BIO-463
Genomics and bioinformatics

Lecture 8:
Bulk gene expression analysis II

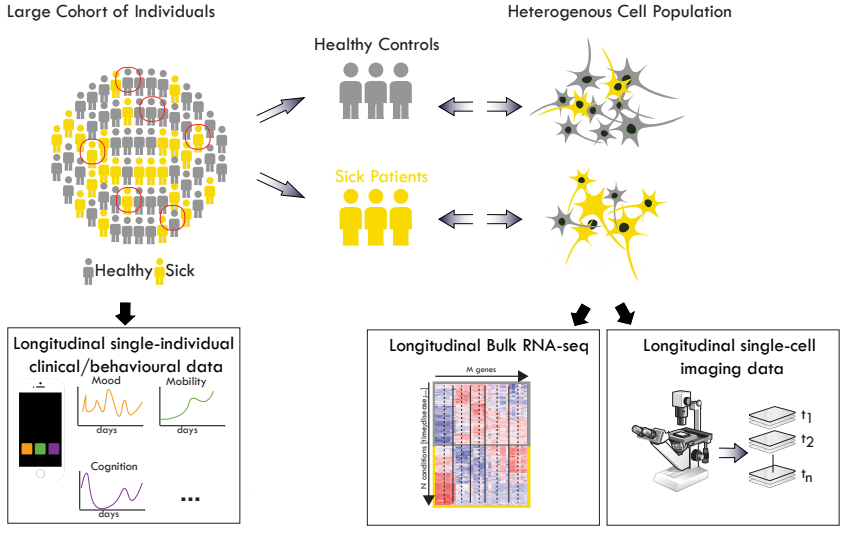
Dr Raphaëlle Luisier

EPFL

Table of Content

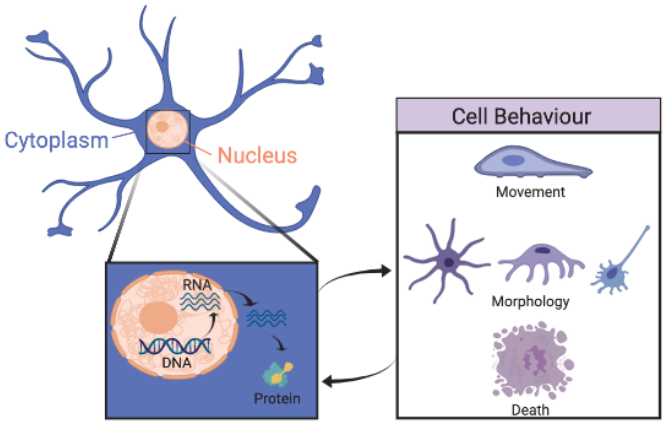
Genomics and Health Informatics Group

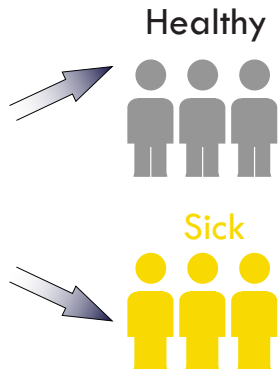
Towards the identification of **early** *mechanisms underlying complex human diseases.*



Genomics and Health Informatics Group

Which molecular perturbations underly disease cellular states?

