BIO-463
Genomics and bioinformatics

Lecture 9:
Single-cell RNA sequencing I

Dr Raphaëlle Luisier

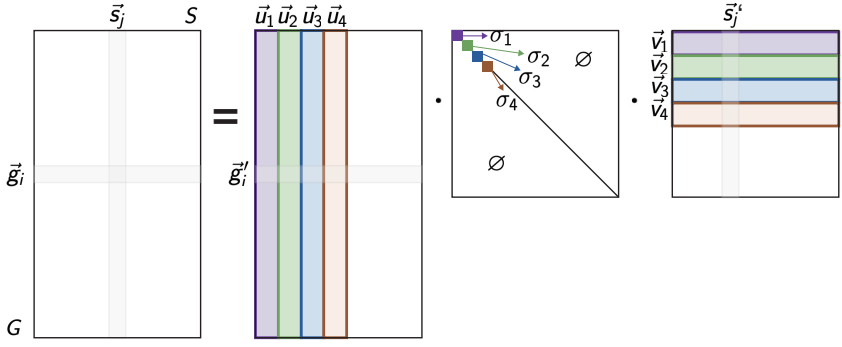
EPFL

Table of Content

Singular Value Decomposition (SVD) Analysis

Definition

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

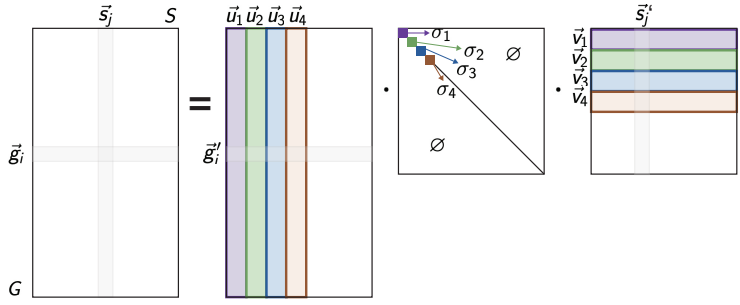


where $\vec{s}_j$ and $\vec{g}_i$ are the expression profiles of **sample** j and **gene** i.

Singular Value Decomposition (SVD)

Definition

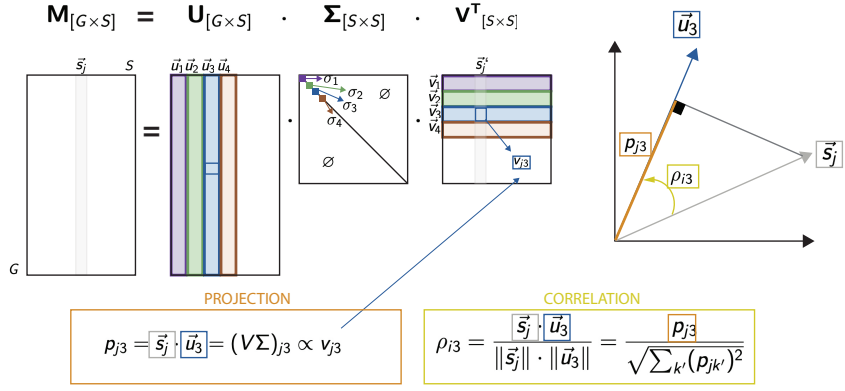
$$\mathbf{M}_{[G\times S]} = \mathbf{U}_{[G\times S]} \cdot \mathbf{\Sigma}_{[S\times S]} \cdot \mathbf{V}^{\mathbf{T}}{}_{[S\times S]}$$



- The 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)

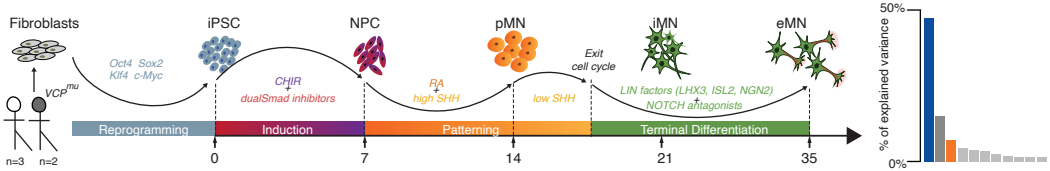
How individual **samples** *relate to singular vectors*

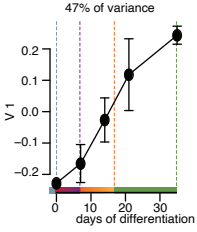


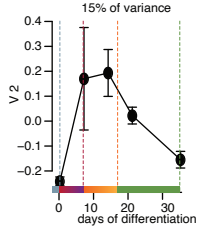
- The projection of $\vec{s_j}$ onto the k^{th} *left* singular vector $\vec{u_k}$ is p_{jk}
- $p_{jk} = \vec{s_j} \cdot \vec{u_k} = (M^T U)_{ik} = (U^T M)_{ik} = (V\Sigma)_{jk} \propto v_{jk}$
- v_{jk} represents how much a sample contributes to the *left* singular vector $\vec{u_k}$.

