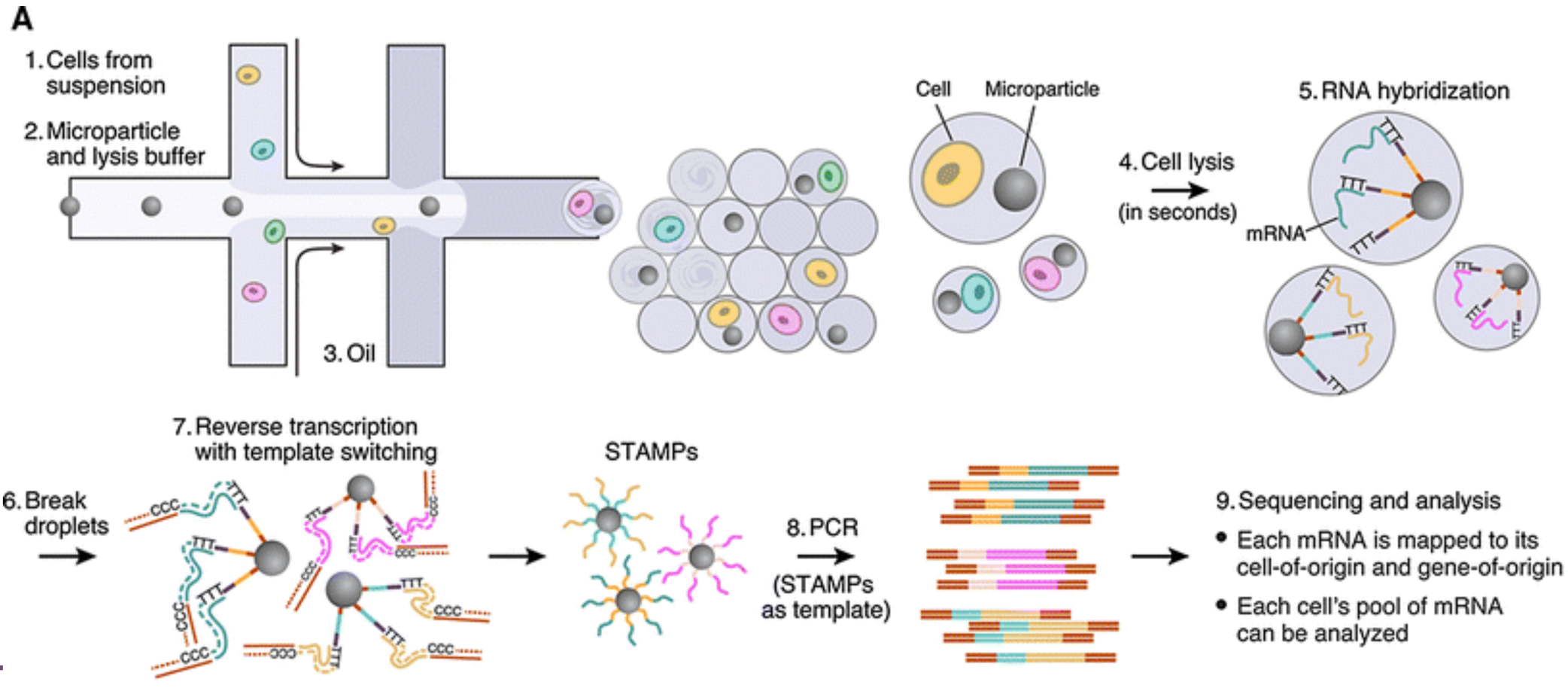
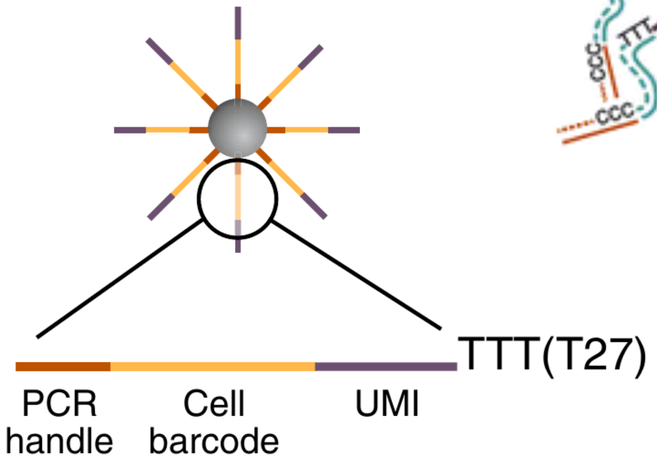
Dimension Reduction Techniques

LALEH HAGHVERDI

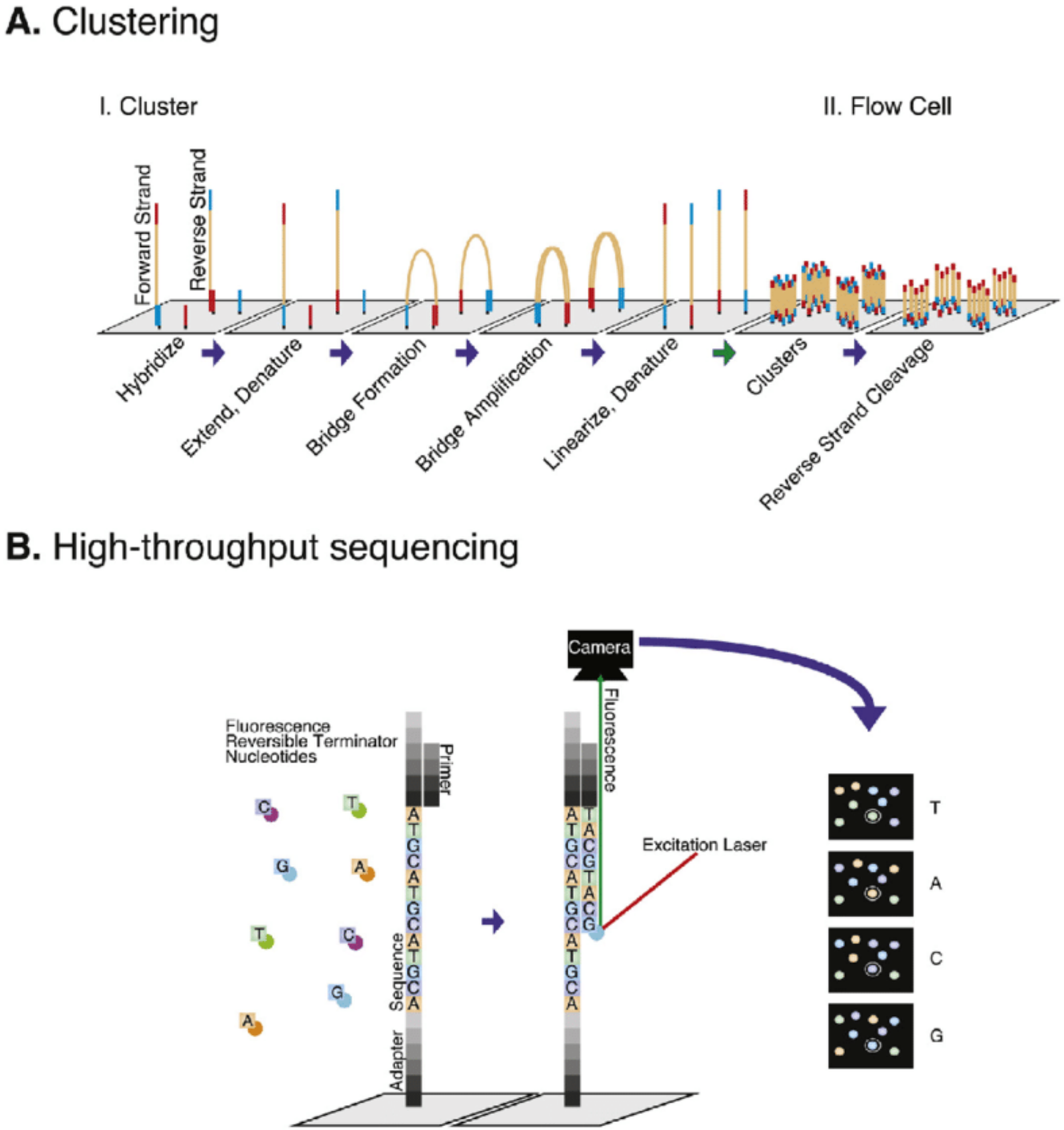
Single-cell omics





Choi JR, Yong KW, Choi JY, Cowie AC. Single-cell RNA sequencing and its combination with protein and DNA analyses. Cells. 2020 May 4;9(5):1130.

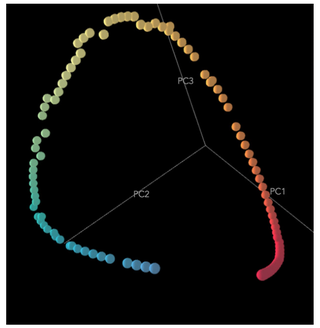
10x Illumina technology



Curse of dimensionality

- Expansion of sampling space and vagueness of distance measures (curse of dimensionality), some algorithms e.g. minimum spanning trees (MST) won't work properly

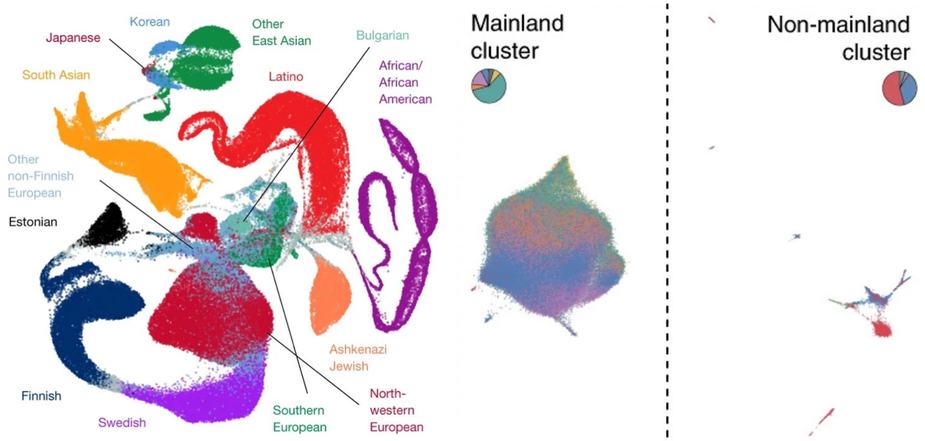
$$\lim_{d\to\infty} E\left(\frac{\mathrm{dist}_{\max}(d) - \mathrm{dist}_{\min}(d)}{\mathrm{dist}_{\min}(d)}\right) \to 0.$$



- Horse-shoe effect in PCA of random points D >>> N
- Weird string shapes in UMAP and t-SNE, overfitting
- especially in population genetics, no. variants >>> no. patients

Diaz-Papkovich, et al. A review of UMAP in population genetics. *J Hum Genet* (2021)

The Genome Aggregation Database (gnomAD, left) and Biobank Japan (BBJ, right) visualized using UMAP.



(a) gnomADv3 data visualized using UMAP. (b) BBJ data visualized using UMAP.

~36 million singleton variants from 3560 whole-genome sequences
Carlson, J. *et al. Nat Commun* (2018)

(a) gnomADv3 data visualized using UMAP.

Exercises:

Take a genomic variants (mutations dataset) and demonstrate that the string and elongated shapes are artefact of curse of dimensionality

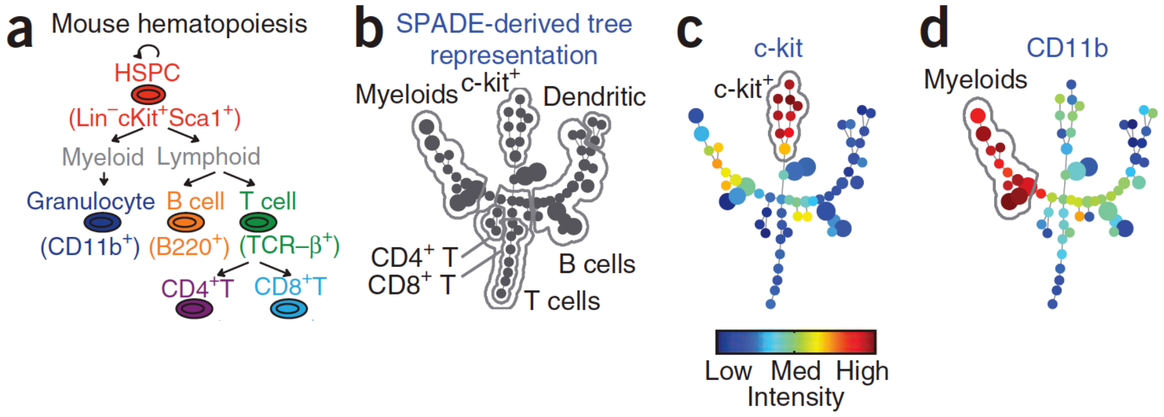
a) e.g. from [A review of UMAP in population genetics, Diaz-Papkovich, Alex and Anderson-Trocm'e, Luke and Gravel, Simon, Journal of Human Genetics2021]

b) Reproduce their UMAP in Figure 4.

c) Take 5 random subsets of the variants with different length. For which length (i.e., number of features) the artefacts of curse of dimensionality disappear from the UMAP?

