Data Visualisation and Clustering

“Single-cell multiomic data analysis”
Summer Semester 2025
Antonio Scialdone

HELMHOLTZ
MUNICH

Visualising “Big” Data

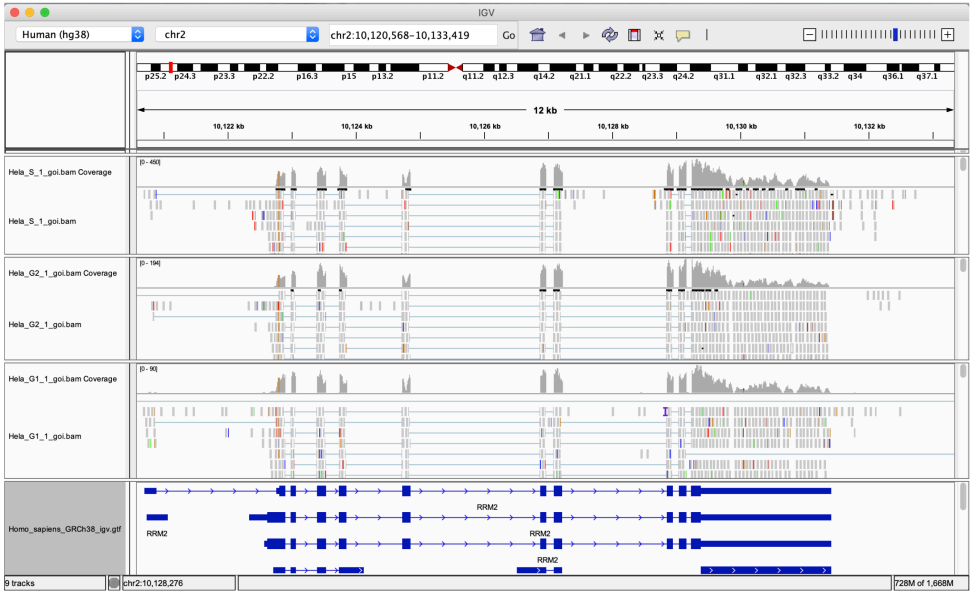
- Data visualisation is important for: exploring data structure, detecting outliers, spotting patterns, evaluating modelling output, …
- “Big data” - such as (single-cell) omics data - presents specific challenges for visualization
- Data typically lie on a *curved (high-dimensional) space*
- *Curse of dimensionality*: Refers to a various challenges arising when dealing with high-dimensional data

Example: data become more sparse!

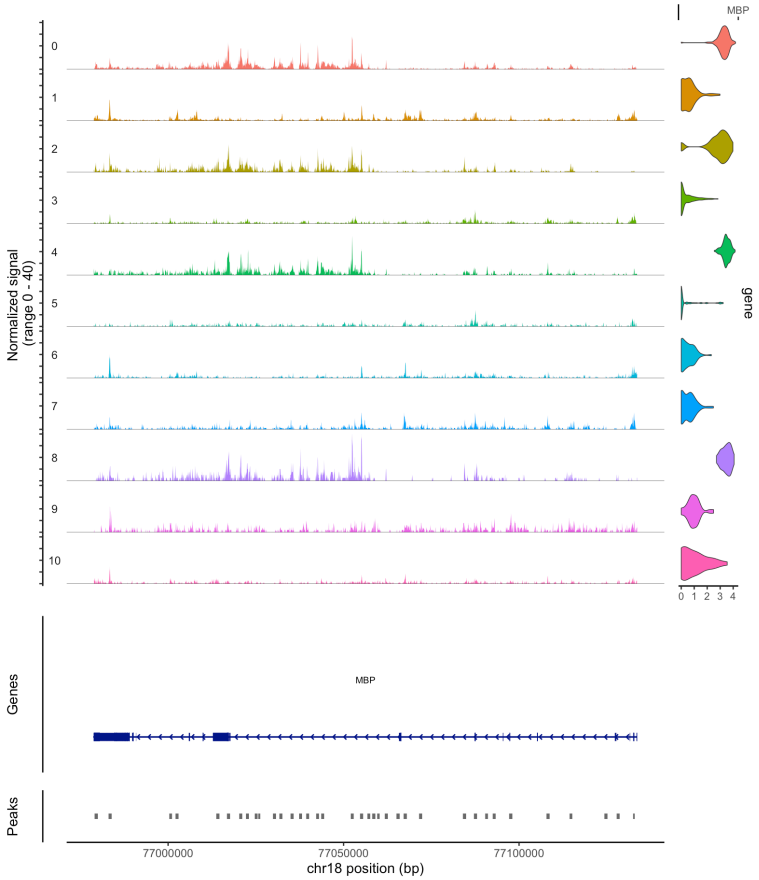
Average (squared) Euclidean distance between two randomly placed points in a unit hypercube in dimension d

$$= \frac{d}{3}$$

Visualising data with IGV and Seurat/Signac functions



- IGV = Integrative Genome Viewer



- Good to explore specific regions in a specific group of cells
- Possibility to visualise multiple types of data as separate tracks

