



Techniques of scRNA seq

2026. 10. 02

HyungGyu Min

Single-cell RNA-seq Analysis – Complete Cheat Sheet

From raw reads to biological insight

1

PRIMARY PROCESSING

How do reads become counts?

GEX library
(transcriptome)

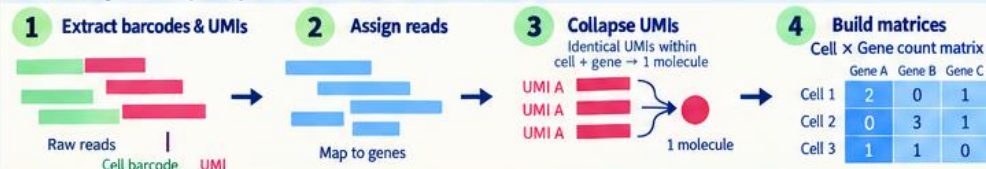
Cell barcode UMI Transcript sequence

Guide library
(CRISPR screen)

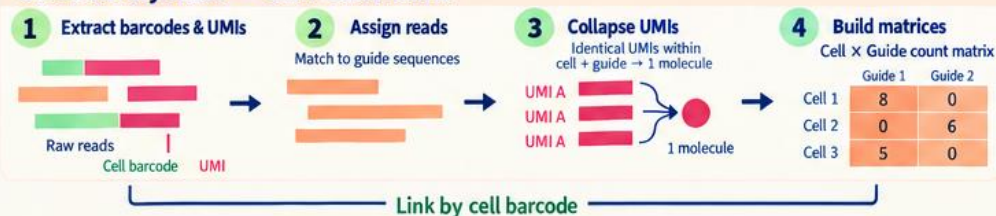
Cell barcode UMI Guide-identifying sequence

Separate guide libraries
depend on preparation.

Transcriptome (GEX) reads → Cell × Gene matrix



Guide library reads → Cell × Guide matrix



Output: Count matrices

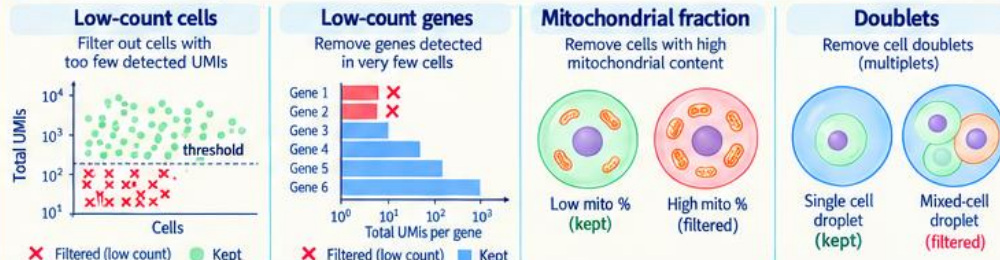
2

QC & PREPROCESSING

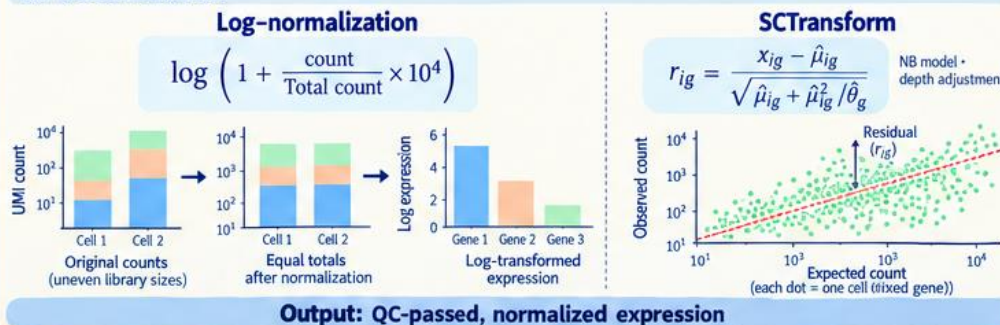
Which cells and counts can we trust?

a. QC Filtering

Context-dependent QC thresholds



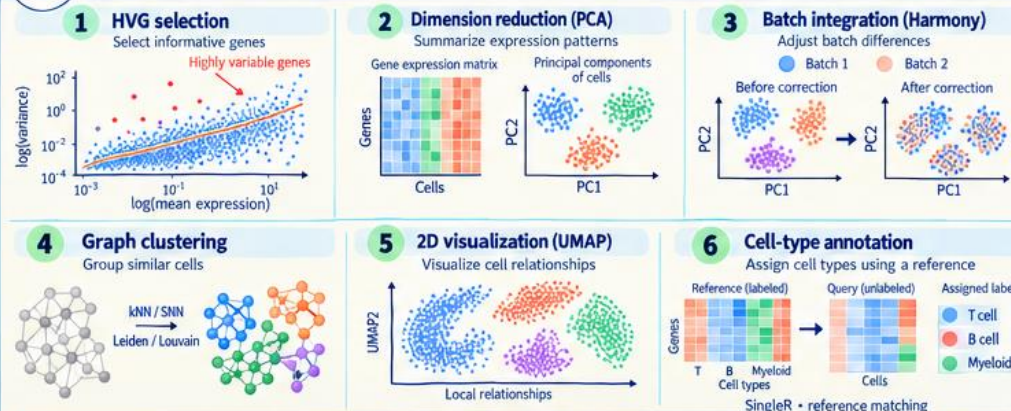
b. Normalization



3

CELL LANDSCAPE

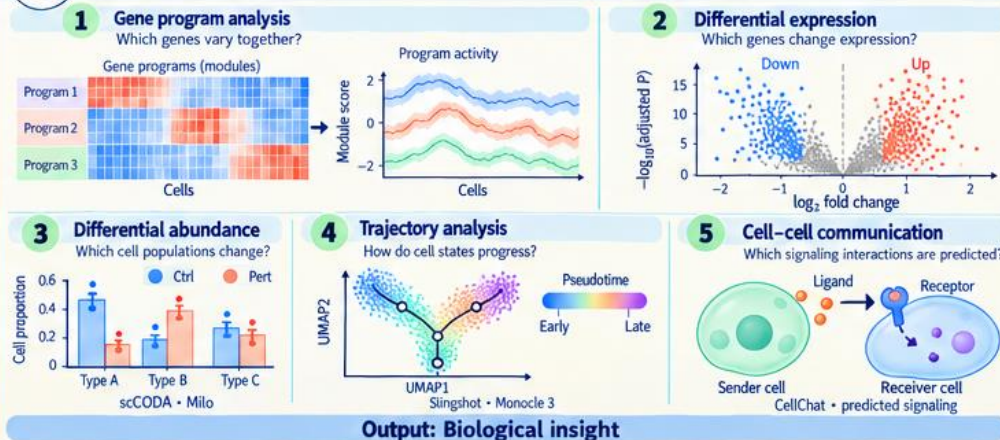
What cell types and states are present?



4

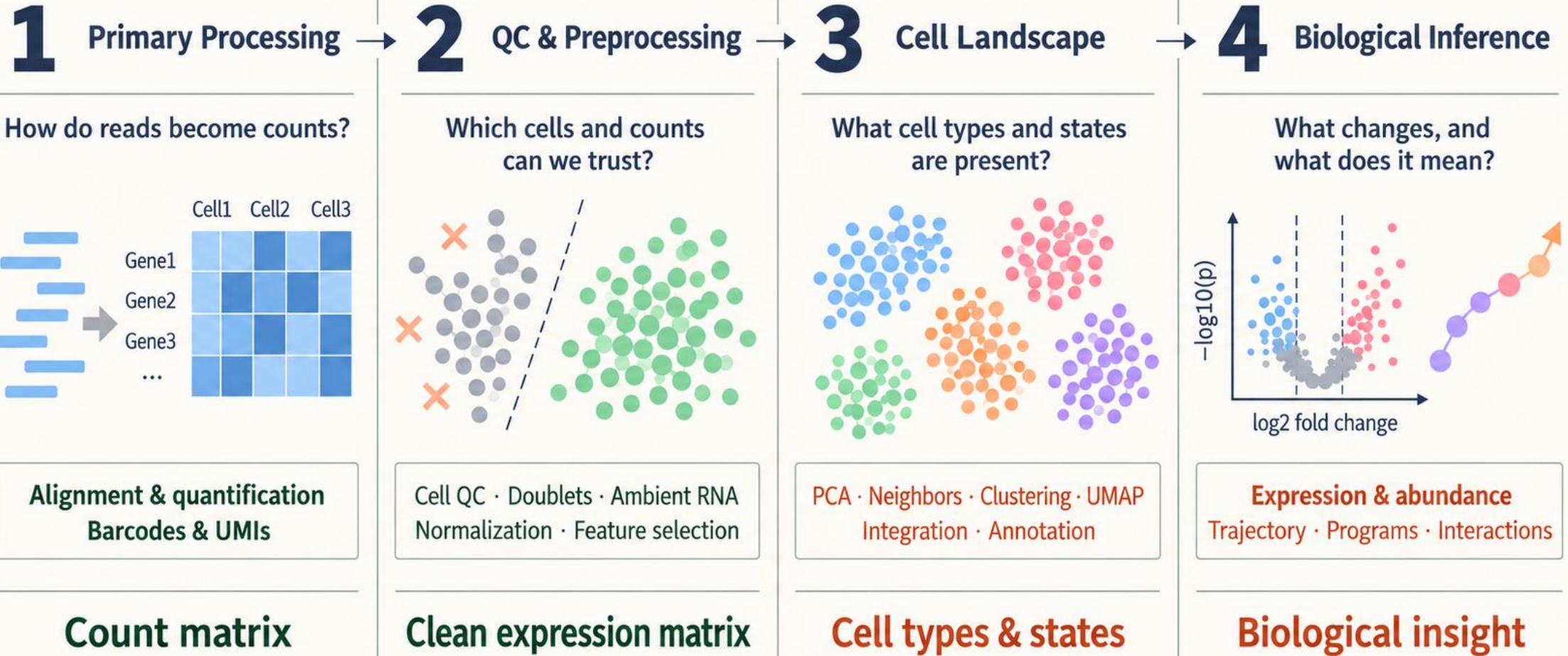
BIOLOGICAL INFERENCE

What changes, and what does it mean?



