PHAROS
THE GREEK AI FACTORY

Single-cell Transcriptomics Analysis

George Georgakilas · Charis Sinnis · Haris Manousaki
Information Management Systems Institute, Athena Research Center
Emails: ggeorgakilas@athenarc.gr , charalampos.sinnis@athenarc.gr , h.manoussaki@athenarc.gr

ATHENA
Research & Innovation
Information Technologies

Webinar path and schedule

11:00–11:05 Introduction to Single-Cell Transcriptomics
11:05–11:20 Pre-processing & HVGs Detection
11:20–11:55 Hands-on 1: Pre-processing & HVGs Detection
11:55–12:10 Dimensionality Reduction & Clustering
12:10–12:45 Hands-on 2: Dimensionality Reduction & Clustering
12:45–13:00 Cell Type Annotation
13:00–13:35 Hands-on 3: Cell Type Annotation
13:35–13:45 Closing remarks

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

What participants will be able to do

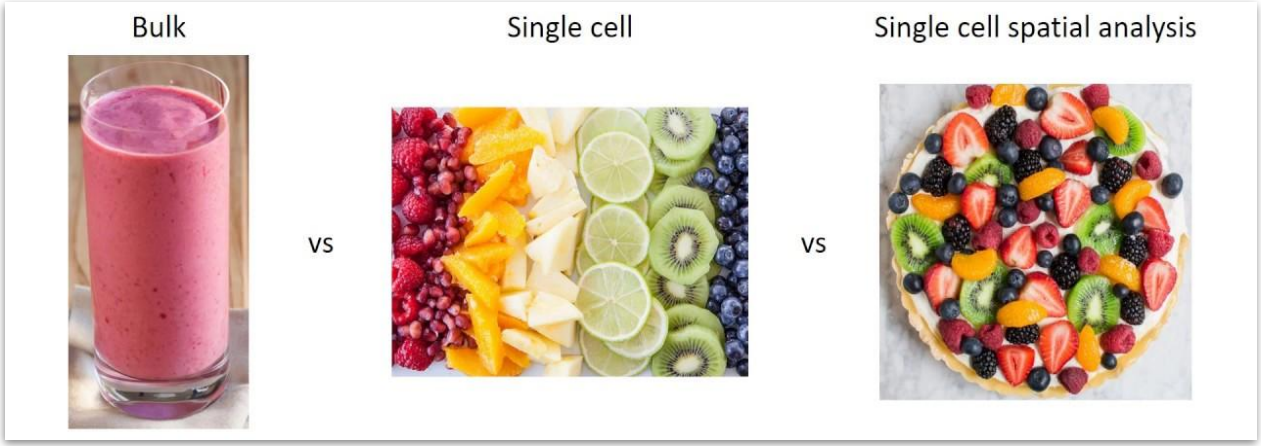
- Process, analyze, and explore single-cell RNA sequencing (scRNA-seq) datasets.
- Identify distinct cell populations and annotate cell types using marker genes and enrichment-based approaches.
- Independently perform, interpret, and critically evaluate single-cell transcriptomics analyses.

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

Introduction to Single-Cell Transcriptomics

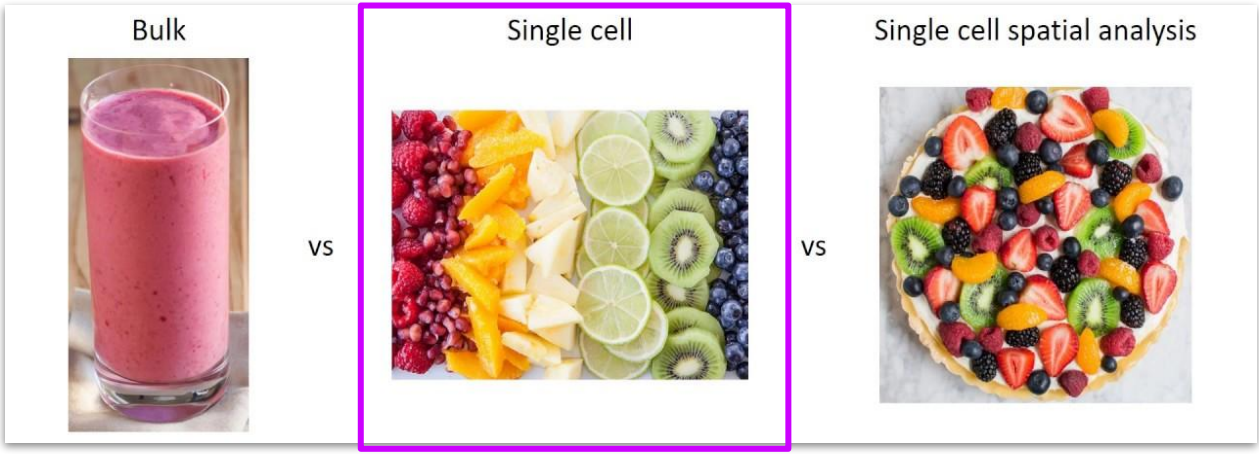
Overview of contemporary transcriptomics



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Overview of contemporary transcriptomics



- What do single-cell transcriptomics offer?
  - Gene expression profile of each individual cell!
  - Cell type heterogeneity and cell-to-cell variability information
- Important because:
  - The developmental processes and disease mechanisms
  - The discovery of rare cell types & transient states

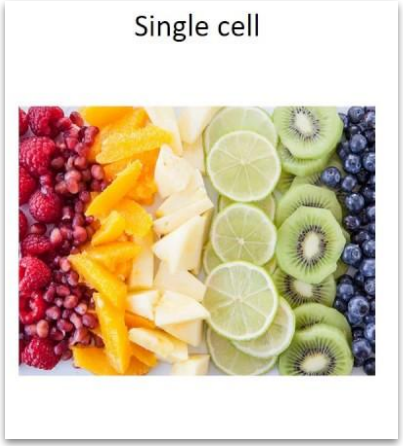
G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Properties of single-cell transcriptomics data

What are the characteristics of single-cell transcriptomics data?

- Count-based expression values as a count matrix (genes × cells)
- High-dimensional (thousands of genes measured per cell)
- Sparse (many zero expression values):
  - True zeros (gene not expressed)
  - Dropouts (technical detection failure, low sequencing depth)
- Noisy
  - Technical variability (library preparation, sequencing)
  - Biological variability (true cell variability)
- Data quality
  - Depends on technology and sequencing depth
  - Common issues:
    - Dead/damaged cells
    - Doublets (multiple cells captured as one)
- Overdispersion (variance depends on mean)

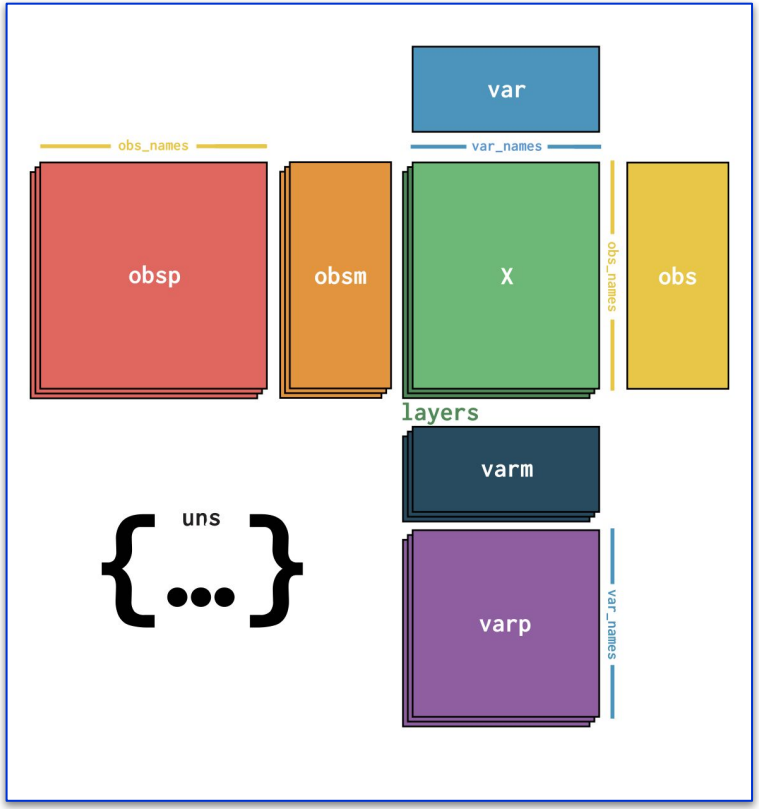


G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Libraries for analyzing single-cell transcriptomics data

- Scanpy (python) & Seurat (R)
- Standard end-to-end analytical workflow:
  - Preprocessing (filtering, normalization)
  - HVGs detection
  - Dimensionality reduction
  - Clustering
  - Differential gene expression analysis
  - Marker gene identification and annotation
- AnnData object for data storage and analysis
- AnnData structure:
  - .X ⟶ expression matrix (counts matrix)
  - .obs ⟶ cell metadata
  - .var ⟶ gene metadata
  - .obsm ⟶ Multidimensional annotation of observations
  - .uns ⟶ unstructured results
  - .layers ⟶ alternative data representations
- Other notable packages: scran, SingleCellExperiment



The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Preparatory work

Clone the GitHub repository before we begin!

https://github.com/GeorgakilasLab/Workshop_scRNAseq_Analysis

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

Pre-processing & HVGs detection

Pre-processing of scRNA-seq data

Preprocessing is essential in **single-cell RNA sequencing (scRNA-seq)** because the raw data contain:

- substantial technical noise
- artifacts
- biases

that can obscure the true biological signal.

