intra-batch)
- projection with PCA, t-SNE, diffusion maps, etc.
- Requires that the reference batch is a super set of all appearing populations

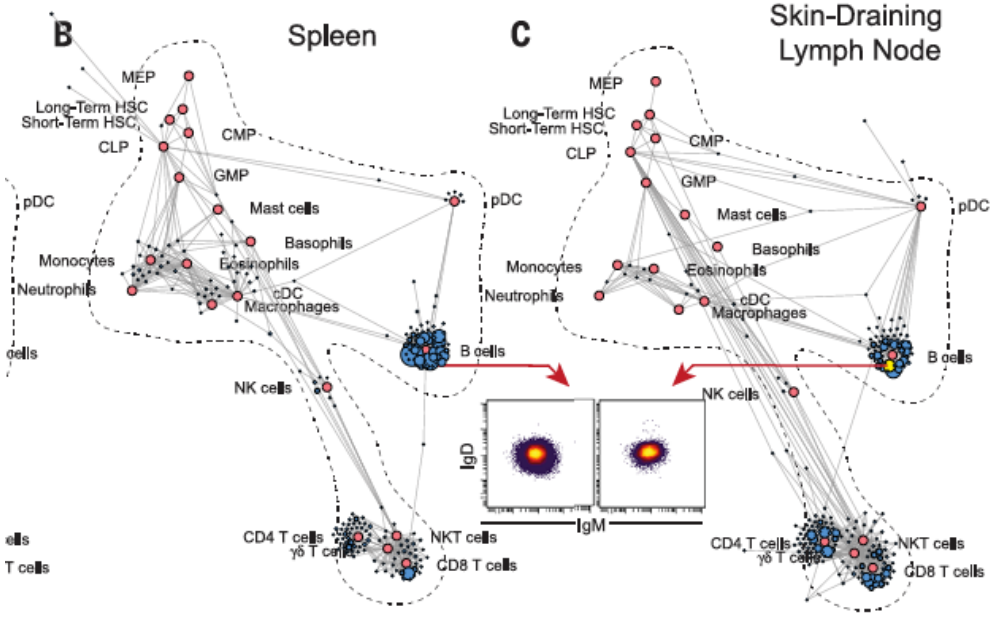
Projection example

Scaffold map of cell types within the immune system for various organs (Spitzer, et al. *Science* 2015)









Embedding of the reference batch (bone marrow)

Combat

(Johnson et al. Bioinformatics 2007)

- Shrunken (empirical Bayesian) inference on a linear model

$$Y_{ijg} = \alpha_g + \underbrace{X\beta_g} + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg},$$

batch **i**
sample (cell) **j**
gene **g**

biological effect

- Centre and scale within each condition

$$Z_{ijg} = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g}{\widehat{\sigma}_g}.$$

- Instead of adjusting each gene independently, barrow information from all other genes

$$Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$$

$$\gamma_{ig} \sim N(Y_i, \tau_i^2) \quad \text{and} \quad \delta_{ig}^2 \sim \text{Inverse Gamma}\,(\lambda_i, \theta_i).$$

RUV, RUVseq

(Risso et al. Nature biotechnology 2014)

$[N, G]$

$$Y = \underbrace{X\beta}_{\text{Biological effects}} + \underbrace{W\alpha}_{\text{Unwanted effects}} + \varepsilon,$$

$[N, K] * [K, G]$ $\qquad$ $[N, L] * [L, G]$

- Supervised approach (e.g. with negative control genes or known biological effects)