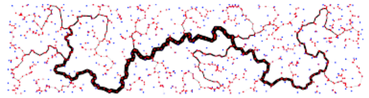
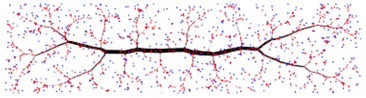
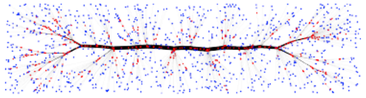
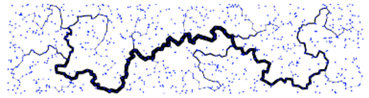
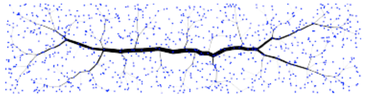
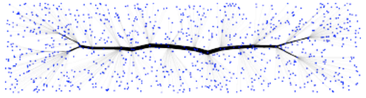
Example: Minimum Spanning Tree (MST) is sensitive to noise and curse of dimensionality

Qiu, Peng, et al. "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE." *Nature biotechnology* (2011).



Edge Centrality Exponent α

Add Steiner points

SP \ α	0	$(0,1)$	1
YES	**Steiner Tree** Gilbert and Pollak (1968) $\min_{X_B,T} \sum_{(i,j)\in E_T} \|\|x_i - x_j\|\|$ 	**Branched Central Spanning Tree** $\min_{X_B,T} \sum_{(i,j)\in E_T} \left(m_{ij}(1-m_{ij})\right)^{\alpha} \|\|x_i - x_j\|\|$ 	**Branched Minimum Routing Cost Tree** $\min_{X_B,T} \sum_{i,j\in V\times V} d_T(i,j)$ 
NO	**Minimum Spanning Tree** Kruskal (1956) $\min_{T} \sum_{(i,j)\in E_T} \|\|x_i - x_j\|\|$ 	**Central Spanning Tree** $\min_{T} \sum_{i,j\in E_T} \left(m_{ij}(1-m_{ij})\right)^{\alpha} \|\|x_i - x_j\|\|$ 	**Minimum Routing Cost Tree** Masone et al. (2019) $\min_{T} \sum_{i,j\in V\times V} d_T(i,j)$ 

Sanmartín, Enrique Fita, Christoph Schnoerr, and Fred A. Hamprecht. "The Central Spanning Tree Problem." (2024).

Exercises:

a) Can you show minimum spanning trees for high dimensions and few samples (i.e., in presence of curse of dimensionality) have similar problems as minimum spanning trees for noisy data?
b) Are any of the regularisations introduced by Sanmartín et al. (previous page) for dealing with noisy data helpful in presence of curse of dimensionality?

Counts matrix processing steps

Standard pipelines (e.g. Seurat, scanpy)

~Data transformation

1. (Library) size factor normalisation
2. Feature selection,
   - Uninteresting genes (mitochondrial, ribosomal, etc) filtering + cells filtering (doublets etc)
   - HVG selection
   - Other, depending on the question e.g. about transitions or separate clusters?
3. Log transform (UMI vs. non-UMI)
4. Metric (Euclidean, Cosine distance, etc) choice
5. Dimension Reduction and visualisation
6. Data integration, (Unsupervised) Clustering and/or Cell type label transfer

❋ Order of steps (1-3)?!

Counts matrix processing steps

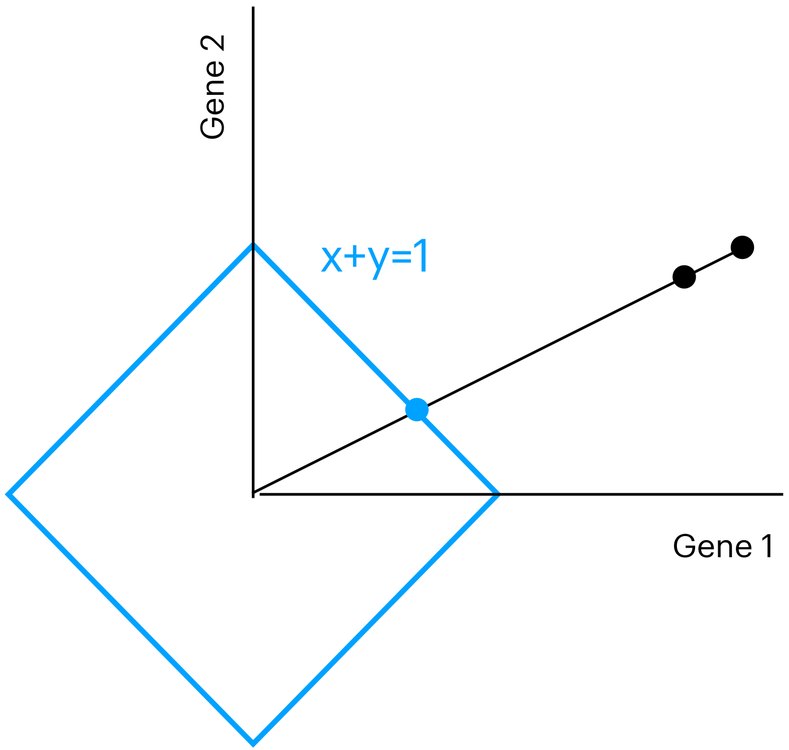
~Data transformation

My suggestion (it is not the consensus!):

1. Feature selection (data and problem specific),
    1. Uninteresting genes (mitochondrial, ribosomal, etc) filtering + cells filtering (doublets etc)
    2. HVG selection (by Pearson residuals)
    3. Other, depending on the question e.g. about transitions or separate clusters?
2. Log transform (both UMI and non-UMI)
3. Size factor normalisation
4. Now Euclidean ~ Cosine distance
5. Dimension Reduction and visualisation
6. Data integration, (Unsupervised) Clustering and/or Cell type label transfer

Size factor normalisation: use only genes' relative expression

- Different cell sizes
- But also different batches and efficiencies
- Absolute count values are not really meaningful
    - Different technologies, different read lengths, amplification and counting procedure
- Only the ratios among my measurements on different genes matters



- Even this is far from perfect:
    - Different genes may have different capture efficiencies in different technologies
    - Dropout effects
    - Distinction of Differentially Expressed genes and fold changes
    - DE analysis is still usually done on raw counts, not transformed data)

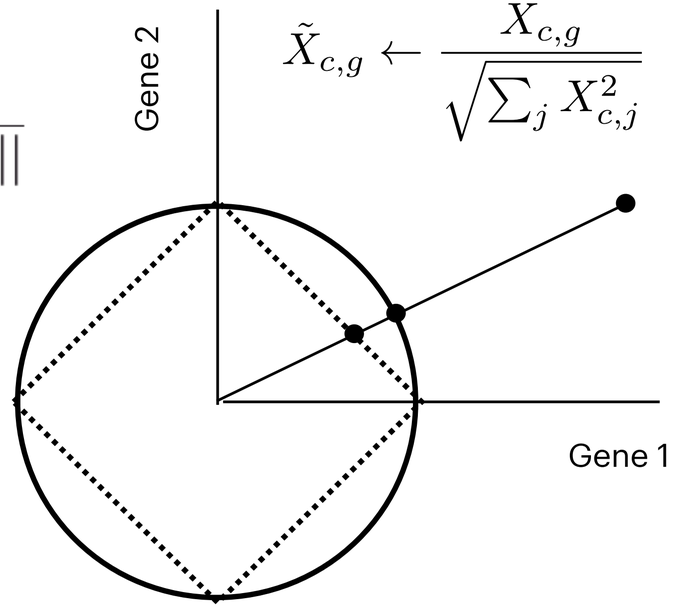
$$\tilde{X}_{c,g} \leftarrow \frac{X_{c,g}}{\sum_j X_{c,j}}$$

Cosine or Euclidean distance for expression data?

- Cosine distance: care only about the angel between two data points
- So after size factor normalisation Euclidean or cosine distance don't make any meaningful difference

- In fact after L2 normalisation, there Euclidean and Cosine distance become exactly the same

$$D^2(x,y) = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} - \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||}\right)^2 + \left(\frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 - 2\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} \cdot \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}$$
$$= 2\left(1 - \frac{\mathbf{Y_x} \cdot \mathbf{Y_y}}{||\mathbf{Y_x}||||\mathbf{Y_x}||}\right) = 2 \cdot \text{cosine.distance}(x,y)$$



Data integration by matching Mutual Nearest Neighbours
Haghverdi,et al., Nature biotechnology 2018 (Supplement!)

scRNA-seq transform Summary

- HVG and other feature selections
- (Library) size factor normalisation
- Log transform (UMI vs. non-UMI)
- Order of steps
- Metric (Euclidean, Cosine distance, etc) choice

- There are sometimes redundant discussions, or doing unnecessary things or doubling, or dirty non-optimal procedures…
- Several (available) works flows can give you valid results
- So don’t worry too much! Just know what are the important properties in your data / question and make sure those make sense

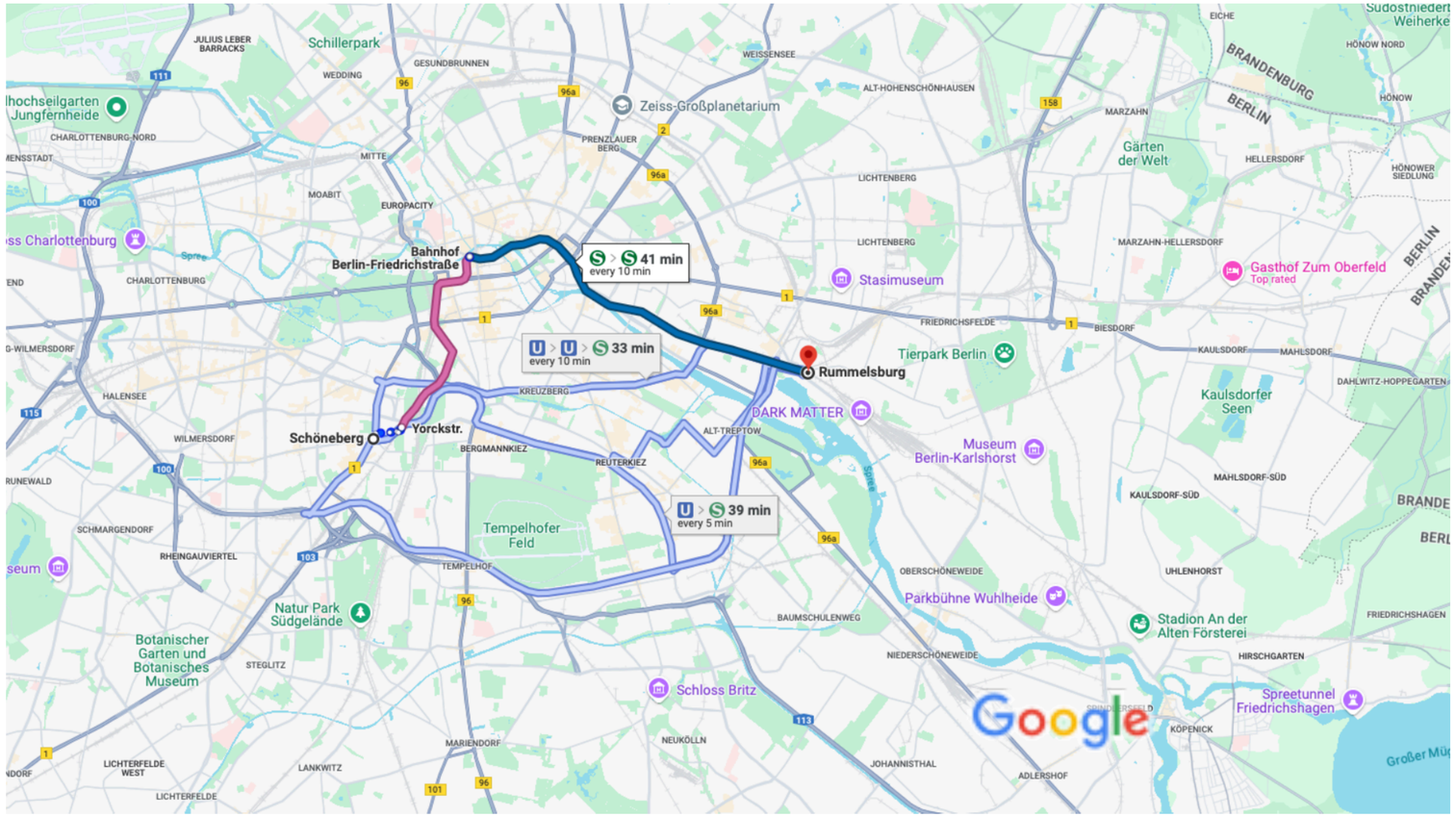
- Nevertheless:
  - We want a standard approach among all researchers for
    - Interpretability
    - Reproducibility
    - Communication
    - Data integration
  - Seurat, scanpy, Human Cell Atlas (HCA) consortium

Exercises:

a) What tests would you do to show a preprocessing pipeline is advantageous over another?