Observed expression=Biological signal+Technical noise



Preprocessing aims to **maximize** the biological component while **minimizing** technical artifacts

Main pre-processing actions:

1. Remove low-quality genes, cells and artifacts
2. Correct for sequencing depth differences
3. Reduce technical noise and dropout effects
4. Remove batch effects
5. Select informative genes

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Sparsity coefficient

Probe ~ 3700 pbmc cells from dataset provided in:

Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. Published 2017 Jan 16. doi:10.1038/ncomms14049

	CGTGATGATTTCAC_1_b_cells	TTTGACTGTTGAGC_1_cd34_filtered	GACAGTACCTGTGA_1_memory_t	CAGACCCTATTCCT_1_b_cells	CAGCTAGATGACTG_1_cd4_t_helper
ENSG00000237683	0	0	0	0	0
ENSG00000230021	0	0	0	0	0
ENSG00000228327	0	0	0	0	0
ENSG00000237491	0	0	0	0	0
ENSG00000225880	0	0	0	0	0
ENSG00000230368	0	0	0	0	0
ENSG00000187634	0	0	0	0	0
ENSG00000188976	0	0	0	0	0
ENSG00000187961	0	0	0	0	0
ENSG00000187583	0	0	0	0	0
ENSG00000188290	0	0	0	0	0
ENSG00000187608	0	3	0	0	0
ENSG00000188157	0	0	0	0	0
ENSG00000273443	0	0	0	0	0
ENSG00000131591	0	0	0	0	0

15 rows × 3774 columns

scRNA-seq data file: pbmc.bead.enriched.sample.zheng

- # genes: 16791
- # cells: 3774

Sparsity coefficient: 96.0%

$$S = \frac{\#\ zero-valued\ elements}{\#\ total\ elements}$$

Pre-processing & HVGs detection

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

QC criteria

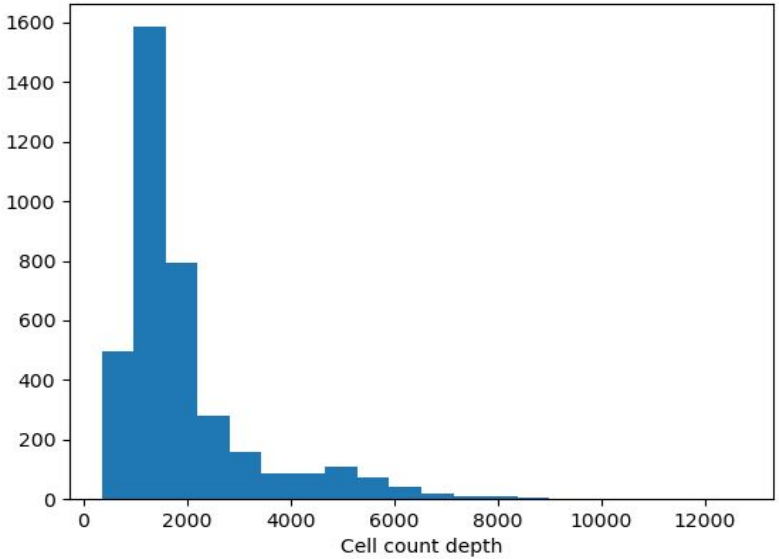
Quality control and removal of low quality genes/cells & artifacts:

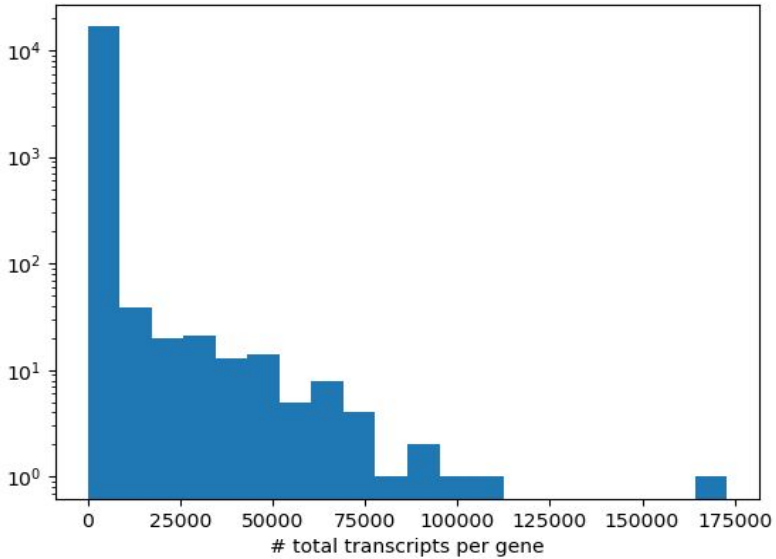
1. Check cell count depth & gene expression levels across cells
   a. Remove **cells** with **zero count depth**
   b. Remove **genes** with **zero total expression across cells**
2. Filter out genes that are rich in:
   a. **Mitochondrial** RNA
   b. **Ribosomal** RNA
   c. **Apoptosis**-related RNA
3. **Genes** that are expressed in a **tiny fraction** of the cell population
4. Cell artifacts:
   a. **Empty droplets** containing ambient RNA
   b. **Doublets/multiplets** where two cells were captured together

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Counts distributions





Number of genes with zero total expression: 0
Number of only non-expressed genes cells: 0

Notice that:

- Small cell count depth
- Majority of genes has small total transcripts count

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

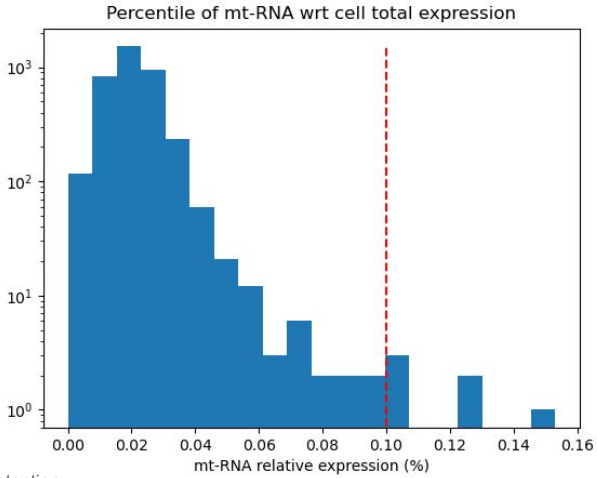
MT count distribution

Mitochondrial RNA (**mt-RNA**) is RNA transcribed directly from the mitochondrial genome rather than the cell nucleus.

In single-cell (scRNA-seq) , we generally do not want a high fraction of mt-RNA over total transcripts because it typically indicates:

- Cell death and stress
- Technical artifacts
- Immune activation (unwanted inflammatory responses during cell stress)

The metric is the fraction of mt-RNA transcripts count over the total transcript counts per cell. The threshold are usually user defined and typical range is [0.05,0.2]



mt-RNA threshold: 10%

Number of cells with excess of mt-RNA: 6

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

rRNA & MT filtering

Ribosomal RNA (**rRNA**) is a specialized, non-coding type of RNA that forms the core structural and catalytic foundation of ribosomes — the cell's protein-building factories.

In single-cell (scRNA-seq) , we generally do not want a high fraction of mt-RNA over total transcripts because it typically indicates:

- Technically poor-quality libraries
- Damaged or dying cells
- Empty/ambient-contaminated droplets
- Low-information transcriptomes

Number of detected rRNA genes in data: 0

Usually, rRNA and mt-RNA filterings are jointly applied.
As a rule of thumb, we aim for simultaneously high rRNA and mt-RNA ratios while the count depth stays relatively small.
A good practice is to look for extreme outliers.

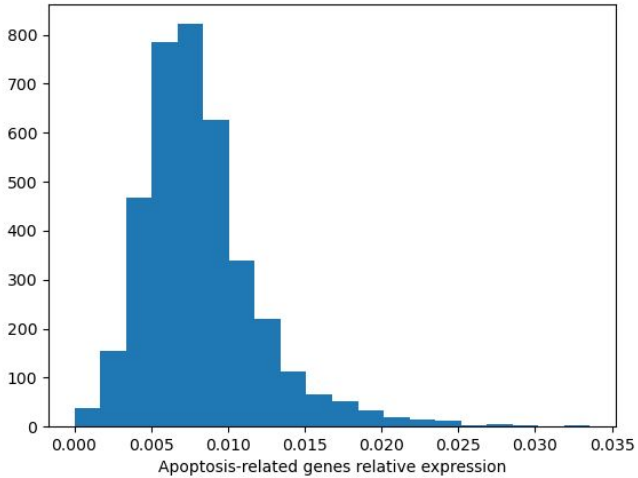
G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Apoptotic counts distribution

In single-cell (scRNA-seq) , we also do not want a high fraction of apoptosis-related genes RNA over total transcripts because it typically indicates:

- Dying or stressed cells
- Technically compromised transcriptomes
- Artifacts introduced during tissue dissociation or handling



Relative expression of apoptosis-related genes under 3.5%.

No cells are filtered based on even under conservative choice of threshold.

Number of cells with excess of apoptosis-related genes: 0

Pre-processing & HVGs detection

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Doublets/Droplets

Filter out:

- Doublets
- Empty droplets

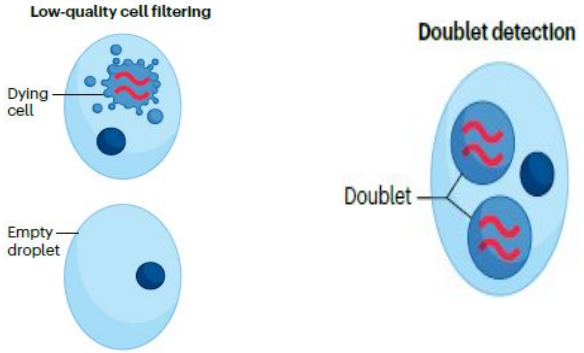
We are going to use 2 different methods:

- Median absolute deviation (MAD) filtering
- Scrublet

The **Median Absolute Deviation** (**MAD**) is a robust, outlier-resistant statistical measure of data variability or spread.