Single-cell RNA-seq Analysis

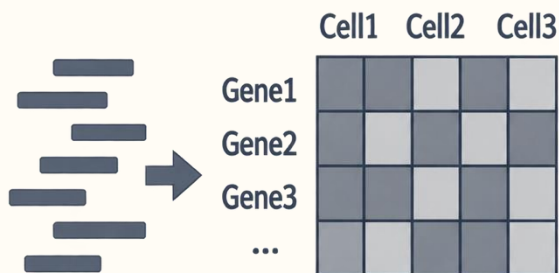
From raw reads to biological insight



1

Primary Processing

How do reads become counts?

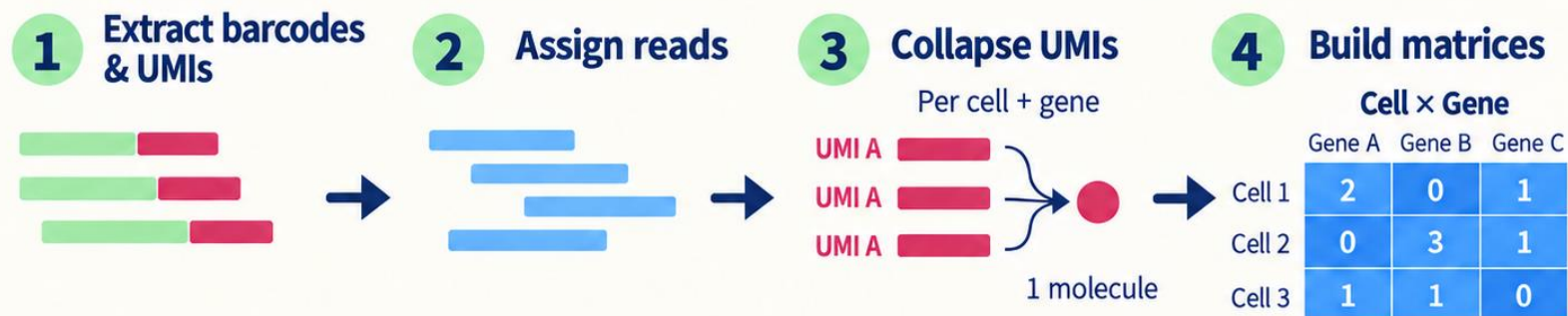
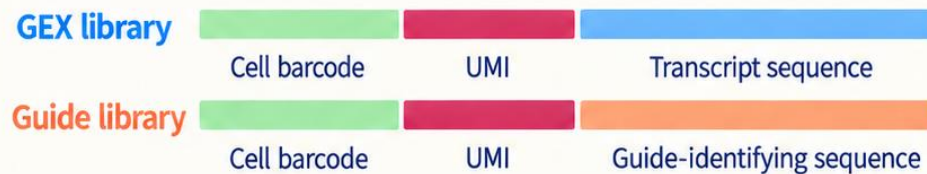


Alignment & quantification
Barcodes & UMIs

Cell Ranger · STARsolo
alevin-fry

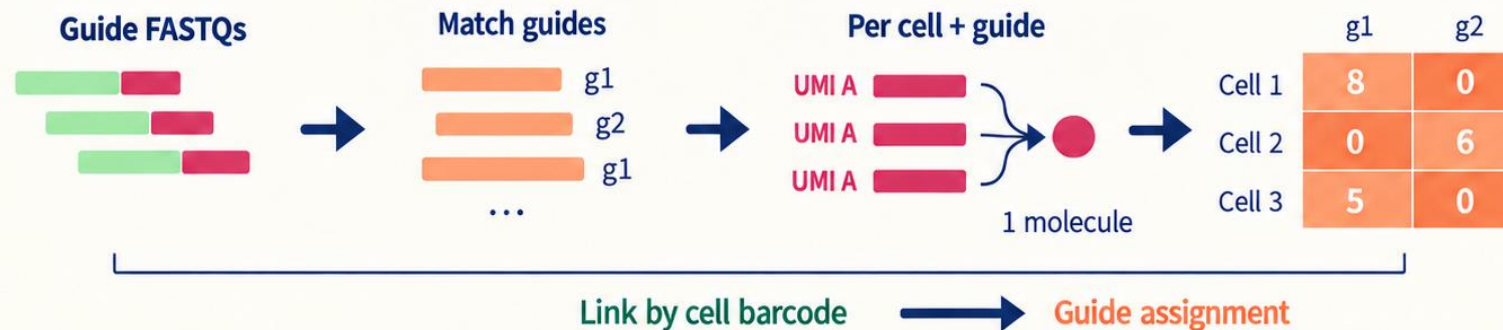
Count matrix

FASTQ to count matrices



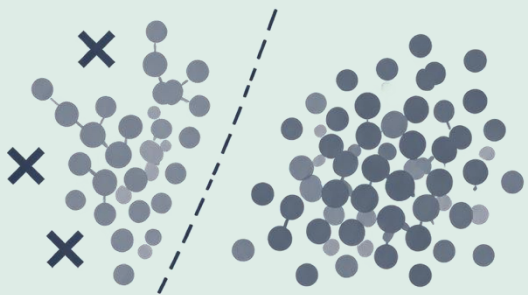
Guide library

Separate-library example



2 QC & Preprocessing

Which cells and counts
can we trust?

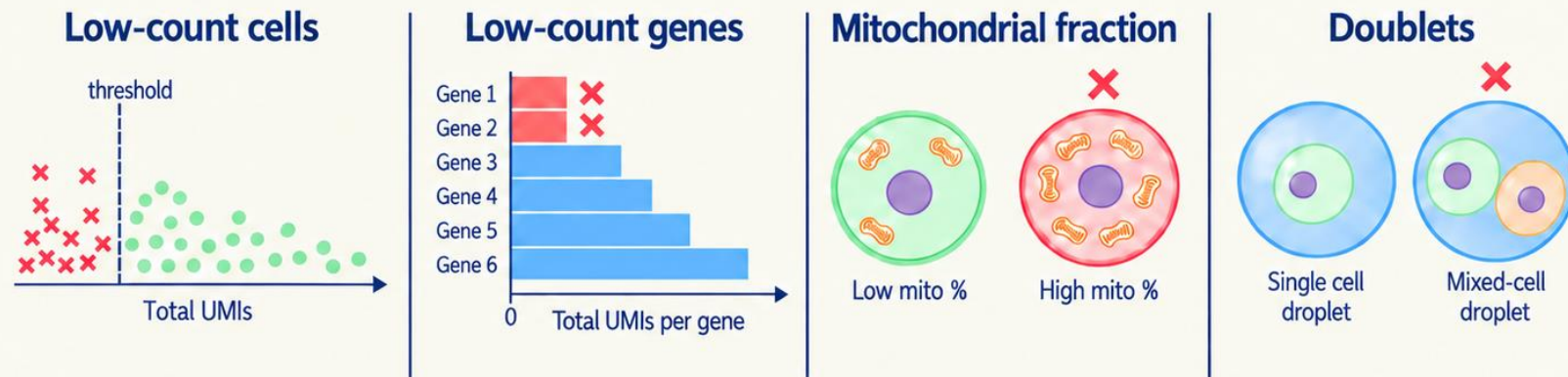


Cell QC · Doublets · Ambient RNA
Normalization · Feature selection

Seurat · Scanpy
SoupX · scDblFinder

Clean expression matrix

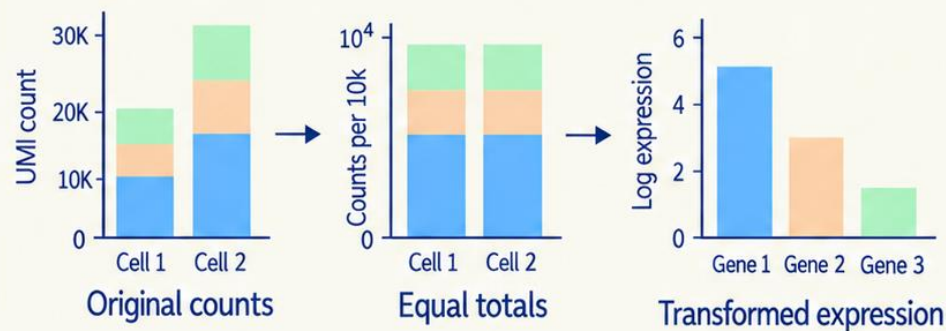
a. QC Filtering



b. Normalization

Log-normalization

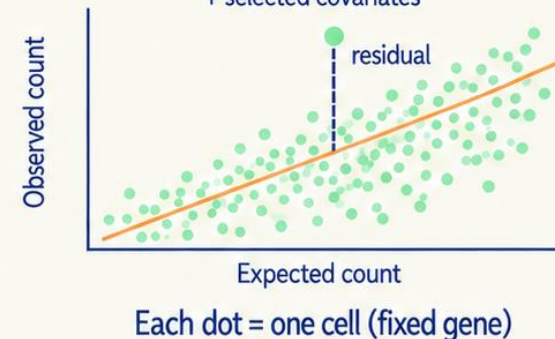
$$\log \left(1 + \frac{\text{count}}{\text{Total count}} \times 10^4 \right)$$



