BIO-463
Genomics and bioinformatics

# Lecture 10:
# Single-cell RNA sequencing II

Dr Raphaëlle Luisier

EPFL

# Table of Content

## 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*

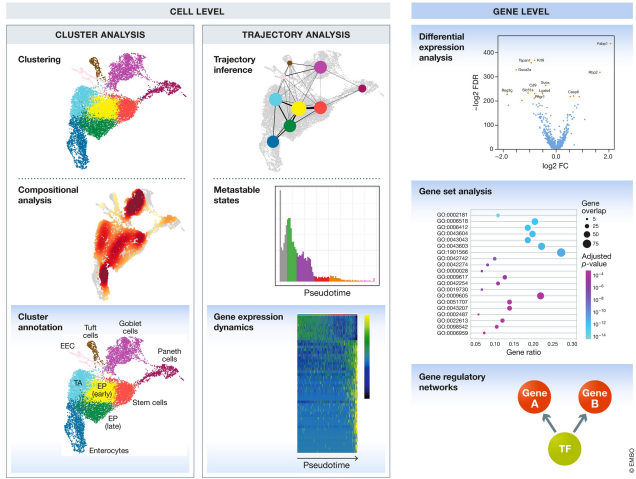
UMAP

+ 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.

# Table of Content

# sc-RNA-seq analysis

*Downstream analysis*




[1]Luecken and Theis 2019.

# sc-RNA-seq analysis

*Downstream analysis*

*" The first step towards identifying cellular populations is to cluster cells into groups with similar expression profiles that explain the heterogeneity in the data."*[2]

[2]Heumos et al. 2023.

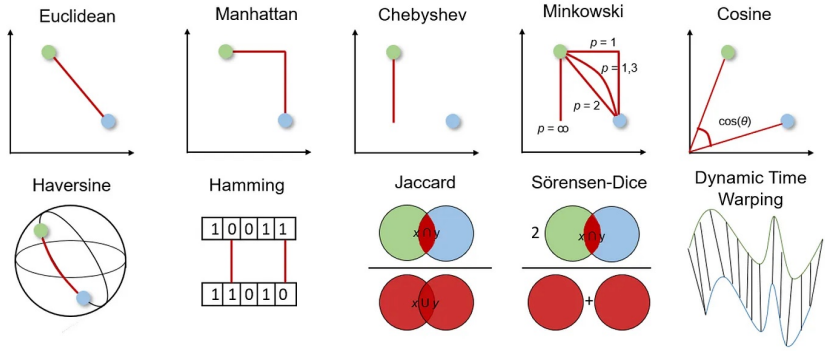
## Single-cell RNA-sequencing

*Unsupervised clustering of single-cell RNA-seq data*

- To group cells based on the *similarity* of their gene expression profiles.
- *Step 1*: Compute distances between cells
  - often based on dimensionality-reduced representations
  - Several distance metrics.

# Single-cell RNA-sequencing

*Unsupervised clustering of single-cell RNA-seq data*






[3] **10 often used distance measures (based on M. Grootendorst).**

# Table of Content

# Single-cell RNA-sequencing

*Distance metrics*

## Minkowski distance

$$d_{\mathbf{xy}} = \Big(\sum_{i=1}^{n} |x_i - y_i|\Big)^{1/p} \tag{1}$$

when $p = 1$, Manhattan distance, which measures the absolute value between two points; when $p = 2$, Euclidean distance, which measures a straight line between two points.

## Cosine similarity[4]

$$d_{\mathbf{xy}} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \cos(\theta) \tag{2}$$

Determines whether two vectors are pointing in roughly the same direction

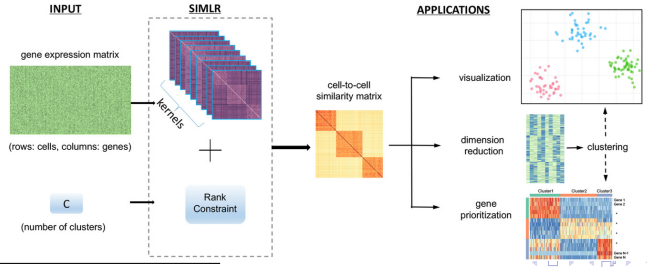
[4] Haghverdi et al. 2018.