1. Find the median: Identify the central value of your dataset.
2. Calculate absolute deviations: Subtract the median from each data point and take the absolute value.
3. Find the median of deviations: Sort these absolute differences and find their median.

$$MAD = median(|X_i - median(X)|)$$



Scrublet works by:

1. Taking the observed single-cell expression matrix.
2. Simulating artificial doublets by combining expression profiles from random pairs of cells.
3. Embedding both observed cells and simulated doublets into a lower-dimensional manifold (typically PCA + k-nearest neighbors).
4. Assigning each real cell a **doublet score** based on similarity to simulated doublets.
5. Predicting likely doublets using an automatically or manually selected threshold.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

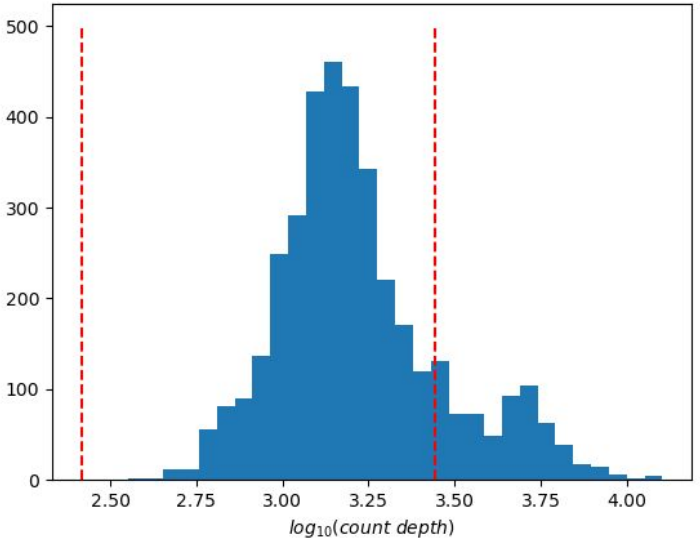
MAD based filtering

The outlier detection using the MAD formulation can be summarised in the equation:

$$X^{\pm}_{MAD,threshold} = median(X) \pm k \cdot MAD$$

k=3 (typical choice)
Lower MAD threshold: 260.0
Upper MAD threshold: 2765.0

Cells above the upper MAD threshold: 637 (!!!!!)
(or ~ 17% of the cell population are doublets !!!!)



The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

MAD based filtering

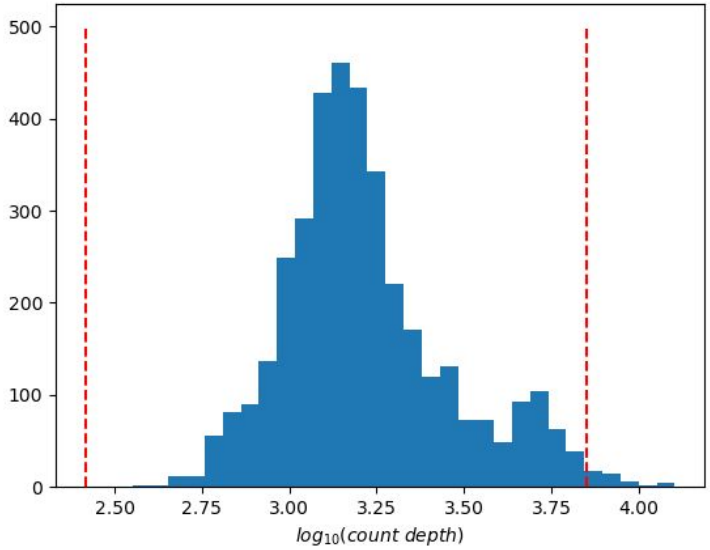
It seems that there is a mixture of cell populations. This could imply that the MAD methodology focused mainly on the most-populated group, thus treating the less-populated counterpart as an outlier group.

Fixing attempt: Use only the higher count depth group for possible doublets detection.

New upper MAD threshold: 7124.0

Cells with count depth above upper MAD threshold: 42

Always check and validate your results!! Strange/non-expected results may indicate to points of interest in your data!!!



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Scrublet based filtering

Scrublet requires the raw count matrix and a random seed number for the initialization of the initial state for doublet simulation and nearest neighbors. We choose two number k = 0,42.

k=42

TATACCACTTTGCT_1_cd34_filtered 6648
TTAGCTACAGTTCG_1_cd56_nk 1342
ATCATCTGGACGAG_1_cd34_filtered 2881
TACGTACTAGCACT_1_cd56_nk 1343
TGACTTACCTTATC_1_cd34_filtered 7567
CGCCGAGAACGGAG_1_cytotoxic_t 951
TAAGTAACCCAGTA_1_cd34_filtered 2160
AAGGTCTGCATGAC_1_cytotoxic_t 2594
TGAAGCTGCAATCG_1_cytotoxic_t 1152

k=0

CGCCGAGAACGGAG_1_cytotoxic_t 951
TGAAGCTGCAATCG_1_cytotoxic_t 1152

Some thoughts on the Scrublet results:

- Random seed should not affect the doublet detection process.
- The count depths of the doublets does not necessarily correspond to high count depths.
- Maybe the population mixture also affect Scrublet.
- We have not explored the effect on the doublet detection process based on the prior decisions we have made thus far.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

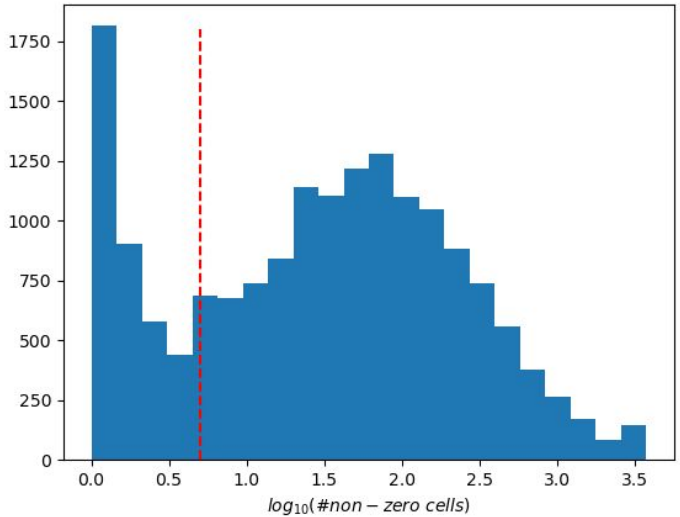
The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Filtering low count genes

A gene detected in only a tiny number of cells is often:

- **Technical noise/dropout** rather than true biology
- Too sparse to provide reliable statistical information
- Computationally expensive to keep without adding much value

We choose a threshold with value 5.



Number of genes to be filtered out: 4105

Important caveat: rare genes can be biologically important

Over-filtering may lead to loss of information for biological processes that include rare genes

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Counts normalization

Finally, after all these steps we have reduced the dimensions of the dataset:

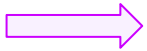
·# genes: 16791
·# cells: 3774



·# genes (filtered): 12559
·# cells (filtered): 3726

Final step: Normalization and variance stabilization

- Normalization: Normalization rescales counts so cells become comparable
  - Counts per Million (Library-size normalization)
  - Counts per 10000 (Library-size normalization)
  - RPKM (Library-size and gene length normalization)
- Variance-stabilization: Even after normalization, scRNA-seq data still have problematic statistical behavior, variance is dependent on the genes mean expression. Genes with higher average expression naturally have larger variance.
  - $x' = \log(1+x)$
  - scTransform (Pearson residuals on NB-fitting)



We opt for the log(1+CPM) option

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Detecting HVGs

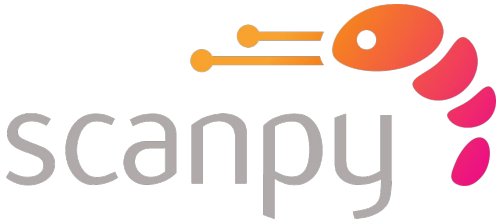
Highly variable genes (**HVGs**) are used in scRNA-seq pipelines because most genes are either:

- not informative for distinguishing cell states/types
- dominated by technical noise
- or nearly constant across cells

HVG selection tries to retain genes whose variation across cells is likely to reflect **biological structure** rather than random noise.

Due to the mean-variance relation, HVGs selection methodologies try to model and alleviate this bias.

For this workshop we focus on the methods 'seurat' & 'seurat_v3' of the scanpy package.



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Detecting HVGs - Frequently used methods

Seurat MeanVarPlot / Dispersion method

- **Implemented in:**
  - Seurat
  - Scanpy (flavor="seurat")
- **Main mechanism:**
  - Compute:
    - mean expression
    - variance / dispersion
  - Bin genes by mean expression
  - Normalize dispersion within each bin
  - Select genes with unusually high dispersion relative to genes of similar abundance
- **Key idea:**
  - Corrects for the fact that highly expressed genes naturally have larger variance

Seurat v3 VST

- **Implemented in:**
  - Seurat
  - Scanpy (flavor="seurat_v3")
- **Main mechanism:**
  - Fit mean–variance trend using regularized regression (LOESS)
  - Standardize gene variance relative to expectation
  - Rank genes by normalized variance
- **Key idea:**
  - Identify genes with variance exceeding technical expectations

Visualization of the conceptual criterion:

$$\text{Normalized Variance} = \frac{\text{Observed Variance}}{\text{Expected Variance}}$$