SCTransform

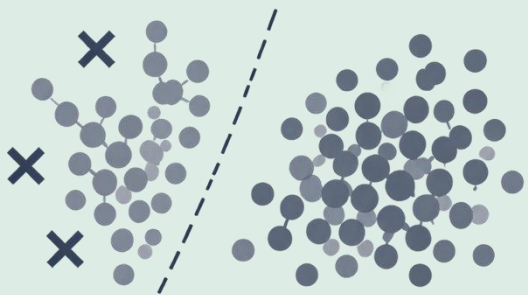
$$r_{ig} = \frac{x_{ig} - \hat{\mu}_{ig}}{\sqrt{\hat{\mu}_{ig} + \hat{\mu}_{ig}^2 / \hat{\theta}_g}}$$

Expected count from sequencing depth
+ selected covariates



2 QC & Preprocessing

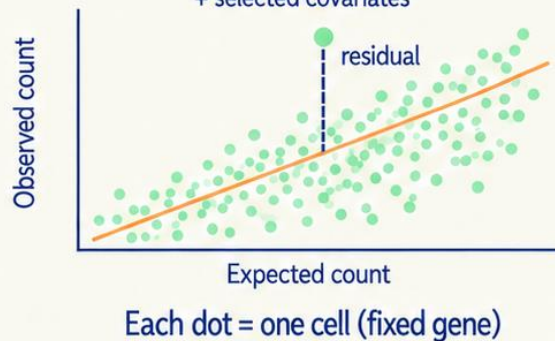
Which cells and counts
can we trust?



SCTransform

$$r_{ig} = \frac{x_{ig} - \hat{\mu}_{ig}}{\sqrt{\hat{\mu}_{ig} + \hat{\mu}_{ig}^2 / \hat{\theta}_g}}$$

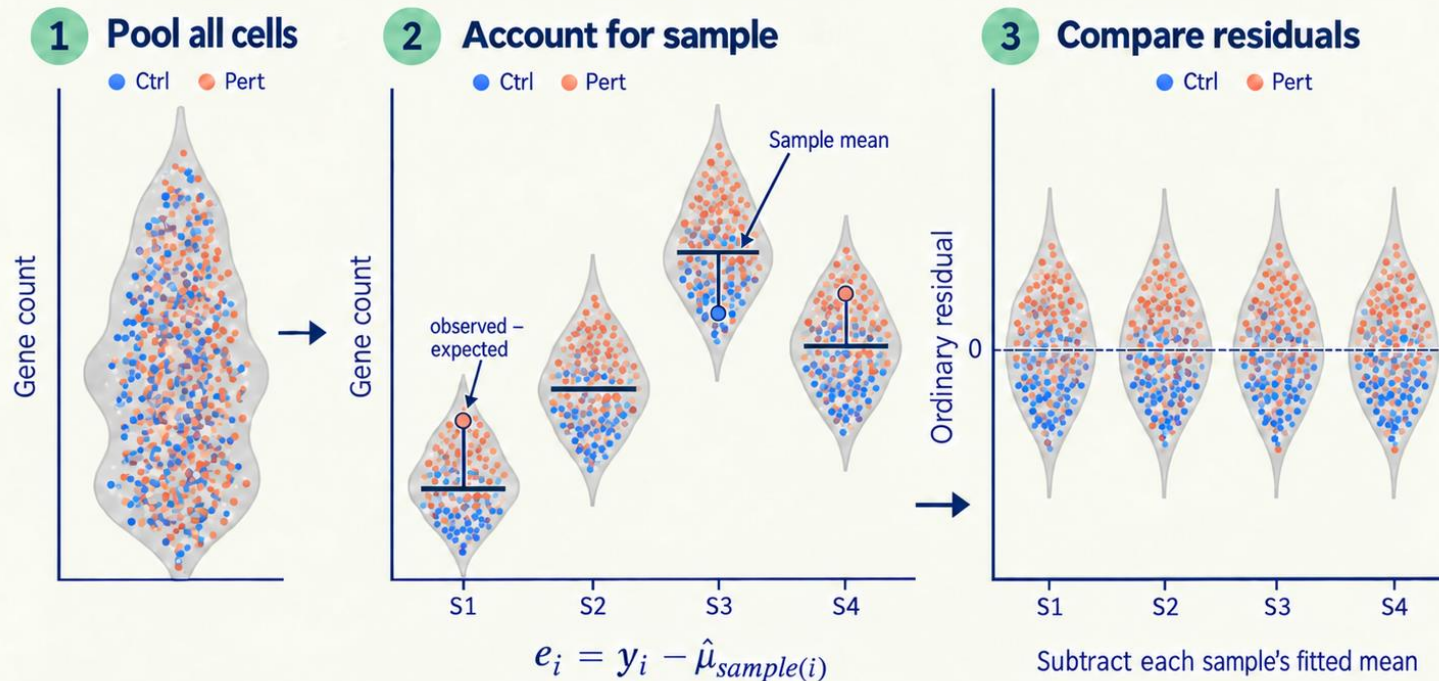
Expected count from sequencing depth
+ selected covariates



Seurat::SCTransform

- To eliminate some confounder

Residuals: removing sample baseline differences



For scRNA-seq: NB model and Pearson residuals

$$x_{ig} \sim \text{NB}(\mu_{ig}, \theta_g), \quad \log \mu_{ig} = \beta_{0g} + \log N_i + \gamma_{g, sample(i)}$$

$$r_{ig} = (x_{ig} - \hat{\mu}_{ig}) / \sqrt{\hat{\mu}_{ig} + \hat{\mu}_{ig}^2 / \hat{\theta}_g}$$

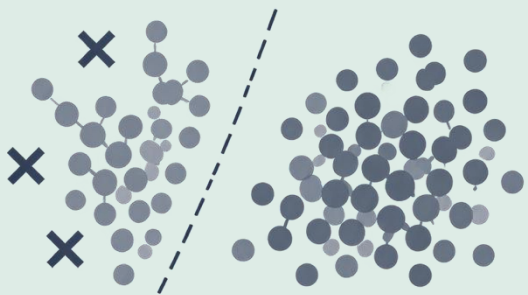
Observed - expected, scaled by expected variability

Conceptual sample-adjusted model; sample adjustment must be explicitly specified. Schematic data.

N_i : total UMIs
 γ : sample effect
 θ : NB dispersion

2 QC & Preprocessing

Which cells and counts
can we trust?



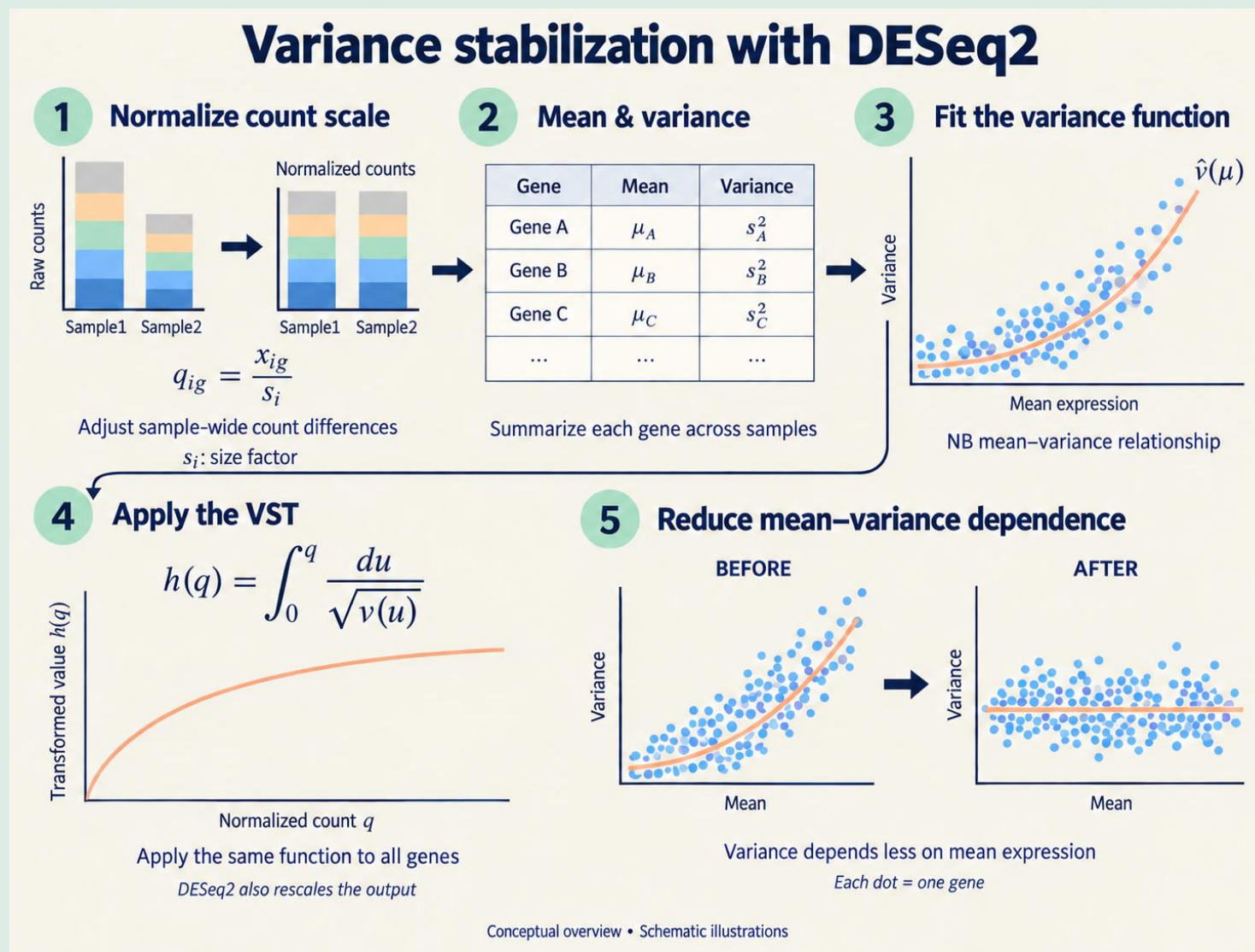
Cell QC · Doublets · Ambient RNA
Normalization · Feature selection

