HELMHOLTZ MUNICH

Basic Machine Learning tools for scRNA-seq data analysis

Antonio Scialdone

Course ‘Single Cell Analysis Techniques in Epigenetics Research’

2025

Lecture outline

- Dimensionality reduction
- Gene selection
- Clustering
- Classifiers

Dimensionality reduction

- Count matrix is huge: up to ~10^6 of cells x ~10^3-10^4 genes
  - How can we visualise the data?
  - Dimensionality reduction: mapping the high-dimensional data into a space of lower dimensionality (for visualisation we need 2-3 dimensions)

High-dimensional space

$$\{\vec{x}_i\}_{i=1}^{N} \rightarrow \begin{pmatrix} x_{1,i} \\ \dots \\ x_{M,i} \end{pmatrix}$$

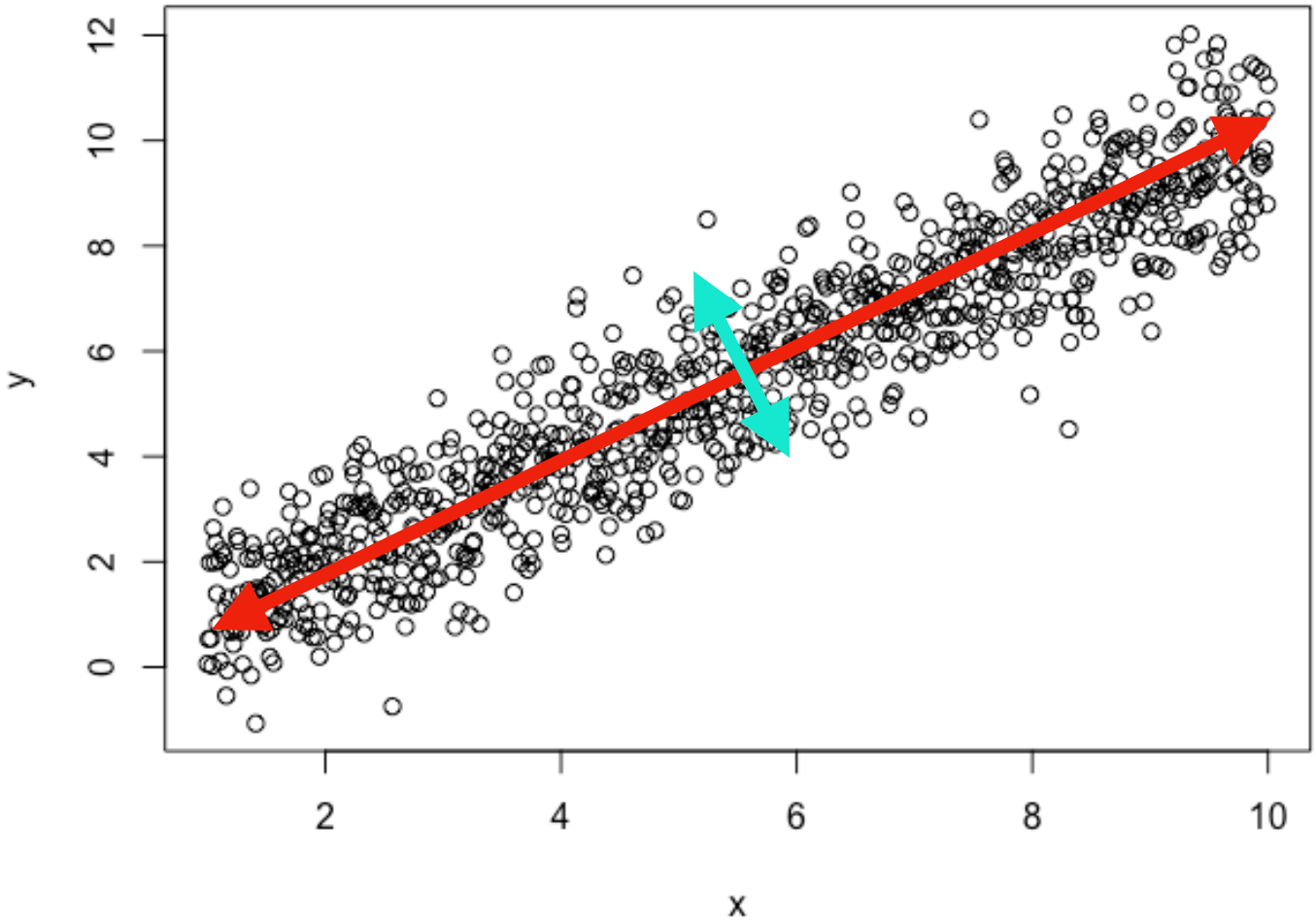
Low-dimensional space

$$\{\vec{y}_i\}_{i=1}^{N} \rightarrow \begin{pmatrix} y_{1,i} \\ \dots \\ y_{L,i} \end{pmatrix}$$