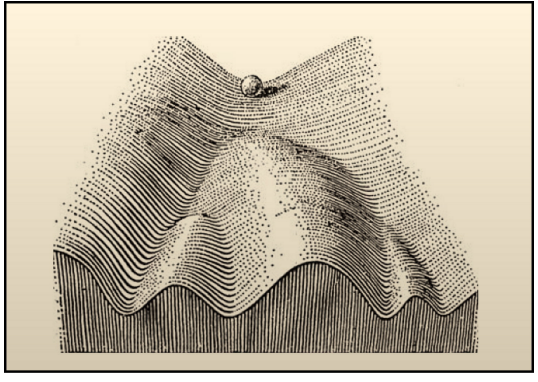




"Cells are residents of a vast landscape of possible states,

over which they travel during **development** *and in* **disease***"*

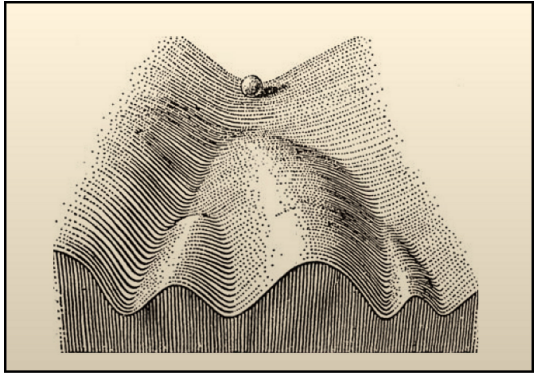
C.H. Waddington



"Cells are residents of a vast landscape of possible states,

over which they travel during **development** *and in* **disease***"*

C.H. Waddington

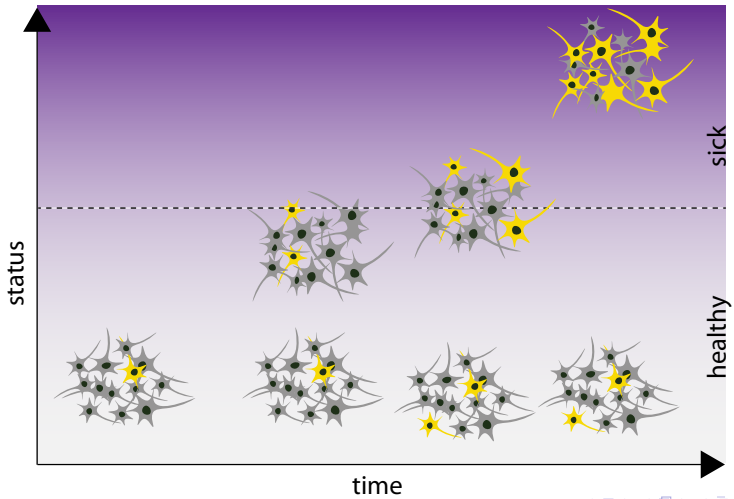


↓

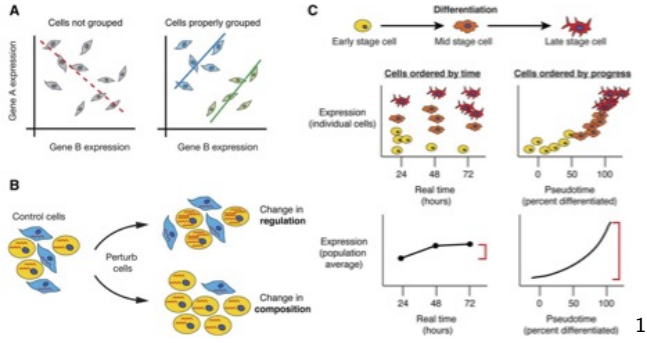
Locating the cells on this landscape == identify cell identity and state.

Genomics and Health Informatics Group

Which molecular perturbations underly disease cellular states?



Simpson's Paradox (Simpson 1951)



[1]

Failing to properly compartmentalize the data by cell type leads to a qualitatively incorrect interpretation.

[1] Trapnell 2015.

One cell type, several cell states

Cell state can be described as the range of cellular phenotypes arising from the interaction of a defined cell type with its environment.

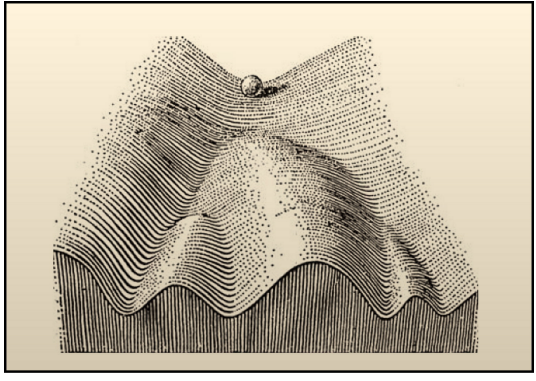
- A meaningful definition of cell types must be associated with what cell types do[2].
- In response to diverse stimuli, the same cell type can exhibit a range of different phenotypes == states.
- Several cell states to address the diverse function of cell type.

[2] Zeng 2022.

"Cells are residents of a vast landscape of possible states,

over which they travel during **development** *and in* **disease***"*

C.H. Waddington

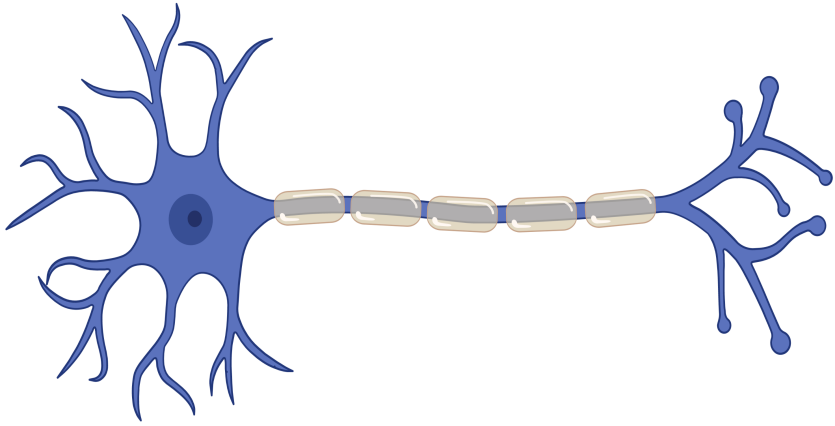


↓

The landscape might indeed change over time!

Decentralised biology of the neurons

Compartment-specific functions



↓

Need to rapidly respond to external cues.

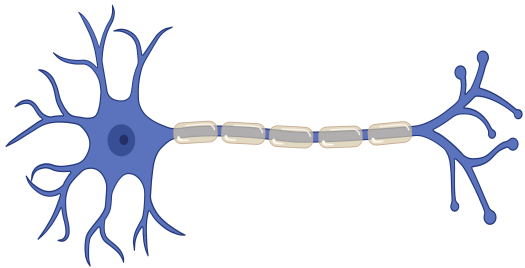
Time-scale underlying transcriptional regulation

- **For most genes, a new steady-state expression level is established by 120 mn.**
- Estimated median mRNA half-life in human cells is 10 h[3].
- mRNA transcription factors have shorter half-life (< 1 h)[4]

[3] Yang et al. 2003.
[4] Sharova et al. 2009.

Decentralised biology of the neurons

Compartment-specific functions



↓

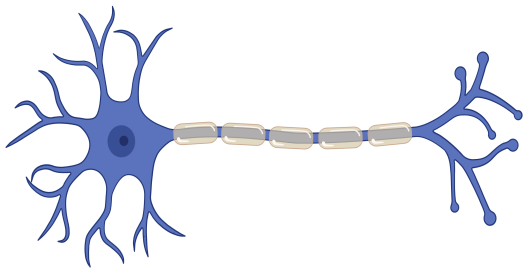
Need to rapidly respond to external cues.

↓

Cannot depend on transcription.