Dimensionality Reduction

- Count matrices are huge: up to ~10^6 of cells x ~10^3-10^4 genes or 10^5 peaks

- Dimensionality reduction: mapping the high-dimensional data into a space of lower dimensionality (for visualisation we need 2-3 dimensions)

High-dimensional space → **Low-dimensional space**

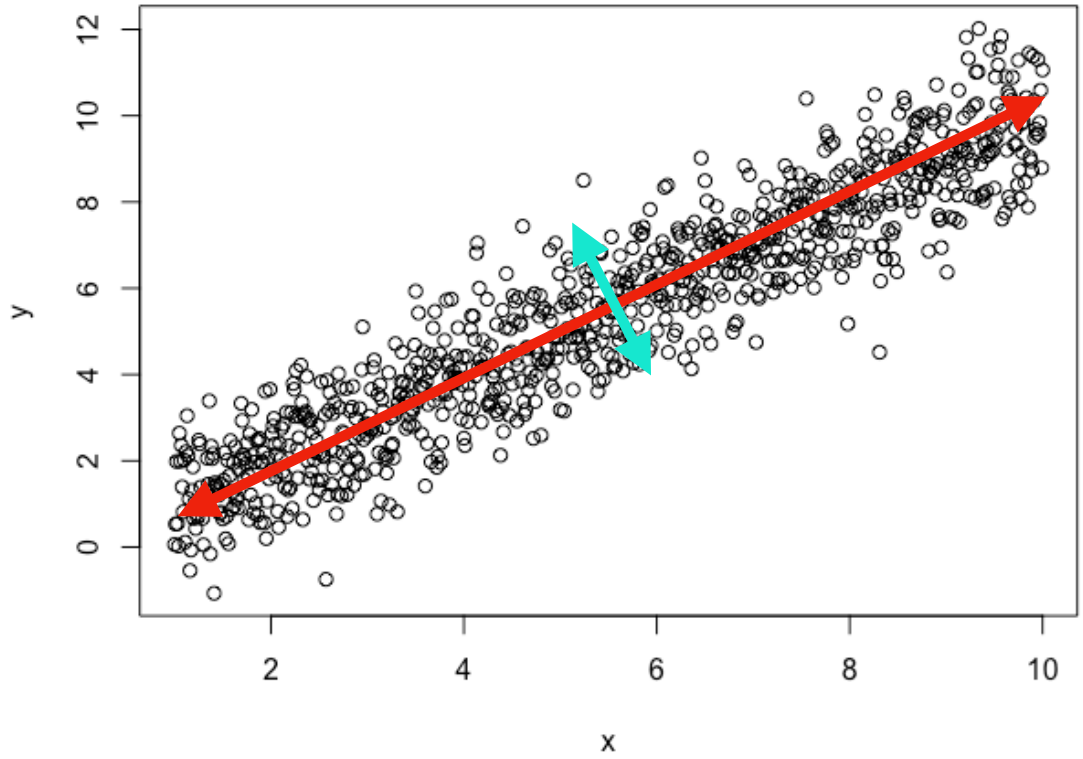
$$\{\vec{x}_i\}_{i=1}^{N} \rightarrow \begin{pmatrix} x_{1,i} \\ \dots \\ x_{M,i} \end{pmatrix} \qquad \{\vec{y}_i\}_{i=1}^{N} \rightarrow \begin{pmatrix} y_{1,i} \\ \dots \\ y_{L,i} \end{pmatrix}$$

...with $L \ll M$

- Many algorithms available; **all of them distort the data and have their pros&cons**

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
- Can easily measure the importance of each variable for each component
- It can be misleading when the data has a non-linear structure

Latent Semantic Indexing

- Technique from Natural Language processing, used to identify patterns and correlations between terms

- Genomic Regions ~ Words in documents; Cells ~ Documents

- Applying Singular Value Decomposition (SVD) to the TF-IDF normalised matrix, M (with m rows and n columns)

$$M = U\Sigma V^T$$

U : $m \times m$ **orthogonal** matrix (left singular vectors)

Σ : $m \times n$ **diagonal matrix** with **non-negative real numbers** on the diagonal (singular values)

V^T: $n \times n$ **orthogonal** matrix (right singular matrix)

- Keep only the top singular values and their corresponding vectors (dimensionality reduction).