…with $L \ll M$

Principal Component Analysis

- Identification of the first *L* orthogonal directions that explain most of the variation in the data



$$\vec{y}_i = \mathbf{W} \cdot \vec{x}_i$$

- Linear transformation
- Clear interpretation of each component
- It can be misleading when the data has a non-linear structure

t-distributed Stochastic Neighbour Embedding

Idea: choose the points in the lower-dimensional space in such a way that the distribution of **local distances** between points is as close as possible to that found in the high-dimensional space

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/(2\sigma_i^2)\right)}{\sum_{k \neq i} \exp\left(-||x_i - x_k||^2/(2\sigma_i^2)\right)}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$$

Cost function = $C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$

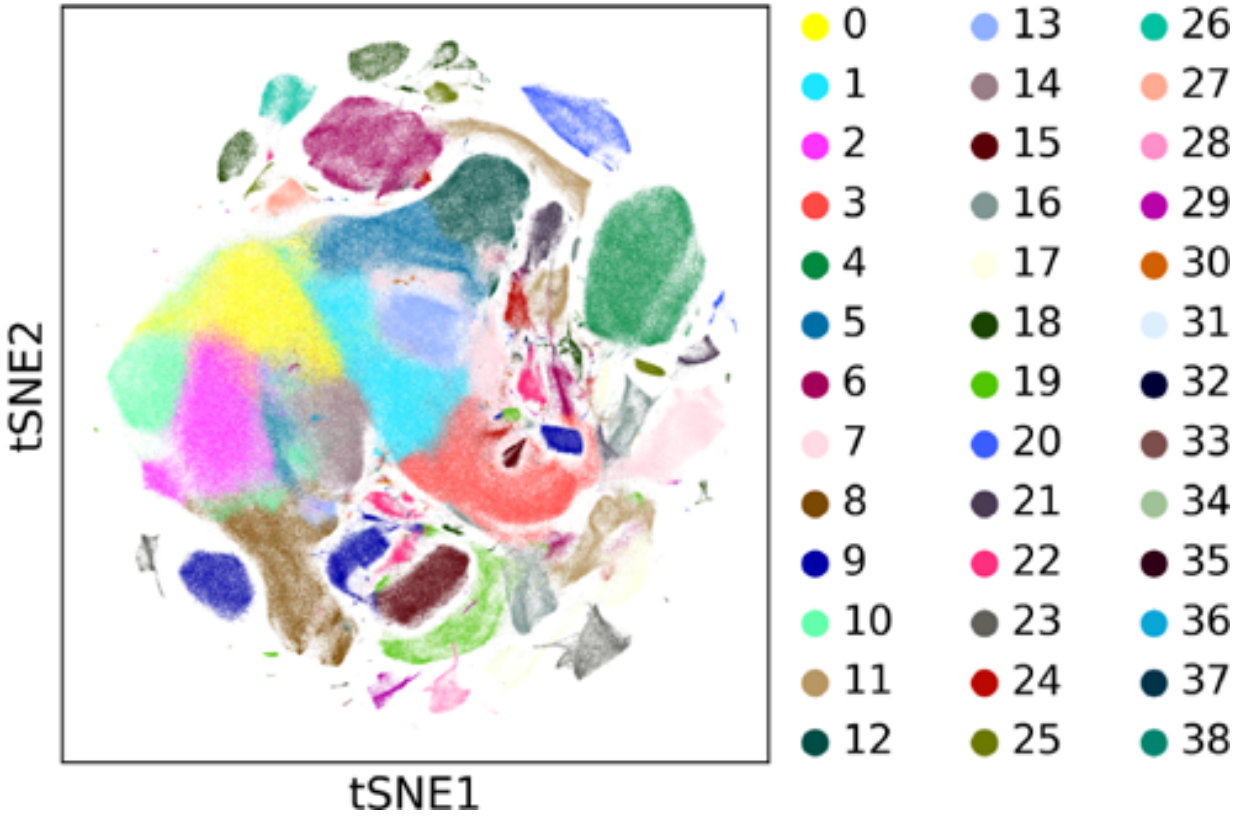
$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}}$$

- *C* is the **Kullback-Leibler divergence** between the joint probability distributions corresponding to p_{ij} and q_{ij}
- y_i are chosen to **minimise** C