Decentralised biology of the neurons

Compartment-specific functions



Neurons homeostasis relies on **local translation** through controlled regulation of axonal mRNA **localization**, **transport**, and **stability**

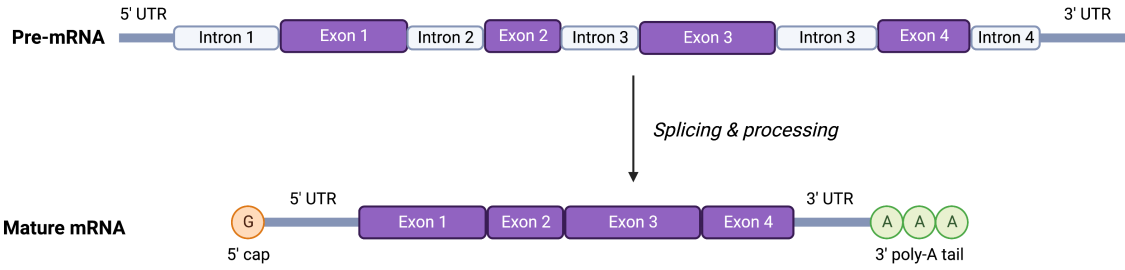
Gene expression analysis has been the major focus to precisely define **cellular states** and catalog them in development and disease, **yet**

- Regulation in gene expression does not enable fast response to environmental cues.
- Alternatively spliced variants can provide further information and help to refine cell types.[5]

[5] Booeshaghi et al. 2021.

One gene, several isoforms

What about their non-coding portions?



$>$ 90% of the transcriptome does not encode for protein.

↓

Key regulatory functions of 5' UTR, introns, and 3' UTR in health and disease.

Regulation by 5' UTR

Key regulatory platform for translational efficiency[6]

- Critical for ribosome recruitment and ultimately initiation of translation
- Switching between 5'UTR isoforms is a way to alter protein synthesis rates
- Can be comprehensively identified using CAGE-seq
- Impact in neuronal biology[7] and disease such as cancer[8]

[6] Hinnebusch, Ivanov, and Sonenberg 2016; Jia et al. 2020.
[7] Biever, Donlin-Asp, and Schuman 2019.
[8] Weber et al. 2023.

Regulation by 3' UTR

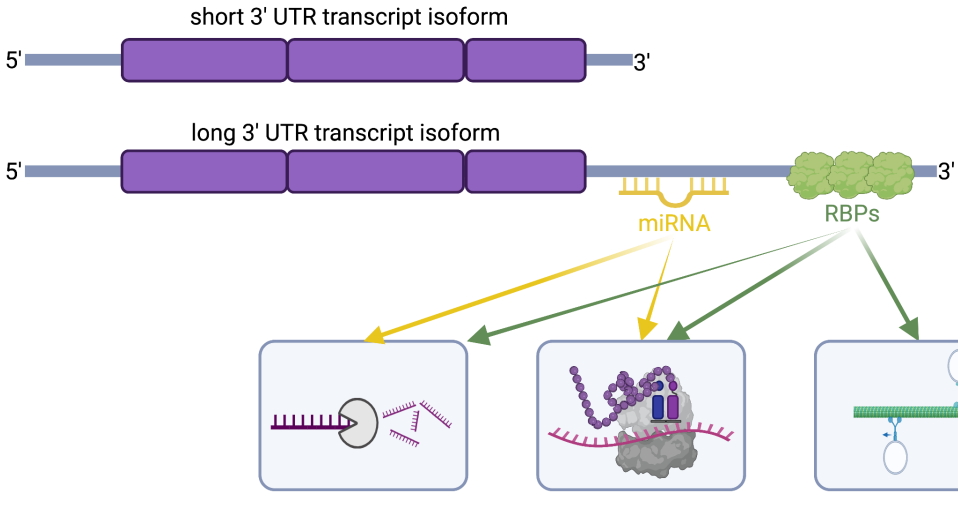
Key regulatory platform for mRNA localisation, storage and translation

- Generated by alternative polyadenylation.
- Tandem 3' UTR: identical protein-coding sequence but different 3' end.
- APA isoforms are cell-type specific and change on activation of signaling pathway[9]
- 3' UTR isoforms can undergo cytoplasmic shorting upon external stimuli[10]

[9] Mayr 2019.
[10] Andreassi et al. 2021.

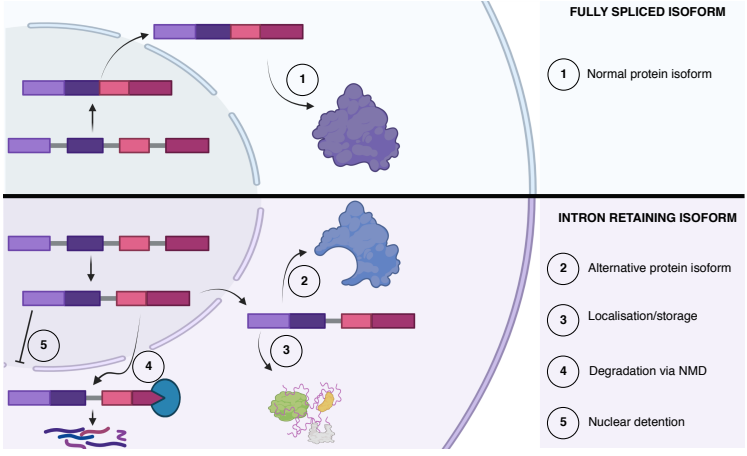
Regulation by 3' UTR

Key regulatory platform for mRNA localisation and translation



Regulation by intronic sequences

Emerging regulatory platform for mRNA and protein localisation



Alternative spliced isoforms enable

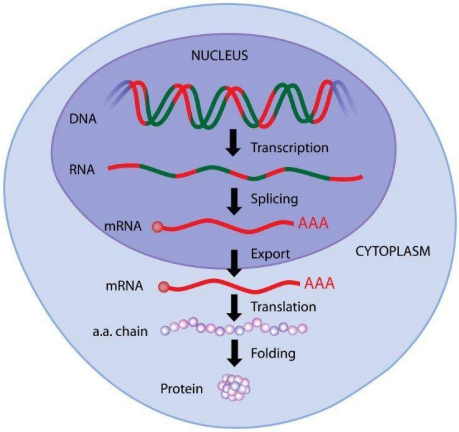
mRNA localisation, storage and translation.