# Single-cell RNA-sequencing

*Distance metrics cont'd*

## SIMLR method[5]

- Learns a distance metric that best fits the structure of the data via combining multiple kernels.
- Employ graph diffusion to overcome high level of drop-out events.
- Discriminative cell-to-cell similarity.
- Constraints the similarity matrix to have an approximate block-diagonal structure with C blocks where the samples of the same populations to be more similar.




[5]Wang et al. 2017

## Single-cell RNA-sequencing

*Distance metrics cont'd*

### Pearson Correlation-based distance

$$d_{\mathbf{xy}} = 1 - \frac{\sum(x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})}{\sqrt{\sum(x_i - \mu_{\mathbf{x}})^2(y_i - \mu_{\mathbf{y}})^2}} \tag{3}$$

- Assumes a Gaussian-like distribution for the data.
- Sensitive to outliers.

## Single-cell RNA-sequencing

*Distance metrics cont'd*

### Spearman Correlation-based distance

$$d_{\mathbf{xy}} = 1 - \rho = \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4}$$

where $d_i$ is the difference between the ranks of $x_i$ and $y_i$ and $n$ is the number of observations.

- For non-linear data.
- Non-parametric.

# Table of Content

## Single-cell RNA-sequencing

*Unsupervised clustering of single-cell RNA-seq data*

- To group cells based on the *similarity* of their gene expression profiles.
- *Step 1*: Compute distances between cells
  - Often based on dimensionality-reduced representations
  - Several distance metrics i.e. Euclidean, cosine similarity, correlation-based, the SIMLR method which learns a distance metric for each dataset using Gaussian kernels
- *Step 2*: group cells accordingly
  - clustering algorithms methods
  - community detection methods

## Single-cell RNA-sequencing

*Unsupervised clustering of single-cell RNA-seq data*

*"Independent benchmarks*[6] *showed that* **community detection** *based on graph modularity optimization via the Louvain algorithm works best for cluster identification."*[7]

[6] Luecken and Theis 2019; Duò, Robinson, and Soneson 2018; Freytag et al. 2018.
[7] Heumos et al. 2023.

# Single-cell RNA-sequencing

*Clustering algorithms methods*

## The general idea

- Based directly on a distance matrix.
- Cells are assigned to clusters by
  1. minimizing intracluster distances
  2. finding dense regions in the reduced expression space

## The algorithms

- Hierarchical clustering algorithm is commonly used with PAGODA[8], SINCERA[9] and bigSCale[10].
- *k*-means clustering algorithm

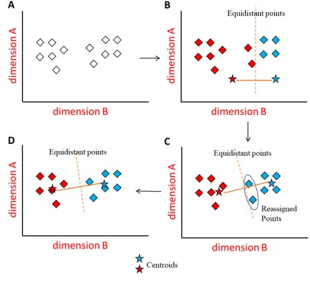
[8] Fan et al. 2016.
[9] Guo et al. 2015.
[10] Iacono et al. 2018.

# Single-cell RNA-sequencing

## *K-means Clustering Algorithm*

### The algorithm

1. estimates $k$ centroids;
2. assign cells to the nearest cluster centroid $k$;
3. recomputes centroids on the basis of the mean of cells in the centroid clusters;
4. reiterate.



# Single-cell RNA-sequencing

*K-means Clustering Algorithm cont'd*

## The algorithm

1. estimates $k$ centroids;
2. assign cells to the nearest cluster centroid $k$;
3. recomputes centroids on the basis of the mean of cells in the centroid clusters;
4. reiterate.

- Parameter: the number of clusters expected; usually unknown and must be calibrated heuristically.
- Correlation-based distances may outperform other distance metrics when used with $k$-means[11].
- Works well in cases where the data is well separated and spherical/circular in shape.
- Struggles to cluster datasets with spiral shapes or varying densities.

[11] Kim et al. 2019.

# Single-cell RNA-sequencing

*K-means Clustering Algorithm cont'd*

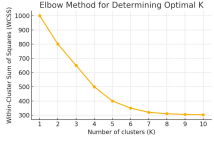
## Determining K

1. Calculate the Within Cluster Sum of Squares (WCSS), also known as Inertia:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

   where $C_k$ is the $k$-th cluster, $x_i$ are data points in cluster $C_k$, and $\mu_k$ is the centroid of cluster $C_k$.

2. Plot WCSS for a range of $K$ values and look for the "elbow" point, where the marginal gain in reduced WCSS drops significantly.