Van Der Maaten and Hinton, Journal of Machine Learning Research, 2008

t-distributed Stochastic Neighbour Embedding

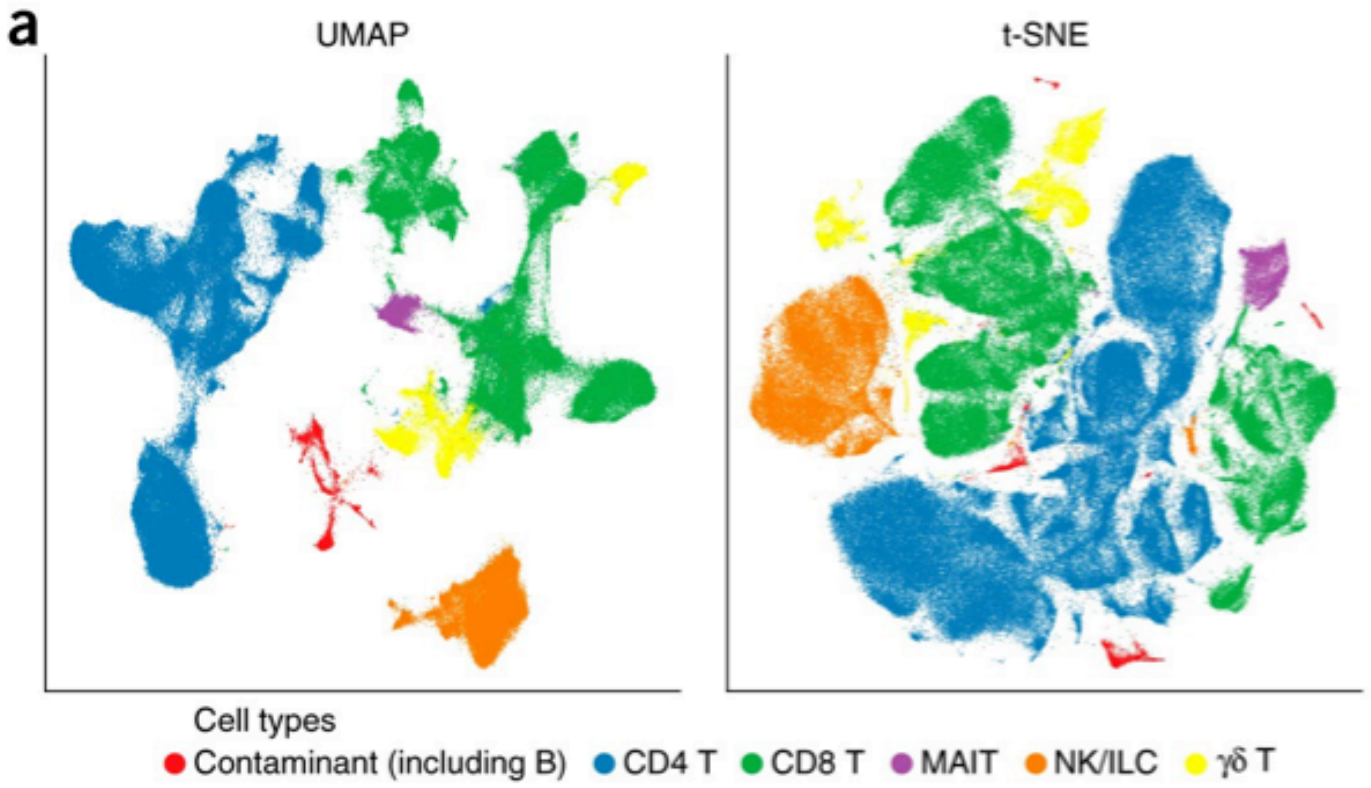
- Non-linear
- Good at clustering data in separate groups
- It could result in artificial breaks of continuous trajectories



Wolf et al, Genome Biology 2018

Uniform Manifold Approximation and Projection (UMAP)

- Based on matching topological representation
- Non-linear
- Fast and scalable



Becht et al, Nature Biotechnology 2019

All dimensionality reduction techniques can produce misleading results!

The specious art of single-cell genomics

Tara Chari, Lior Pachter

Published: August 17, 2023 • https://doi.org/10.1371/journal.pcbi.1011288

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Dimensionality reduction is standard practice for filtering noise and identifying relevant features in large-scale data analyses. In biology, single-cell genomics studies typically begin with reduction to 2 or 3 dimensions to produce "all-in-one" visuals of the data that are amenable to the human eye, and these are subsequently used for qualitative and quantitative exploratory analysis. However, there is little theoretical support for this practice, and we show that extreme dimension reduction, from hundreds or thousands of dimensions to 2, inevitably induces significant distortion of high-dimensional datasets. We therefore examine the practical implications of low-dimensional embedding of single-cell data and find that extensive distortions and inconsistent practices make such embeddings counter-productive for exploratory, biological analyses. In lieu of this, we discuss alternative approaches for conducting targeted embedding and feature exploration to enable hypothesis-driven biological discovery.