And certainly much more!

Table of Content

RNA-sequencing data

What are we looking at?



12

[12] Image fromhttps://chem.libretexts.org/ (CC BY 3.0).

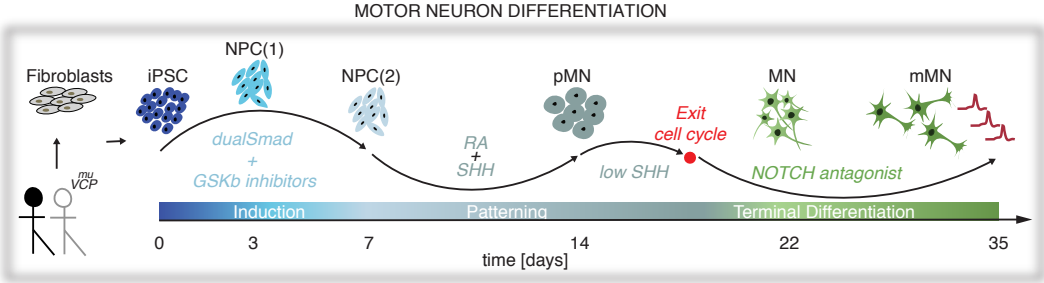
RNA-sequencing

Overview of a standard bioinformatic pipeline

1. Quality control (QC) of the fastq files (FASTQC)
2. Mapping of the reads onto a reference genome and/or transcriptome to obtain a count data table.
3. QC of the library
4. Pre-processing (statistical modelling)
5. Down-stream analysis
   - Unsupervised analysis (clustering)
   - Supervised analysis (DGE, pathway analysis)

Bulk RNA-sequencing Analysis of Differentiating Motor Neurons

Data-set for practical session

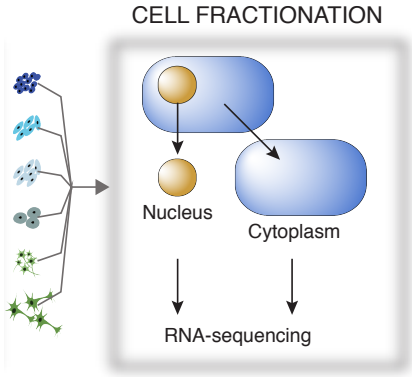


13

[13] Howe et al. 2022.

Bulk RNA-sequencing Analysis of Differentiating Motor Neurons

Data-set for practical session



14

[14] Howe et al. 2022.

RNA-sequencing

Pre-processing of the read count table

1. Log-transformation
2. Filter-out lowly expressed genes
3. Normalisation

Table of Content

Table of Content

Unsupervised clustering analysis

Hierarchical agglomerative clustering

Compute pair-wise distance between samples

- Flexible distance metrics between samples.
- Euclidean, correlation-based (Pearson or Spearman).

Agglomerative clustering of the samples

- Linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.
- Linkage criterion influences shape of the clusters.
- The definition of shortest distance is what differentiates between the different agglomerative clustering methods.
- Complete-linkage tends to produce more spherical clusters than single-linkage.
- Single-linkage tends to produce long thin clusters in which nearby elements of the same cluster have small distances

Table of Content

Unsupervised clustering analysis

Principal Component Analysis

- To transform a set of possibly correlated variables (genes) into "some more fundamental set of independent variables".
- To project a dataset from **many correlated** coordinates onto **fewer uncorrelated** coordinates
- The orthogonal coordinates are called principal components (PCs).
- PC retain most variability in the data.

Unsupervised clustering analysis

Principal Component Analysis

- To transform a set of possibly correlated variables (genes) into "some more fundamental set of independent variables".
- To project a dataset from **many correlated** coordinates onto **fewer uncorrelated** coordinates
- The orthogonal coordinates are called principal components (PCs).
- PC retain most variability in the data.

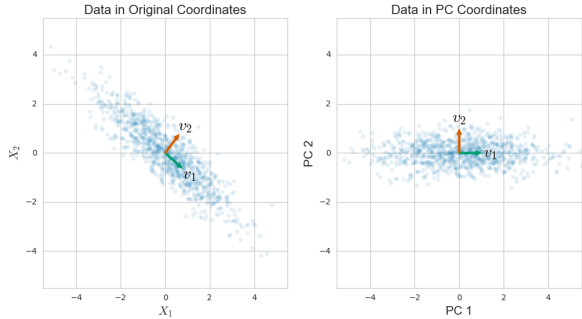
↓

- To reduce the dimensionality of data
- To extract essential information
- To characterize the structure of the data.

Principal Component Analysis

Objective

- Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix, mean-centered.
- Find $\mathbf{v}_1$ such that:
  1. $\|\mathbf{v_1}\| = 1$
  2. Variance of projection onto $\mathbf{v}_1$ is maximized



16

Principal Component Analysis

Covariance and projection

- The projection of a vector $\mathbf{x} \in \mathbb{R}^n$ onto $\mathbf{v}_1$ is given by $\mathbf{v}_1^T \mathbf{x}$.