Seurat · Scanpy
SoupX · scDblFinder

Clean expression matrix

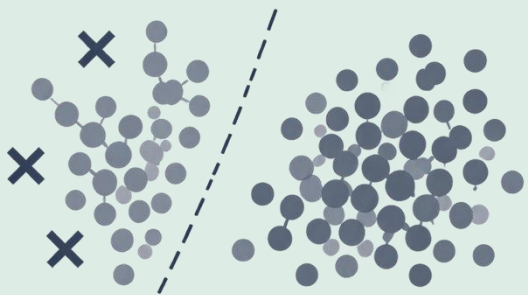
DESeq2::vst

- To stabilize the variance in between the genes



2 QC & Preprocessing

Which cells and counts
can we trust?



Cell QC · Doublets · Ambient RNA
Normalization · Feature selection

Seurat · Scanpy
SoupX · scDblFinder

Clean expression matrix

DESeq2::vst

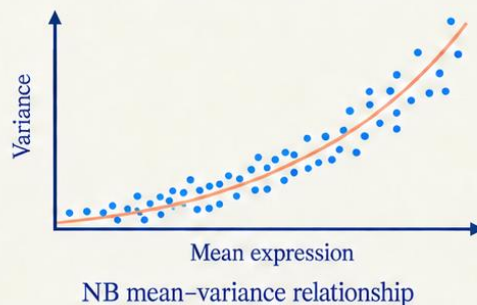
- To stabilize the variance in between the genes

DESeq2 VST: from variance to transformation

1 Estimate the variance function

$$v(\mu) = A\mu + B\mu^2$$

$$A = 1 + a_1, \quad B = a_0 \quad (A, B > 0)$$



2 Apply the delta method

$$\text{Var}[h(X)] \approx [h'(\mu)]^2 v(\mu)$$

$$h'(\mu) = \frac{1}{\sqrt{v(\mu)}}$$

$$\text{Var}[h(X)] \approx 1$$

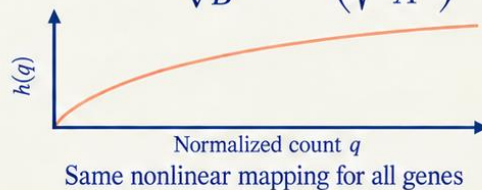


Choose a slope that compensates for variance

3 Integrate the slope

$$h(q) = \int_0^q \frac{du}{\sqrt{v(u)}}$$

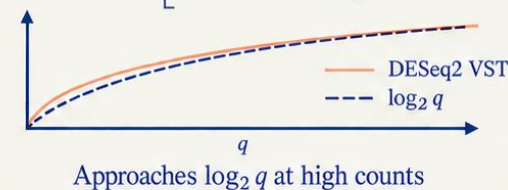
$$= \frac{2}{\sqrt{B}} \operatorname{asinh}\left(\sqrt{\frac{Bq}{A}}\right)$$



4 Rescale to the log₂ scale

$$h_{\text{DESeq2}}(q) = \frac{\sqrt{B}}{\ln 2} h(q) + \log_2\left(\frac{A}{4B}\right)$$

$$h_{\text{DESeq2}}(q) = \log_2\left[\frac{A + 2Bq + 2\sqrt{Bq(A + Bq)}}{4B}\right]$$



h : count-transforming function

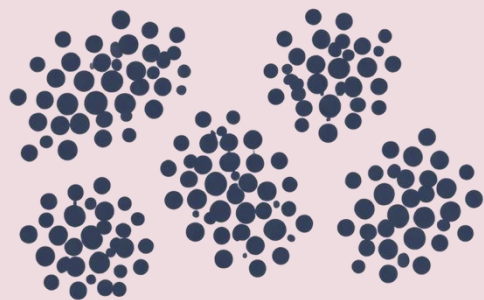
v : mean $\rightarrow$ variance function

q : raw count

3

Cell Landscape

What cell types and states are present?



PCA · Neighbors · Clustering · UMAP
Integration · Annotation

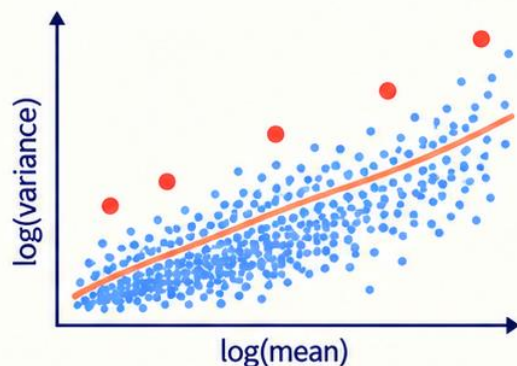
Seurat · Scanpy
Harmony · scVI · CellTypist