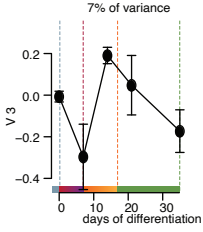
Singular Value Decomposition (SVD)

The analysis of the samples loadings onto v_k



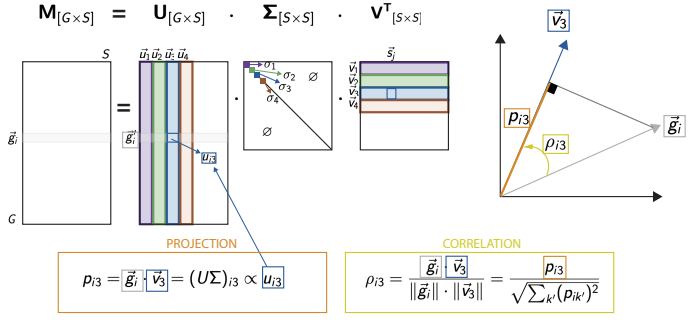






Singular Value Decomposition (SVD)

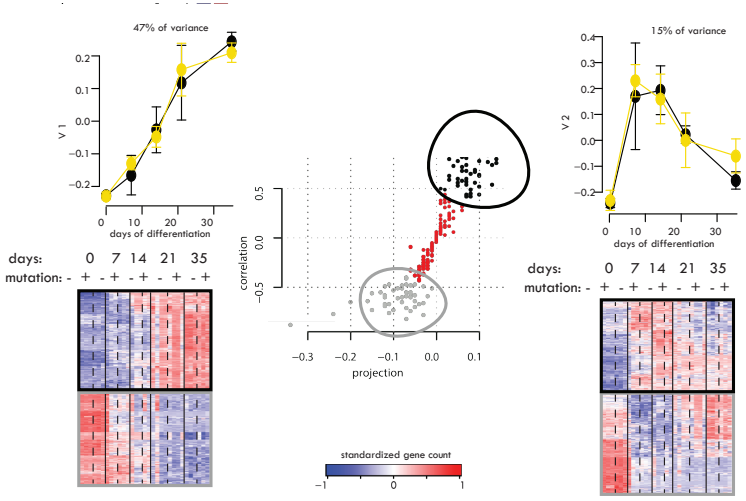
How individual **genes** *relate to singular vectors*