Uniform Manifold Approximation and Projection (UMAP)

- Very popular **non-linear technique**
- Focus on preserving **local distances**

$\{\mathbf{x}_i\}_{i=1}^{N}; \quad \mathbf{x}_i \in \mathbb{R}^D$ $\qquad$ $\{\mathbf{y}_i\}_{i=1}^{N}; \quad \mathbf{y}_i \in \mathbb{R}^L$

$p(\mathbf{x}_i, \mathbf{x}_j)$ = probability that the points i and j are connected in the high-dimensional space → Use of a locally defined metric; it decays with distance.

$q(\mathbf{y}_i, \mathbf{y}_j)$ = probability that the points i and j are connected in the low-dimensional space → Use of a **globally defined euclidean metric**.

Task: find the $\{\mathbf{y}_i\}_{i=1}^{N}$ in such a way that $q(\mathbf{y}_i, \mathbf{y}_j)$ is *as close as possible to* $p(\mathbf{x}_i, \mathbf{x}_j)$

The similarity of p and q is measured with the **cross-entropy:**

$$CE = \sum_i \sum_j \left[p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) \log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right) \right]$$

UMAP: more reading and examples

- Some more details, interactive examples and comparison with another popular algorithm (t-SNE): https://pair-code.github.io/understanding-umap/index.html

- Two main parameters: n_neighbors, min_dist.

- The topic of dimensionality reduction and use of UMAP (or t-SNE) with omics datasets is hotly debated

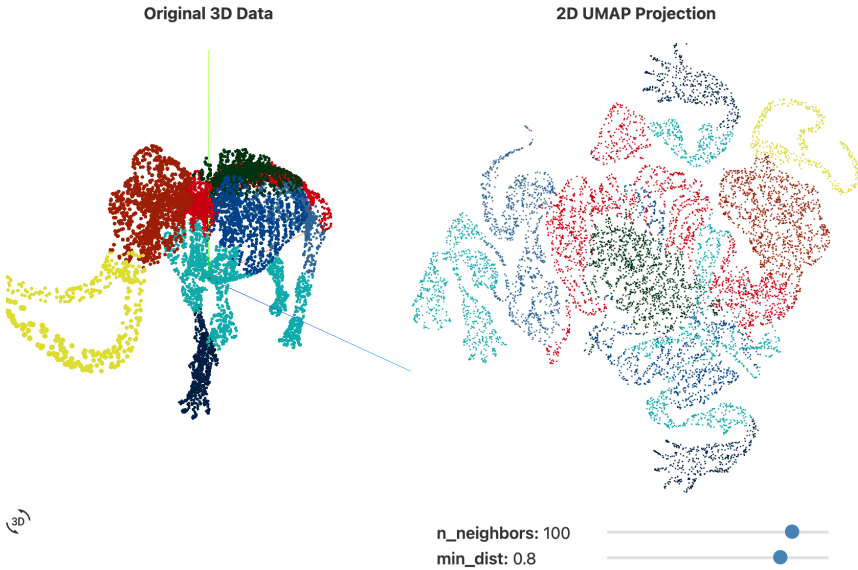
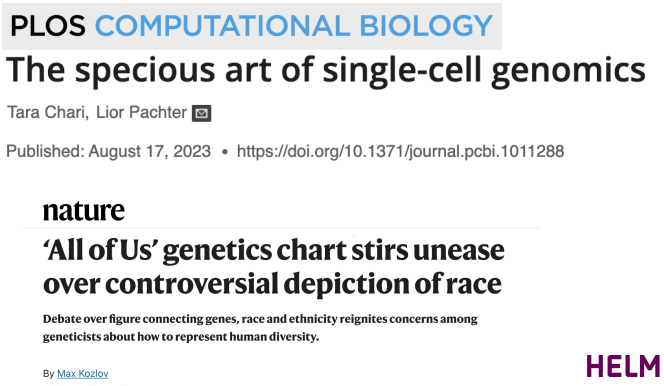


Figure 5: UMAP projections of a 3D woolly mammoth skeleton (50k points, 10k shown) into 2 dimensions, with various settings for the `n_neighbors` and `min_dist` parameters.