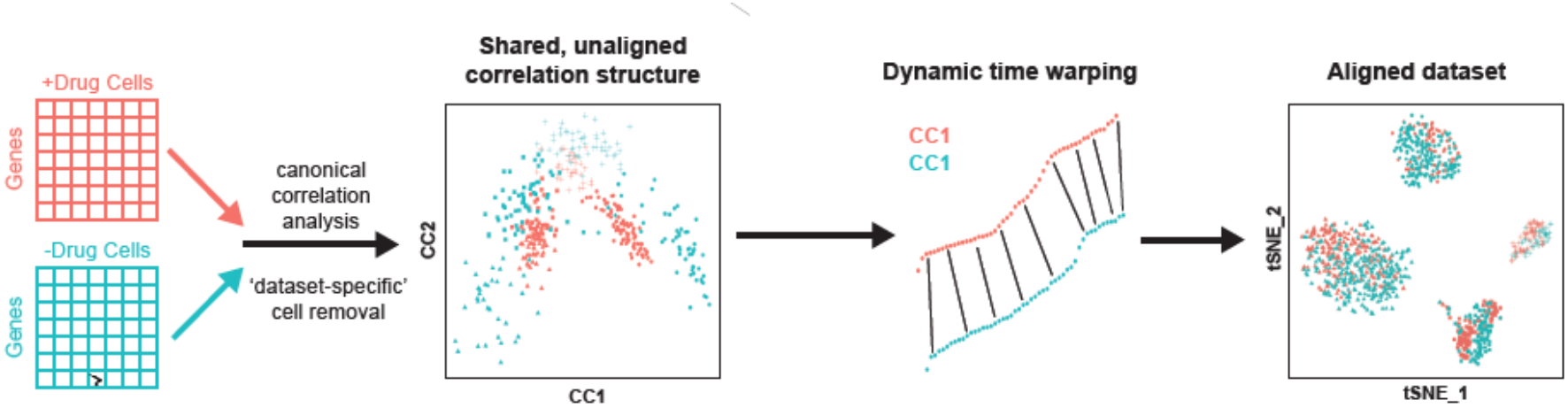
$$Y_c = W\alpha_c + \varepsilon_c,$$

- How to know which genes will not be affected?

CCA based approach

Butler, et al. *Nature biotechnology* 2018

- Find directions of maximal correlation between batches; Canonical Correlation analysis (Bhatia,1998)
- Align cells on the canonical correlation components