## Single-cell RNA-sequencing

*Community detection methods*

- Graph-partitioning algorithms as implemented in SNN-Cliq[12] and in Seurat[13].
- Graph representation of the data obtained using a K-Nearest Neighbour.
- Nodes represent cells and edges indicating similar expression.
- Partitions the graphs into interconnected 'quasi-cliques' or 'communities'.
- Cluster stability is measured via
  - resampling methods (e.g. bootstrapping);
  - on the basis of cell similarities within assigned clusters (e.g., silhouette index).

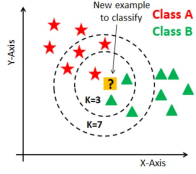
[12] Xu and Su 2015.
[13] Satija et al. 2015.

# Community detection methods

## *K-Nearest Neighbour (KNN)*

### Overview

- Similar cells obtained using Euclidean distances on the PC-reduced expression space
- Each cell is connected to its K most similar cells.
- K is commonly set to be between 5 and 100 nearest neighbours.
- Defining K can be a balancing act:
  - Lower values of K can have high variance, but low bias i.e. overfitting.
  - Higher values of K may lead to high bias and lower variance i.e. underfitting.
- The accuracy can be severely degraded by the presence of noisy or irrelevant features.



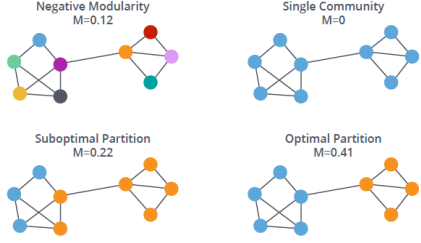
# Community detection methods

## *Louvain Clustering Algorithm*

### Goal

Detects communities as groups of cells that have more links between them than expected from the number of links the cells have in total.

- Works on a graph of cells, often constructed using nearest neighbors (kNN).
- Widely used after dimensionality reduction (e.g., PCA, UMAP).
- Maximize **modularity**
- Greedy optimization method with running in time $O(n \cdot \log n)$ where $n$ is the number of nodes.
- Default clustering method implemented in Seurat.
- Can lead to arbitrarily poorly connected communities.

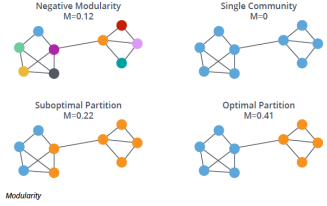


# Community detection methods

## *Louvain Clustering Algorithm*

### Modularity

- Modularity is a measure of the density of links inside communities compared to links between communities.
- Modularity is a scale value between -0.5 (non-modular clustering) and 1 (fully modular clustering).
- Optimizing this value theoretically results in the best possible grouping of the nodes of a given network.
- The optimized modularity function includes a resolution parameter, which allows the user to determine the scale of the cluster partition.



## Community detection methods

*Louvain Clustering Algorithm*

Equation of Modularity $Q$

$$Q = \frac{1}{2m} \sum_{i,j}^{N} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

- $A_{ij}$: weight of edge between nodes $i$ and $j$
- $k_i$: sum of the weights of the edges attached to nodes $i$
- $m$: the sum of all of the edge weights in the graph
- $N$: the total number of nodes in the graph;
- $\delta(c_i, c_j)$: Kronecker delta function; 1 if nodes $i$ and $j$ are in the same community, 0 otherwise

## Community detection methods

*Louvain Clustering Algorithm*

### Two-phase approach

1. **Modularity Optimization:**
   - Each node starts in its own community.
   - Iteratively move nodes to neighboring communities to increase modularity.
2. **Community Aggregation:**
   - Communities are contracted into "super-nodes."
   - Build a new graph with these super-nodes and repeat phase 1.

Repeat until modularity no longer increases.

## Community detection methods

*Louvain Clustering Algorithm*

### Why Louvain in Single-cell RNA-seq?

- Efficient for large cell graphs (millions of nodes).
- One of the fastest modularity-based algorithms .
- reveals a hierarchy of communities at different scales, which is useful for understanding the global functioning of a network.
- Captures non-convex, non-spherical clusters better than K-means.
- Compatible with nearest-neighbor graphs derived from PCA or UMAP.
- Implemented in popular tools: Seurat, Scanpy.