Different distance measures for different data/problems

- Example:
  - Euclidean distance
  - Great-circle (on earth) distance
  - S-bahn distance
  - S-+U-bahn distance
  - Drive distance



Map data ©2024 GeoBasis-DE/BKG (©2009), Google 2 km

Metrics: properties and examples

- Valid metric (distance):

$$D(x,x) = 0$$
$$D(x,y) = D(y,x)$$
$$D(x,z) \leq D(x,y) + D(y,z)$$

- Distance between data points: Hamming distance, Euclidean, Cosine, geodesic, Diffusion distance,...
- Distance between distributions: Wasserstein (Earth movers distance), (application in Variational methods, image processing,...), Jensen-Shannon Divergence (JSD)

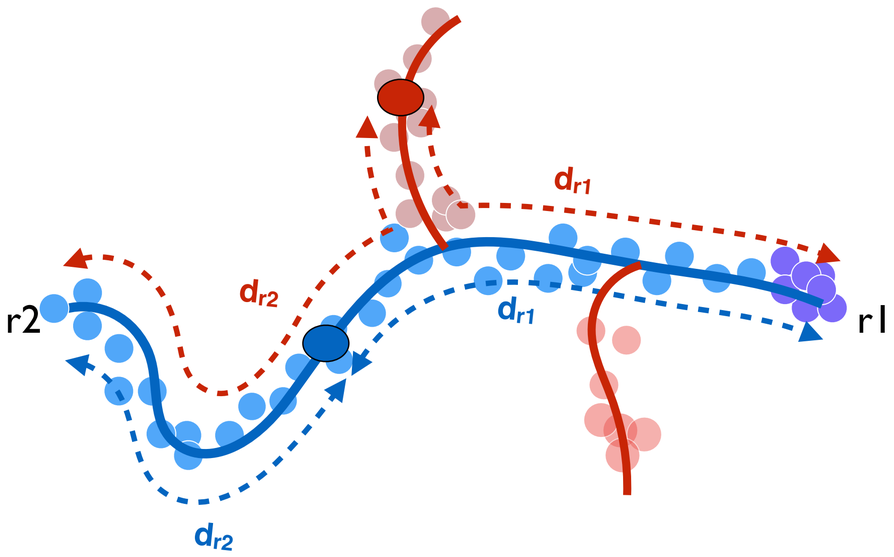
$$JSD(P,Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$
$$M = \frac{1}{2}P + Q$$

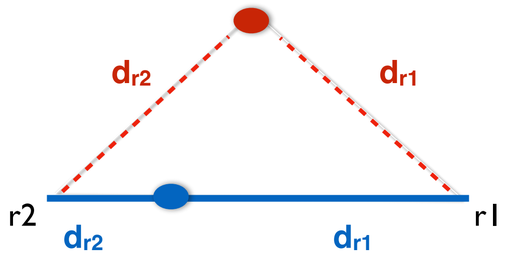
- Note Kullback–Leibler divergence is not a distance itself!

$$\mathrm{KL}(P||Q) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Trajectory analysis: pseudotime ordering and branch identification

- Pseudotime: on-manifold distance from the root cell
- Branch identification: generalised triangle inequality





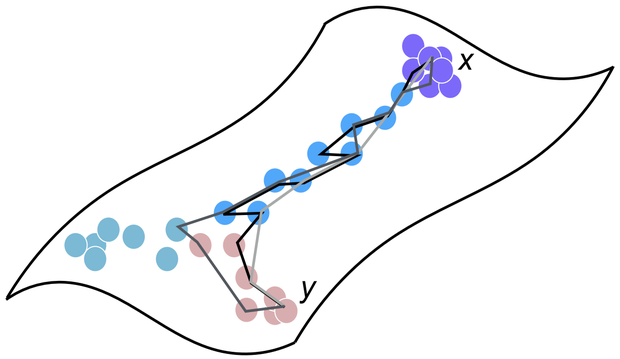
We need a continuity keeping DR method

Haghverdi *et al. Nature Methods* 2016 (Supplement)

(Geometrical) Pseudotime

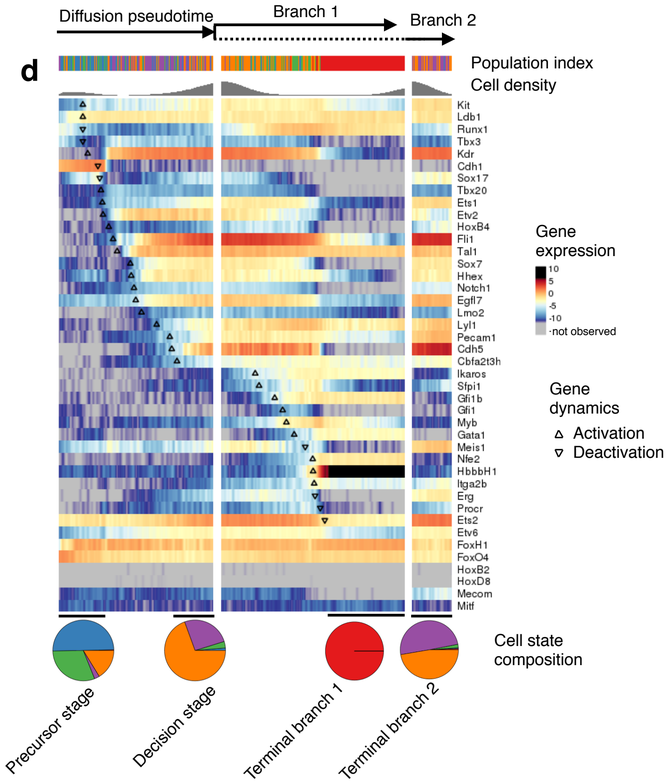
"Diffusion pseudotime robustly reconstructs lineage branching" Haghverdi et al. Nature methods 2016

- Leverage asynchrony of cells progress to reconstruct the differentiation manifold
- Use distances on some reduced dimension (latent space) as PT
- Among DR methods, diffusion maps/pseudotime are especially suitable for differentiation data analysis
    - Robust to noise
    - Biological interpretation of diffusion
- Limited complexity of the manifold (no. branchings)