Figures

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

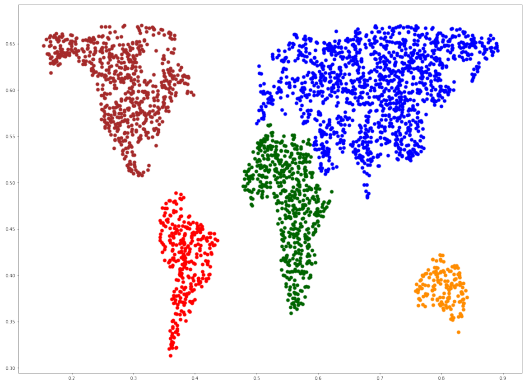
NEWS | 23 February 2024

'All of Us' genetics chart stirs unease over controversial depiction of race

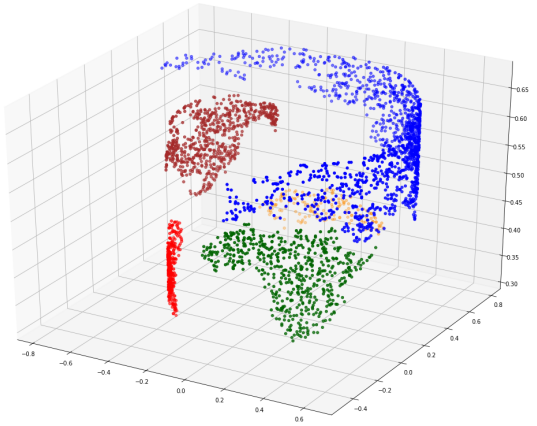
Debate over figure connecting genes, race and ethnicity reignites concerns among geneticists about how to represent human diversity.

By Max Kozlov

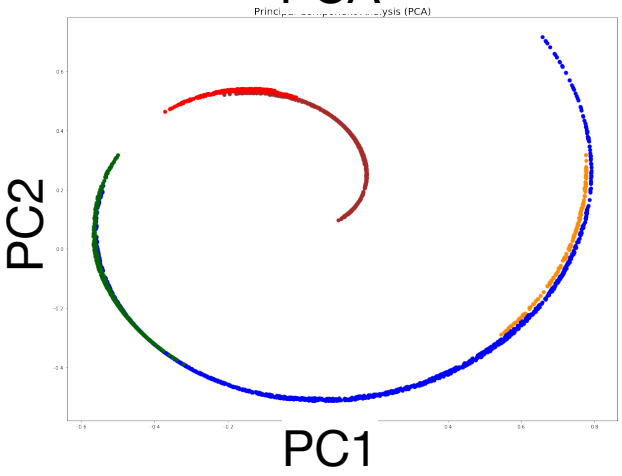
An example: the “world map” dataset



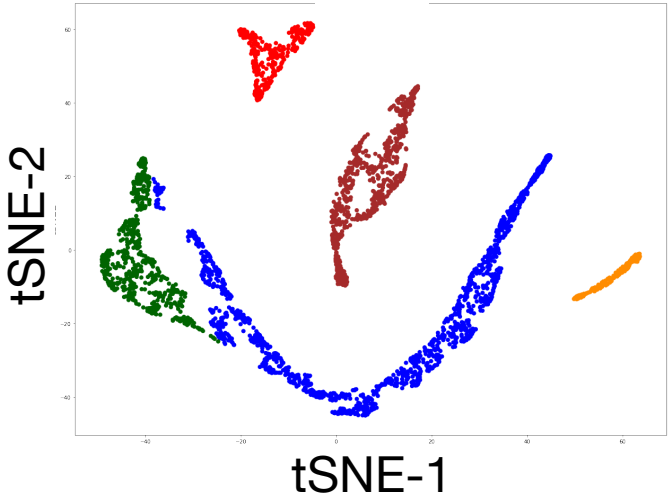
Embedded on a “swiss roll” in 3D



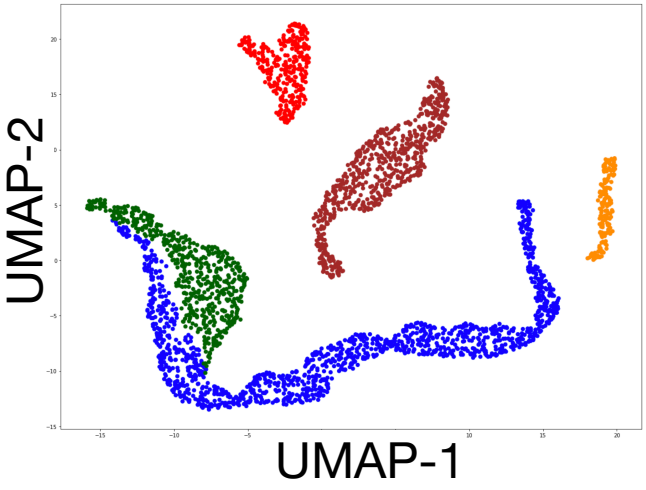
PCA



tSNE



UMAP



Nikolay Oskolkov, SciLifeLab and Lund University
github.com/NikolayOskolkov

Diffusion maps

- Data ~ network of connections
- “Random walks” through the data
- Diffusion distance ~ how easily one can travel from A to B with these random walks
- Diffusion maps use eigenvalue decomposition to extract the most important patterns from the random walks