# Table of Content

# Single-cell RNA-sequencing

*Identification of cell clusters*

- Process of giving detected cell clusters a biological interpretation such as cell type.
- Clustered data are analysed by finding the gene signatures of each cluster.
- Manual or automatic approaches.
- A three-step approach is recommended[14]:
  1. automated annotation;.
  2. expert manual annotation;
  3. verification to obtain the ideal annotation result.
- Step 3 is especially relevant for data sets with high complexity or studies that involve rare cell subpopulations for which references might not be available

[14] Clarke et al. 2021.

# Single-cell RNA-sequencing

*Mapping cell clusters to cell identities*

## Step 1: Automated cell-type annotation

- classifier-based methods
  - train a supervised machine learning model using a reference single-cell dataset where cell types are known.
  - Once trained, the classifier predicts the cell types of new, unlabeled cells.
  - strongly affected by the classifier type;
  - strongly affected by the quality of the training data;
  - clustifyr[15], CellTypist[16]

[15] Fu et al. 2020.

[16] Domínguez Conde et al. 2022.

# Single-cell RNA-sequencing

*Mapping cell clusters to cell identities*

## Step 1: Automated cell-type annotation

- reference mapping-based methods
  - Instead of training a classifier, you directly match your query cells to a pre-annotated reference dataset.
  - perform label transfer on the resulting joint embedding;
  - references can be either individual samples of the data set or well-curated existing atlases;
  - the quality of the annotations depends on the quality of the reference data, the model and the suitability to the data set
  - scArches[17] Azimuth[18] Symphony[19]

[17] Lotfollahi et al. 2022.
[18] Hao et al. 2021.
[19] Kang et al. 2021.

# Single-cell RNA-sequencing

*Mapping cell clusters to cell identities*

## Step 2: Manual Annotation

- Leverages gene signatures of each cluster to annotate cell clusters.
- Gene signatures are commonly known as marker genes.
- Marker genes characterize the cluster.
- Marker genes can be found by applying differential expression testing between two groups: the cells in one cluster and all other cells in the dataset.
- Use t-tests or Wilcoxon rank-sum tests.
- The obtained markers are then compared with marker genes from well-annotated references to annotate cell clusters.

# Single-cell RNA-sequencing

*Mapping cell clusters to cell identities*

## Marker Genes

- Effective and useful marker genes have specific characteristics that are not shared by all DE genes![20]
- Good marker genes typically exhibit
  - a large difference in expression between cell types
  - are strongly up-regulated in a cell type of interest
  - exhibit high expression in that cell type
  - no or low expression in other cell types.

[20] Pullin and McCarthy 2024.

## Single-cell RNA-sequencing

*Mapping cell clusters to cell identities cont'd*

The most recent methods

- ScType[21]
- Cell BLAST[22]
- scGPT[23]
- scCATCH[24]

[21] Ianevski, Giri, and Aittokallio 2022.
[22] Cao et al. 2020.
[23] Cui et al. 2024.
[24] Shao et al. 2020.

## Single-cell RNA-sequencing

*Identification of marker genes*

- Most methods use some form of differential expression (DE) testing (Seurat, Scanpy, scran findMarkers(), presto, edgeR, limma).
- Other methods use ideas of feature selection (RankCorr), predictive performance (NSForest, SMaSH) or alternative statistics (Cepo, scran scoreMarkers(), Venice) to select marker genes.
- Among DE-based methods a variety of multiple-hypothesis correction methods are used.
- RankCorr and NSForest return only a specific set of genes determined to be marker genes.

## Single-cell RNA-sequencing

*Identification of marker genes cont'd*

- To account for the difficulty of selecting a fixed set of markers, follow convention and select only a fixed-size set of the top $n = 5, 10, 15, 20$ marker genes as ranked by the method.
- The best performing methods were those based on the Wilcoxon rank-sum test, Student's t-test or logistic regression.
- Methods that only selected a subset of genes (RankCorr and NSForest) had excellent specificity they did not show strong predictive performance.
- SMaSH and methods designed for bulk RNA-seq data were particularly memory intensive, while Seurat's methods were unexpectedly slow.
- Scanpy and Seurat have issues and inconsistencies with their implemented methods.