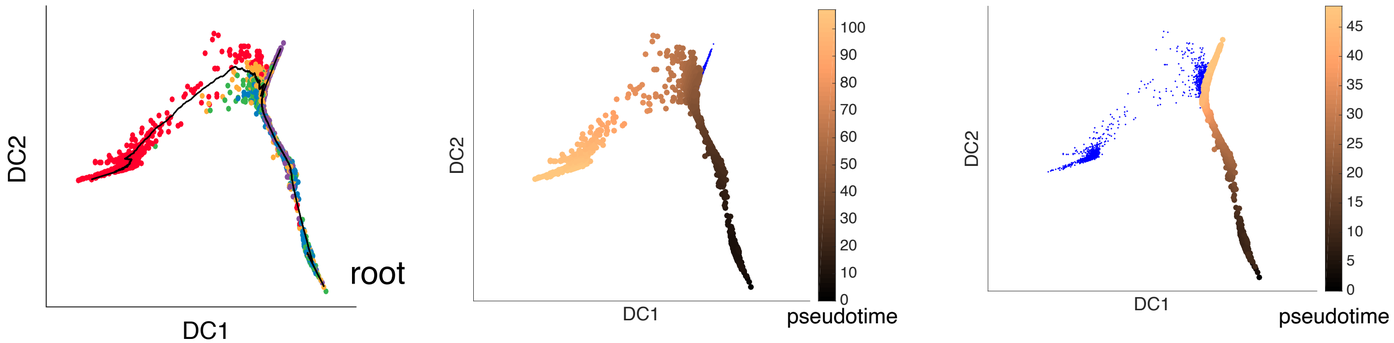


Accumulated probability transition matrix M

$$\mathrm{dpt}^2(x,y) = xM^2x + yM^2y - 2xM^2y$$

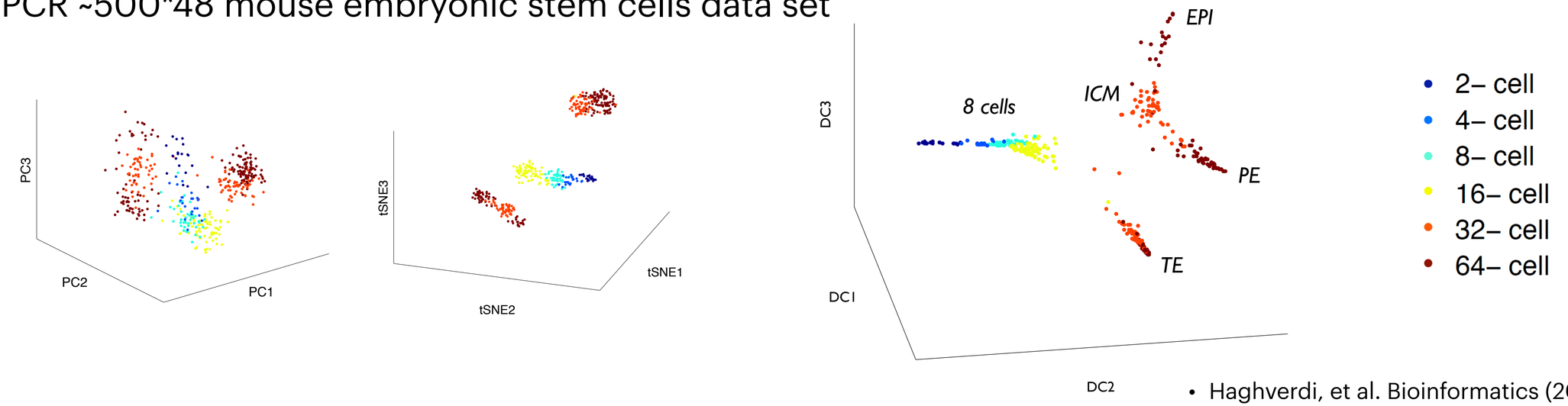


sc-qPCR ~4000*42 mouse early blood development



Importance of feature selection

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



- Haghverdi, et al. Bioinformatics (2015)
- Guo *et al. Developmental cell 2010*

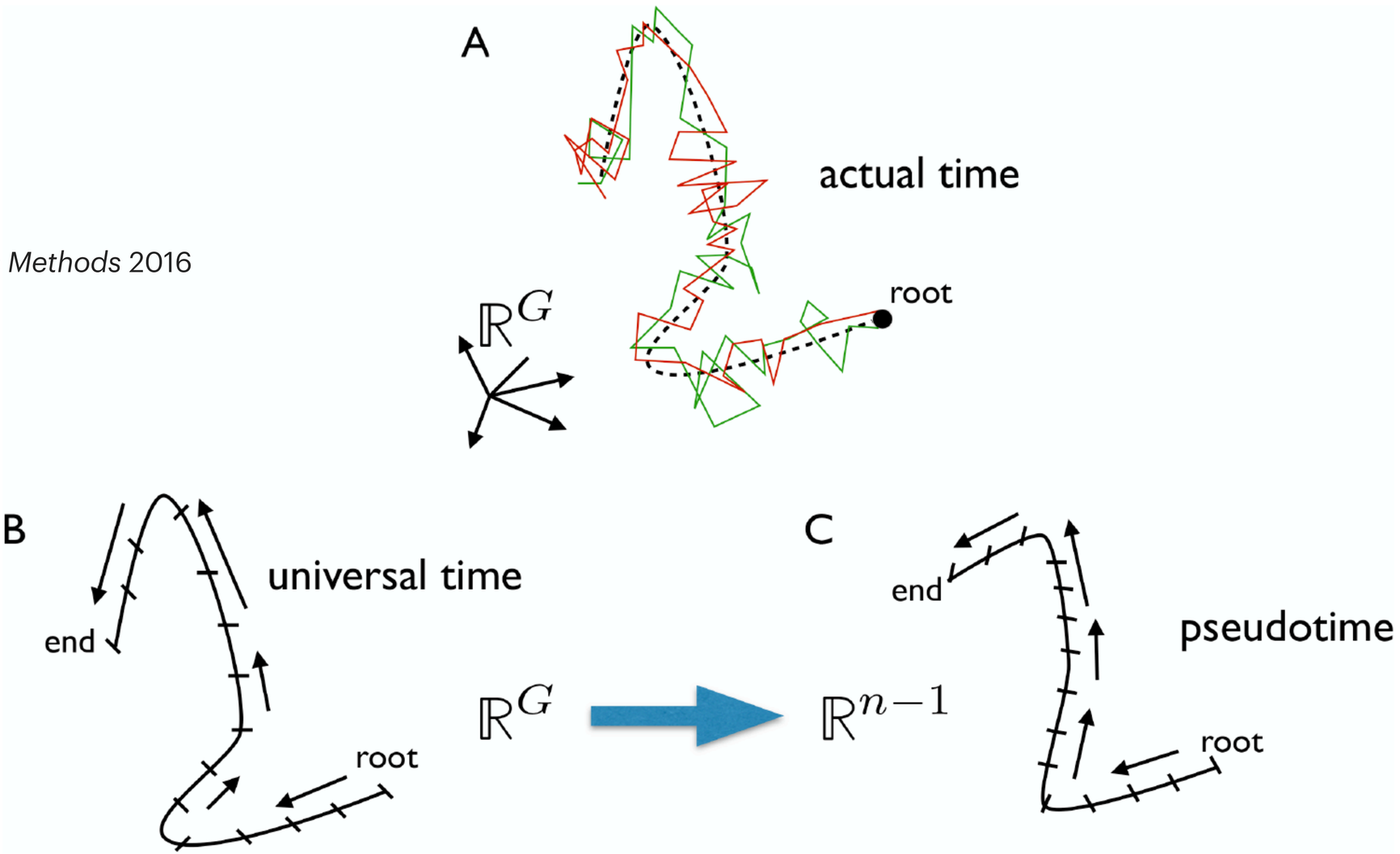
- Avoid weird high-D effects (curse of dimensionality)
- Even randomly selecting fewer features helps!
- Sufficiently covered data manifold
- But also keeping more interesting features, e.g. for better separation of specific clusters (cell type? Or patients? Or ...) or for capturing trajectories

- Ranek JS, et al. DELVE: feature selection for preserving biological trajectories in single-cell data. Nature Communications, 2024.
- Cui, S., et al. Mcadet: a feature selection method for fine-resolution single-cell RNA-seq data based on multiple correspondence analysis and community detection. *PLOS Computational Biology,* 2024.

- ~2000 genes for how many cells? Don't take HVGs too seriously!

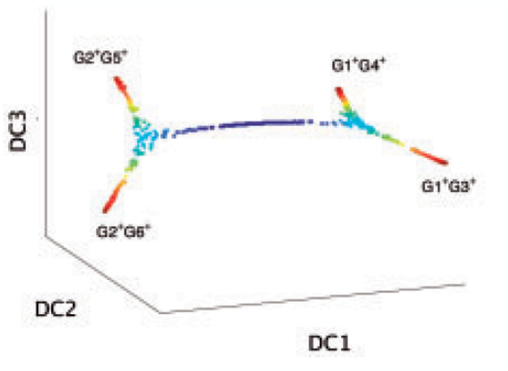
How pseudotime is defined

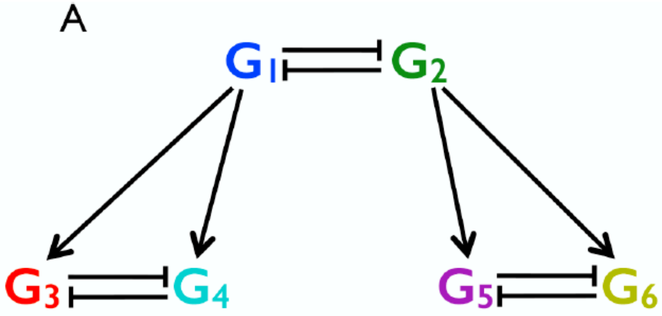
Haghverdi *et al*. *Nature Methods* 2016
(Supplement)

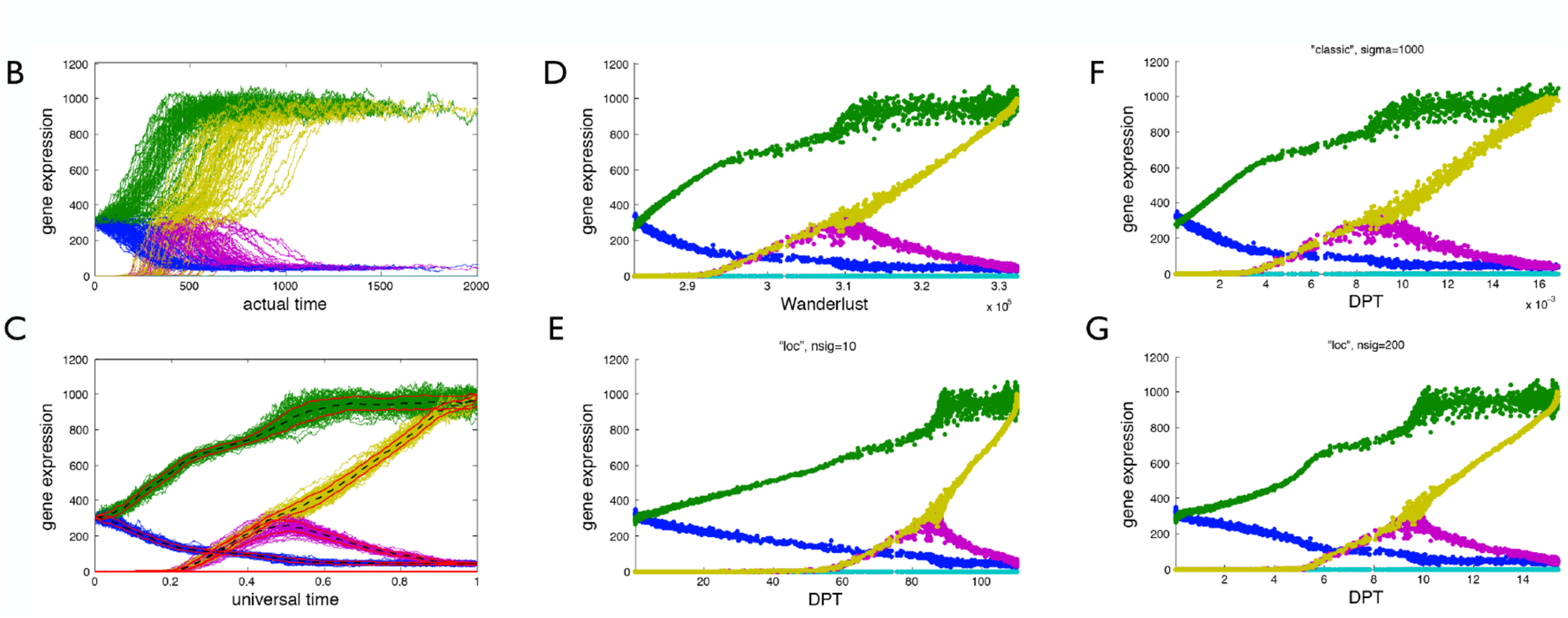


Supplementary Figure N7: A) Two (red and green) actual time single cell trajectories in gene expression space($\mathbb{R}^G$). Each jump on a trajectory happens in an (equidistant) unit of actual time. B) Universal time is defined as *arc* length on the data manifold starting from the root. This manifold $C \subset \mathbb{R}^G$ remains the same for several single cell trajectories, as well as for snapshot samples of single cells. C) Pseudotime (in

Actual time, universal time, pseudotime





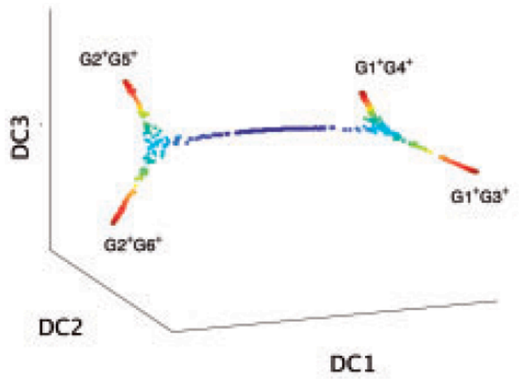


universal time

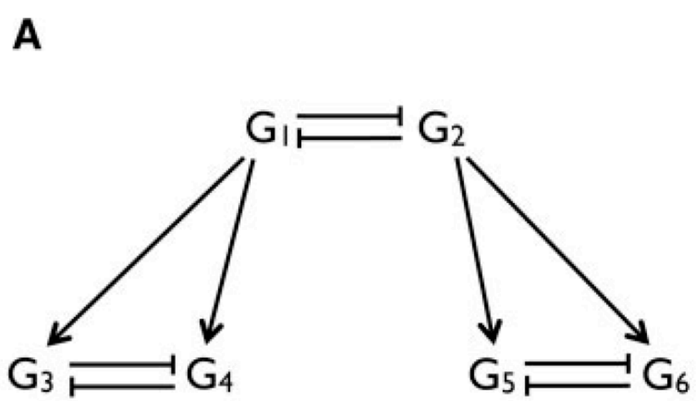
$$u(t) = \int_0^t \frac{1}{\rho(t')} dt'$$

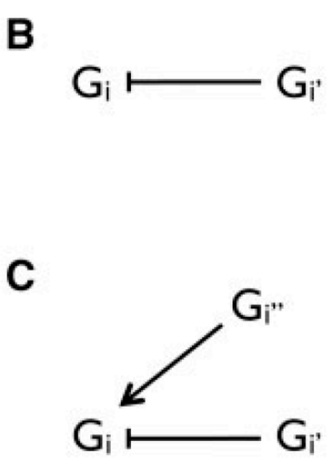
Exercises:

a) Use the package BeeLine and BoolODE to simulate data and differentiation trajectory as below, then share the notebook with the class on Github. Specify the gene-gene interaction strength as well as the degradation rates. How can you increase the noise level? As we will need this simulation in future exercises as well, could you share your code with everyone?