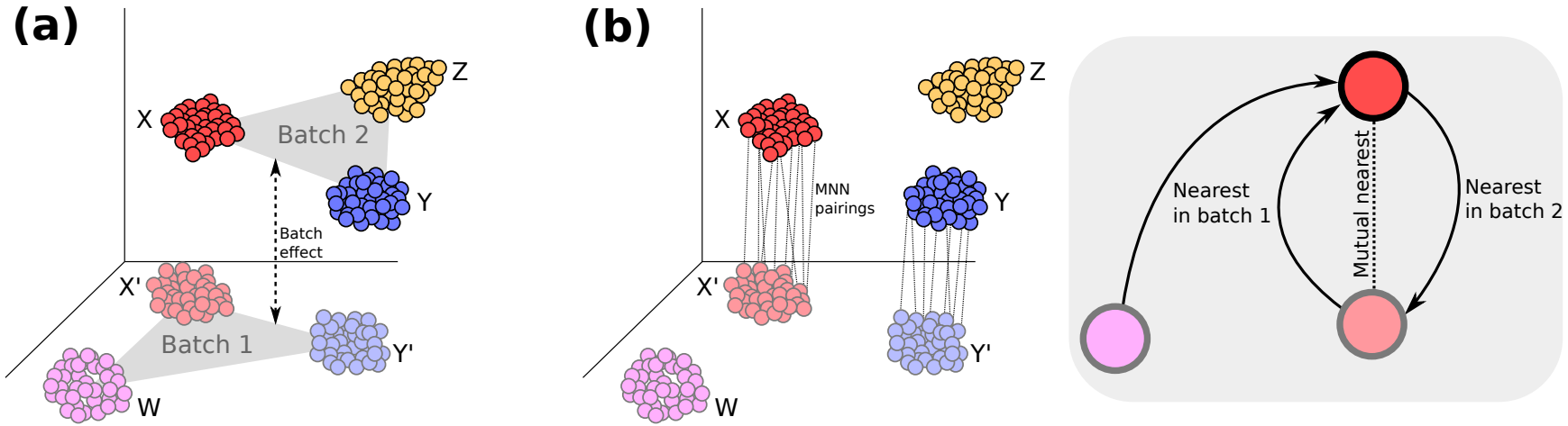


Mutual Nearest Neighbours Analysis

Haghverdi et al. *Nature biotechnology* 2018

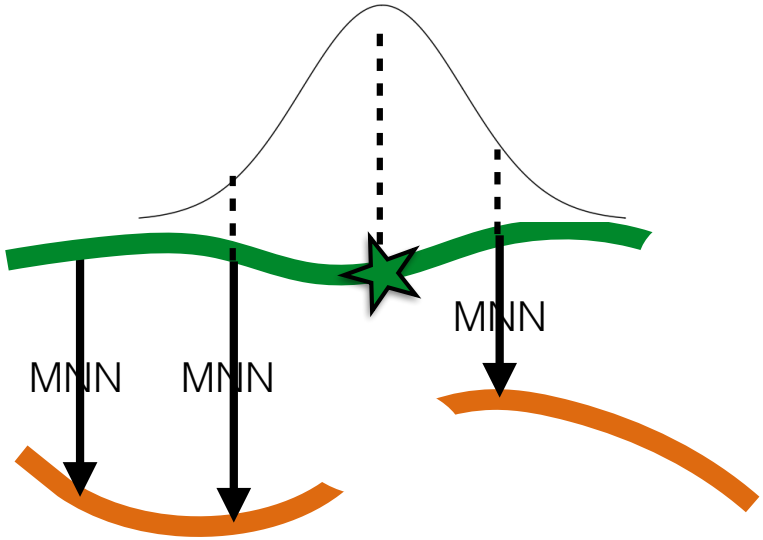
- Batch effect is almost orthogonal to the biological hyper plane (Central limit theorem)