- Assuming $\mathbf{X}$ is mean-centered, the sample variance of the projection onto $\mathbf{v}_1$ is:

$$\mathrm{Var}(\mathbf{X}\mathbf{v}_1) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{v}_1^T \mathbf{x}_i)^2 = \mathbf{v}_1^T \left(\frac{1}{n-1} \mathbf{X}^T \mathbf{X} \right) \mathbf{v}_1$$

- The covariance matrix is defined as:

$$\mathrm{Cov}(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}$$

Principal Component Analysis

Solution

The variance of $\mathbf{X}$ projected onto $\mathbf{v}_1$ must be maximized.

$$\mathrm{Var}(\mathbf{X}\mathbf{v}_1) = \mathbf{v}_1^T \left(\frac{1}{n-1} \mathbf{X}^T \mathbf{X} \right) \mathbf{v}_1 = \mathbf{v}_1^T \mathrm{Cov}(\mathbf{X}) \mathbf{v}_1$$

$$\downarrow$$

$$\mathrm{Cov}(\mathbf{X})\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \quad \Rightarrow \quad \mathbf{v}_1^T \mathrm{Cov}(\mathbf{X}) \mathbf{v}_1 = \lambda_1$$

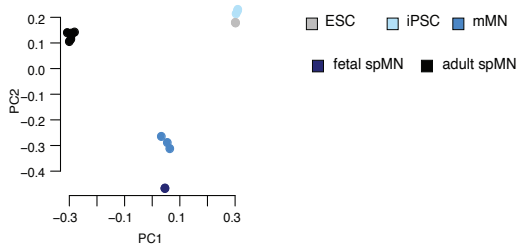
$\mathbf{v}_1$ and λ_1 are an eigenvector and its associated eigenvalue of the covariance matrix.

$$\downarrow$$

The optimal $\mathbf{v}_1$ is the eigenvector corresponding to the largest eigenvalue of $\mathrm{Cov}(\mathbf{X})$.

$$\mathrm{Cov}(\mathbf{X}) = \frac{1}{n-1} \mathbf{X}^T \mathbf{X} = V \Sigma V^T$$

Principal Component Analysis



- Most biologists focus on the clustering of the samples in PC1 and PC2.
- This is ok for simple experimental set-up (2 or 3 covariates).
- However what if more complex experiment?

↓

More subtle but biologically relevant signal might be captured in other components.

Table of Content

Singular Value Decomposition (SVD)

Let $\mathbf{M} \in \mathbb{R}^{g \times s}$ be a real-valued gene expression matrix, with g genes and s samples ($g \geq s$).

Each entry m_{ij} represents the expression of gene i in sample j.

- $\vec{g}_i \in \mathbb{R}^s$: expression profile of gene i across samples
- $\vec{s}_j \in \mathbb{R}^g$: expression profile of sample j across genes

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

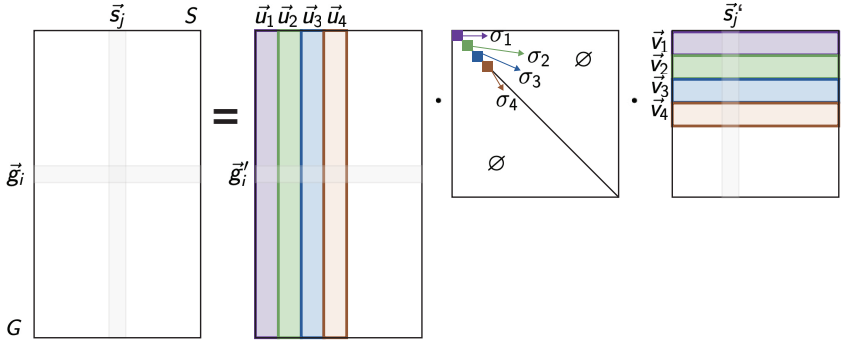
where:

- $\mathbf{U} \in \mathbb{R}^{g \times s}$: orthonormal basis for genes
- $\mathbf{\Sigma} \in \mathbb{R}^{s \times s}$: diagonal matrix of singular values
- $\mathbf{V}^T \in \mathbb{R}^{s \times s}$: orthonormal basis for samples

Singular Value Decomposition (SVD) Analysis

Definition

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$

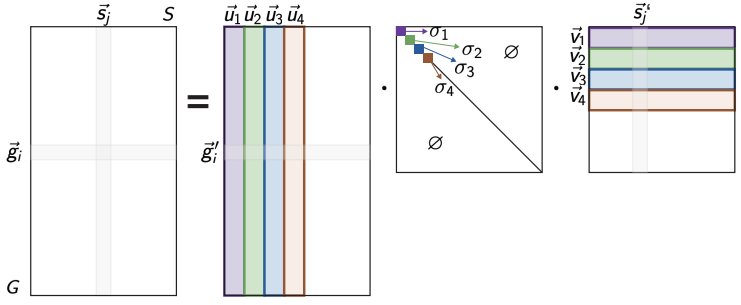


where $\vec{s}_j$ and $\vec{g}_i$ are the expression profiles of **sample** j and **gene** i.

Singular Value Decomposition (SVD)

Definition

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$

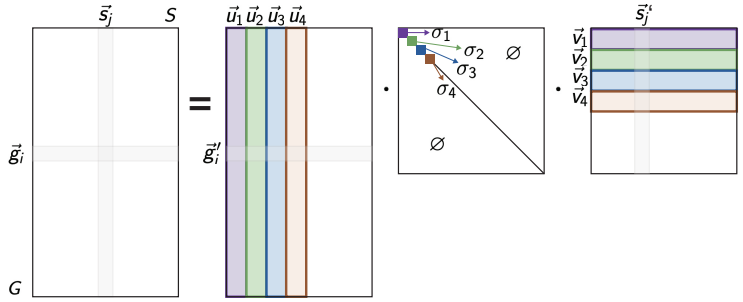


- the **left** singular vectors $\{\vec{u}_k\}$ form an orthonormal basis of the space of **sample** expression profiles so that $\mathbf{u}_i \cdot \mathbf{u}_j = 1$ for $i = j$ and 0 otherwise.
- the **right** singular vectors $\{\vec{v}_k\}$ form an orthonormal basis of the space of **gene** expression profiles so that $\mathbf{v}_i \cdot \mathbf{v}_j = 1$ for $i = j$ and 0 otherwise.