Haghverdi, et al. Bioinformatics (2015)





$$\varnothing \xrightarrow{\alpha_i} \text{Protein}_i \xrightarrow{\gamma} \varnothing$$

$$\alpha_i = k_i \frac{1}{1 + (\frac{\text{Protein}_{i'}}{k_{i'}})^{m_1}}$$

$$\varnothing \xrightarrow{\alpha_i} \text{Protein}_i \xrightarrow{\gamma} \varnothing$$

$$\alpha_i = k_i \frac{1}{1 + (\frac{\text{Protein}_{i'}}{k_{i'}})^{m_1}} \cdot \frac{1}{1 + (\frac{k_{i''}}{\text{Protein}_{i''}})^{m_2}}$$

Matlab Code for toggle switch data simulation with 6 genes

Page 1

```
decay         = 0.25;
synth         = 250;
K_d_inhibition = 200;
K_d_activation = 400;
n_inhibition   = 2;
n_activation   = 20;
init = [300 300 0 0 0 0]';
sigma   = 100e-0;
dt      = 1;
T       = 240;
a(:,1)  = init;
nRealiz = 400;
% ---------------------------------------------------------------------
% ---------------------------------------------------------------------
MM = 100000;
for i = 1:MM
   rp = randperm(2); rp = rp(1);
   if rp==1
      np = randn/8 + 0.22;
   else
      np = randn/8 + 0.77;
   end
   Np(i) = np;
end
k = 1;
for i = 1:length(Np)
   if (Np(i)<1)&&(Np(i)>0)
      SampleDistribution(k) = Np(i);
      k = k+1;
   end
end
figure, hist(SampleDistribution,100), title('Sampling strategy')
MM = length(SampleDistribution);
```

Matlab Code for toggle switch data simulation with 6 genes

Page 2

```
% ---------------------------------------------------------------------
% ---------------------------------------------------------------------
for j = 1:nRealiz

   clear a
   a(:,1) = init;

   for i = 1:T-1

      a(1,i+1) = a(1,i) + dt*(synth*(1/(1+ (a(2,i)/K_d_inhibition)^n_inhibition)) - decay*a(1,i)) + sqrt(sigma)*randn;
      a(2,i+1) = a(2,i) + dt*(synth*(1/(1+ (a(1,i)/K_d_inhibition)^n_inhibition)) - decay*a(2,i)) + sqrt(sigma)*randn;

      a(3,i+1) = a(3,i) + dt*(synth*(1/(1+(K_d_activation/a(1,i))^n_activation))*(1/(1+ (a(4,i)/K_d_inhibition)^n_inhibition)) - decay*a(3,i)) + sqrt(sigma)*randn;
      a(4,i+1) = a(4,i) + dt*(synth*(1/(1+(K_d_activation/a(1,i))^n_activation))*(1/(1+ (a(3,i)/K_d_inhibition)^n_inhibition)) - decay*a(4,i)) + sqrt(sigma)*randn;

      a(5,i+1) = a(5,i) + dt*(synth*(1/(1+(K_d_activation/a(2,i))^n_activation))*(1/(1+ (a(6,i)/K_d_inhibition)^n_inhibition)) - decay*a(5,i)) + sqrt(sigma)*randn;
      a(6,i+1) = a(6,i) + dt*(synth*(1/(1+(K_d_activation/a(2,i))^n_activation))*(1/(1+ (a(5,i)/K_d_inhibition)^n_inhibition)) - decay*a(6,i)) + sqrt(sigma)*randn;

      if sum(a([3 4 5 6],i+1)>950)
         ind = ceil(i*SampleDistribution(round(rand*MM)));
         DataDiff(1:6,j) = a(1:6,ind);
         break
      end

   end

end

figure, plot(a'), title('Single realisation from toggle switch network')

% save diffusionInput DataDiff
```

Grading system

- 10 points: (~5) Exercises and seminars
- 10 points: final project
  - Groups of 2-3 people
  - 5-6 pages report including Abstract, Intro, Methods, Results, Discussion
  - In the Abstract specify what problem you are addressing and based on which lecture(s)/exercise(s) it is
  - Share coding scripts and data (on GitHub) for reproducing your plots

Dimension reduction techniques

Why dimension reduction?

- Interpreting high-D data (even after feature selection) is confusing
- Visualisation
- Facilitate biological interpretation and downstream analysis,
  - Cell type annotation
  - Trajectory and pseudotime inference
- Denoising
- Data compression (latent space)

Linear DRs

$$V : \mathbb{R}^G \to \mathbb{R}^K$$

$$Y = XV \qquad\qquad Y_{ik} = \sum_{j=1}^{G} X_{ij} V_{jk}$$

$$[N, K] = [N, G][G, K]$$

- Sometimes non-linear DRs needed especially with lower dim. data
- In higher dims more likely there is a good transformation for what you look for
- Maybe not the maximum variance (PCA) though...
- Examples: SVD, CCA, ICA, supervised regression, ...

Principal Component Analysis (PCA)

Standardised (gene-wise) the data matrix $X \quad [N, G]$ i.e., mean = 0 and variance=1 for each dimension

$$\text{argmax}_V \sum_{k=1}^{K} \underbrace{V^T X^T}_{Y^T} \underbrace{XV}_{Y}$$

$$V \quad [G, K], \quad K \leq G$$

$$V^T V = I$$

The solution for V is: eigen-vectors of $X^T X$ (corresponding to largest eigen values)

$$Y \quad [N, K]$$

$$V^T X^T X V = V^T \Sigma V = \Sigma \longrightarrow \text{Diagonal matrix}$$

The rotation W diagonalises $Y^T Y$

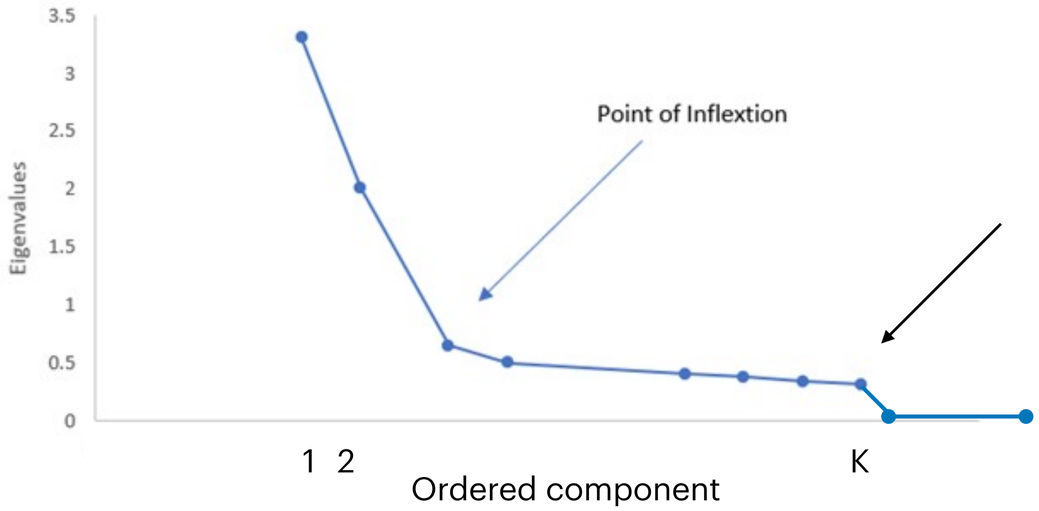
A [G,G] sized V as such (orthonormal matrix) is in fact a rotation matrix

so we are maximizing the sum of variance of data points in the embedding space 1:K

Eigenvalues of X^T X == explained variance of each component

- PCA on X [N,G] can have maximally K nonzero eigenvalues

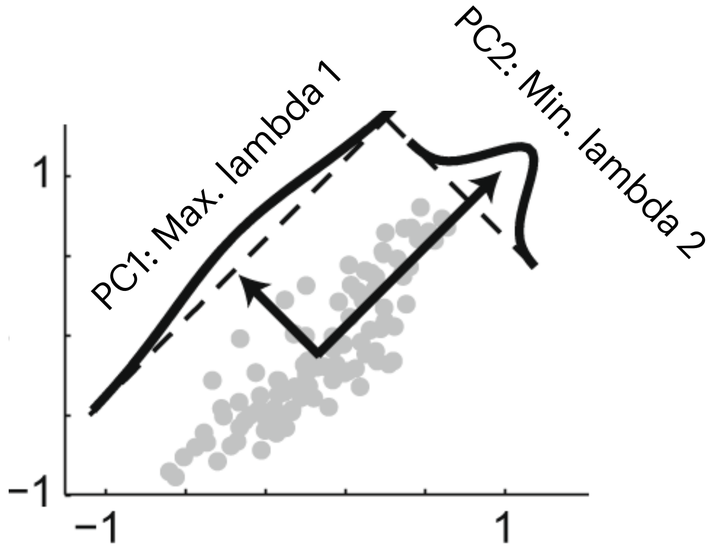
$$K = \min(N, G)$$



- With real valued X, X^TX will be positive definite thus Sigma is positive and we can write:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \sigma_G^2 \end{bmatrix}$$

- PCA as a regression model



$$\text{argmax}_V(\overbrace{\sum_{k=1}^{K} \sigma_k^2}^{\text{Explained variance}}) \equiv \text{argmin}_V(\overbrace{\sum_{k=K+1}^{G} \sigma_k^2}^{\text{Sum of errors^2}})$$

$$\sigma^2 = \sum_{k=1}^{G} \sigma_k^2$$ Explained variance

Exercises:

1) What is Singular Value Decomposition (SVD) and how does it relate with PCA?
2) What is Independent Component Analysis (ICA) and how does it relate with PCA?

Nonlinear DR techniques

Challenges of non-linear DR

Which intrisinc 3D manifolds and structures can be presented in 2D with diffeomorphism?

- Line?
- Arc?
- Sphere?
- Cylinder?
- Piramid?
- Tree? (E.g. phylogenetic, some differentiation paths)

→

- Crowding
- Breaks and cuts
- Altered areas and aspect ratios

World map nonlinear 3D —> 2D

- Distortion of distances and areas?
- Some cuts (discontinuity) in the data?
- Other (multiple) embeddings possible?
- It is diffeomorphism?
- Is this useless?

Exercises:

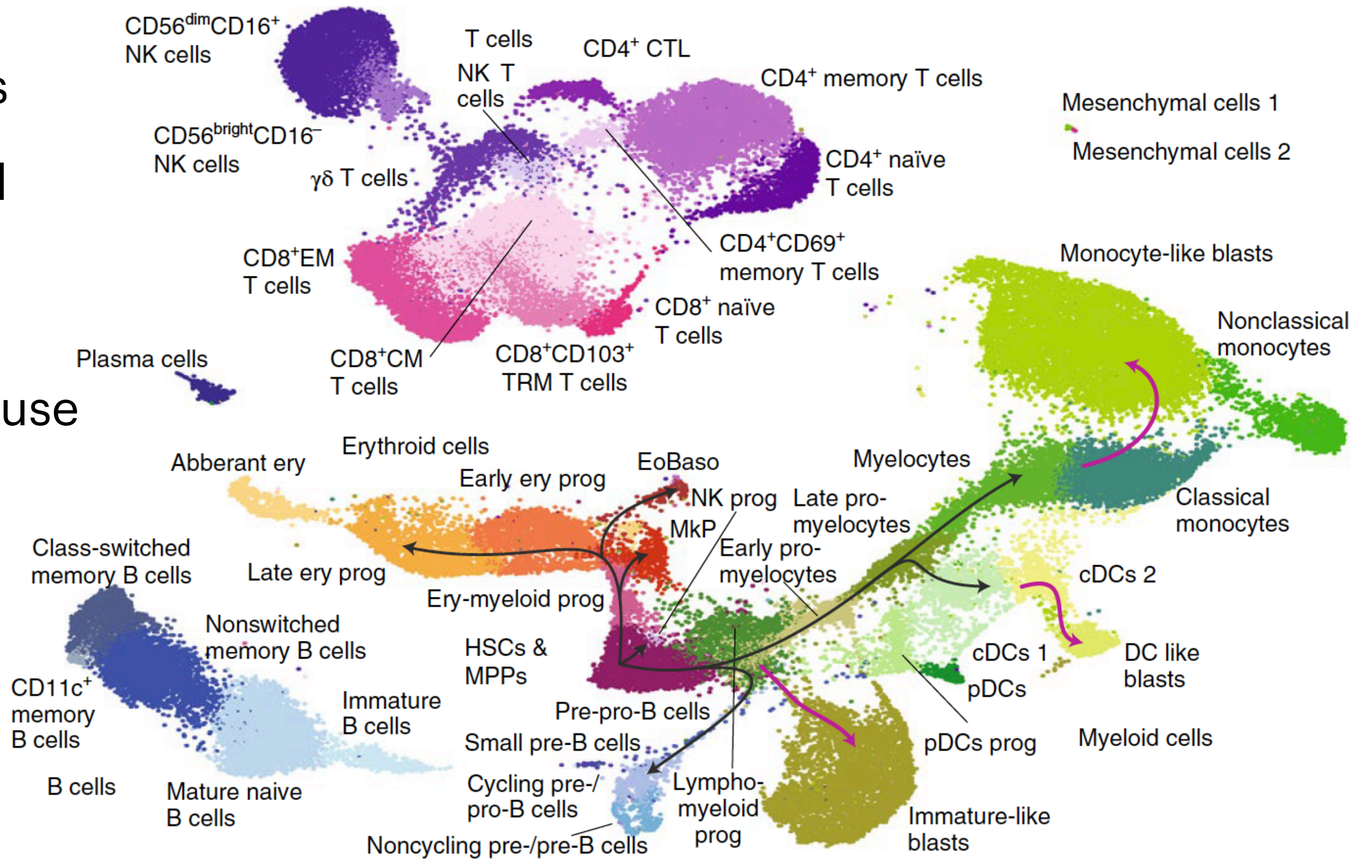
a) List some practices for notifying break lines or non-conserved areas on nonlinear maps.

What is a good metric in gene expression space?