$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

$$P = \textit{NxN}\ \text{matrix} = \left[p(x_i, x_j) = \frac{k(x_i, x_j)}{\sum_l k(x_i, x_l)}\right]$$

$$P^t \psi_k = \lambda_k^t \psi_k \qquad 1 = \lambda_0 > \lambda_1 \geq ... \geq \lambda_{N-1} \geq 0$$

$$\psi_0 = \mathbb{1}$$

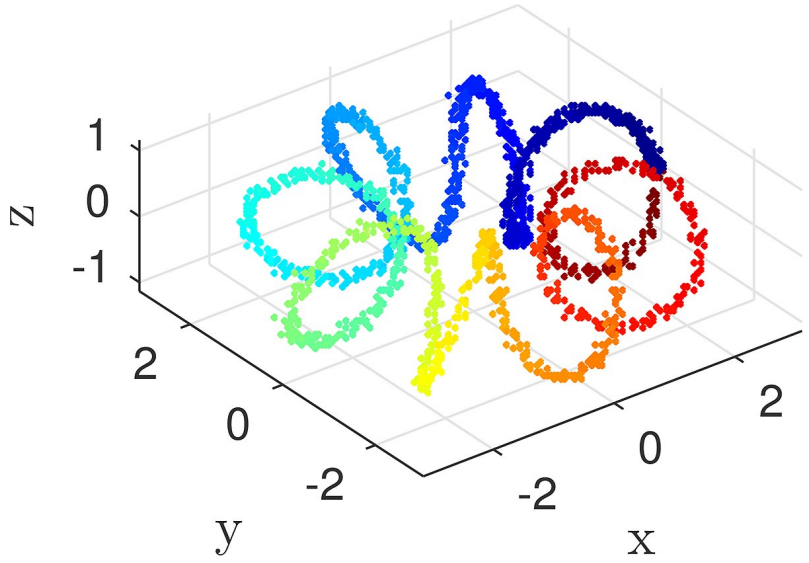
$$D_t^2(x_i, x_j) = \begin{matrix}\text{Diffusion} \\ \text{distance}\end{matrix} = \sum_l (p_t(x_i, x_l) - p_t(x_j, x_l))^2 \sim \sum_{k=1}^{q} \lambda_k^{2t} (\psi_k(x_i) - \psi_k(x_j))^2$$

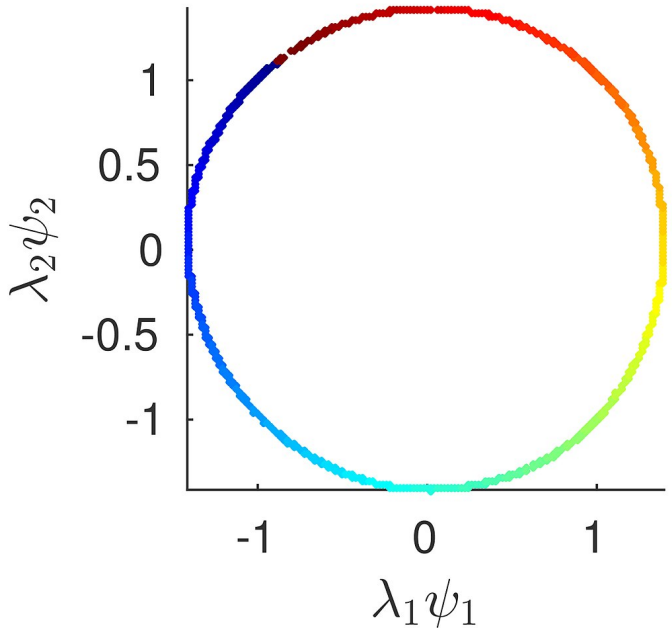
$$C = \sum_{i,j} (D_t^2(x_i, x_j) - ||y_i - y_j||^2)$$

$$x_i \rightarrow y_i = \begin{bmatrix} \lambda_1^t \psi_1(x_i) \\ . \\ . \\ . \\ \lambda_q^t \psi_q(x_i) \end{bmatrix}$$

Coifman et al, PNAS,102:7426, 2005
Haghverdi et al, Bioinformatics, 31:2989, 2015