Cell types & states

Cell landscape: six core methods

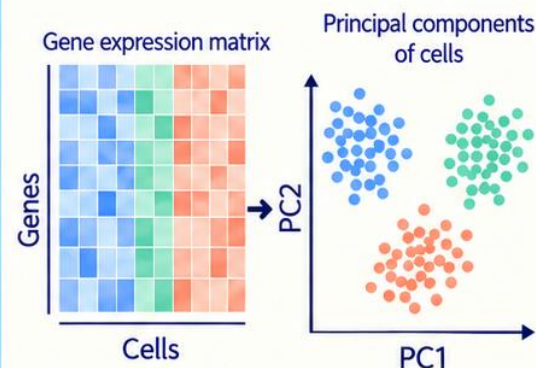
HVG Selection

Select informative genes



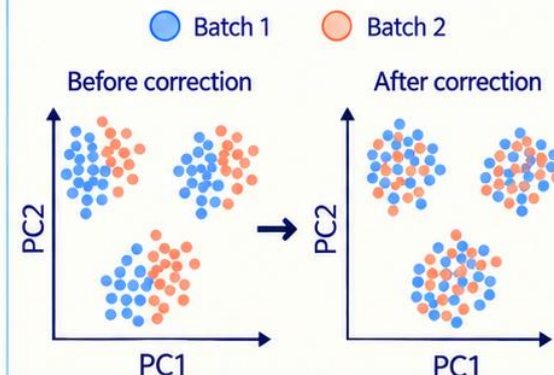
Dimension Reduction

Summarize expression patterns



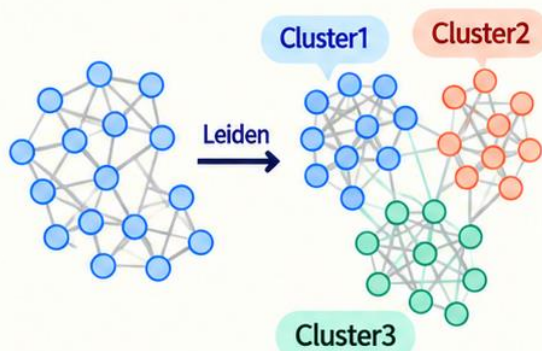
Batch Integration

Adjust batch differences



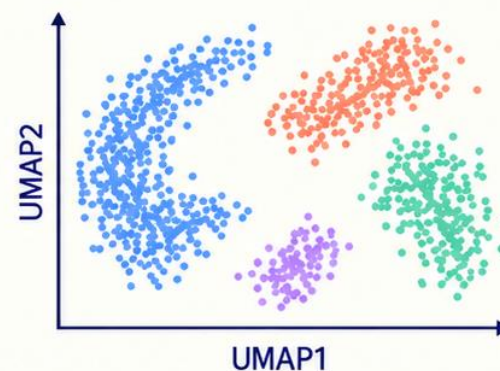
Clustering

Group similar cells



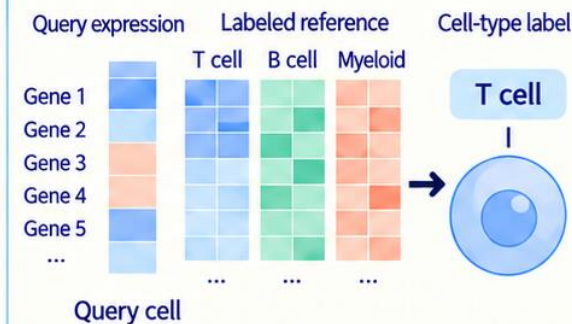
2D Visualization

Visualize cell relationships



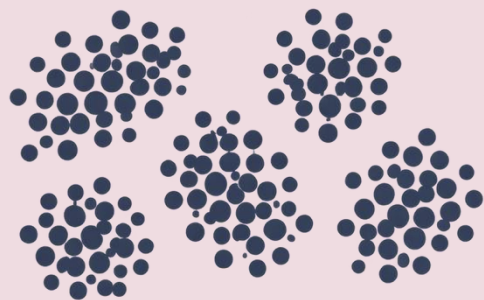
Celltype Annotation

Assign cell types using a reference



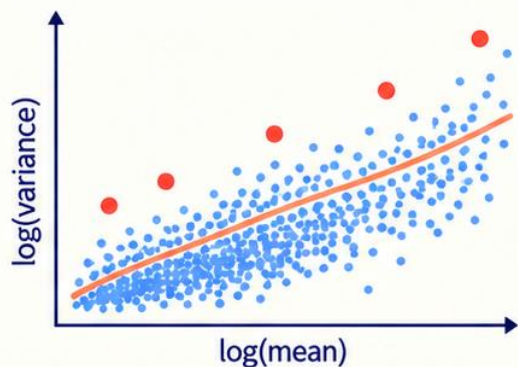
3 Cell Landscape

What cell types and states are present?



HVG Selection

Select informative genes

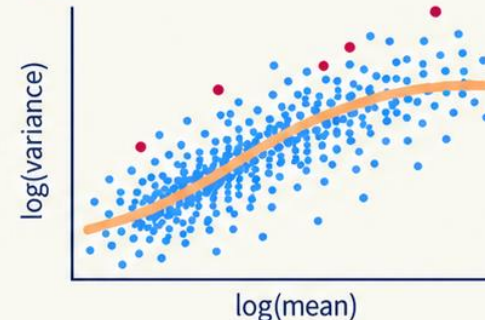


HVG selection with Seurat vst

1 Mean & variance

Gene	Mean	Variance
Gene A	μ_A	s_A^2
Gene B	μ_B	s_B^2
Gene C	μ_C	s_C^2

2 LOESS fit

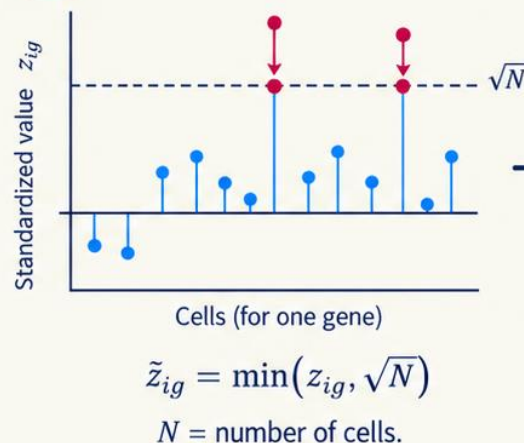


3 Standardize counts

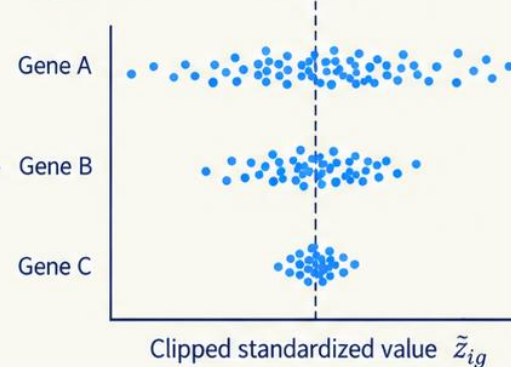
$$z_{ig} = \frac{x_{ig} - \bar{x}_g}{\sqrt{\hat{v}_g}}$$



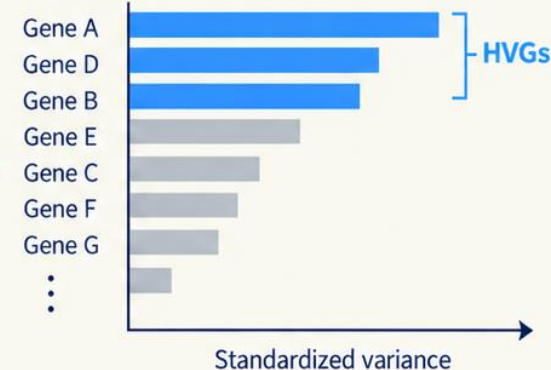
4 Clip extreme values



5 Calculate variance after clipping



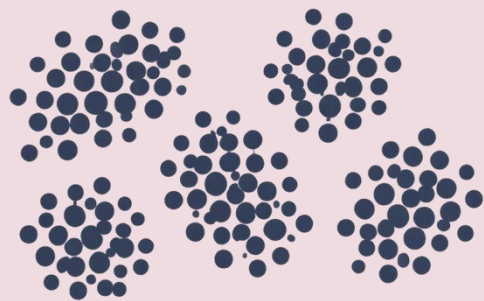
6 Select top genes



FindVariableFeatures(..., selection.method = "vst")

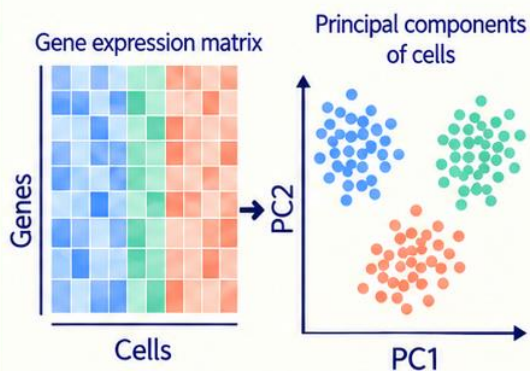
→ 3 Cell Landscape —

What cell types and states are present?



Dimension Reduction

Summarize expression patterns

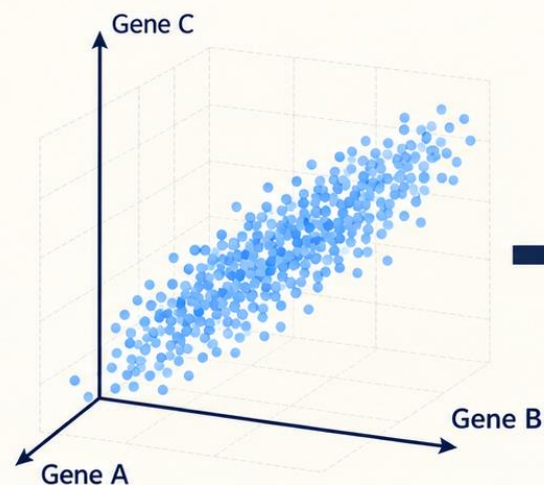


PCA

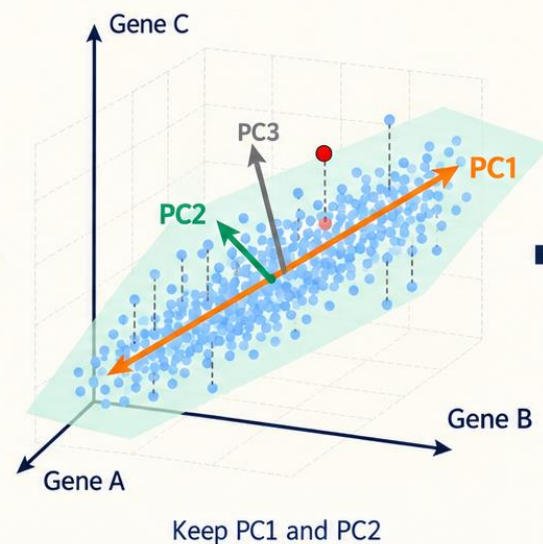
- Compress information from tens of thousands of genes into a vector of around 50 numbers.

PCA: from 3D expression to 2D

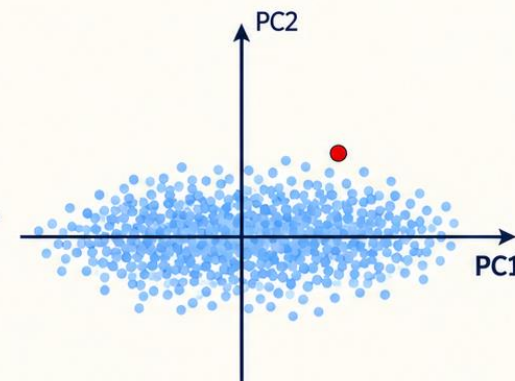
1 Three-dimensional expression



2 Find the principal plane



3 Project onto two PCs



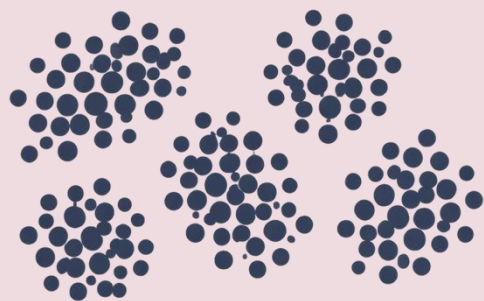
Each dot = one cell

3 genes → 2 principal components

Illustration; real data contain thousands of genes

3 Cell Landscape

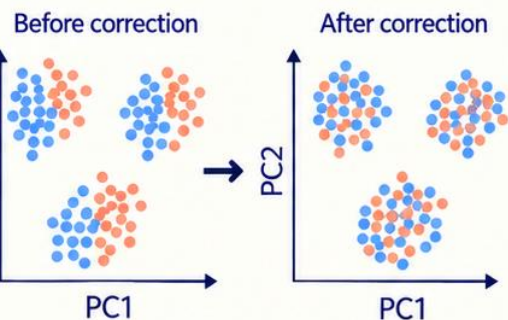
What cell types and states are present?