$$\vec{X} \cdot \vec{Y} = \sum_{i=1}^{D} x_i y_i$$

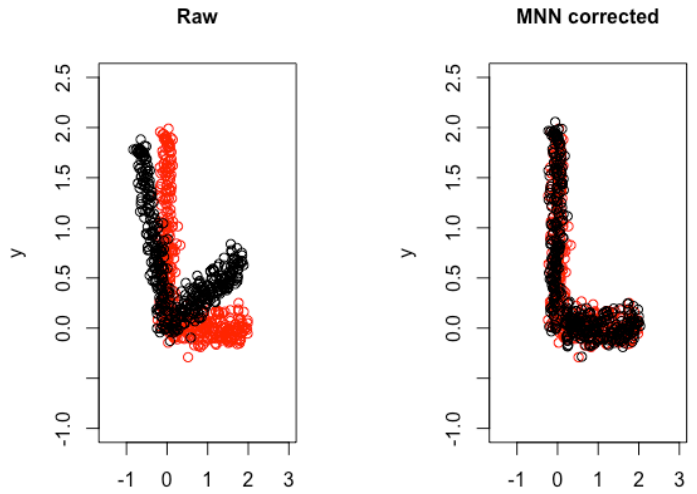


Mutual Nearest Neighbours Analysis

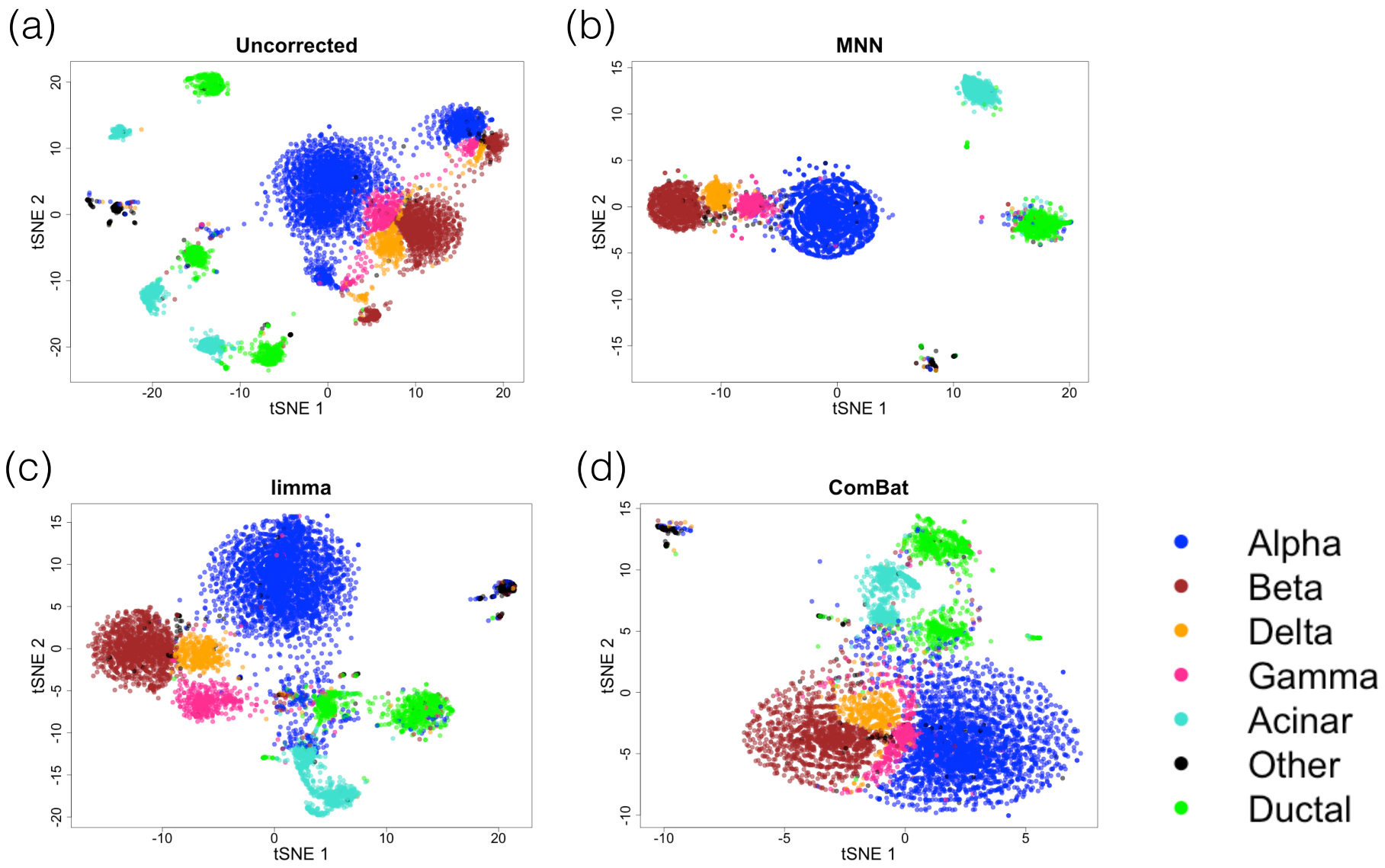
- Once a few MNN pairs have been identified at a few locations, batch effects for everywhere else is estimated using a smoothing Gaussian kernel