Diffusion maps



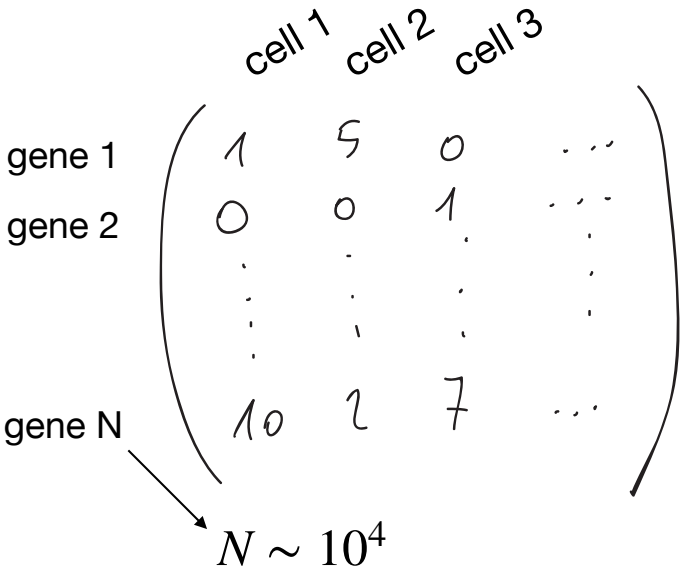


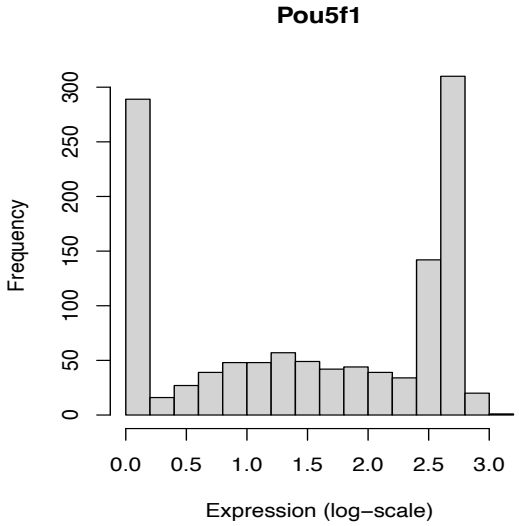
- Robust to noise
- Good with continuous data structures

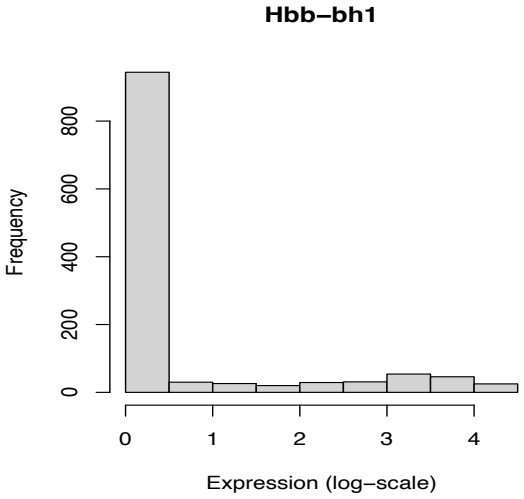
Lecture outline

- Dimensionality reduction
- Gene selection
- Clustering
- Classifiers

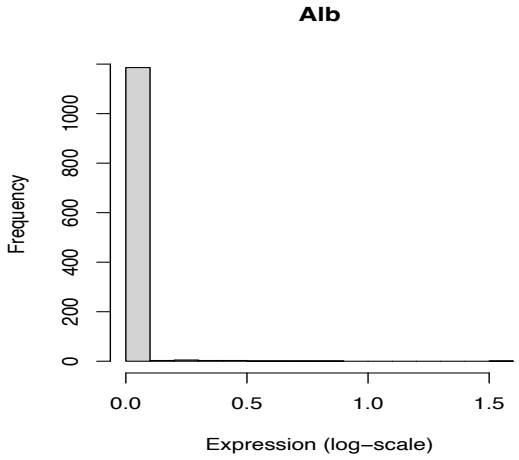
How do we select “interesting” genes?

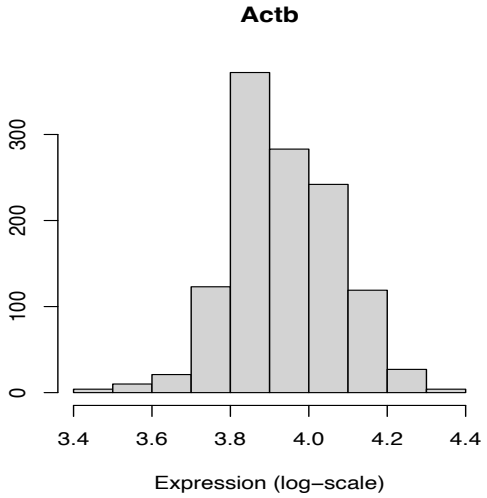






…some genes look interesting

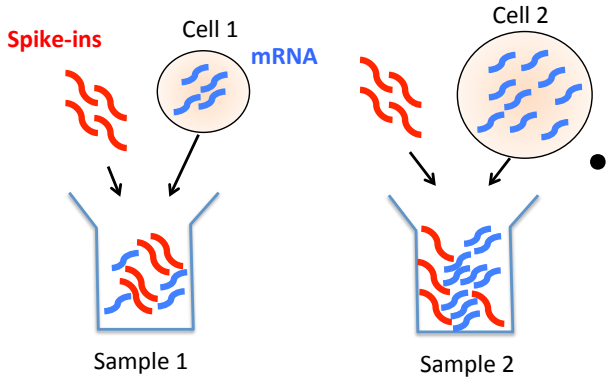




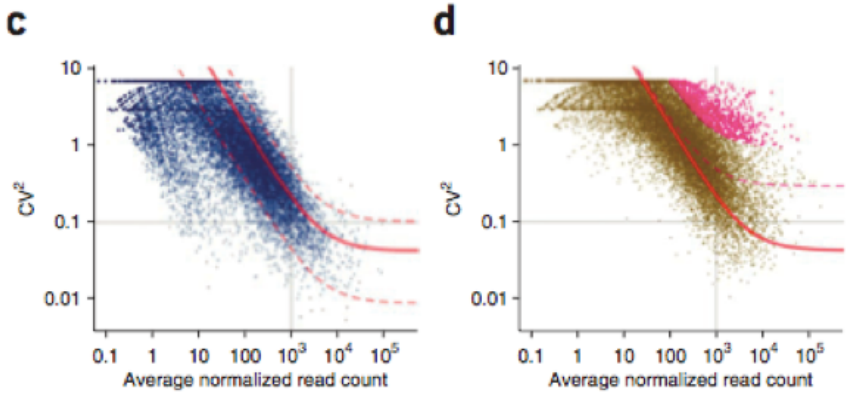
…others less so

Gene selection - spike ins

IDEA: Select genes that are more variable than expected by chance…
But what is the "null model"?