- We can only talk about the sampled regions
- Distance measures along the manifolds and trajectories would be ideal
- For completely separate cell clusters, who knows what is a good metric?! So better to use transition probabilities rather than raw distances

UMAP on bone marrow scRNA-seq & surface markers data

Triana *et al. Nature Immunology 2021*



Non-linear dimension reduction methods

- Minimize the cost function, either

Original (high-D) space distances

Embedding space distances

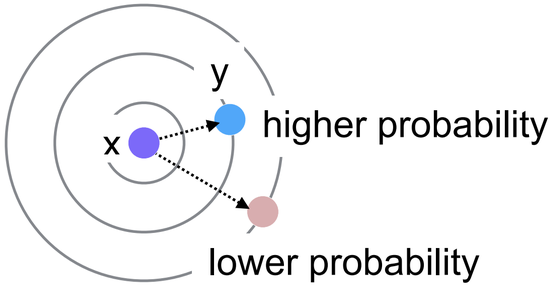
$$L = \sum_{i,j} g(d_{ij}) \cdot f(\tilde{d}_{ij})$$

$$F_i = \frac{\partial L}{\partial Y_i} = 0$$

  - Analytical
    - Laplacian eigenmap ~ Diffusion map, MDS, ...
  - Numercal e.g. (stochastic) gradient descent
    - t-SNE, UMAP, pacmap,...

Laplacian Eigen-maps (analytical solution)

Gaussian kernel



- Find a mapping that minimises the cost function

$$\mathbb{R}^G \rightarrow \mathbb{R}^K$$

Distance in low-D

Gaussian Kernel in high-D

Cost = $$\sum_{i,j} ||Y_i - Y_j||^2 W_{ij} = trace(Y^T DY - Y^T WY) = trace(Y^T LY)$$

$$\text{argmin}_Y \quad trace(Y^T LY)$$

$$\text{subject to} \quad Y^T DY = I$$

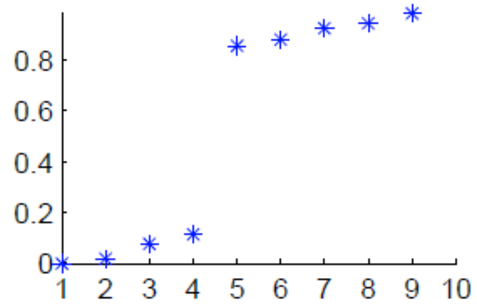
$$L = D - W \qquad D_{ii} = \sum_j W_{ij}$$

Laplacian matrix

$$Ly = \lambda' Dy \qquad \lambda'_i \geq 0$$

$$D^{-1} LY = \Lambda' Y$$

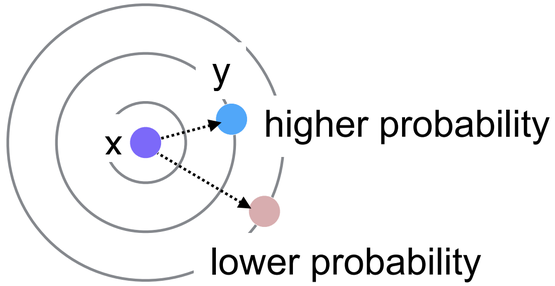
Eigenvalues



Diffusion maps

Coifman *et al. PNAS 2005*

W: Gaussian kernel

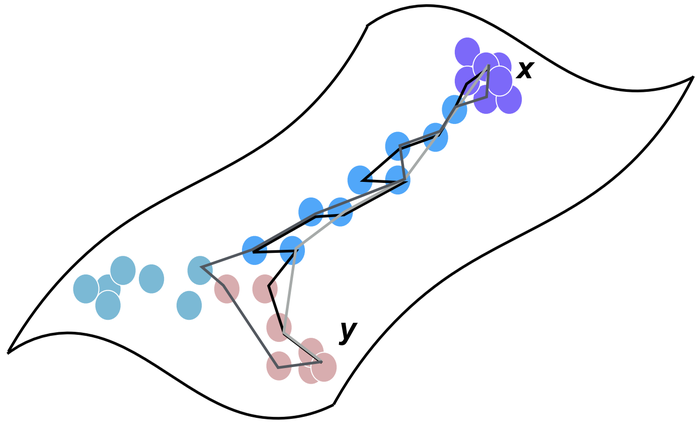


$$P = D^{-1}W \qquad D_{ii} = \sum_j W_{ij}$$

$P_{n\times n}$ row normalised transition matrix (from a Gaussian kernel)

P^t transition probabilities in t steps

$$\phi_i P = \lambda_i \phi_i, \quad P\psi_i = \lambda_i \psi_i, \quad \lambda_i \leq 1$$



$$D_t^2(x,y) = ||P^t(x,.) - P^t(y,.)||^2_{1/\phi_0}$$

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

For large t we can cut the sum at i=K for low-dimensional approximation

Laplacian eigen-map vs. Diffusion map

For Laplacian eigenmap we had:

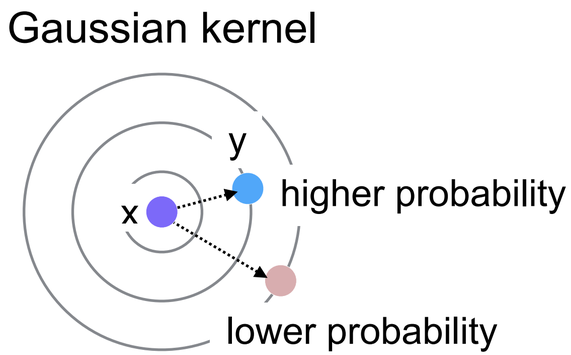
$$D^{-1}LY = D^{-1}(D-W)Y = \Lambda' Y \qquad 1 \geq \lambda'_i \geq 0$$

$$P = D^{-1}W$$

$$D^{-1}LY = (I-P)Y$$

$$P\Psi = \Lambda\Psi \qquad 0 \leq \lambda_i \leq 1$$

: For Diffusion map

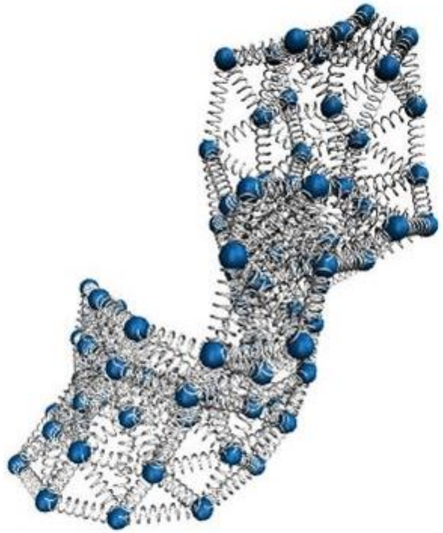


- So the eigen vectors matrix **Y** from the Laplacian eigenmap are the same as eigenvectors **\Psi** from diffusion map
- And for the Eigenvalues $\lambda'_i = 1 - \lambda_i$

Exercises:

a) Show that the eigenvalues of P are positive real values, even though P is not Hermitian.

Numerical solution non-linear dimension reduction methods



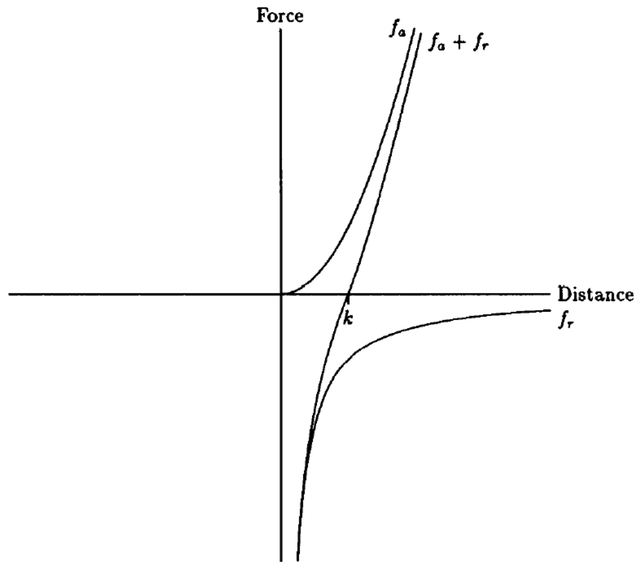
- Minimize the cost function

Original (high-D) space

Embedding space

$$L = \sum_{i,j} g(d_{ij}) \cdot f(\tilde{d}_{ij})$$

Move towards: $$F_i = \frac{\partial L}{\partial Y_i} = 0$$ (stochastic) gradient descent



- Force directed maps (t-SNE-, UMAP, ...)

$$F_i = \sum_j \mathbf{Repulsive} \cdot \tilde{d}_{ij} - \sum_j \mathbf{Attractive} \cdot \tilde{d}_{ij}$$

t-SNE

Van der Maaten, "Visualizing data using t-SNE.", 2008
Van Der Maaten, Laurens. "Accelerating t-SNE using tree-based algorithms."2014

Original (high-dim) space X

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

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

embedding (e.g. 2D) space Y

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

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

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

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

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

t-SNE

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

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

  Learning rate Momentum

  (Move Y in the direction of the largest gradient)

- The final solution for Y depends on the initialisation Y^0

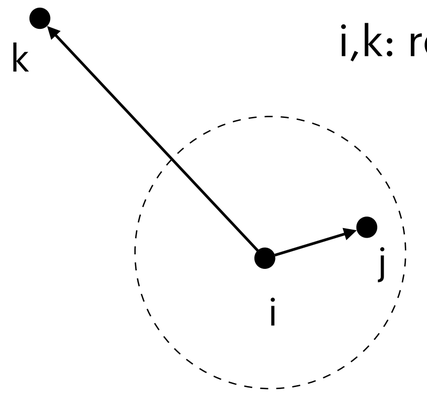
- Does not respect the global structure nor continuity of the data manifold

Computational cost and efficiency of optimization

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

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

- Stochastic gradient descent
- Negative sampling
- Contrastive learning
- Pre-defining attraction-repulsion zones

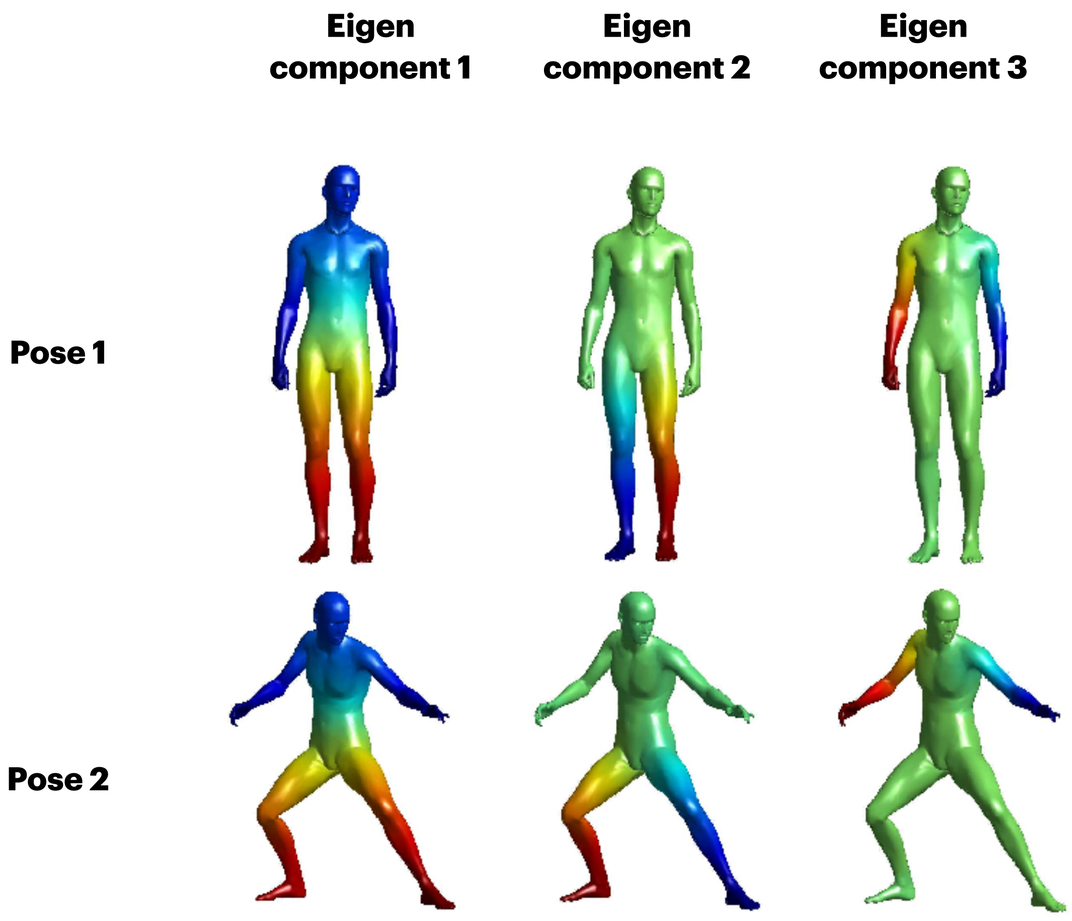


i,k: repulsive pair

i,j: attractive pair (i.e. n. neighbours in high-D)

Keeping continuity vs. global structure

- What are your objectives in your application of a metric space and dimension reduction?
- e.g. Laplacian eigen maps keep continuity but not the global structure
- For sc transcriptomics we care more about continuity than the global structure



Dubrovina, Anastasia, and Ron Kimmel. "Matching shapes by eigendecomposition of the Laplace-Beltrami operator." *Proc. 3DPVT*. Vol. 2. No. 3. 2010.