Batch Integration

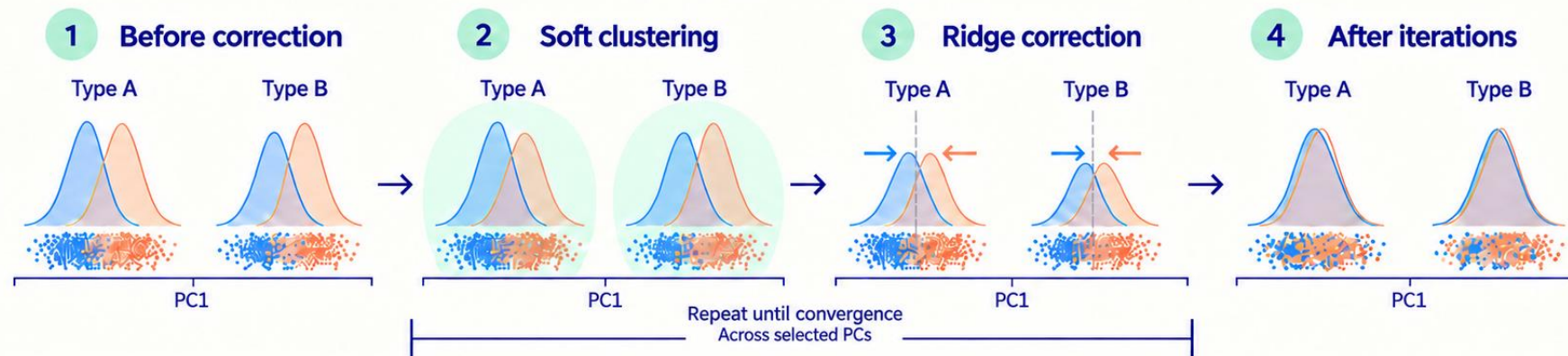
Adjust batch differences

● Batch 1 ● Batch 2

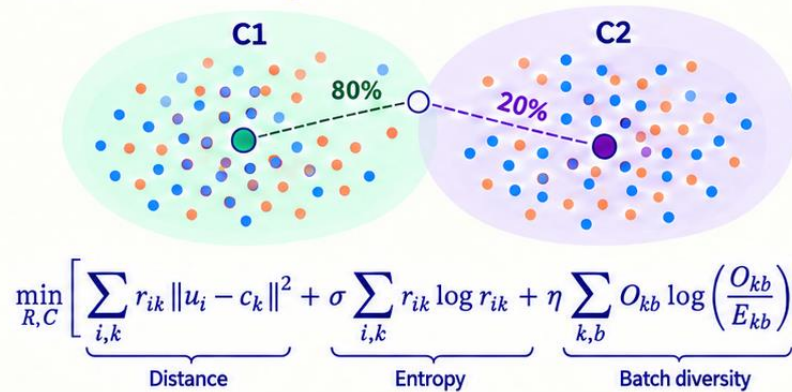


Harmony: iterative batch correction

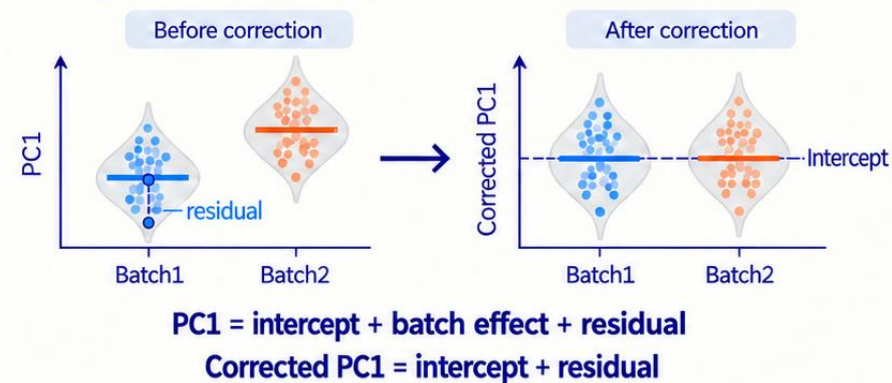
● Batch1 ● Batch2



Soft clustering

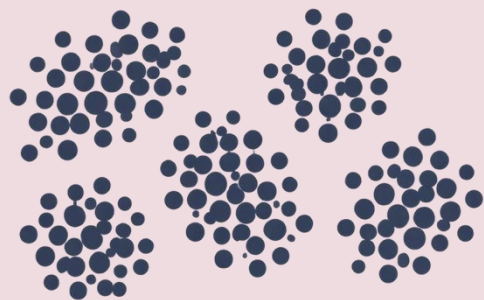


Weighted ridge regression



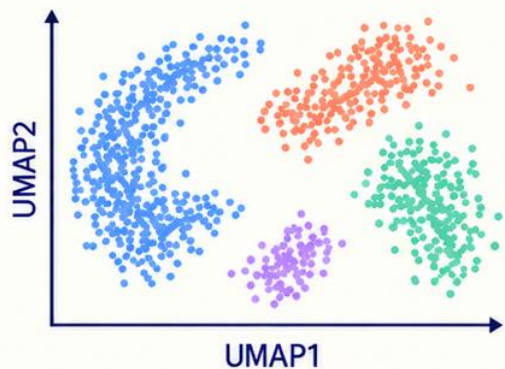
3 Cell Landscape

What cell types and states are present?



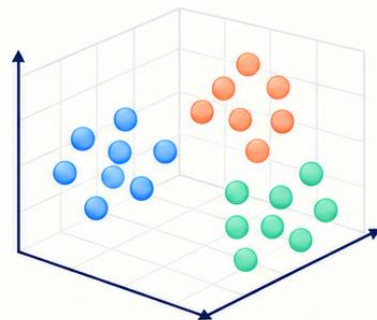
2D Visualization

Visualize cell relationships



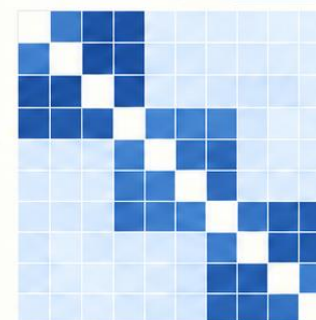
t-SNE & UMAP

High-dimensional data

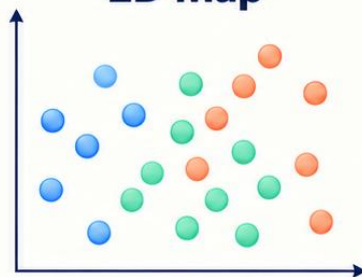


f_1

Similarity Fixed

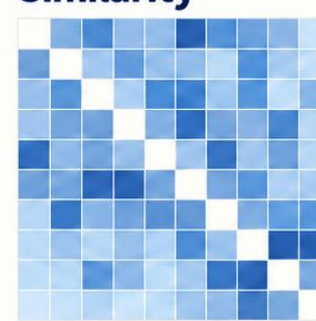


2D map



f_2

Similarity

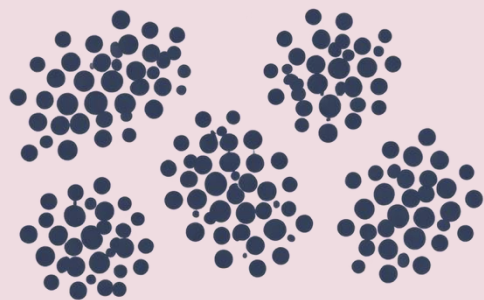


Update positions

Compare
 C

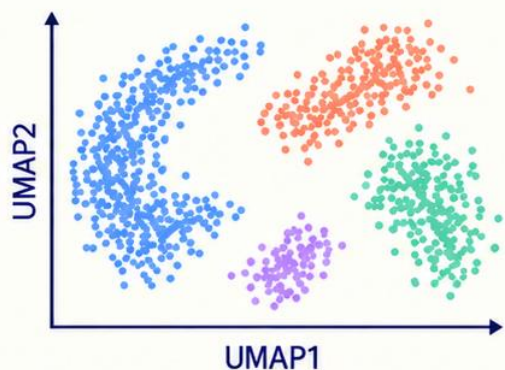
→ 3 Cell Landscape —

What cell types and states are present?