- 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

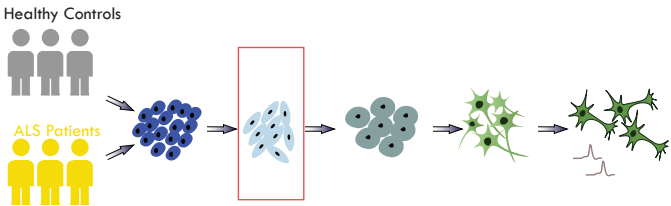


Table of Content

Table of Content

Single-cell RNA-sequencing

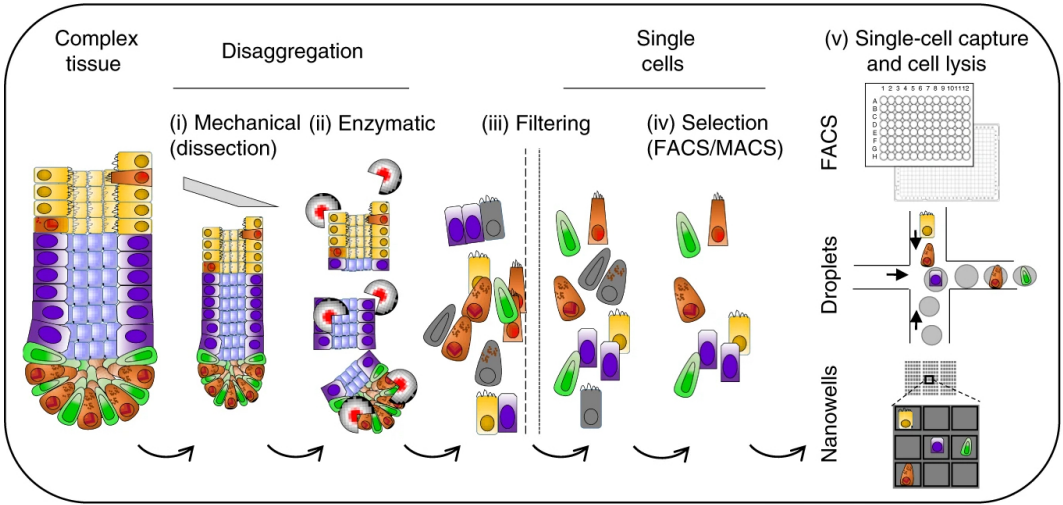
A transformative technology

- Cell states are now known to be more flexible than previously thought.
- Sc-RNA-seq enable to identify a variety of cell types/subpopulations that were invisible with traditional experimental techniques.
- Enable to characterize complex tissues

Single-cell research has become one of the fastest-growing fields in life science, **yet** sometimes at the cost of the quality of the data analysis which is not as mature as we would like.

Single-cell RNA-sequencing

Step 1: sample preparation



[1]

Single-cell RNA-sequencing

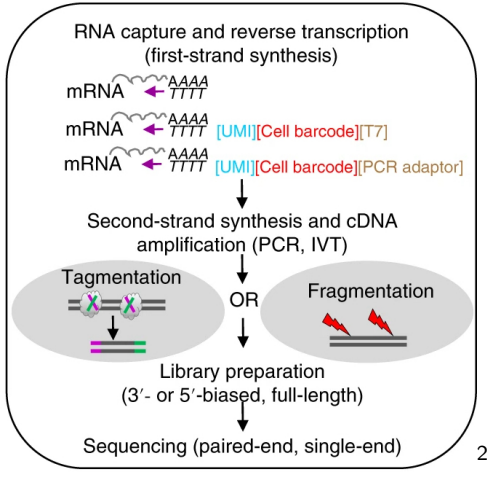
Step 1: sample preparation

- Starting material of various sources:
  - **fresh** viable single cells
  - preserved sample (**paraffin blocks**)
  - nuclear RNA from **frozen** tissue
- Preparation of single-cell suspensions
  - Primary cells, stem cells and other sensitive cell types may require washing and suspension in alternative buffers to ensure viability.
  - Sample processing might introduce variation in the gene expression profile for. ex. activation of stress-related genes.
  - Physical isolation of cells by microdissection or pipetting.
  - Target populations can be selected by FACS and MACS with appropriate labeling.

The number sequenced cells depend on the type of experiment and the expected frequency of the population of interest: https://satijalab.org/howmanycells/

Single-cell RNA-sequencing

Step 2: sequencing



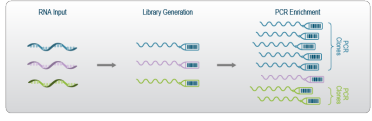
[2]

[2] Lafzi et al. 2018.

Single-cell RNA-sequencing

Step 2: sequencing

1. Most frequent **capture** of poly(A) RNA molecules by poly(T) oligo
2. **Reverse Transcription** (RT) and transcriptome **amplification**:
   - Captured RNA is reverse-transcribed into stable cDNA.
   - Most methods add single-cell-specific **barcodes**
   - Allows multiplexed processing of pooled samples.
   - Random-nucleotide-sequence stretches in the poly(T) oligonucleotide serve as unique **molecule** identifiers (UMIs)
   - UMI allow the user to correct for amplification biases and reduce technical noise



lexogen

Single-cell RNA-sequencing

Step 2: sequencing

1. Most frequent **capture** of poly(A) RNA molecules by poly(T) oligo
2. **Reverse Transcription** (RT) and transcriptome **amplification**
3. For short-read sequencing applications, the amplified cDNA (PCR) or RNA (IVT) is **fragmented** before sequencing adaptors are added.
4. Full-length versus 3'- or 5'-end transcript sequencing:
   - Full-length protocols do not allow the introduction of UMIs
   - Microtiter plates by FACS: few cells (hundreds) but full length.
   - Microfluidics: many cells (thousands) but 3'- or 5'-end biaised

Single-cell RNA-sequencing

Step 2: sequencing

Cell doublets

- Intrinsic problem for most microfluidics-based methods
- Two cells can be captured per reaction site (nanowell or droplet)