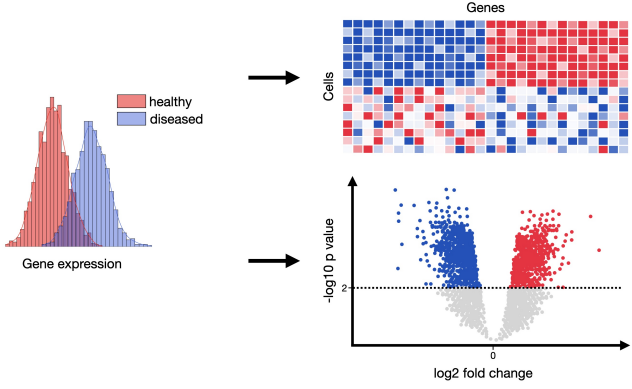
# Table of Content

# Differential gene expression analysis

*Noise model in bulk RNA-sequencing*

## Statistical question

Are the read counts for a given gene drawn from similar distribution between two samples?




https://www.sc-best-practices.org/conditions/differential_gene_expression.html

# Differential gene expression analysis

*Why noise matters so much*

## Effect size

- Most of the time, changes in gene expression are quantified in $log2FC$.
- A change should be of sufficient magnitude to be considered biologically significant.

## P-value

- Most approaches to testing for DGE test against the null hypothesis of zero log2FC.
- Probability of the difference in gene expression between two samples to be greater than or equal to the observed difference, given the null hypothesis that the observed read count originate from same distribution.
- Can be obtained empirically by shuffling samples or analytically using the form of the null distribution.

# Differential gene expression analysis

*Noise model in bulk RNA-sequencing*

Method

Open Access

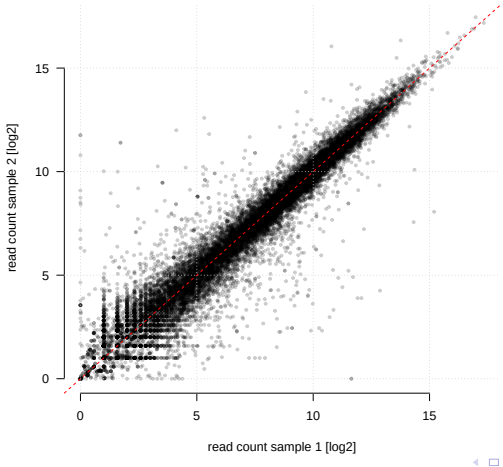
**Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data**

Piotr J Balwierz*, Piero Carninci†, Carsten O Daub†, Jun Kawai†, Yoshihide Hayashizaki†, Werner Van Belle‡, Christian Beisel‡ and Erik van Nimwegen*

Addresses: *Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland. †RIKEN Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho Tsurumi-ku Yokohama, Kanagawa, 230-0045 Japan. ‡Laboratory of Quantitative Genomics, Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zurich, Mattenstrasse 26, 4058 Basel, Switzerland.

# Differential gene expression analysis

*Noise model in bulk RNA-sequencing*



## Differential gene expression analysis

*Noise model in bulk RNA-sequencing*

### Technical noise

From library preparation to sequencing, induces systematic biases between samples that can have specific impact on different genes.

### Biological noise

Transcription is a stochastic process and variation naturally occurs within samples; dominant noise for strongly expressed genes.

# Differential gene expression analysis

*Noise model in bulk RNA-sequencing*

## Poisson Noise

Assuming that a particular gene has fraction $f$ in the read pool, this gene is expected to be sequenced $< n >= fN$ times with $N$ the total number of reads. The actual number of times $n$ that this genes is sequenced will be Poisson distributed according to:

$$P(n|f, N) = \frac{(fN)^n}{n!} e^{-Nf} \tag{5}$$

## Gaussian Noise

Additional noise in the log-count whose size is approximately independent of the total log-count. The noise that is introduced in going from the biological input sample to the final library that goes into the sequencer.

## Differential gene expression analysis

*Total Noise: Negative Binomial*

The total noise can be modelled as a convolution of multiplicative noise, specifically a Gaussian distribution of log-count with variance $\sigma^2$, and Poisson sampling. The probability to obtain $n$ reads for a gene is approximately:

$$P(n|\sigma, f, N) = \frac{exp(-\frac{(log(n/N)-log(f))^2}{2\sigma(n)^2})}{n\sqrt{(2\pi)}\sigma(n)} \tag{6}$$

where $\sigma^2(n) = \sigma^2 + \frac{1}{n}$, $f$ is the original concentration of mRNA for a given gene in the original pool, and $N$ total number of.