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

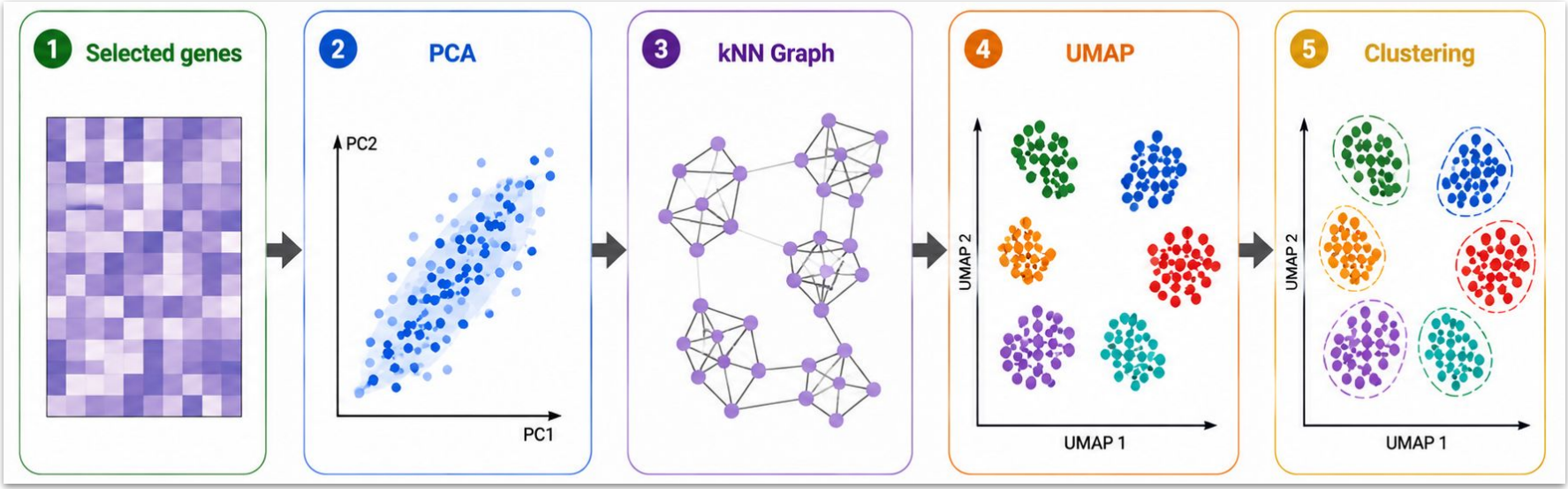
Hands-on 1

Pre-processing & HVGs Detection

PHAROS
THE GREEK AI FACTORY

Dimensionality Reduction & Clustering

Overview



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Linear Dimensionality Reduction

Problem:

~2000 HVGs = ~2000 dimensions ⟶ impossible to visualize directly

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Linear Dimensionality Reduction

Problem:
~2000 HVGs = ~2000 dimensions ⟶ impossible to visualize directly

Solution ⟶ Linear DR
Principal Component Analysis (PCA) projects the data into new orthogonal components (principal components) that capture the main sources of variation between cells.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Linear Dimensionality Reduction

Problem:
~2000 HVGs = ~2000 dimensions ⟶ impossible to visualize directly

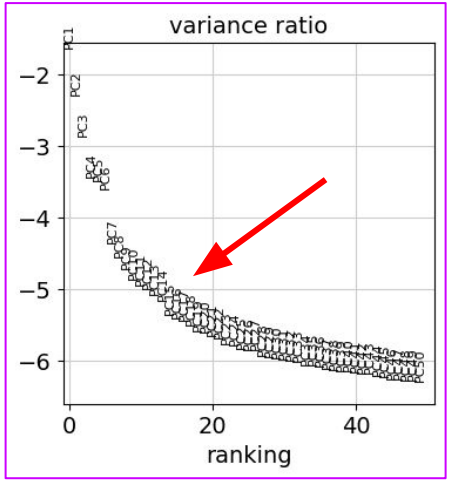
Solution ⟶ Linear DR
Principal Component Analysis (PCA) projects the data into new orthogonal components (principal components) that capture the main sources of variation between cells.

Advantages

- Preserves Euclidean distance between cells well.
- Reduces computational cost/complexity for downstream analysis.

Disadvantages

- Assumes linear relationships.
- Sensitive to sparsity, dropouts, and technical noise.



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction

Ok so PCA 2000 genes → 20-50 PCs….
However, this space is still too high-dimensional to visualize directly in 2D or 3D →Non-Linear DR!

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction

Ok so PCA 2000 genes → 20-50 PCs….
However, this space is still too high-dimensional to visualize directly in 2D or 3D →Non-Linear DR!

Goal of nonlinear dimensionality reduction?
Create a low-dimensional representation that preserves biologically meaningful relationships between cells.

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

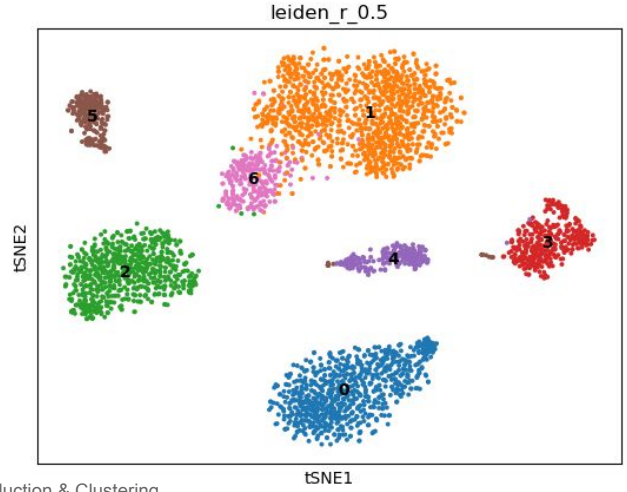
Non-Linear Dimensionality Reduction

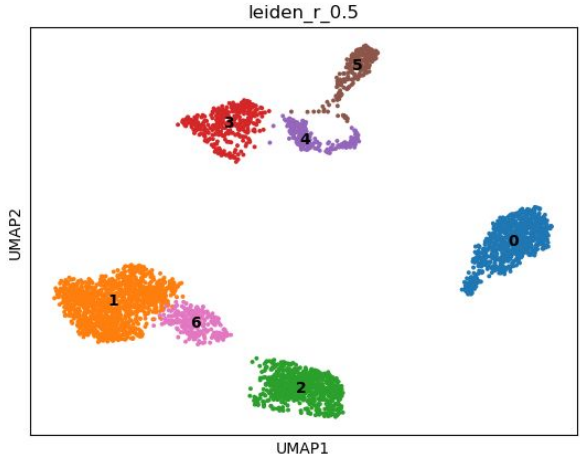
Ok so PCA 2000 genes ⟶ 20-50 PCs….
However, this space is still too high-dimensional to visualize directly in 2D or 3D ⟶Non-Linear DR!

Goal of nonlinear dimensionality reduction?
Create a low-dimensional representation that preserves biologically meaningful relationships between cells.

Common methods:

- t-distributed stochastic neighbor embedding (t-SNE) - *focuses on local similarity*
- Uniform manifold approximation and projection (UMAP) - *preserves global topology*



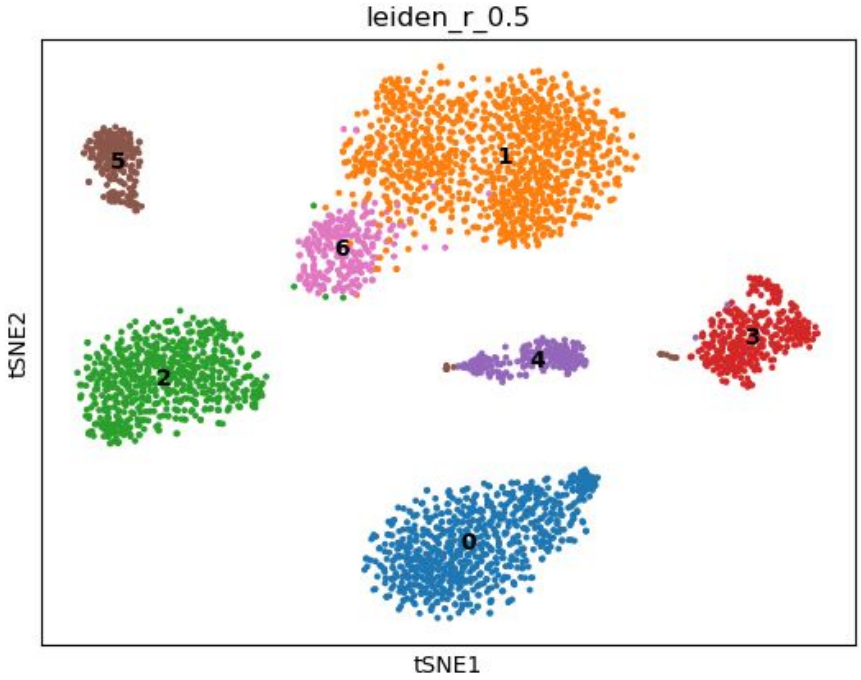


G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction/*t-SNE*

t-SNE preserves local relationships between cells by placing transcriptionally similar cells close together in 2D space.



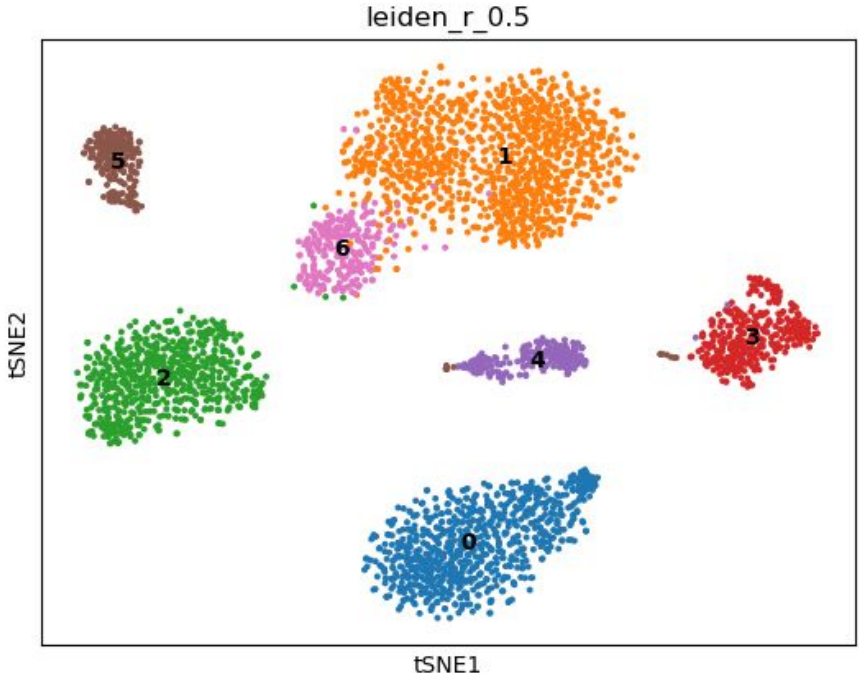
G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction/*t-SNE*

t-SNE preserves local relationships between cells by placing transcriptionally similar cells close together in 2D space.