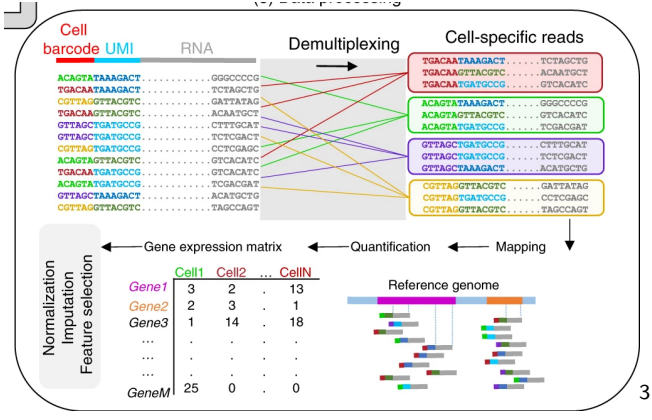
Cell capture efficiency

- Highly relevant in experiments involving primary or rare samples.
- The number of cells captured directly relates to the proportion of sample that enters downstream analysis

Table of Content

Single-cell RNA-sequencing

Step 3: pre-processing



[3]

[3] Lafzi et al. 2018.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
    - FASTQ reads are quality checked with FASTQC.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping to reference genomes.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping to reference genomes.
4. Quantification to create a transcript/gene expression matrix.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping to reference genomes.
4. Quantification to create a transcript/gene expression matrix.

- The scRNA-tools database provides a comprehensive list of available computational tools for data processing and analysis.
- Steps $2, 3, 4$ can be done with Kallisto bustools.

Single-cell RNA-sequencing

Bioinformatics libraries for single-cell RNA-seq analysis

- Seurat (*R*)[4] and Scanpy (*python*)[5] are widely used packages for scRNA-seq analysis.
- Main difference between Seurat and Scanpy in marker gene selection and DEG analysis
- These methods use different formulas to calculate fold-change values, based on the raw count and mean-log values.
- We will use Seurat given that we are working in R.

[4] Satija et al. 2015.
[5] Wolf, Angerer, and Theis 2018.

Single-cell RNA-sequencing

Data-set for tutorial

6

- 451Lu melanoma cell line
- Two conditions:
  - untreated cell line called **parental**.
  - cell line treated for 6 weeks with BRAF inhibitors; as still proliferating these are called **resistant.**

[6] Ho et al. 2018.

Single-cell RNA-sequencing

Seurat object

```
#Create Seurat object
GE <- CreateSeuratObject(
          counts = expression_matrix,
          meta.data = as.data.frame(cell_metadata),
          min.cells = 5,
          min.features = 0)

#Access raw counts matrix:
GE@assays$RNA@layers$counts

#Access normalised count matrix:
GE_qn@assays$RNA@layers$data
GE_sct@assays$SCT@data

#Access scaled count matrix:
GE_qn@assays$RNA@layers$data
```

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping.
4. Quantification.
5. **Filtering of low-quality cells**

Single-cell RNA-sequencing

Low-quality cells[7]

- Low QC cells might be due to dying cells, cells whose membranes are broken or doublets.
- Cells with a low count depth, few detected genes, and a high fraction of mitochondrial counts can indicate **cytoplasmic mRNA leakage** through a broken membrane, and thus, only mRNA located in the mitochondria is still conserved.
- Cells with unexpectedly high counts and a large number of detected genes may represent doublets.
- Cells with a high fraction of mitochondrial counts may be involved in respiratory processes.
- Cells with low counts and/or genes may correspond to quiescent cell populations.
- Cells with high counts may be larger in size.

QC covariates should be considered jointly and with expert eyes, being aware of the system we are working on.

[7] Luecken and Theis 2019.

Single-cell RNA-sequencing

Filtering of low-quality cells

To ensure that all cellular barcode data correspond to viable cells,
Seurat object has two relevant features:

nFeature_RNA

- number of genes detected in each cell.
- low *nFeature_RNA* for a cell indicates that it may be dead/dying or an empty droplet.
- High *nFeature_RNA* indicates that the cell may in fact be a doublet.

nCount_RNA

- total number of molecules (UMI) detected within a cell.
- Can be high yet all coming from the same molecule.
- nCount values that are too low could possibly mean an empty droplet
- nCount values that are too high could possibly mean a doublet.

Single-cell RNA-sequencing

Filtering of low-quality cells

Cells that are poor quality are likely to have[8]

- low # genes detected per cell (*nFeature_RNA*)
- low UMI counts per cell (*nCount_RNA*)
- high mitochondrial counts ratio

[8] Ilicic et al. 2016; Griffiths, Scialdone, and Marioni 2018.