Non-linear dimension reduction methods

General form of the cost function: $$L = \sum_{i,j} g(d_{ij}) \cdot f(\tilde{d}_{ij})$$

t-SNE

$$L = \sum_{i,j} \frac{p_{ij} \log(p_{ij})}{\log(q_{ij})}$$

$$p_{ij} = \frac{1}{z_{X_i}} \exp\left(\frac{-d_{ij}^2}{2\sigma_i^2}\right)$$

$$q_{ij} = \frac{1}{z_Y}\left(\frac{1}{1+\tilde{d}_{ij}^2}\right)$$

UMAP

$$L = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-p_{ij}) \log \frac{(1-p_{ij})}{(1-q_{ij})}$$

(According to the paper, not the code!)

$$p_{ij} = \frac{1}{z_{X_i}} \exp\left(\frac{-max(0, d_{ij} - \rho_i)}{\sigma_i}\right)$$

$$q_{ij} = \frac{1}{z_Y} \cdot \left(\frac{1}{1+a\tilde{d}_{ij}^{2b}}\right)$$

Diffusion map
(Laplacian eigen map)

$$L = \sum_{i,j} p_{ij} \cdot \tilde{d}_{ij}^2$$

$$p_{ij} = \frac{1}{z_{X_i}} \exp\left(\frac{-d_{ij}^2}{2\sigma^2}\right)$$

Exercises:

a) Write the PCA problem as a least squares optimisation problem. Can it also acquire the form below?

$$L = \sum_{i=1}^{N} g(x_i) \cdot f(\tilde{x_i})$$

b) Why do you think the general form below is sufficient/insufficient form for all nonlinear DR methods?

$$L = \sum_{i,j} g(d_{ij}) \cdot f(\tilde{d_{ij}})$$

c) Can there be any advantage to a DR method that also considers interaction terms?

$$L = \sum_{i=1}^{N} h(d_{ij} \cdot \tilde{d_{ij}})$$

Comparison of single-cell omics datasets

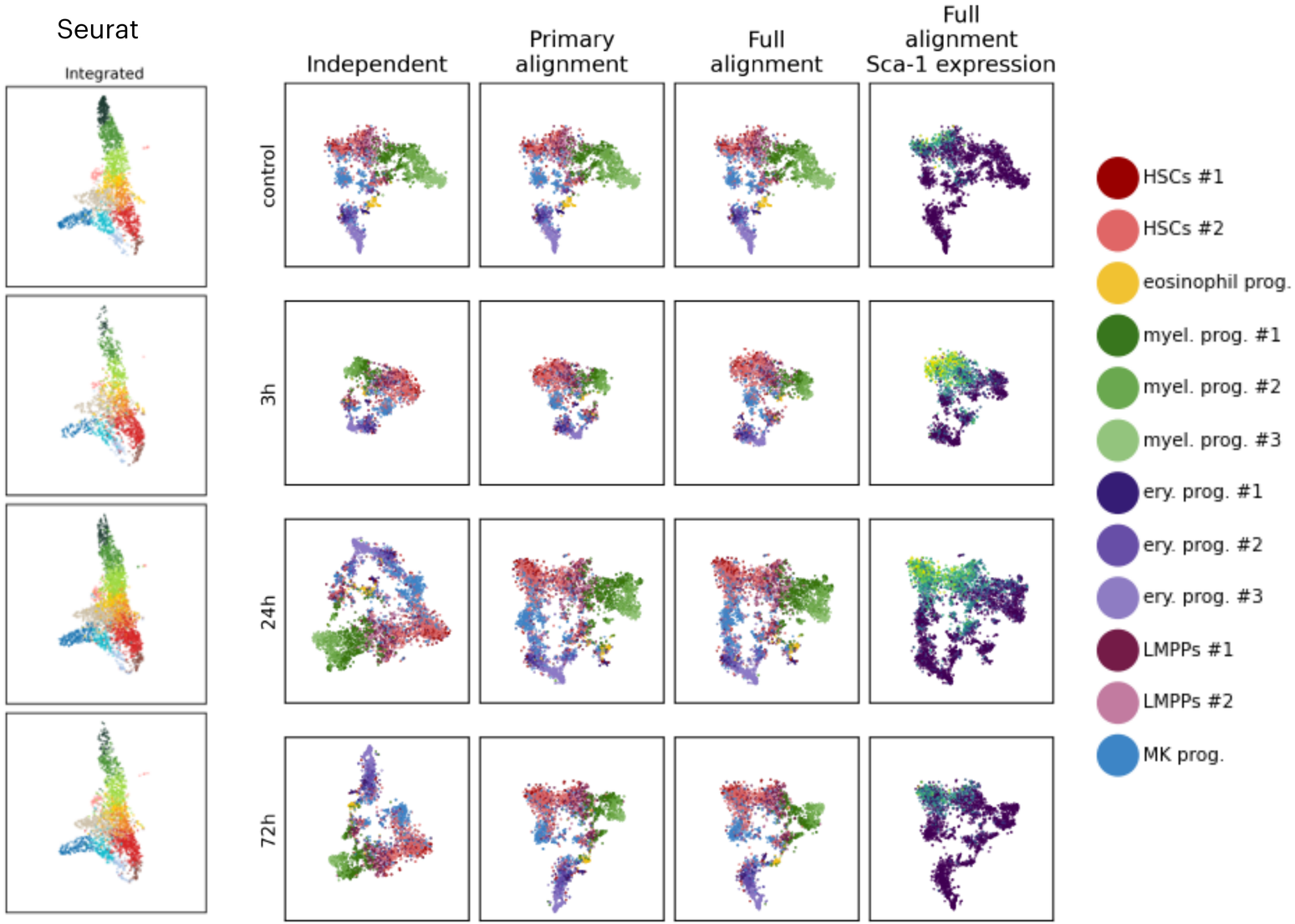
Compound-SNE soft alignment of t-SNEs for visual comparison

- Unlike data integration preserves data specific manifold structures
- Soft alignment of via
  - Initialisation (Primary)
  - Coupling forces (Full)
- ~ modifies cost function

$$d_i = (Y_{r,center_i} - Y_{s,center_i})^2$$

$$L_{total} = L_{tsne} + \sum_{i=1}^{K} \lambda_i d_i$$

Cess, C. G., & Haghverdi, L. *Bioinformatics (2024).*



Compound-SNE aligning different time points of HSPCs response to inflammation

Exercises:

a) Why does t-SNE use a Gaussian distribution function in the original high dimensions but a t-distributed function in the embedding low-dimensional space?
b) What is Multi-dimensional scaling (MDS) and how does it compare with Diffusion map (Or Laplacian Eigenmaps)? When is it better to use transition probabilities (p_{ij}) than distances (d_{ij})?
c) Why doesn't t-SNE preserve data continuity but diffusion map does?
d) Can a) initialisation by PCA and b) annealing improve t-SNEs tendency to preserve data continuity better?
e) Write a pseudo-code (and code) for the stochastic gradient descent version of t-SNE. Also how would you choose the learning rate and momentum?
f) Use Compound-SNE for alignment of two (or more) scRNA-seq data sets across species (e.g. blood lineages in the bone marrow of mouse and human)

Thank you for your attention!

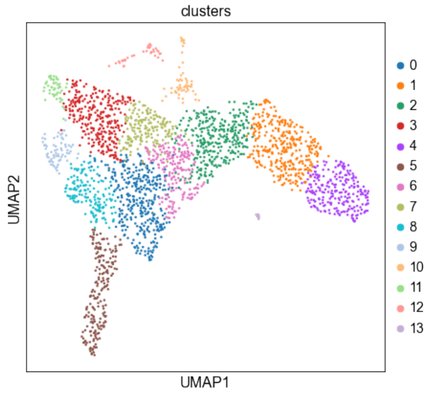
Optional

Cell type labelling

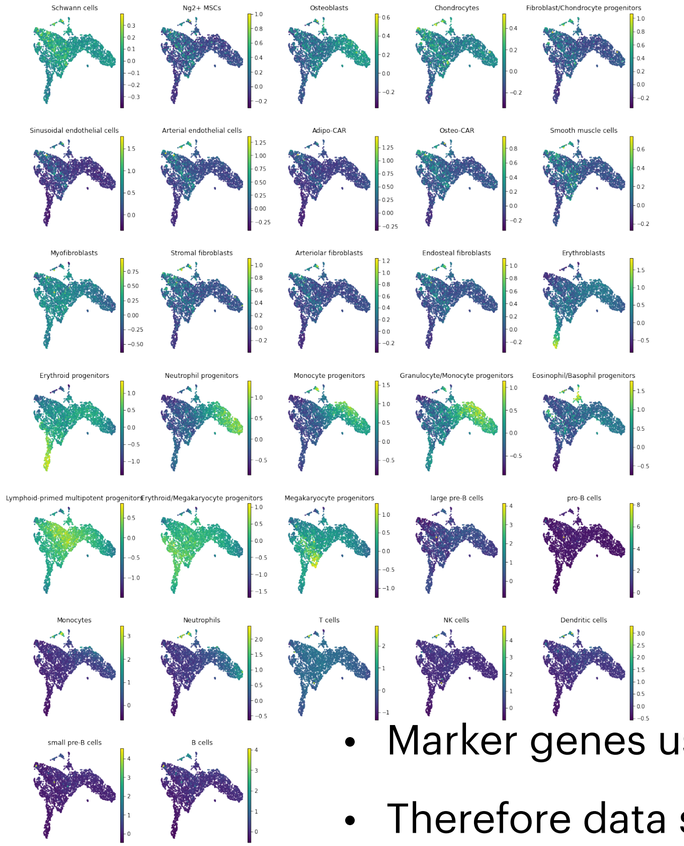
Unsupervised clustering, annotation using marker genes

- Louvain, Leiden, Hierarchical clustering, K-means
- In original hvgs space?
- In reduced linear or nonlinear dimensions? Data dependent, generally don’t use nonlinear
- How many PCs to keep? Also data dependent; if you have rare cell types they may not be captured by the first PCs

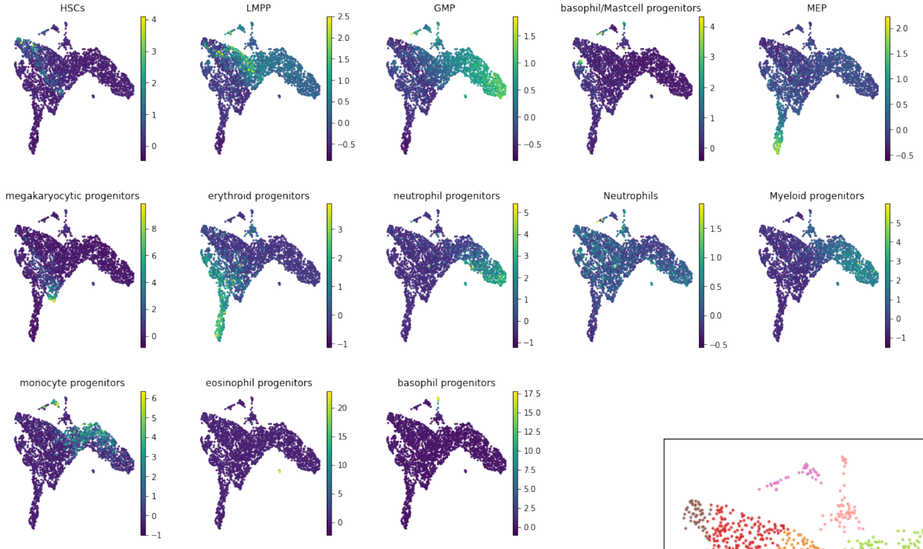
Clustering and cell type annotation

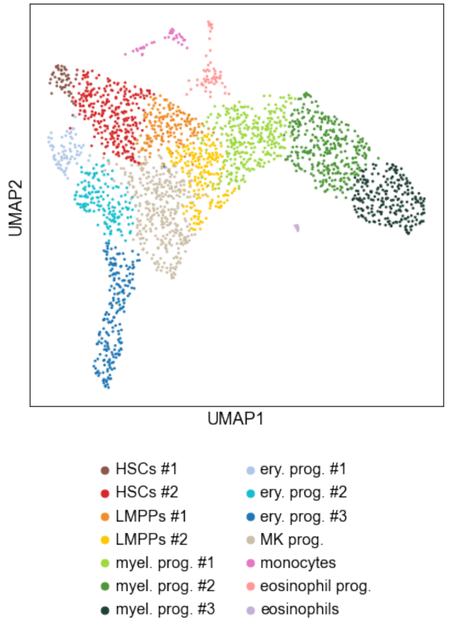


Using marker gene set 1



Using marker gene set 2





https://github.com/bjbouman/prj_HSPC/

- Marker genes usually identified relative to other clusters
- Therefore data specific
- No good state of automation
- Pick a reference that is closer to your dataset composition, use THEIR HVGs
- Consult a biologist about you results!