- Computes cell-to-cell similarities in high dimensions and reconstructs these similarities in low dimensions.
- Preserves local cellular neighborhoods well, but distances between clusters might not mean anything.
- Stochastic method: random initialization can produce slightly different embeddings across runs.



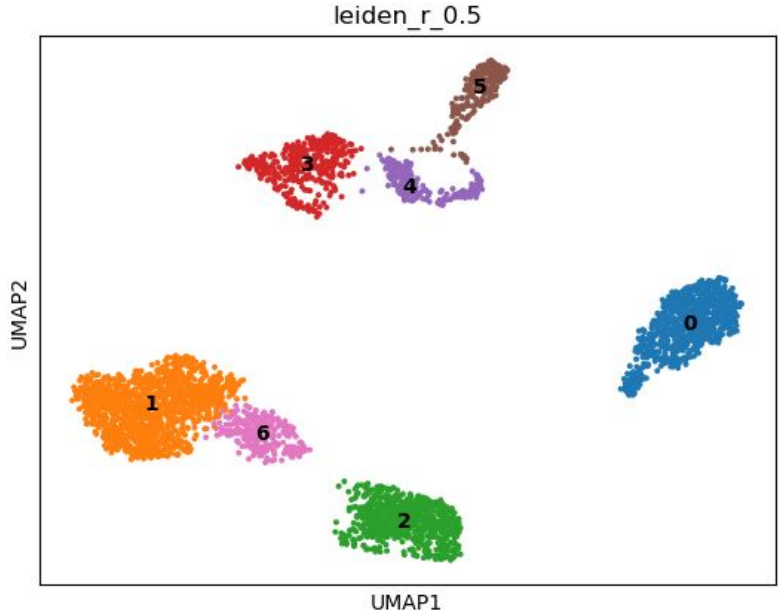
G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction/*UMAP*

UMAP on the other hand preserves both local and global structure why?

- UMAP assumes that cells lie on a low-dimensional manifold embedded in high-dimensional gene-expression space.
- Constructs a fuzzy graph of neighboring cells.
- Learns the manifold structure of the data.
- Projects cells into low dimensions while preserving data topology.



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Non-Linear Dimensionality Reduction/UMAP

UMAP on the other hand preserved both local and global structure why?

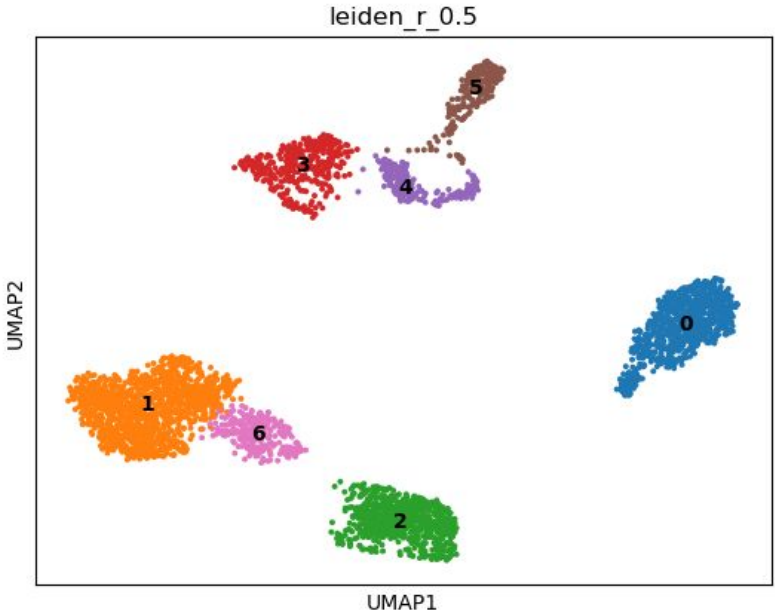
- UMAP assumes that cells lie on a low-dimensional manifold embedded in high-dimensional gene-expression space.
- Constructs a fuzzy graph of neighboring cells.
- Learns the manifold structure of the data.
- Projects cells into low dimensions while preserving data topology.

Why is UMAP preferred?

- Better preservation of global relationships
- Faster and more stable than t-SNE

Important parameters

- n_neighbors
  - small ⟶ emphasizes local structure
  - large ⟶ emphasizes global structure
- min_dist
  - small ⟶ tighter, more separated clusters
  - large ⟶ smoother organization of the manifold



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

What is the purpose of all this?

So far, we have:

- reduced the dimensionality of the data
- reduced computational complexity
- we learned how to visualize our data

However, our main biological goal is to identify distinct cell populations…how?

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

What is the purpose of all this?

So far, we have:

- reduced the dimensionality of the data
- reduced computational complexity
- we learned how to visualize our data

However, our main biological goal is to identify distinct cell populations…how?

While some genes are housekeeping genes expressed across many cells, other genes determine cellular identity and function.
For example, different expression programs distinguish T cells, B cells → Thus, each cell type has a characteristic transcriptomic profile.

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

What is the purpose of all this?

So far, we have:

- reduced the dimensionality of the data
- reduced computational complexity
- we learned how to visualize our data

However, our main biological goal is to identify distinct cell populations…how?

While some genes are housekeeping genes expressed across many cells, other genes determine cellular identity and function.For example, different expression programs distinguish T cells, B cells → *Thus, each cell type has a characteristic transcriptomic profile.*

Cells with similar gene expression profiles represent similar cell types or cellular states.
Therefore, similar cells should cluster together in transcriptomic space!
➡ Clustering identifies and organizes cells into groups that share similar gene expression profiles.



Note: Clusters are computationally defined and do not automatically correspond to true biological cell types!

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Clustering

Ok but how can we cluster our cells based on their transcriptomic profiles?

This is usually done using distance metrics in the PCA-reduced expression space.
A common approach in single-cell analysis is the use of community detection methods.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Clustering

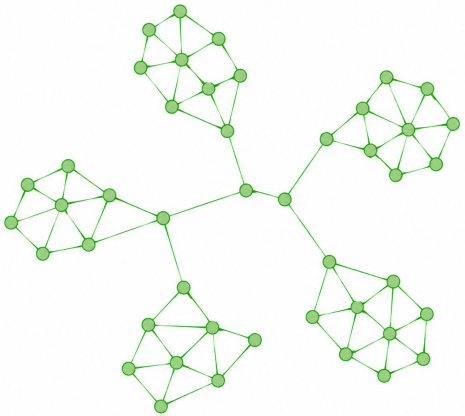
Ok but how can we cluster our cells based on their transcriptomic profiles?

This is usually done using distance metrics in the PCA-reduced expression space.
A common approach in single-cell analysis is the use of community detection methods.

These methods are graph-partitioning algorithms that rely on a graph representation of the data using a K-Nearest Neighbor (KNN) graph.

KNN graph
Each cell is connected to its k nearest neighbors (k is user-defined)
& cells with similar transcriptomic profiles form dense regions of connections

Thus:
dense regions → highly connected cells → biologically similar populations
Community detection algorithms identify these dense regions as clusters.



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Clustering

Louvain and Leiden are widely used community detection methods in single-cell analysis.
Both methods include a resolution parameter, which controls clustering granularity.

Higher resolution → more, smaller clusters
Lower resolution → fewer, broader clusters

The resolution is often adjusted based on:

- expected number of cell types
- the output of the annotation step.

Louvain can sometimes produce poorly connected communities. For this reason, Leiden is often preferred in recent single-cell workflows.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

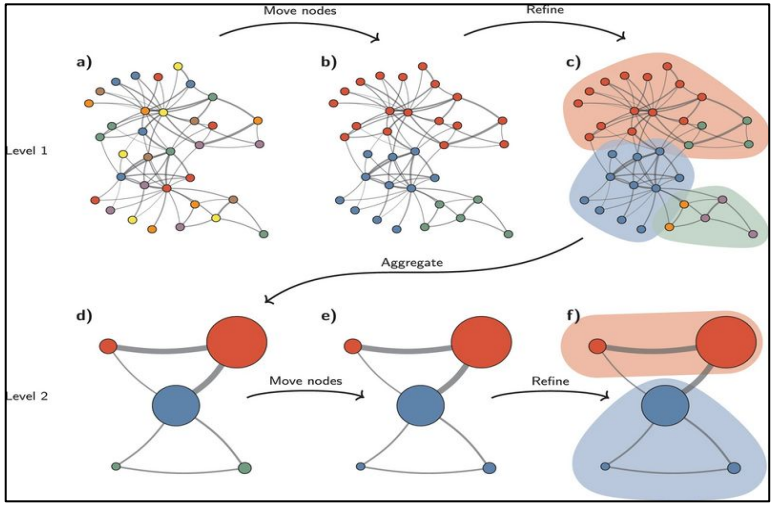
The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Leiden algorithm

The **Leiden algorithm** is a community detection algorithm developed by Traag *et al.* at Leiden University.

Leiden identifies groups of highly connected cells by optimizing modularity within the KNN graph.

A. Initializes communities in the graph
B. Moves cells between neighboring communities to maximize modularity
C. Refines the partition to improve community connectivity
D. Aggregates communities into larger nodes
E. Repeats the process iteratively until stable clusters are obtained



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Clustering Evaluation

Internal Metrics

Metrics that evaluate clustering structure **without** known labels.

- Silhouette Coefficient (SC) ⟶ Is each point closer to its own cluster than to others?
- Davies–Bouldin Index (DB) ⟶ measures the average similarity between each cluster and its most similar cluster.
- Calinski–Harabasz Index (CH) ⟶ measures the ratio between the between-cluster separation and the within-cluster dispersion.

External Metrics

Metrics that compare clustering results **to known/reference labels**.