Single-cell RNA-sequencing

Filtering of low-quality cells

Low quality cells are identified with the following criteria:

- UMI counts (*nCount_RNA*) per cell
- Genes (*nFeature_RNA*) detected per cell
- UMIs vs. genes detected
- Mitochondrial counts ratio
- Complexity of the library

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping.
4. Quantification.
5. Filtering of low-quality cells
6. **Filtering of lowly expressed genes**

Single-cell RNA-sequencing

Filtering of genes

- Single-cell data have a very low sensitivity.
- scRNA-seq data are sparse.
- The majority of cells do not express more than 3000 genes
- Remove genes not detected in any cells
- Many studies select genes with read count above **a pre-defined threshold** for example selecting genes which are expressed in 10 or more cells.
- An alternative is to select for genes reliably expressed in specific conditions by creating pseudo-bulk expression matrix.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping.
4. Quantification.
5. Filtering of low-quality cells
6. Filtering of lowly expressed genes
7. **Normalisation**

Single-cell RNA-sequencing

Technical challenges in scRNA-seq data

- Methods developed for bulk RNA-seq tend to neglect prominent features of scRNA-seq data.
- High level of noise such as dropout events due to stochastic RNA loss, biased amplification and incomplete library sequencing.
- zero inflation, i.e., excess of zero read counts observed in some single-cell protocols.
- transcriptome-wide nuisance effects (e.g., batch) comparable in magnitude to the biological effects of interest.
- uneven sample quality, e.g., in terms of alignment rates and nucleotide composition.

Single-cell RNA-sequencing

Normalisation

Scaling

- The goal is to compare concentrations rather than absolute amount of mRNAs per cell.
- Multiply each UMI count by a cell specific factor.
- To get all cells to have the same UMI counts.

Simple transformation

- Apply the same function to each individual measurement such as log transform or square root transform.
- Genes with different abundances are affected differently[9].
- Effective normalization using the log transform is only observed with low/medium abundance genes.
- Substantial imbalances in variance were observed with the log-normalized data.

Source

[9] Hafemeister and Satija 2019.

Single-cell RNA-sequencing

Normalisation cont'd

Pearson residuals for transformation

- We cannot treat all genes the same[10].
- Use Pearson residuals for transformation as implemented in Seurat's **SCTransform** function.
- Measurements are multiplied by a gene-specific weight.
- Each gene is weighted based on how much evidence there is that it is non-uniformly expressed across cells.
- More evidence == more of a weight; Genes that are expressed in only a small fraction of cells will be favored (useful for finding rare cell populations).
- Not just a consideration of the expression level is, but also the distribution of expression.

Source

[10] Hafemeister and Satija 2019.

Single-cell RNA-sequencing

Normalisation cont'd

Normalisation choices determines DE-analysis performance[11]!

SCONE[12] is an R package for comparing and ranking the performance of different between-samples normalization schemes for single-cell RNA-seq.

[11] Vieth et al. 2019.
[12] Cole et al. 2019.

Single-cell RNA-sequencing

Existing methods

- Quantile Normalisation: has been previously proposed in protocols lacking UMI counts[13].
- Methods developed for scRNA-seq: SCnorm[14], SCRAN[15].
- Methods developed for bulk RNA-seq: Census[16], MR[17], TMM[18], Linnorm[19].

[13] Townes and Irizarry 2020.
[14] Bacher et al. 2017.
[15] Lun, Bach, et al. n.d.
[16] Qiu et al. 2017.
[17] Anders and Huber 2010.
[18] Robinson and Oshlack 2010.
[19] Yip et al. 2017.

Single-cell RNA-sequencing

Normalisation in Seurat

- **LogNormalize**: gene counts for each cell are divided by the total counts for that cell and multiplied by the scale.factor. This is then natural-log transformed.
- **CLR**: $clr(x) := \left(\ln x_i - \frac{1}{D} \sum_{j=1}^{D} \ln x_j \right)_i$.
- **RC**: gene counts are divided by the total counts per cell and multiplied by the scale.factor. *No log-transformation* is applied. For counts per million (CPM) set scale.factor = 1e6.
- **SCTransform**= modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiments. This procedure omits the need for heuristic steps including pseudocount addition or log-transformation.

Single-cell RNA-sequencing

Step 3: pre-processing

1. QC of the FASTQ reads.
2. De-multiplexing of reads using cell barcodes.
3. Mapping.
4. Quantification.
5. Filtering of low-quality cells
6. Filtering of lowly expressed genes
7. Normalisation
8. **Imputation**