- Spike-in genes to estimate technical noise



c: spike-in genes; **d**: plant genes

Brennecke et al., Nature Methods (2013)

(c) noise fit: CV^2 vs mean

$$CV_i^2 = \frac{\alpha}{m_i} + \beta$$

(d) highly variable genes: CV^2 exceeds technical prediction

- Possible issue: technical noise might affect differently spike-ins and endogenous genes
- Alternative: hypothesis that most genes do not vary

Lecture outline

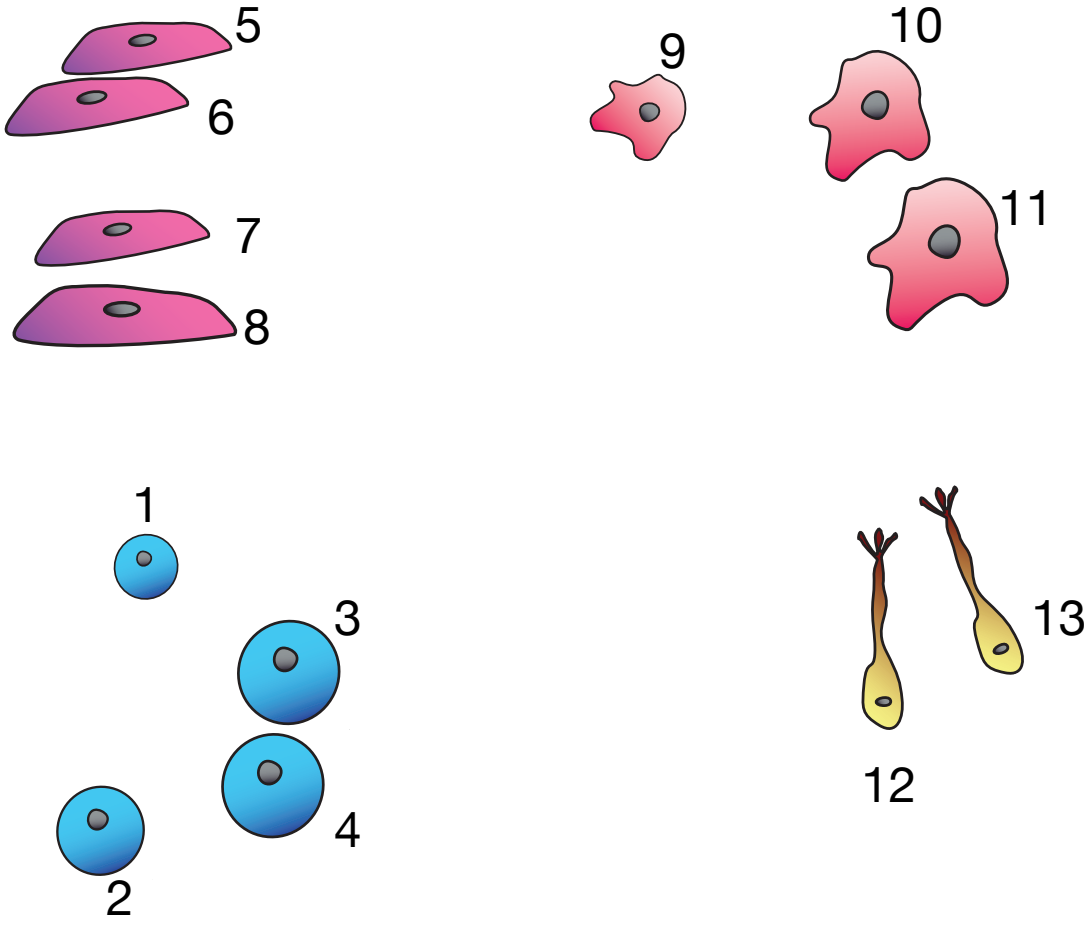
- Dimensionality reduction
- Gene selection
- Clustering
- Classifiers