- Adjusted Rand Index (ARI) ⟶ measures agreement between predicted clusters and known labels while correcting for random agreement.
- Normalized Mutual Information (NMI) ⟶ measures how much information is shared between clustering assignments and reference labels.
- Jaccard ⟶ measures the overlap between the predicted and true sets of clustered pairs.



Note: Good visualization with UMAP does not guarantee meaningful clustering!

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

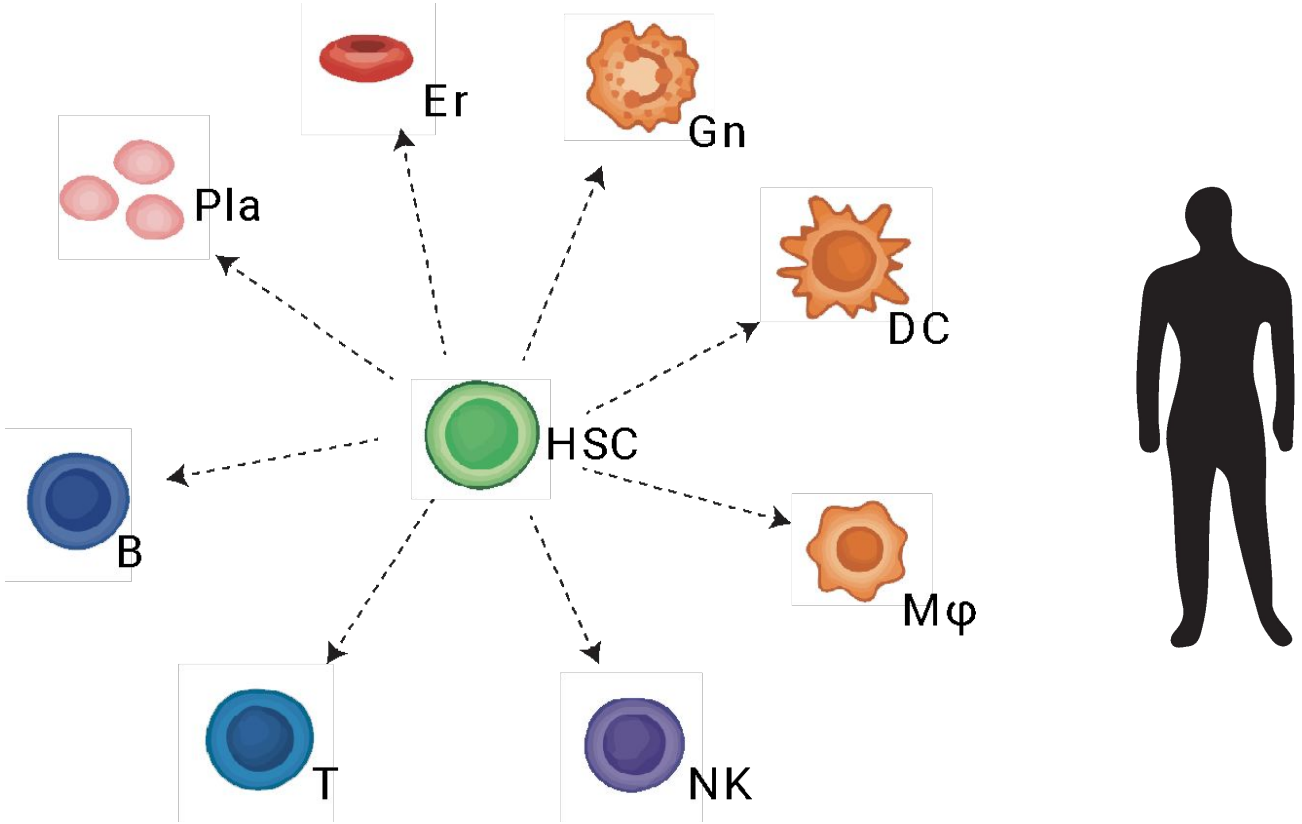
Hands-on 2

Dimensionality Reduction & Clustering

PHAROS
THE GREEK AI FACTORY

Cell Type Annotation

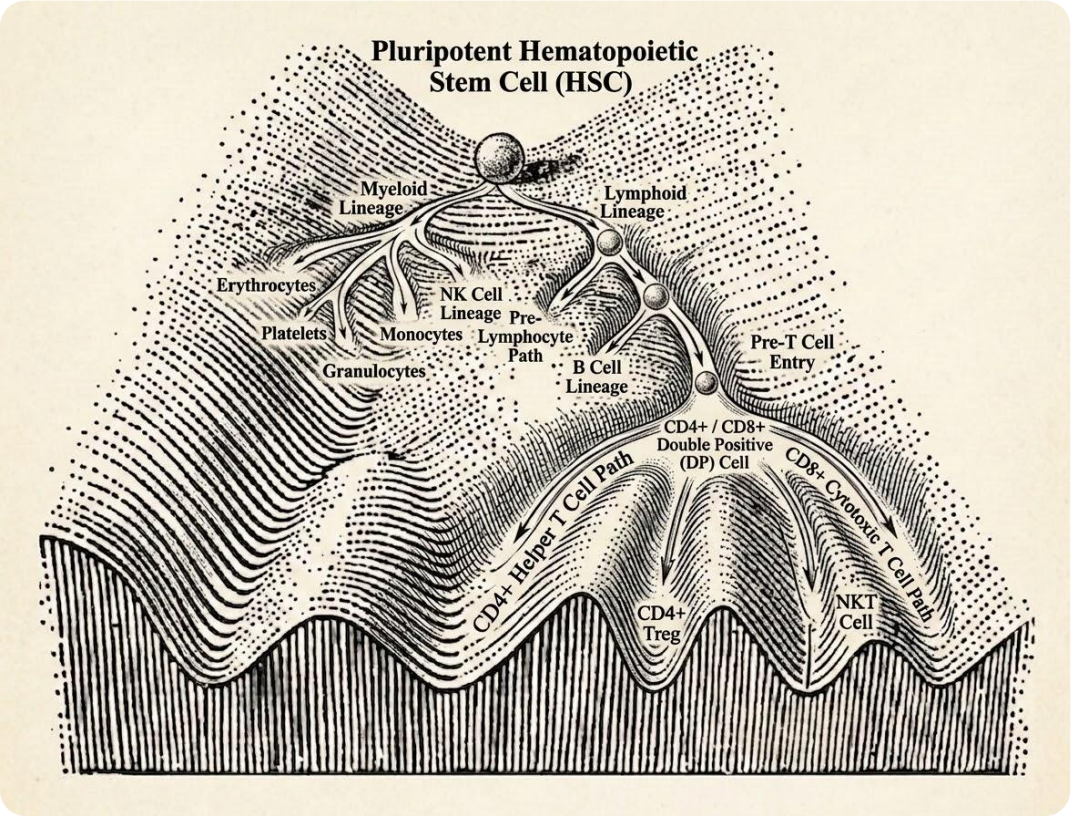
From a common ancestor to different phenotype and function



Lara-Astiaso, et al., Science, (2014)

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Lineages depend on the selected trajectories within differentiation

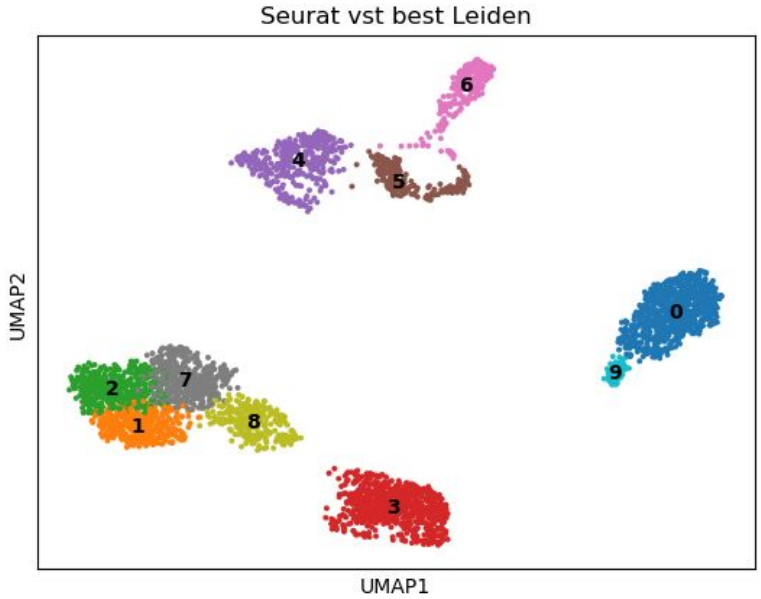


- Although (nearly) all cells in an organism contain the same genome, **different subsets of genes are activated or repressed in different cellular contexts**.
- At the core of differentiation lies the regulation of gene expression.
- Cell populations at each stage **share similar gene expression profiles**.
- There are intrinsic and extrinsic mechanisms that drive differentiation.
- Single-cell transcriptomics offer a unique opportunity to study such mechanisms at the highest resolution possible.

Cell Type Annotation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Well! We ended up with cell clusters. Now what?



- We ended up with some reasonably well separation of cells.
- Cells that are grouped together exhibit similar gene expression profiles.
- It is reasonable to assume that such cells possess similar morphological and functional characteristics.
- Thus, such cells will belong in the same cell type.

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters

The process of assigning biological meaning to detected cell clusters!