Bouman BJ, Demerdash Y, Sood S, Grünschläger F, Pilz F, Itani AR, Kuck A, Marot-Lassauzaie V, Haas S, Haghverdi L, Essers MA. Single-cell time series analysis reveals the dynamics of HSPC response to inflammation. Life Science Alliance. (2024)

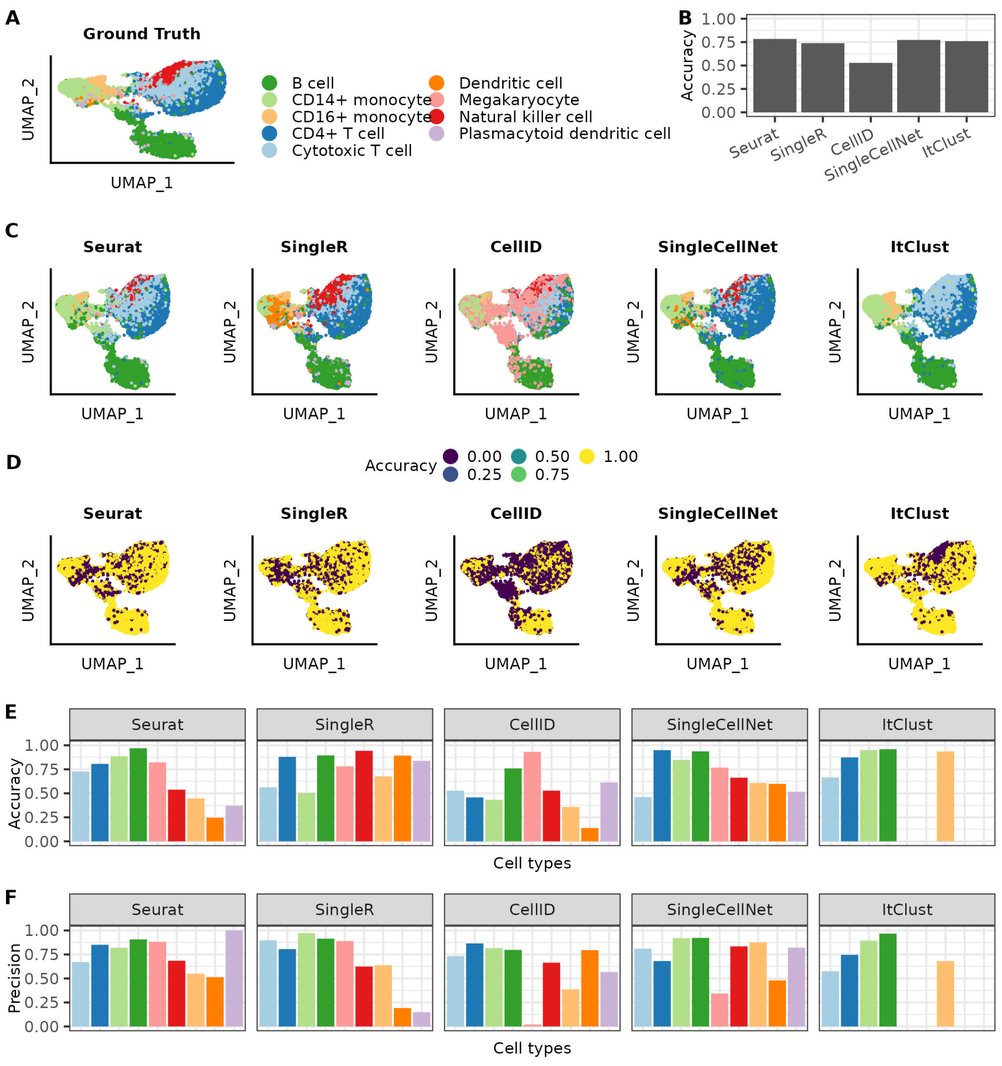
SingleR for annotation

Aran, D., Looney, A.P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).

- SingleR:
  - using only the variable genes in the reference data set (bulk).
  - top correlated cell types
  - Update hvg list with remaining cell types
- Correctly treats HVGs (the reference matters) and their relativistic nature (hvg among which clusters?)
- Because of using reference HVGs, might not be good for detecting new cell types

Label transfer

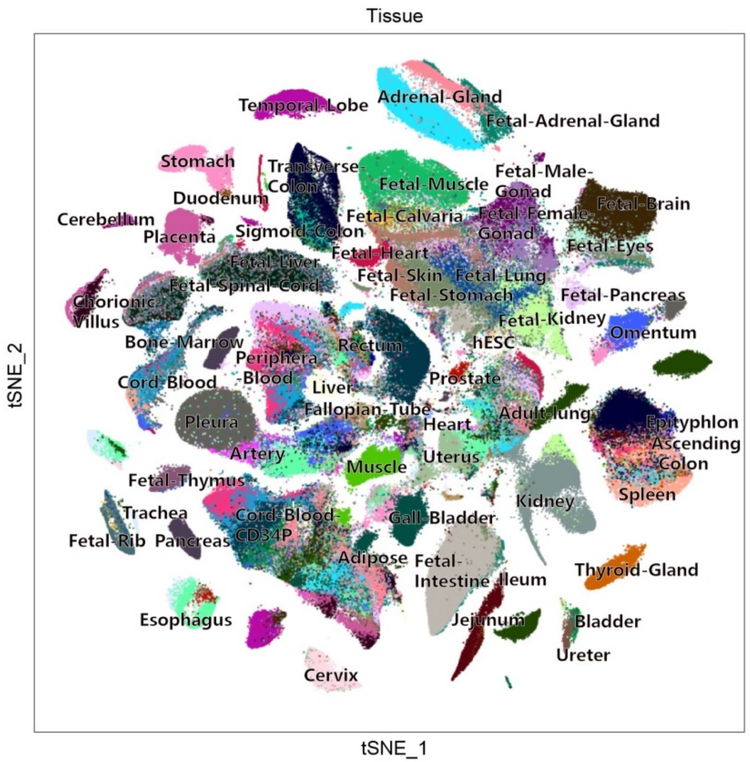
- Transfer via data integration (use the HVGs of your reference!)
    - Seurat: CCA + time warping —> CCA + MNN —> RPCA + MNN
    - Scanorama: PCA +MNN
- Overall accuracies comparable (B)
- Different methods —> different predictions (C)
- Specify model, more importantly specify and reuse the feature set used for the reference, not your own hvgs!
- Error regions near cell type boundaries and rare cell types of the reference (D)
- Per cell type Accuracy ↑, Precision ↓ (E-F)
- So depending if Accuracy or Precision of a certain cell type is important for you, you may want to subsample a more balanced reference set

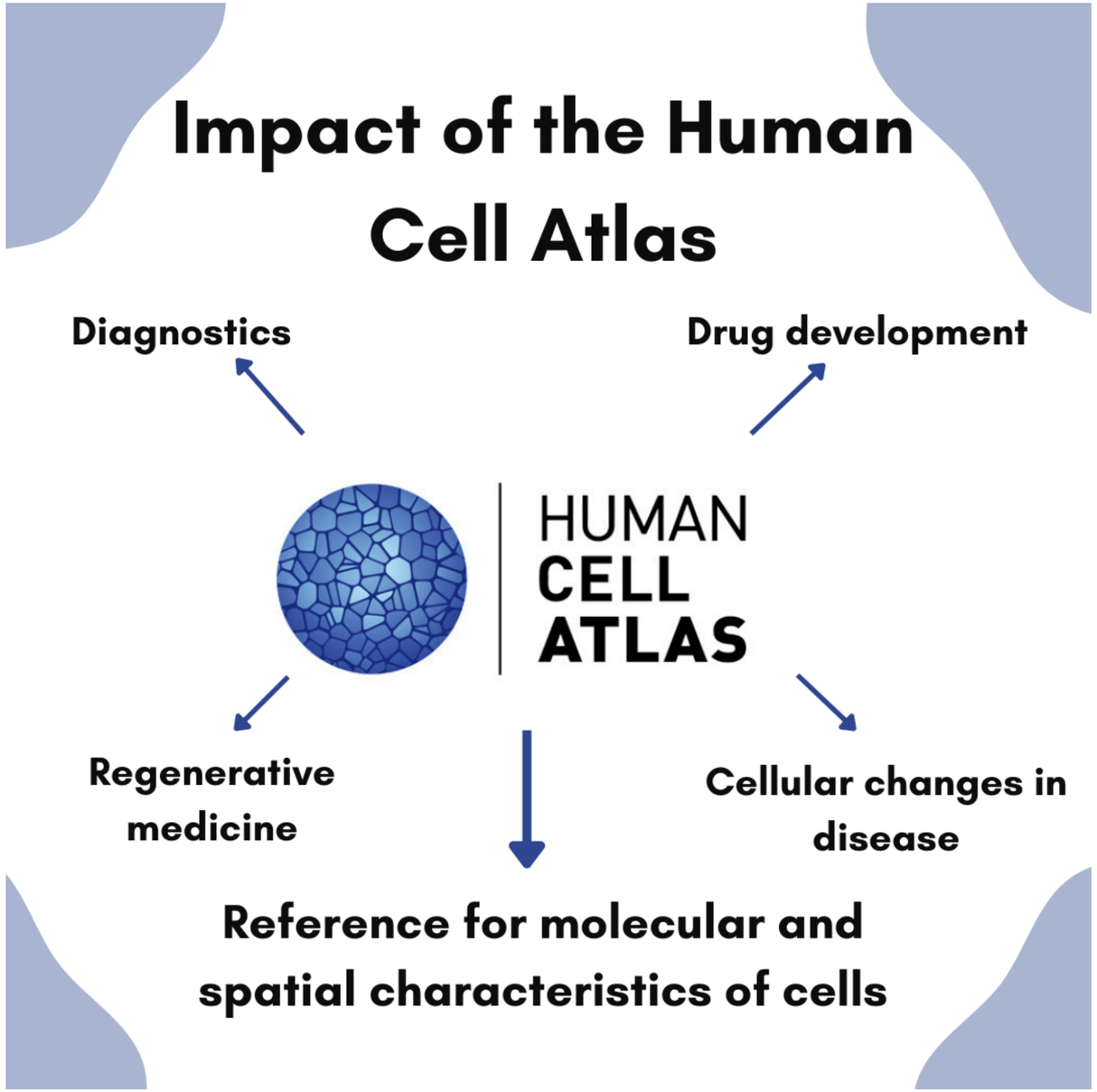


Mölbert, C., & Haghverdi, L. Adjustments to the reference dataset design improve cell type label transfer. *Frontiers in Bioinformatics 2023*.

HCA consortium for standardising data collection and processing

- An SI system for omics data





Exercises:

a) What is contrastive learning? What is its loss function?
b) Why might it be a good model for cell type classification?
c) Write a pseudo-code (and code) for contrastive learning and test it on a previously annotated scRNA-seq data from “Azimuth annotation”
d) https://satijalab.github.io/azimuth/articles/run_azimuth_tutorial.html#scrna-seq-queries