Singular Value Decomposition (SVD)

Important results

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$



- $M^{(l)} = \sum_{k=1}^{l} \mathbf{u}_k s_k \mathbf{v}_k^T$ is the closest rank-(l) matrix to M i.e. $M^{(l)}$ minimizes $\sum_{ij} \|m_{ij} - m_{ij}^{(l)}\|^2$

Singular Value Decomposition (SVD)

How to get U, Σ and V

$$M^T M = V \Sigma^2 V^T \tag{5}$$

$$U = XV\Sigma^{-1} \tag{6}$$

Relation Between SVD and PCA

When PCA is performed via the covariance matrix, it is closely related to the SVD of the data matrix $\mathbf{M}$.

Assuming $\mathbf{M}$ is column-centered:

- $\mathbf{M}^T\mathbf{M} \propto \text{Cov}(\mathbf{M})$
- Diagonalizing $\mathbf{M}^T\mathbf{M}$ gives $\mathbf{V}$, the right singular vectors , which are the principal directions (PCs).
- The matrix $\mathbf{U\Sigma}$ gives the principal component scores (coordinates of genes in PC space).

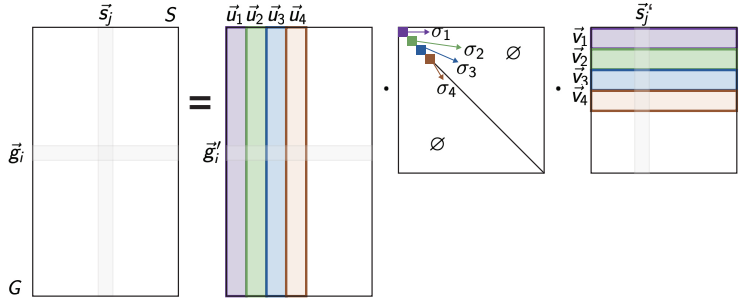
Alternatively, if rows of $\mathbf{M}$ are centered:

- $\mathbf{M}\mathbf{M}^T \propto \text{Cov(genes)}$
- The left singular vectors $\mathbf{U}$ then correspond to the principal components of gene expression profiles.

Singular Value Decomposition (SVD)

Definition

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$

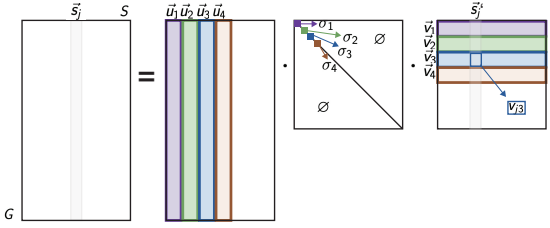


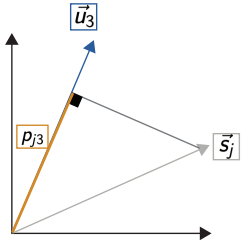
- The columns vectors $\{\vec{v}_k\}$ with $k \in [1 : S]$ are orthogonal and compose the right singular vectors.
- Each $\vec{v}_k$ can be thought as a linear combination of $\{\vec{g}_i\}$
- Each $\vec{g}_i$ can be thought as a linear combination of $\{\vec{v}_k\}$.

Singular Value Decomposition (SVD)

How individual **samples** *relate to singular vectors*

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$





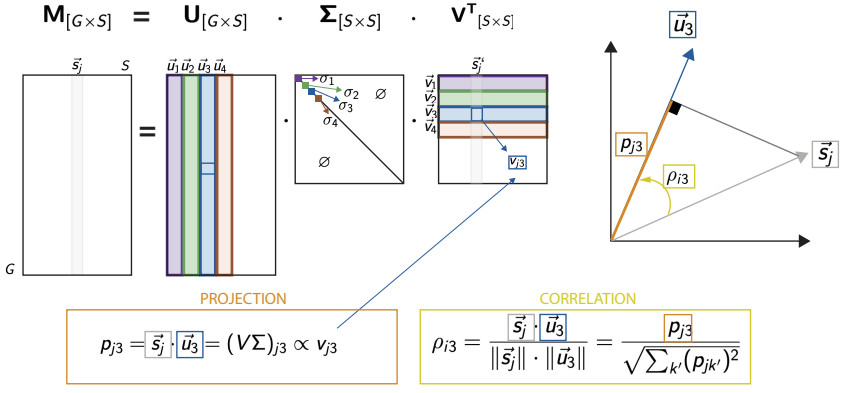
PROJECTION

$$p_{j3} = \vec{s_j} \cdot \vec{u_3}$$

- The projection of $\vec{s_j}$ onto the k^{th} *left* singular vector $\vec{u_k}$ is p_{jk}

Singular Value Decomposition (SVD)

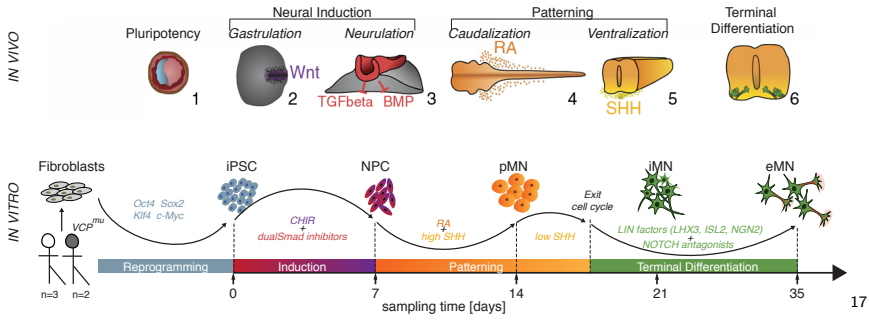
How individual **samples** *relate to singular vectors*



- The projection of $\vec{s_j}$ onto the k^{th} *left* singular vector $\vec{u_k}$ is p_{jk}
- $p_{jk} = \vec{s_j} \cdot \vec{u_k} = (M^T U)_{ik} = (U^T M)_{ik} = (V\Sigma)_{jk} \propto v_{jk}$
- v_{jk} represents how much a sample contributes to the *left* singular vector $\vec{u_k}$.