PLOS COMPUTATIONAL BIOLOGY

The specious art of single-cell genomics

Tara Chari, Lior Pachter

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

nature

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

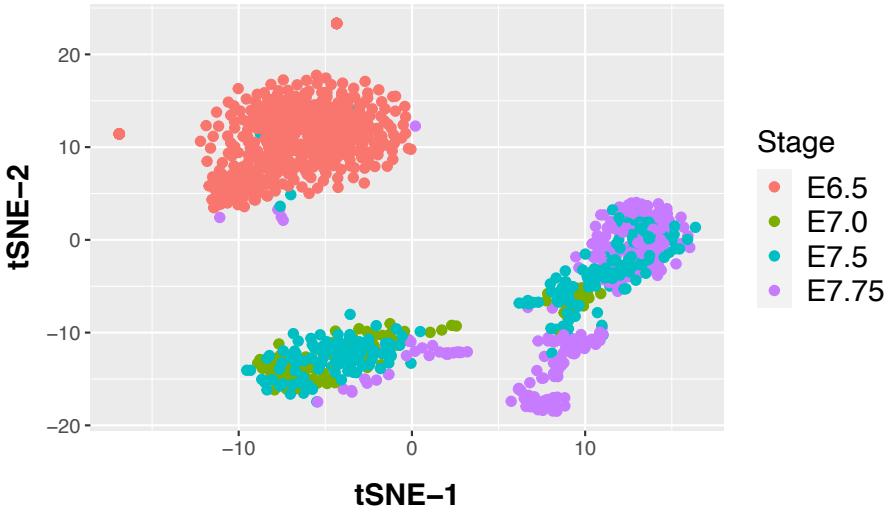
Debate over figure connecting genes, race and ethnicity reignites concerns among geneticists about how to represent human diversity.

By Max Kozlov

HELMHOLTZ MUNICH

Clustering

Clustering: identifying cell types/states



What cell types/states do we have in our data?

First step is grouping cells based on the similarity of their transcriptomes, measured by some **distance metric** (eg, euclidean, correlation-based)

Algorithms: mainly two types, either work on distance matrix directly or on graphs

Number of clusters: could be decided based on some robustness criteria (eg, gene sub-sampling)

Clustering: Graph-based algorithms

Clustering on a graph: find densely connected regions

Building k-nearest-neighbour graph

Goal is to optimize modularity

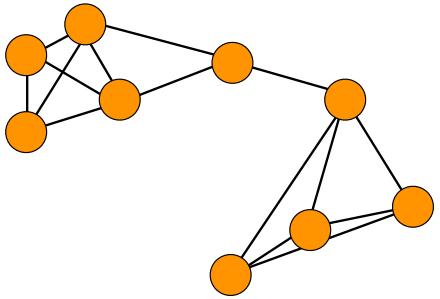
$$Q = \frac{1}{2m} \sum_c \left(e_c - \gamma \frac{K_c^2}{2m} \right)$$

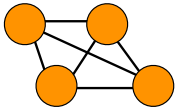
Sum of degrees of nodes in community c

Tot edges In the network

Number of edges in community c

Resolution





Significantly faster than hierarchical clustering on large datasets

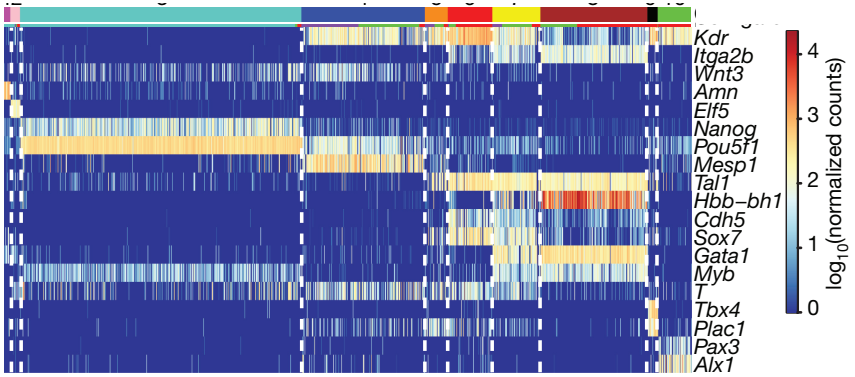
Examples: Louvain algorithm, Leiden algorithm