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters

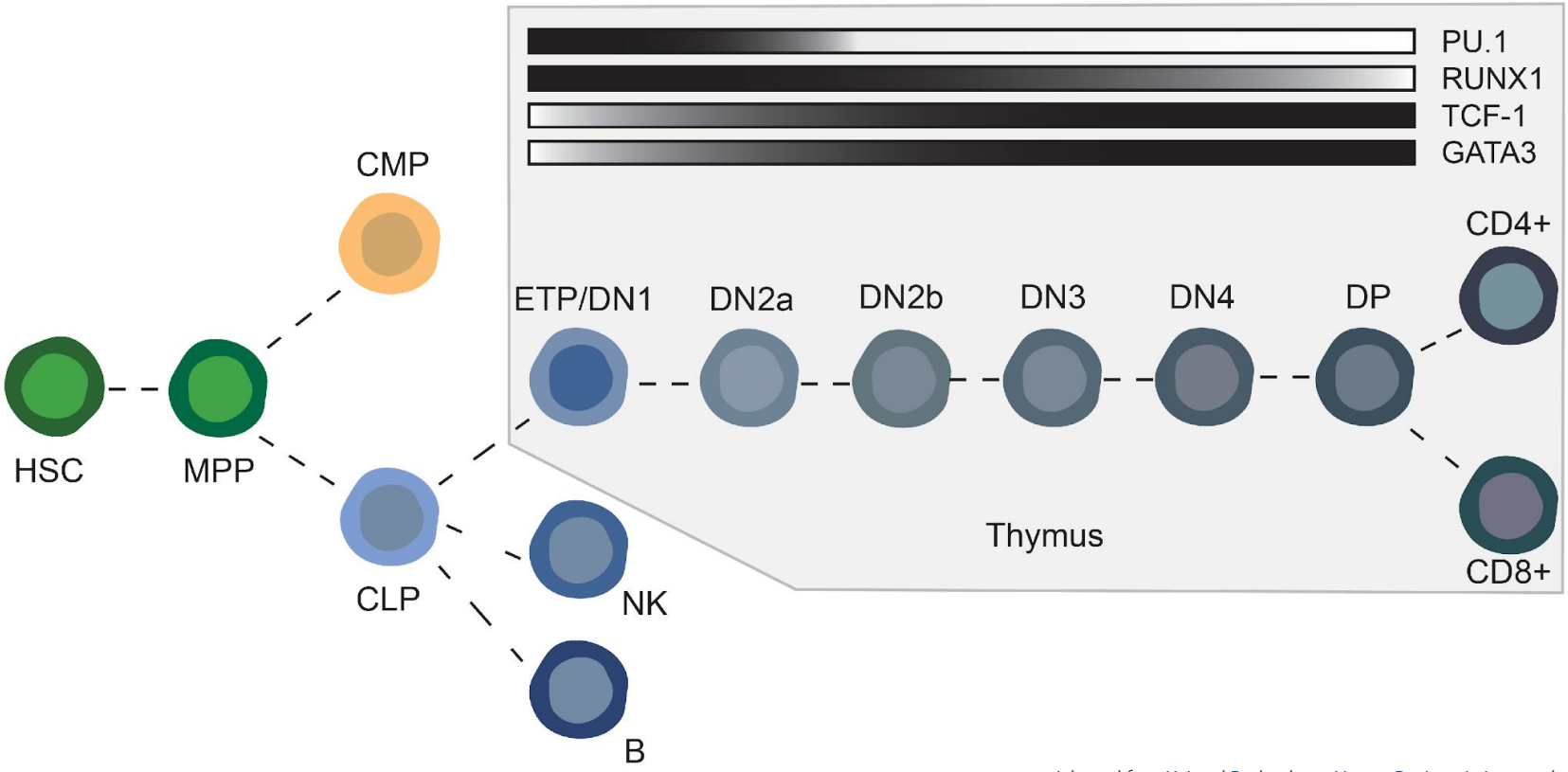
The process of assigning biological meaning to detected cell clusters!

How do we proceed doing this? Any thoughts?

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Certain genes exhibit a gradient in their expression profile



Adapted from Yui and Rothenberg, Nature Reviews in Immunology, (2014)

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters - Manual annotation

The process of assigning biological meaning to detected cell clusters!

Manual annotation:

- Leverage gene signatures (commonly known as marker genes)
- Such genes are reported in the literature
- *E.g.*, CD_4 (CD_4+ T cells), CD8 (CD8+ T cells) *etc*
- Manually inspect signature gene expression across clusters

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

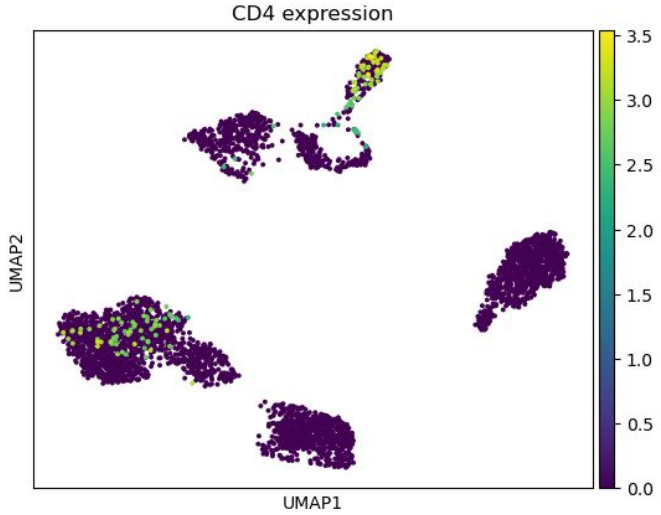
The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters - Manual annotation

The process of assigning biological meaning to detected cell clusters!

Manual annotation:

- Leverage gene signatures (commonly known as marker genes)
- Such genes are reported in the literature
- *E.g.*, CD4 (CD4+ T cells), CD8 (CD8+ T cells) *etc*
- Manually inspect signature gene expression across clusters



G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

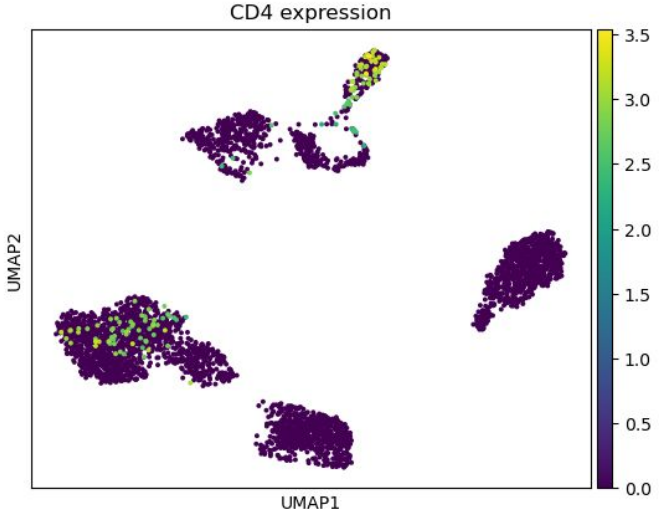
The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters - Manual annotation

The process of assigning biological meaning to detected cell clusters!

Manual annotation:

- Leverage gene signatures (commonly known as marker genes)
- Such genes are reported in the literature
- *E.g.*, CD4 (CD4+ T cells), CD8 (CD8+ T cells) *etc*
- Manually inspect signature gene expression across clusters



Cons:

- Manual process
- Does not provide a holistic view

Pros:

- Can sometimes help identify rare subpopulations

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters - Manual annotation with DEGs

The process of assigning biological meaning to detected cell clusters!

Manual annotation with DEGs:

- Leverage gene signatures lists per cell type
- Such lists are created by genes reported in the literature
- *E.g.*, [GNLY, NKG7, …] for NK cells *etc*
- Apply DEG analysis on clusters (*e.g.*, DESeq2, Wilcoxon *etc*)
- Identify gene signatures per cluster
- Compare with the literature lists

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Cluster A	Others
5	0
7	1
4	0
6	2
8	1

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Cluster A	Others		Assign ranks
5	0		0 -> 1.5
7	1		0 -> 1.5
4	0	➡	1 -> 3.5
6	2		1 -> 3.5
8	1		2 -> 5
			4 -> 6
			5 -> 7
			6 -> 8
			7 -> 9
			8 -> 10

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Cluster A	Others	Assign ranks	Sum ranks
5	0	0 -> 1.5	$R_A = 40$
7	1	0 -> 1.5	
4	0	1 -> 3.5	$R_O = 15$
6	2	1 -> 3.5	
8	1	2 -> 5	
		4 -> 6	
		5 -> 7	
		6 -> 8	
		7 -> 9	
		8 -> 10	

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Cluster A	Others	Assign ranks	Sum ranks	Mann-Whitney statistic
5	0	0 -> 1.5	$R_A = 40$	$U_A = 25$
7	1	0 -> 1.5		
4	0	1 -> 3.5	$R_O = 15$	$U_O = 0$
6	2	1 -> 3.5		
8	1	2 -> 5		
		4 -> 6		
		5 -> 7		
		6 -> 8		
		7 -> 9		
		8 -> 10		

$$U_A = R_A - \frac{n_A(n_A+1)}{2}$$

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Cluster A	Others	Assign ranks	Sum ranks	Mann-Whitney statistic	Null hypothesis
5	0	0 -> 1.5	$R_A = 40$	$U_A = 25$	H_0: Both groups come from the same distribution
7	1	0 -> 1.5			• Labels are arbitrary
4	0	1 -> 3.5	$R_O = 15$	$U_O = 0$	• Observed separation can happen by chance
6	2	1 -> 3.5		$U_A = R_A - \frac{n_A(n_A+1)}{2}$	• Any assignment of 5 observations to A is equally likely
8	1	2 -> 5			
		4 -> 6			
		5 -> 7			
		6 -> 8			
		7 -> 9			
		8 -> 10			

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Null hypothesis

H_0: Both groups come from the same distribution

- Labels are arbitrary
- Observed separation can happen by chance
- Any assignment of 5 observations to A is equally likely

Under null

Any 5 observation could belong to A

$$Total = \binom{10}{5} = 252$$

Each assignment produces a different U statistic

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Null hypothesis

H_0: Both groups come from the same distribution

- Labels are arbitrary
- Observed separation can happen by chance
- Any assignment of 5 observations to A is equally likely

Under null

Any 5 observation could belong to A

$$Total = \binom{10}{5} = 252$$

Each assignment produces a different U statistic

Exact p-value

U under null: How many assignments produce each U

$$p = \frac{2}{252} = 0.00794$$

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Wilcoxon rank sum test

- Non-parametric test
- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Rank all observations together
  - Test whether one group tends to have systematically higher ranks