Singular Value Decomposition (SVD)

The analysis of the samples loadings onto v_k

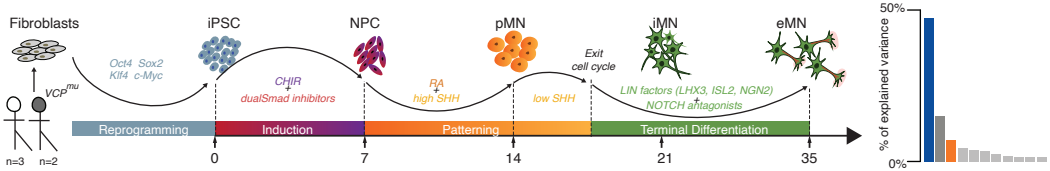


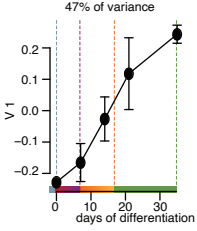
17

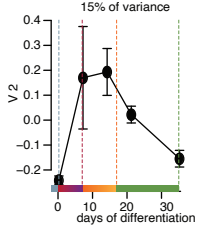
[17] Adapted from "Modeling ALS with motor neurons derived from human induced pluripotent stem cells" by Samuel Sances, Lucie I Bruijn, Siddharthan Chandran et al. 2016, Nat Neurosci, 19(4). Copyright 2016 Nature America, Inc. All rights reserved.

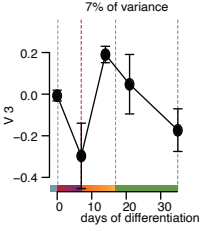
Singular Value Decomposition (SVD)

The analysis of the samples loadings onto v_k





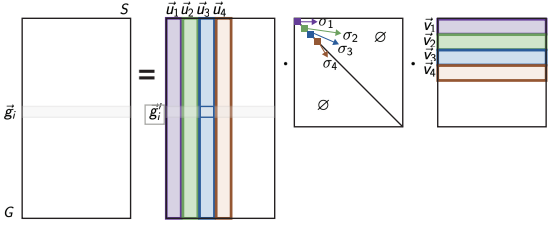


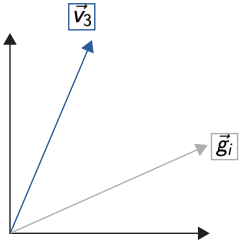


Singular Value Decomposition (SVD)

How individual **genes** *relate to singular vectors*

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$

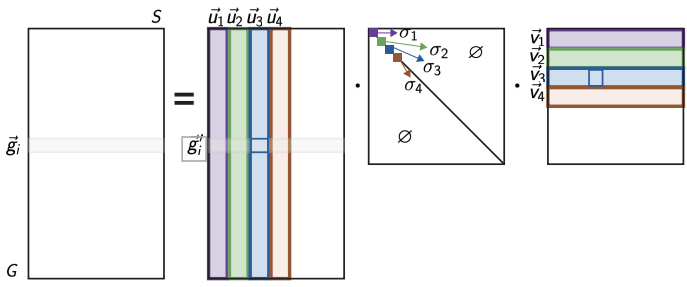


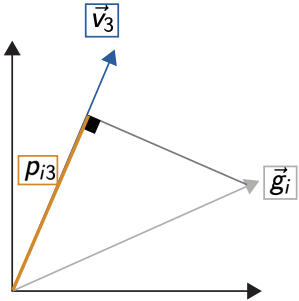


Singular Value Decomposition (SVD)

How individual **genes** *relate to singular vectors*

$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$



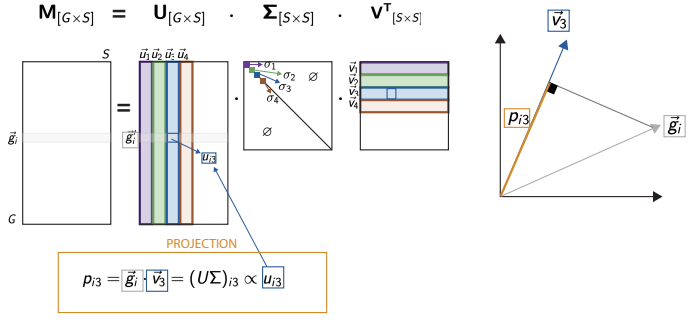


PROJECTION

$$p_{i3} = \vec{g}_i \cdot \vec{v}_3$$

Singular Value Decomposition (SVD)

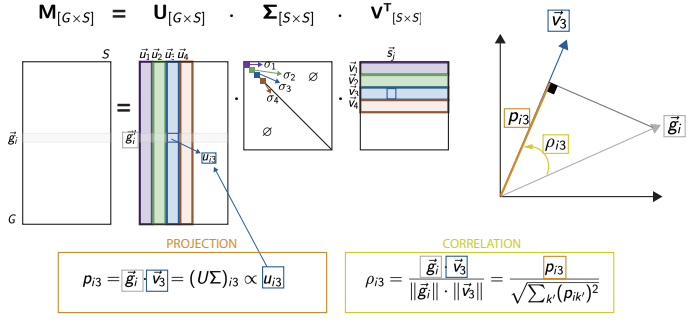
How individual **genes** *relate to singular vectors*



- The projection of $\vec{g}_i$ onto the k^{th} *right* singular vector $\vec{v}_k$ is p_{ik}
- $p_{ik} = \vec{g}_i \cdot \vec{v}_k = (MV)_{ik} = (U\Sigma)_{ik} \propto u_{ik}$
- u_{ik} represents how much a gene contributes to the *right* singular vector $\vec{v}_k$.