2D Visualization

Visualize cell relationships



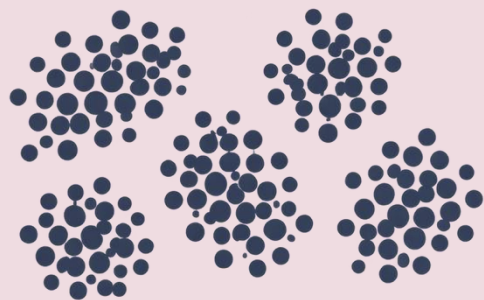
t-SNE vs UMAP

	t-SNE	UMAP
f_1 High-dimensional similarity	Gaussian-based neighbor affinities Symmetrized and normalized	Locally scaled kNN graph Fuzzy connection strengths
f_2 2D similarity	Student-t kernel $q_{ij} \propto (1 + \ y_i - y_j\ ^2)^{-1}$	Distance-based kernel $q_{ij} = \frac{1}{1 + a\ y_i - y_j\ ^{2b}}$
$\mathcal{C}$ Objective	KL divergence $\sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$	Fuzzy-set cross-entropy $- \sum_{i < j} [p_{ij} \log q_{ij} + (1 - p_{ij}) \log(1 - q_{ij})]$
Philosophy	Match a distribution of neighbor similarities. Strongly penalize separating similar points.	Match a weighted neighborhood graph. Balance attraction and repulsion.

Both optimize 2D coordinates to reproduce high-dimensional neighborhood relationships.

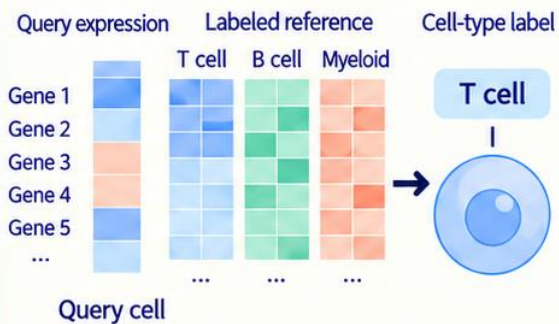
3 Cell Landscape

What cell types and states are present?



Celltype Annotation

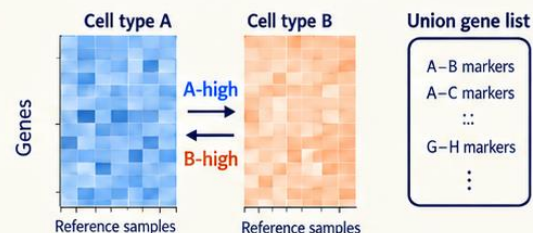
Assign cell types using a reference



SingleR: from reference profiles to cell labels

1 Select marker genes

Reference-only marker selection

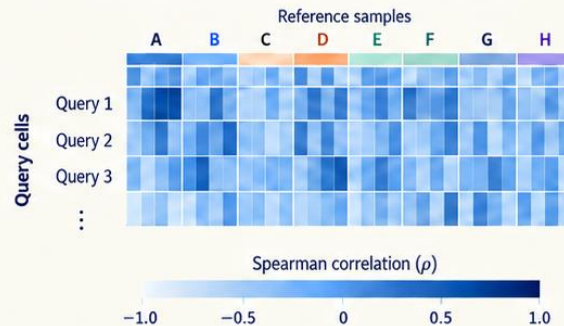


$$m = 500(2/3) \log_2 K$$

$K = 8$: ~148 genes per direction
Union across all type pairs

2 Compute correlations

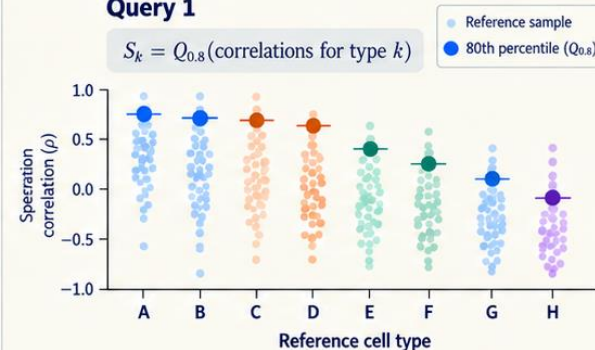
Spearman correlation across marker genes



3 Score each cell type

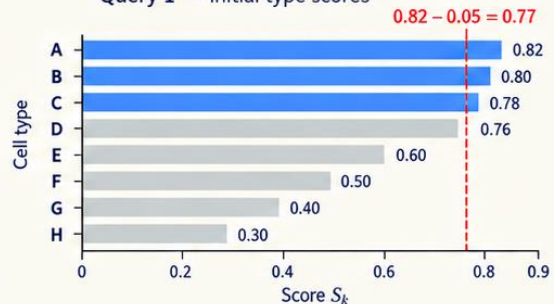
Query 1

$$S_k = Q_{0.8}(\text{correlations for type } k)$$



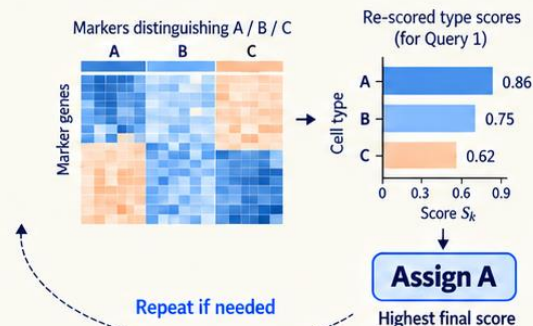
4 Select candidate types

Query 1 — initial type scores



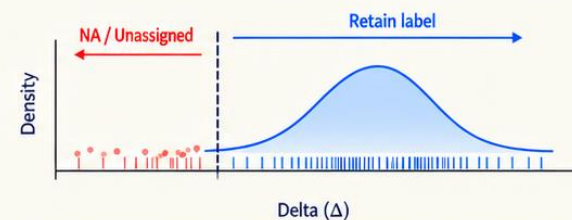
5 Fine-tune & assign

Re-score against reference samples for candidate types



6 Prune uncertain labels

Distribution of Δ for cells with the same initial label L



$$\Delta_i = S_{i, \text{assigned}} - \text{median}_k(S_{ik})$$

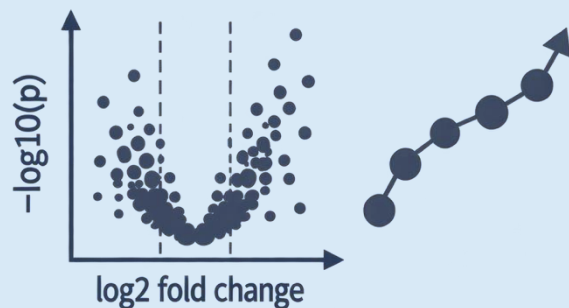
Prune if $\Delta_i < \text{median}(\Delta_L) - 3 \text{MAD}(\Delta_L)$

Also prune if final best — next ≤ 0

Classic defaults • MAD uses the 1.4826 scaling factor

→ 4 Biological Inference

What changes, and what does it mean?



Expression & abundance
Trajectory · Programs · Interactions

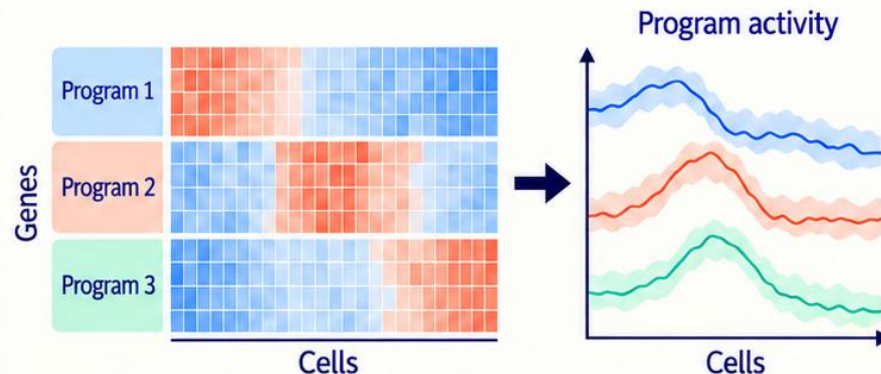
DESeq2 · Milo · Slingshot
cNMF · CellChat

Biological insight

Biological inference: five core analyses

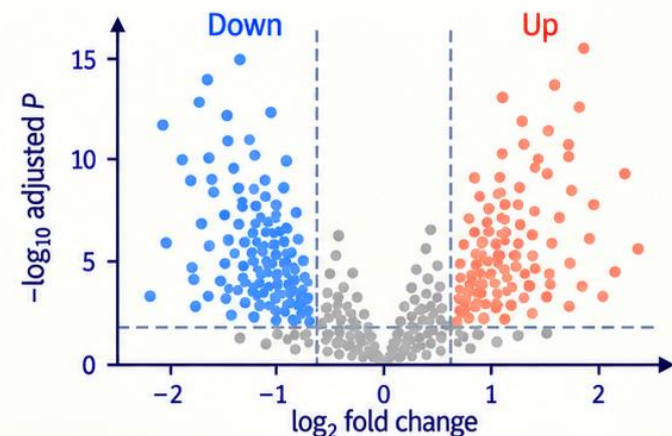
Gene Program Analysis

Which genes vary together?



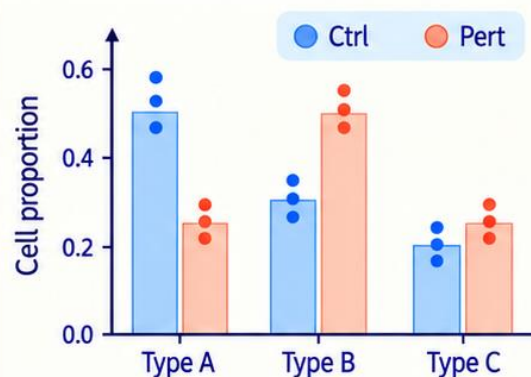
Differential Expression

Which genes change expression?