Null hypothesis

H_0: Both groups come from the same distribution

- Labels are arbitrary
- Observed separation can happen by chance
- Any assignment of 5 observations to A is equally likely

Under null

Any 5 observation could belong to A

$$Total = \binom{10}{5} = 252$$

Each assignment produces a different U statistic

Exact p-value

U under null: How many assignments produce each U

$$p = \frac{2}{252} = 0.00794$$

Multiple testing correction

gene 1 -> 0.001
gene 2 -> 0.009
gene 3 -> 0.02
gene 4 -> 0.04
gene 5 -> 0.2

$$q_i = \frac{p_i m}{i}$$

gene 1 -> 0.005
gene 2 -> 0.0225
gene 3 -> 0.033
gene 4 -> 0.05
gene 5 -> 0.2

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Log fold change calculation

- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Find mean expression value in A
  - Find mean expression value in others
  - Calculate log fold change

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Log fold change calculation

- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Find mean expression value in A
  - Find mean expression value in others
  - Calculate log fold change

Cluster A	Others
5	0
7	1
4	0
6	2
8	1

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Log fold change calculation

- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Find mean expression value in A
  - Find mean expression value in others
  - Calculate log fold change

Cluster A	Others
5	0
7	1
4	0
6	2
8	1

Mean expression

$$A = \frac{5+7+4+6+8}{5} = 6$$

$$Others = \frac{0+1+0+2+1}{5} = 0.8$$

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Log fold change calculation

- For a given gene and cluster A
  - Collect its expression values in cluster A
  - Collect its expression values in all other clusters
  - Find mean expression value in A
  - Find mean expression value in others
  - Calculate log fold change

Cluster A	Others
5	0
7	1
4	0
6	2
8	1

Mean expression

$$A = \frac{5+7+4+6+8}{5} = 6$$

$$Others = \frac{0+1+0+2+1}{5} = 0.8$$

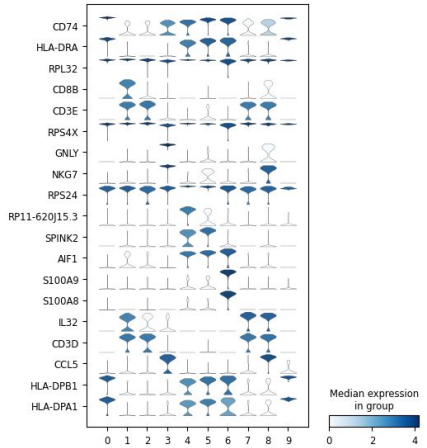
Log fold change

$$logFC = log_2(\frac{6+1}{0.8+1}) = 1.96$$

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

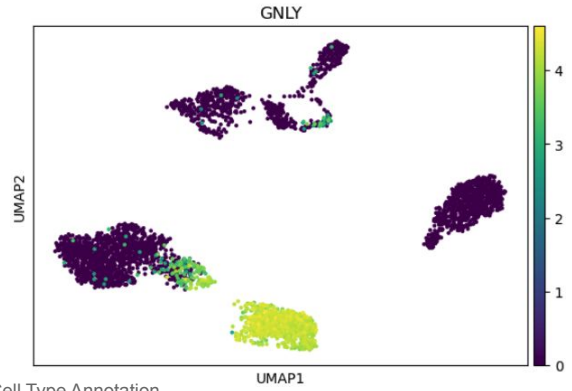
Manual annotation with DEGs

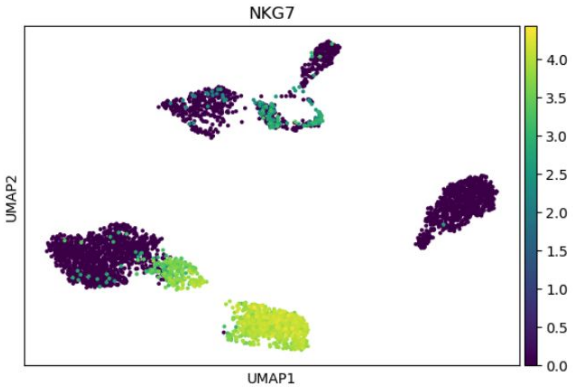


The process of assigning biological meaning to detected cell clusters!

Manual annotation with DEGs:

- Leverage gene signatures lists per cell type
- Such lists are created by genes reported in the literature
- *E.g.*, [GNLY, NKG7, …] for NK cells *etc*
- Apply DEG analysis on clusters (*e.g.*, DESeq2, Wilcoxon *etc*)
- Identify gene signatures per cluster
- Compare with the literature lists





Cons:

- Might miss rare subpopulations
- Depends on literature lists
- Depends on clustering quality

Pros:

- Fully automated
- Statistically robust
- Tailored on the dataset

Cell Type Annotation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Enrichment on DEGs

The process of assigning biological meaning to detected cell clusters!

Enrichment based annotation with DEGs:

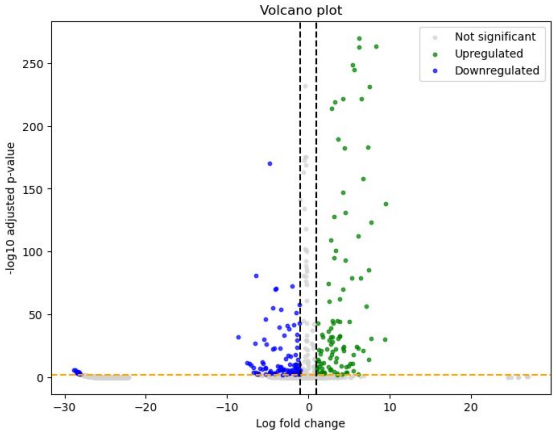
- Leverage gene signatures lists per cell type
- Such lists are created by genes reported in the literature
- *E.g.*, [GNLY, NKG7, …] for NK cells *etc*
- Apply DEG analysis on clusters (*e.g.*, DESeq2, Wilcoxon *etc*)
- Identify gene signatures per cluster
- Find which literature gene signatures are enriched in cluster specific gene signatures

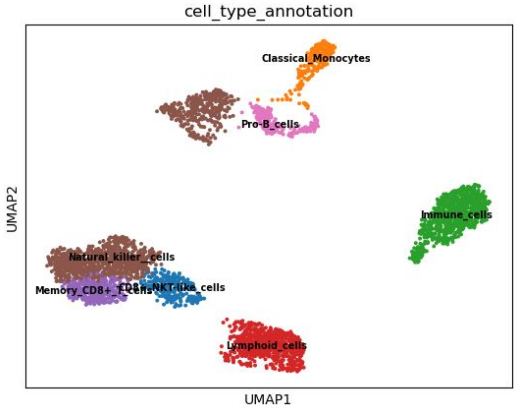
Cons:

- Might miss rare subpopulations
- Depends on literature lists
- Depends on clustering quality

Pros:

- Fully automated
- Statistically robust
- Tailored on the dataset





Cell Type Annotation

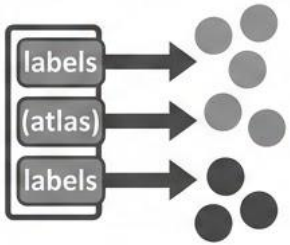
The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Cell type annotation of clusters - Other approaches

- Pre-trained classifiers
- Mapping to existing, annotated single-cell references and performing label transfer on the resulting joint embedding. References can be either
  - Individual samples of the data set
  - Well-curated existing atlases



1. PRE-TRAINED CLASSIFIERS:
Models trained on existing data to predict cell identities.



2. MAPPING & LABEL TRANSFER:
Projecting data onto annotated references for consistent biological interpretation.

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

Hands-on 3
Cell Type Annotation

PHAROS
THE GREEK AI FACTORY

Closing remarks

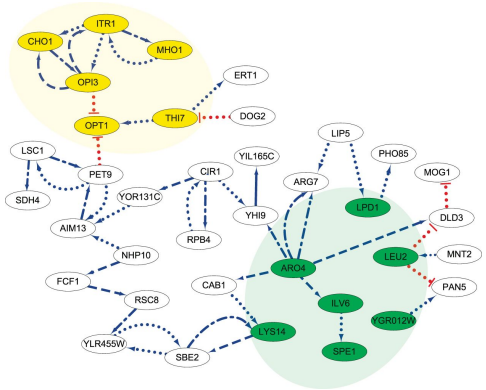
Beyond cell type annotation of clusters

Trajectory analysis



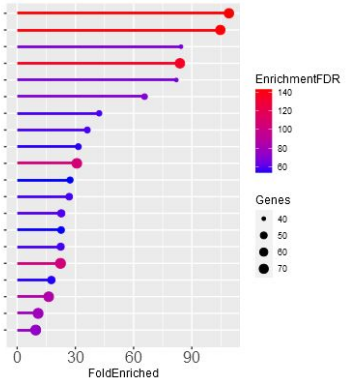
Jin-Hong Du, PNAS, (2024)

Gene Regulatory Network



Chen Chen, Scientific Reports, (2019)

Pathway Analysis



https://bioinformatics.sdstate.edu/go/

G.Georgakilas & C. Sinnis & H. Manousaki, IMSI / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Useful resources

- Practical bioinformatics pipelines for single-cell RNA-seq data analysis : https://pmc.ncbi.nlm.nih.gov/articles/PMC10189648/
- Single-cell RNA sequencing technologies and applications: A brief overview : https://pmc.ncbi.nlm.nih.gov/articles/PMC8964935/#ctm2694-sec-0030
- Current best practices in single-cell RNA-seq analysis: a tutorial: https://link.springer.com/article/10.15252/msb.20188746
- https://medium.com/@tudoupengkai/how-to-understand-the-anndata-structure-bcf4e423a641

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Access Our Research

Project Repository



← *Scan to explore*

Researcher Profiles



G.Georgakilas



H. Manousaki



C. Sinnis

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

Thank you!

Questions and discussion

www.pharos-aifactory.eu

EuroHPC Joint Undertaking
Co-funded by the European Union
AI Factories PHAROS GREECE
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Ψηφιακής Διακυβέρνησης και Τεχνητής Νοημοσύνης