Blondel et al, *J. Stat. Mech. Theory Exp,* 10008:6, 2008

Traag et al, *Sci Rep,* 9:5233, 2019

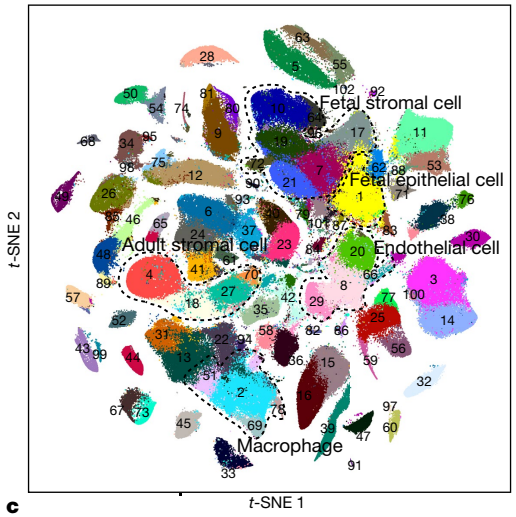
Cluster annotation

Use previously known **marker genes**



It can be quite laborious and not always straightforward…

Use **reference datasets** with automated cluster annotation

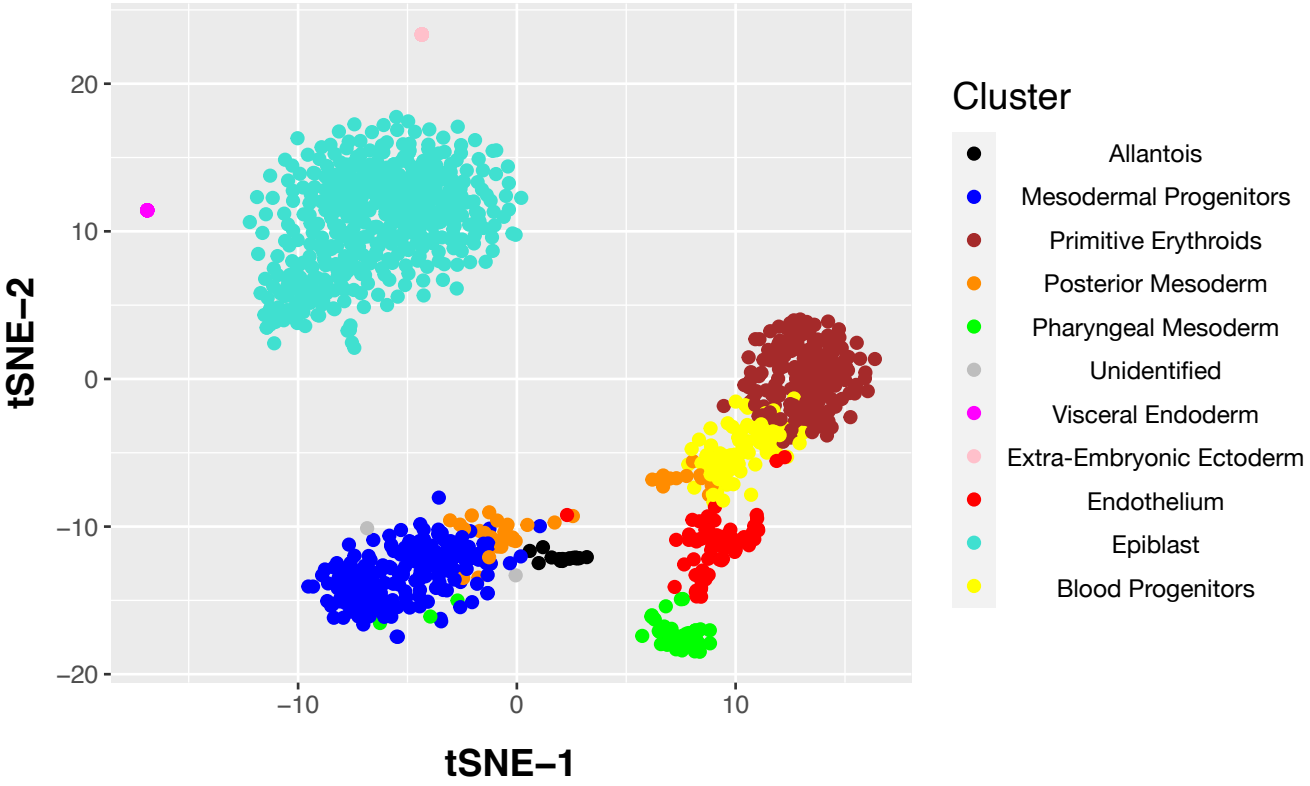


Need to be careful about possible differences between reference and test datasets (eg, experimental conditions)

Han et al, Nature, 581:303, 2020

Something to bear in mind about clusters…

- Clustering is a more or less good description of the signal/variability present in the data. There are very often additional layers of variability not described by the clustering



- Level of detail in the annotation (resolution) should be decided by the user