Singular Value Decomposition (SVD)

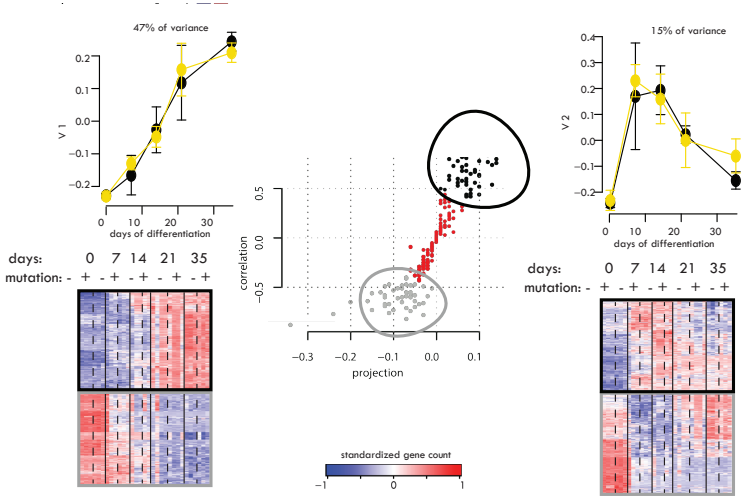
How individual **genes** *relate to singular vectors*



- The projection of $\vec{g}_i$ onto the k^{th} *right* singular vector $\vec{v}_k$ is p_{ik}
- $p_{ik} = \vec{g}_i \cdot \vec{v}_k = (MV)_{ik} = (U\Sigma)_{ik} \propto u_{ik}$
- u_{ik} represents how much a gene contributes to the *right* singular vector $\vec{v}_k$.

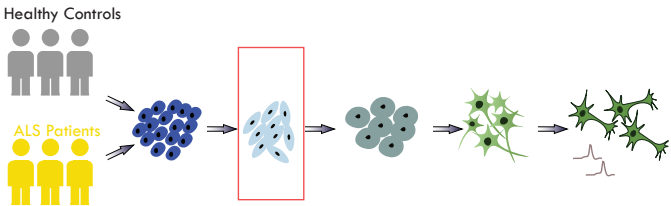
Singular Value Decomposition (SVD)

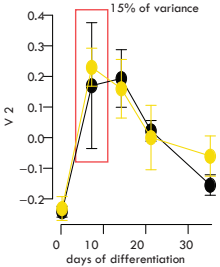
Extract most contributing and correlating genes

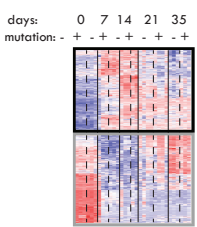


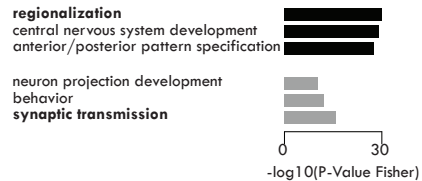
Singular Value Decomposition (SVD)

Biological Pathway Enrichment Analysis









Bibliography I

- Andreassi, Catia et al. (2021). "Cytoplasmic cleavage of IMPA1 3 UTR is necessary for maintaining axon integrity". In: *Cell Reports* 34.8, p. 108778.
- Biever, Anne, Paul G Donlin-Asp, and Erin M Schuman (2019). "Local translation in neuronal processes". In: *Current opinion in neurobiology* 57, pp. 141–148.
- Booeshaghi, A Sina et al. (2021). "Isoform cell-type specificity in the mouse primary motor cortex". In: *Nature* 598.7879, pp. 195–199.
- Hinnebusch, Alan G, Ivaylo P Ivanov, and Nahum Sonenberg (2016). "Translational control by 5-untranslated regions of eukaryotic mRNAs". In: *Science* 352.6292, pp. 1413–1416.
- Howe, Marija Petrić et al. (2022). "Physiological intron retaining transcripts in the cytoplasm abound during human motor neurogenesis". In: *Genome Research* 32.10, pp. 1808–1825.
- Jia, Longfei et al. (2020). "Decoding mRNA translatability and stability from the 5 UTR". In: *Nature structural & molecular biology* 27.9, pp. 814–821.
- Mayr, Christine (2019). "What are 3 UTRs doing?" In: *Cold Spring Harbor perspectives in biology* 11.10, a034728.
- Sharova, Lioudmila V et al. (2009). "Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells". In: *DNA research* 16.1, pp. 45–58.

Bibliography II

- Trapnell, Cole (2015). "Defining cell types and states with single-cell genomics". In: *Genome research* 25.10, pp. 1491–1498.
- Weber, Ramona et al. (2023). "Monitoring the 5 UTR landscape reveals isoform switches to drive translational efficiencies in cancer". In: *Oncogene* 42.9, pp. 638–650.
- Yang, Edward et al. (2003). "Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes". In: *Genome research* 13.8, pp. 1863–1872.
- Zeng, Hongkui (2022). "What is a cell type and how to define it?" In: *Cell* 185.15, pp. 2739–2755.