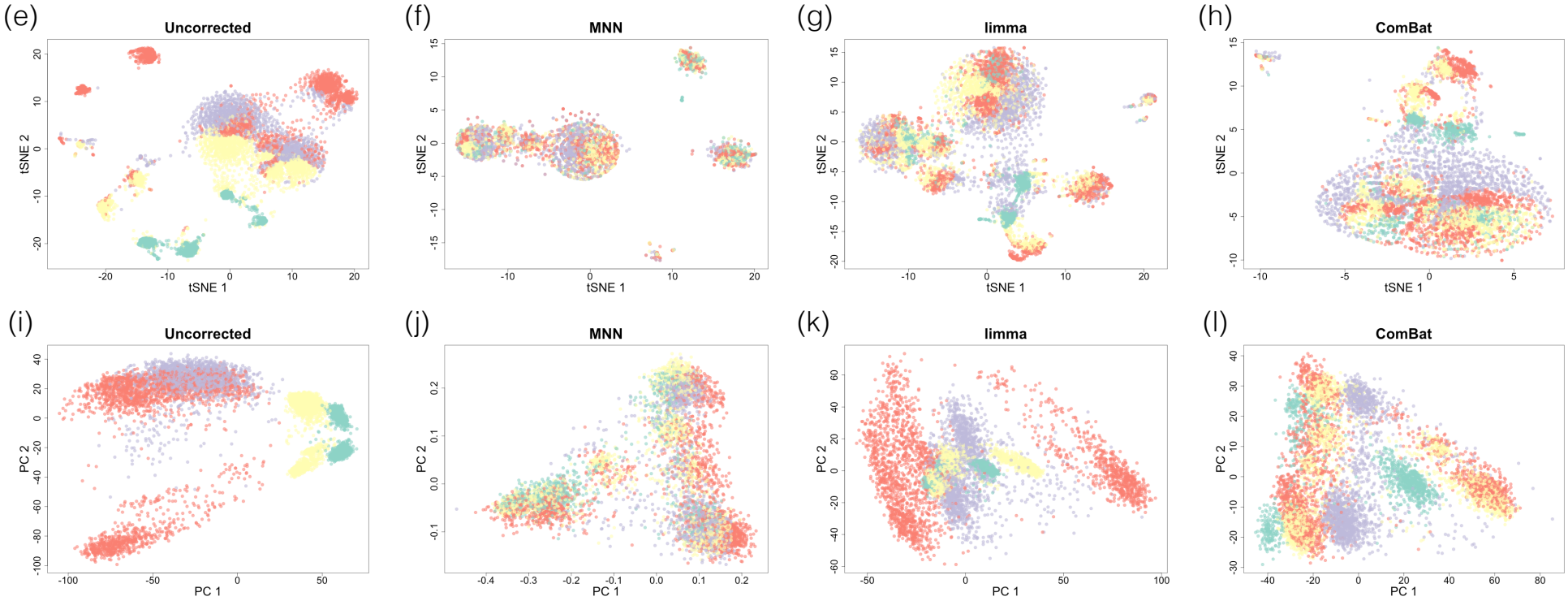
- Locally linear (overall nonlinear)
- Tolerates small non-orthogonalities

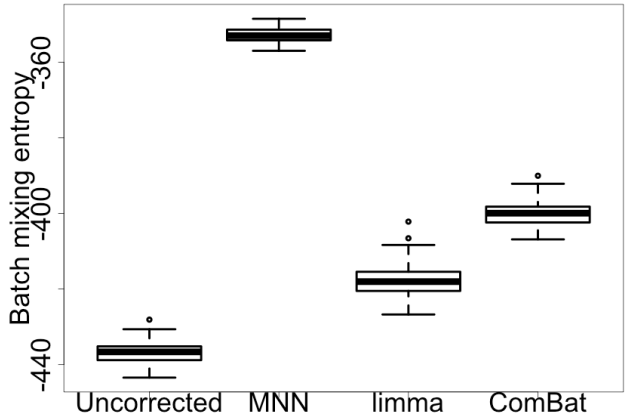


MNN on four Pancreas data sets

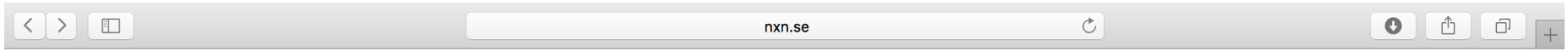


MNN on four Pancreas data sets





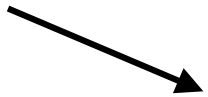
With autoencoders



WHAT DO YOU MEAN "HETEROGENEITY"?

VALENTINE SVENSSON

BLOG ARCHIVE NOTES PHOTOGRAPHY PUBLICATIONS ABOUT



2018

Count based autoencoders and the future for scRNA-seq analysis
Apr 20, 2018

Illustrating spatially variable genes with a 1-dimensional zebrafish
Apr 5, 2018

Actionable scRNA-seq clusters
Mar 5, 2018

The effect of Poisson zeros on OLS regression results
Feb 27, 2018