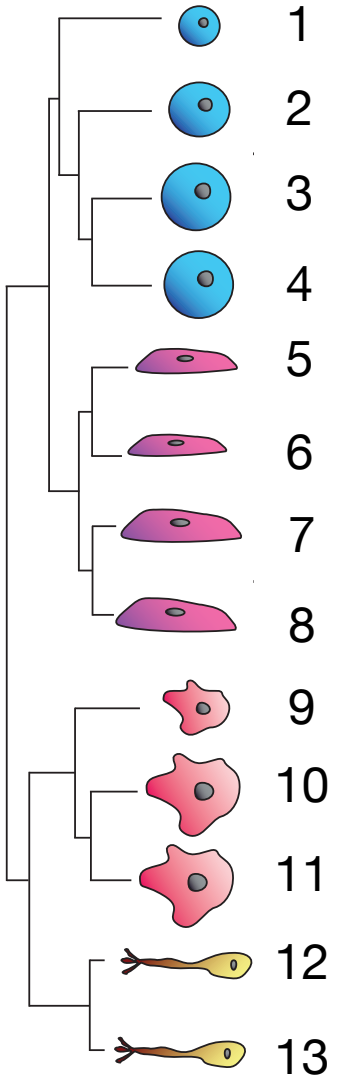
Clustering

- Identification of **groups** of cells that share **similar** transcription profiles
- Many different algorithms available

- How many clusters? What genes to use? How shall we measure the distance? …

Example: **Hierarchical clustering**



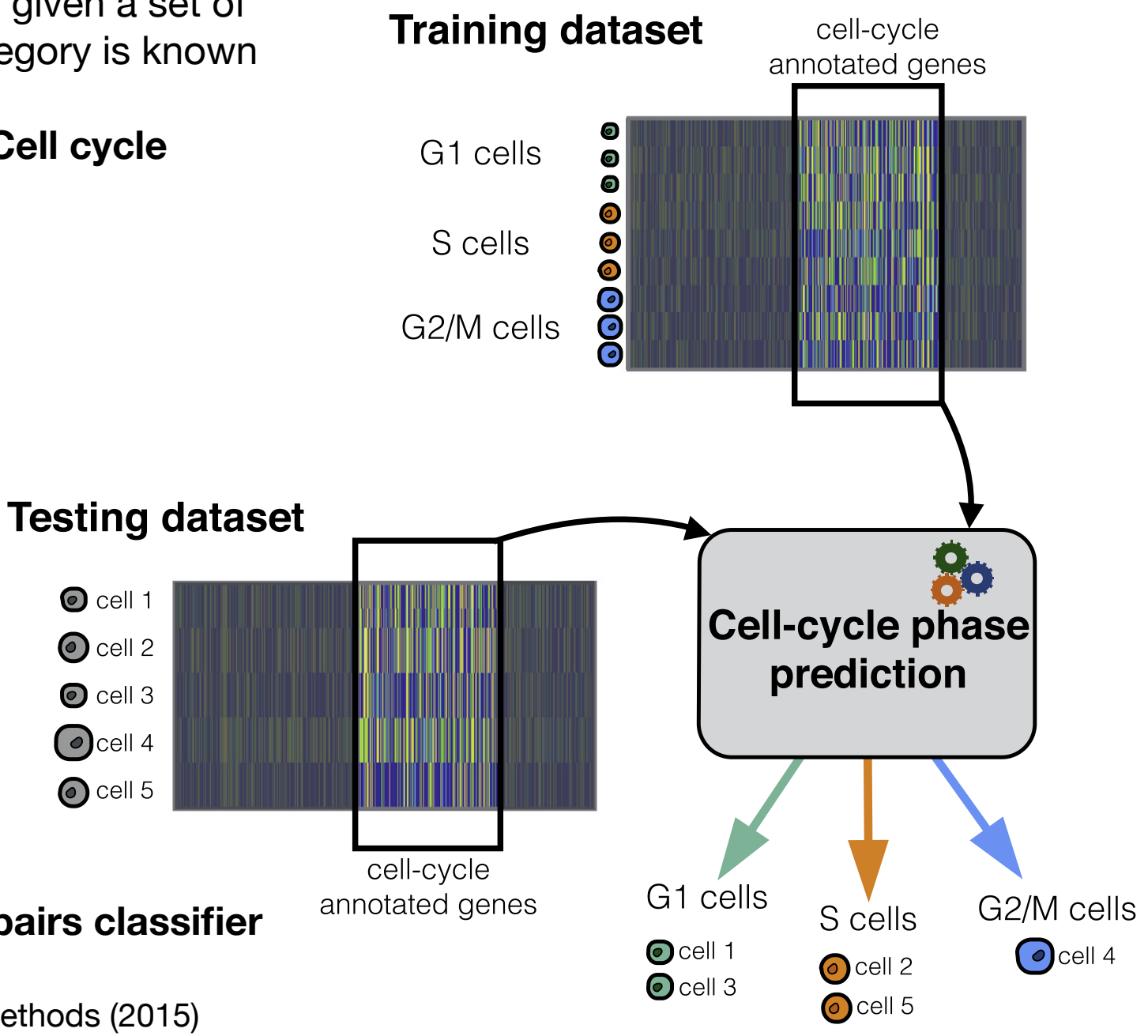


Lecture outline

- Dimensionality reduction
- Gene selection
- Clustering
- Classifiers

Classifiers

- Identification of the category a cell belongs to, given a set of cells whose category is known
- Example: **Cell cycle**



- We will use the **pairs classifier**

Scialdone et al, Methods (2015)

Questions?