Single-cell RNA-sequencing

Imputation

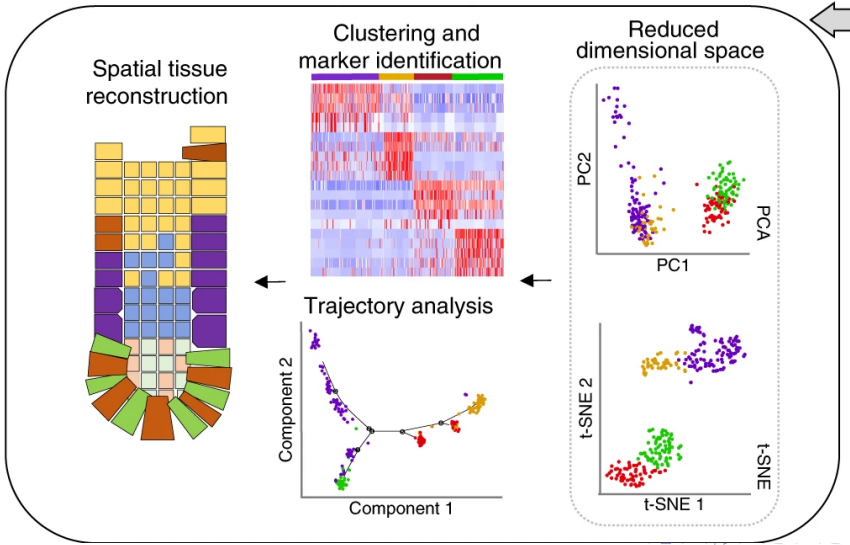
- scRNA-seq datasets are also very sparse (many zeros).
- Non-expressed genes and technical shortcomings, such as dropout events (unsequenced transcripts).
- Missing transcript values can be computationally inferred by imputation, for example, with MAGIC[20], scImpute[21], scIGAN[22].
- See ▸ Hou et al. (2020, Genome Biology) for a complete review of the methods.

This step is not performed in this tutorial.

[20] Van Dijk et al. 2018.
[21] Li and Li 2018.
[22] Xu et al. 2020.

Single-cell RNA-sequencing

Step 4: down-stream analysis



[23]

Single-cell RNA-sequencing

Step 4: down-stream analysis

- Dimensionality reduction analysis
- Unsupervised clustering
- Differential gene expression analysis
- Cell type identification
- Trajectory analysis

Table of Content

Single-cell RNA-sequencing

Dimensionality reduction analysis[24]

- By reducing the dimensionality, the computational complexity can be significantly reduced, making the algorithms more efficient.
- Can help in filtering out noisy/irrelevant features to achieve better generalization and predictive accuracy.
- To enable the representation of data in two or three dimensions, allowing for easier visualization and interpretation.
- Can assist in feature engineering by creating new, derived features that capture the most relevant information in the original dataset.

[24] source:https://aurigait.com/blog/blog-easy-explanation-of-dimensionality-reduction-and-techniques/

Single-cell RNA-sequencing

Dimensionality reduction analysis

- To visually inspect cellular subpopulation structures.
- To enable unsupervised clustering analysis.
- Most frequent methods: PCA, t-SNE and UMAP.
- For PCA:
  - Choose the most variable features then scale the data.
  - Highly expressed genes exhibit the highest amount of variation
  - Do not want highly variable genes only to reflect high expression
  - Need to scale the data to scale variation with expression level.

Single-cell RNA-sequencing

Selection of highly informative genes

- Dimensionality reduction analysis challenging with too may genes
- A common strategy is to analyze highly variable genes across datasets.
- Increases the signal-to-noise ratio.
- Reduces the computational complexity.
- With the Seurat *FindVariableFeatures*() function.

Single-cell RNA-sequencing

Data Scaling in Seurat

- With the Seurat *ScaleData*() function.
- Adjust the expression of each gene across D cells $x_i \rightarrow \tilde{x}_i$ such that $\mu_i = \frac{1}{D} \sum_{i=1}^{D} \tilde{x}_i = 0$.
- Scaling expression of each gene to give a variance across cells of 1.

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)[25]

The aims

- To perform non-linear scaling to represent changes at different levels.
- To find optimal separation in 2-dimensions.

[25] Maaten and Hinton 2008.

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)[26]

The idea

- Based around all-vs-all table of pairwise cell to cell distances.
- Convert similarities between data points in the high-dimensional space into probabilities.
- Map these probabilities to a lower-dimensional space in a way that preserves the relationships between the data points as much as possible.
- Minimize the divergence between the pairwise similarities in the high versus the low-dimensional spaces.

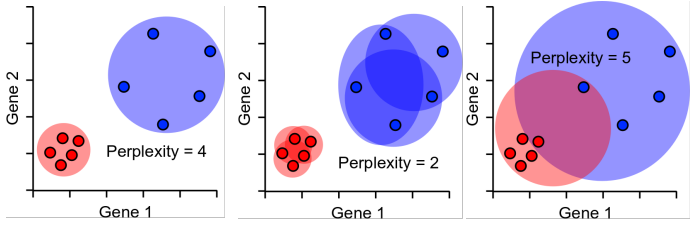
[26] Maaten and Hinton 2008.