## Differential gene expression analysis

*Negative Binomial Distribution*

$$NB(n, p) = NB(\alpha, \frac{1}{1+\beta}) \tag{7}$$

mean$= \frac{\alpha}{\beta}$ and variance$= \frac{\alpha(1+\beta)}{\beta^2}$

$\alpha$ =dispersion and $\beta$ =parameter that best fits the data.

A dispersion value of 0.01 means that the gene's expression tend to different by typically $\sqrt{0}.01 = 10\%$ between samples of the same treatment group.

# Differential gene expression analysis

*Design Matrix*

```
   B.lactating B.pregnant B.virgin L.lactating L.pregnant L.virgin
1            0          0        1           0          0         0
2            0          0        1           0          0         0
3            0          1        0           0          0         0
4            0          1        0           0          0         0
5            1          0        0           0          0         0
6            1          0        0           0          0         0
7            0          0        0           0          0         1
8            0          0        0           0          0         1
9            0          0        0           0          1         0
10           0          0        0           0          1         0
11           0          0        0           1          0         0
12           0          0        0           1          0         0
attr(,"assign")
[1] 1 1 1 1 1 1
```

# Differential gene expression analysis

*Multi-test correction*

## Why is then an issue with P-values

- P-value is only statistically valid when a single score is computed.
- When 20,000 genes are tested, the chance to obtained small P-values are higher that predicted P-values.

## Bonferroni adjustment

- If a significance threshold of $\alpha$ is used, but $n$ separate tests (genes$\times$contrasts) are performed, then $\alpha \rightarrow \alpha/n$ or BH adjusted P-value(BH)=P$\times n$.
- With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.
- If error rate equals 0.05, expects 0.05 genes to be significant by chance.

## Differential gene expression analysis

*Multi-test correction cont'd*

### Benjamini and Hochberg False Discovery Rate

1. Sort the P-values in ascending order.
2. Divid each observed P-value by its percentile rank.
3. Tolerates more false positives. There will be also less false negative genes.
4. If error rate equals 0.05, 5% of genes considered statistically significant will be identified by chance (false positives).

## Single-cell RNA-sequencing

*Differential expression analysis*

- In general, agreement among the tools in calling DE genes is not high.
- Methods with higher true positive rates tend to show low precision.
- Methods with high precision show low true positive rates due to identifying few DE genes.
- Methods designed for scRNAseq data do not tend to show better performance compared to methods designed for bulk RNAseq data.
- Seurat and Scanpy use a "one-vs-rest" cluster comparison strategy.
  - creates a situation with highly imbalanced sample sizes and increased biological heterogeneity in the pooled "other" group;

# Single-cell RNA-sequencing

*Differential expression analysis cont'd*

- Model-based approaches or zero-inflated model
  - fit a probabilistic model that explicitly tries to capture how the data is generated, instead of just using simple statistical tests (like Wilcoxon rank-sum).
  - account for both the zeros and the nonzero counts
  - use two-part joint model and separately handles:
    - The first part models the probability of detection (e.g using logistic regression) i.e. whether the gene is detected at all.
    - The second part models the level of expression (often using a Gaussian or a log-normal model after transformation).
  - ▸ SCDE [25], ▸ MAST [26], ▸ scDD [27]

[25] Kharchenko, Silberstein, and Scadden 2014.
[26] Finak et al. 2015.
[27] Korthauer et al. 2016.

## Single-cell RNA-sequencing

*Differential expression analysis cont'd*

- Nonparametric methods
  - Do not model the distributions of gene expression values nor estimate their parameters.
  - Identify DE genes by employing a distance metric between the distributions of genes in two conditions.
  - SigEMD[28], EMDomics[29], and D3E[30].

[28] Wang and Nabavi 2018.
[29] Nabavi et al. 2016.
[30] Delmans and Hemberg 2016.

# Public repositories of high-content molecular data

## Sequence Read Archive (SRA)

- Public repository of high throughput sequencing data
- Raw sequencing data and alignment information Link

## Gene Expression Omnibus

- Public functional genomics data repository
- Array- and sequence-based raw data, processed data and metadata
- For high-throughput sequencing, GEO arrange the transfer of the raw data to SRA Link

## jPOSTrepo[31]

- Public repository of sharing MS raw/processed data Link

[31] Okuda et al. 2017.