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE)[27]

The parameters in *RunTSNE()* from Seurat

- *dims* the dimensions of the PCA to use as input features.
- *perplexity* the expected number of neighbours within a cluster.
- Distances scaled relative to perplexity neighbours.



source: here

[27] Maaten and Hinton 2008.

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE) cont'd[28]

The procedure

1. Pairwise similarities between data points are computed using a Gaussian kernel that measures the similarity based on the Euclidean distance between data points.
2. From the pairwise similarities, probability distributions are constructed for each data point that represent the similarities between a data point and all other points in the high-dimensional space.
3. An initial embedding of the data points in the low-dimensional space is randomly generated.
4. Pairwise similarities between data points are computed in the low-dimensional space.

[28] Maaten and Hinton 2008.

Dimensionality Reduction

t-distributed Stochastic Neighbor Embedding (t-SNE) cont'd

The resulting embedding

- *Unlike PCA*, X and Y do not mean anything.
- *Unlike PCA*, distance does not mean anything.
- Cannot rationalise distances, or add in more data.
- Gives reliable information on the closest neighbours.
- Large distance information is almost irrelevant.

Pros and cons

+ Can reduce to 2D.
+ Can cope with non-linear scaling.
- Slow and does not scale well to large numbers of cells (10k+)
- Does not cope well with noisy data.

source: here

Dimensionality Reduction

Uniform Manifold Approximation and Projection (UMAP)[29]

The idea

- To construct a low-dimensional representation of the data that preserves **global** and **local** structure.
- Uses a **graph-based approach** to build a topological representation of the data..
- Minimize the divergence between the pairwise similarities in the high-dimensional space versus in the low-dimensional space.

[29] Becht et al. 2019.

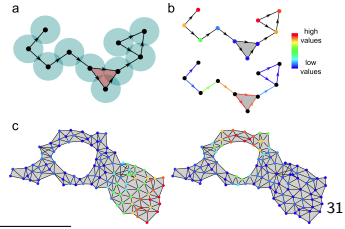
Dimensionality Reduction

Uniform Manifold Approximation and Projection (UMAP)[30]

The procedure

1. Calculate the pairwise distances between data points in the high-dimensional space using Euclidean distance, cosine distance, or correlation distance.
2. Based on the pairwise distances, a fuzzy simplicial set is constructed representing the local neighborhood structure of the data points.

Simplical sets are higher-dimensional generalizations of directed graphs, partially ordered sets and categories.



31

[30] Becht et al. 2019.

[31] Goyek, Yamajala, and Camara 2019.

Dimensionality Reduction

Uniform Manifold Approximation and Projection (UMAP)[32] *cont'd*

Key parameters

- *n.neighbors* the number of expected nearest neighbours; the same concept as perplexity; will affect the influence given to global vs local information.
- *min.dist* how tightly UMAP packs points which are close together; will affect how compactly packed the local parts of the plot are.
- *metric* determines the choice of metric used to measure distance in the input space.

[32] Becht et al. 2019.

Dimensionality Reduction

Uniform Manifold Approximation and Projection (UMAP)[33] *cont'd*

Pros and cons

- \+ Can reduce to 2D.
- \+ Can cope with non-linear scaling.
- \+ Allow new data to be added to an existing projection
- +/- A bit quicker than tSNE
- +/- *Is supposed* to preserve more global structure than tSNE
- +/- *It is claimed* that it can be run on raw data without PCA preprocessing

▸ source: here

[33] Becht et al. 2019.

Single-cell RNA-sequencing

Compare methods

PCA

+ Linear transformation that preserves Euclidean distances between cells in the full PCA space.
+ Interpretable.
+ Effective for capturing global patterns.
+ Computationally efficient.
- The structure of most scRNA-seq datasets cannot be captured by 2 or 3 PCs.

Single-cell RNA-sequencing

Compare methods cont'd

t-SNE

+ Focus on capturing **local similarity** at the expense of global structure. .
- Non-linear i.e. **the interpretability** of the reduced dimensions is **sacrificed**.
- May exaggerate differences between cell populations.
- t-SNE graphs may show strongly different numbers of clusters depending on perplexity parameter.
- Computationally intensive.

Single-cell RNA-sequencing

Compare methods cont'd

UMAP

+ Supposed to better preserve large-scale structures than t-SNE.
+ Fast and able to scale to large numbers of cells.
- Tends to favor fully connected representations of the data rather than the discrete clusters favored by t-SNE.
- Non-linear i.e. **the interpretability** of the reduced dimensions is **sacrificed**.
- Computationally intensive.

Single-cell RNA-sequencing

Reproducibility issue with tSNE and UMAP

- t-SNE and UMAP require user-defined hyperparameters
- Result are sensitive to the value chosen.
- t-SNE and UMAP are stochastic.
- Results significantly depend on initialization.
- Faithful representation of local and/or global structure in low dimensions not always true[34].

[34] Kobak and Berens 2019; Cooley et al. 2019.

Single-cell RNA-sequencing

Optimality of linear embedding

- The Johnson-Lindenstrauss lemma on the optimality of linear embedding shows that preservation of pairwise distances with a margin of error of at most 20% for a sized dataset of 10,000 cells would require at least 1,842 dimensions..
- Extreme dimension reduction inevitably induces significant distortion of high-dimensional datasets[35].
- Poor preservation of local neighborhoods by both PCA and the nonlinear reduction methods.

It discouraged to blindly apply such heuristic procedures.

[35] Chari and Pachter 2023.

Bibliography I

- Anders, S and W Huber (2010). *Differential expression analysis for sequence count data. Nat Prec.*
- Bacher, Rhonda et al. (2017). "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nature methods* 14.6, pp. 584–586.
- Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1, pp. 38–44.
- Chari, Tara and Lior Pachter (2023). "The specious art of single-cell genomics". In: *PLOS Computational Biology* 19.8, e1011288.
- Cole, Michael B et al. (2019). "Performance assessment and selection of normalization procedures for single-cell RNA-seq". In: *Cell systems* 8.4, pp. 315–328.
- Cooley, Shamus M et al. (2019). "A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data". In: *Biorxiv*, p. 689851.
- Govek, Kiya W, Venkata S Yamajala, and Pablo G Camara (2019). "Clustering-independent analysis of genomic data using spectral simplicial theory". In: *PLoS computational biology* 15.11, e1007509.
- Griffiths, Jonathan A, Antonio Scialdone, and John C Marioni (2018). "Using single-cell genomics to understand developmental processes and cell fate decisions". In: *Molecular systems biology* 14.4, e8046.

Bibliography II

- Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1, p. 296.
- Ho, Yu-Jui et al. (2018). "Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations". In: *Genome research* 28.9, pp. 1353–1363.
- Ilicic, Tomislav et al. (2016). "Classification of low quality cells from single-cell RNA-seq data". In: *Genome biology* 17, pp. 1–15.
- Kobak, Dmitry and Philipp Berens (2019). "The art of using t-SNE for single-cell transcriptomics". In: *Nature communications* 10.1, p. 5416.
- Lafzi, Atefeh et al. (2018). "Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies". In: *Nature protocols* 13.12, pp. 2742–2757.
- Li, Wei Vivian and Jingyi Jessica Li (2018). "An accurate and robust imputation method scImpute for single-cell RNA-seq data". In: *Nature communications* 9.1, p. 997.
- Luecken, Malte D and Fabian J Theis (2019). "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular systems biology* 15.6, e8746.
- Lun, ATL, K Bach, et al. (n.d.). "(2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome Biol* 17 (), p. 75.

Bibliography III

- Maaten, Laurens Van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE.". In: *Journal of machine learning research* 9.11.
- Qiu, Xiaojie et al. (2017). "Single-cell mRNA quantification and differential analysis with Census". In: *Nature methods* 14.3, pp. 309–315.
- Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome biology* 11, pp. 1–9.
- Satija, Rahul et al. (2015). "Spatial reconstruction of single-cell gene expression data". In: *Nature biotechnology* 33.5, pp. 495–502.
- Townes, F William and Rafael A Irizarry (2020). "Quantile normalization of single-cell RNA-seq read counts without unique molecular identifiers". In: *Genome biology* 21, pp. 1–17.
- Van Dijk, David et al. (2018). "Recovering gene interactions from single-cell data using data diffusion". In: *Cell* 174.3, pp. 716–729.
- Vieth, Beate et al. (2019). "A systematic evaluation of single cell RNA-seq analysis pipelines". In: *Nature communications* 10.1, p. 4667.
- Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19, pp. 1–5.

Bibliography IV


Xu, Yungang et al. (2020). "scIGANs: single-cell RNA-seq imputation using generative adversarial networks". In: *Nucleic acids research* 48.15, e85–e85.


Yip, Shun H et al. (2017). "Linnorm: improved statistical analysis for single cell RNA-seq expression data". In: *Nucleic acids research* 45.22, e179–e179.