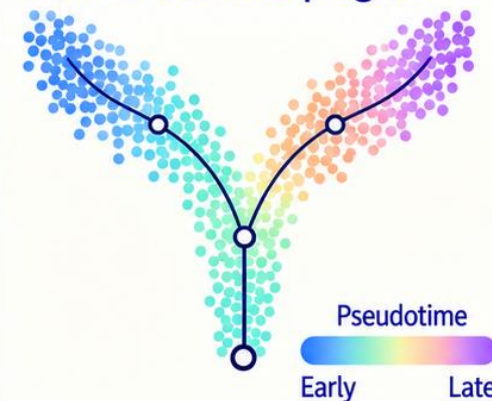
Differential Abundance

Which cell populations change?



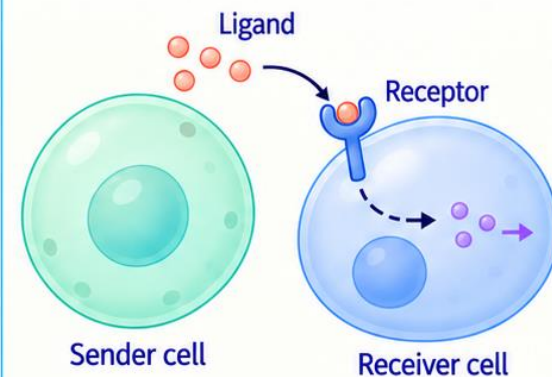
Trajectory Analysis

How do cell states progress?



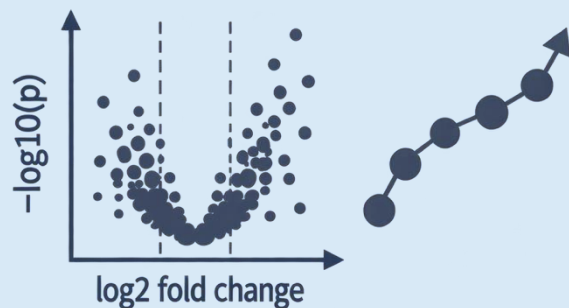
Cell-Cell Communication

Which signaling interactions are predicted?



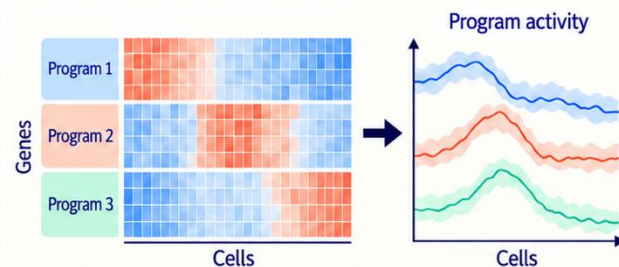
→ 4 Biological Inference

What changes, and
what does it mean?



Gene Program Analysis

Which genes vary together?



Convergence of coronary artery disease genes onto endothelial cell programs

Schnitzler, G.R., Kang, H., Fang, S. et al.

Nature 626, 799–807 (2024).

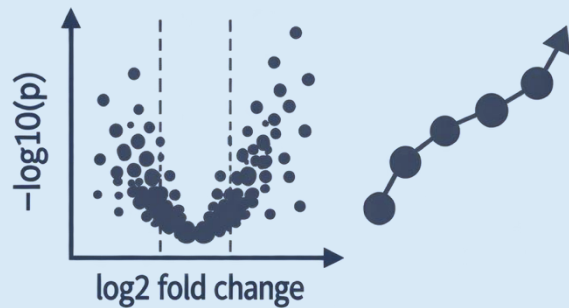
<https://doi.org/10.1038/s41586-024-07022-x>

FA & cNMF

- <https://thewonlab.slack.com/archives/CCQTR2UNM/p1726769274196319>

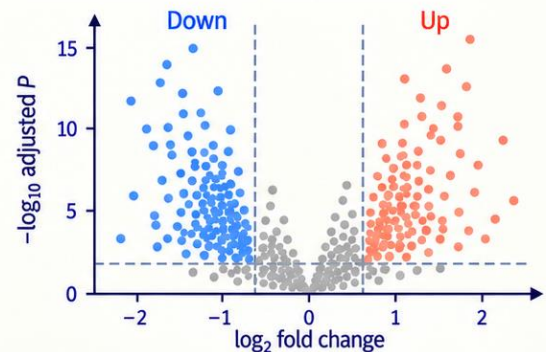
→ 4 Biological Inference

What changes, and
what does it mean?



Differential Expression

Which genes change expression?



muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data

Helena L. Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra & Mark D. Robinson

Nature Communication (2020)

Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

6 models for DE analysis

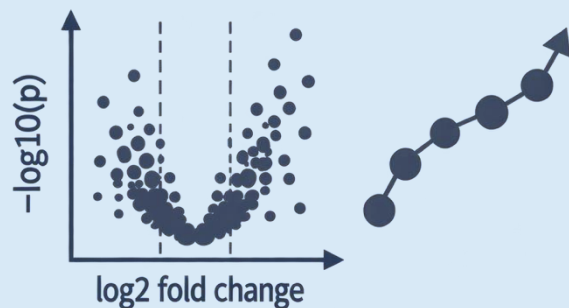
- <https://thewonlab.slack.com/archives/CCQTR2UNM/p1774023602138449>



ZiPert ; <https://thewonlab.slack.com/archives/CCQTR2UNM/p1782500083480739>

→ 4 Biological Inference

What changes, and
what does it mean?



Differential Expression Analysis (DE)

Gene expression $\sim$ Condition + Covariates

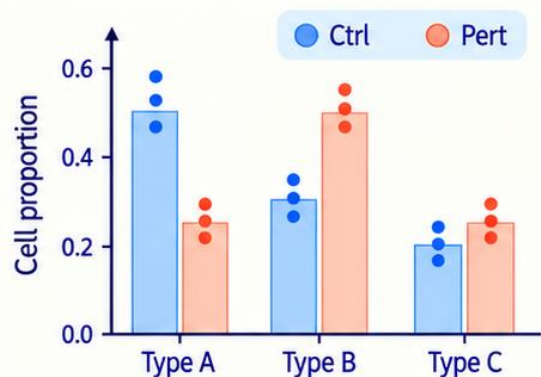
e.g., Identifying genes differentially expressed following perturbation.

Differential Abundance Analysis (DA)

Cell count $\sim$ Condition + Covariates

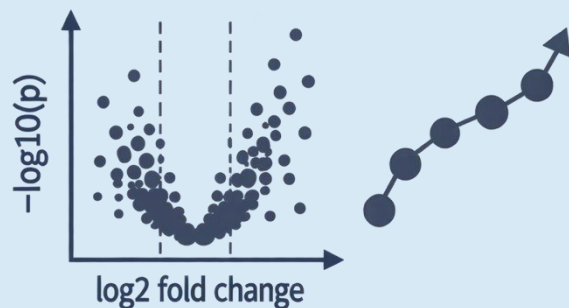
e.g., Identifying cell types differentially abundant following exposure.

Differential Abundance Which cell populations change?



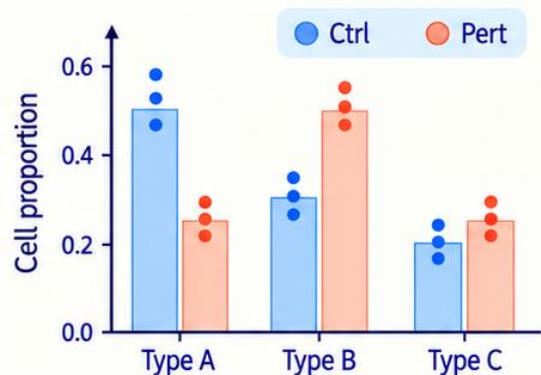
→ 4 Biological Inference

What changes, and what does it mean?



Differential Abundance

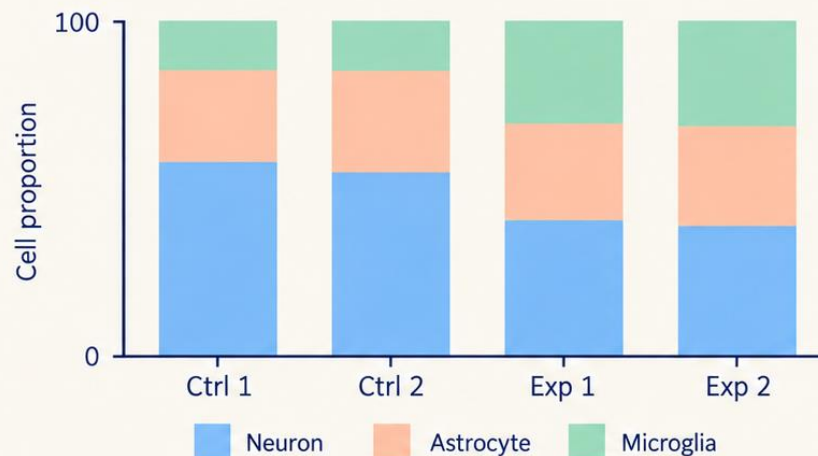
Which cell populations change?



Differential Abundance: what do we compare?

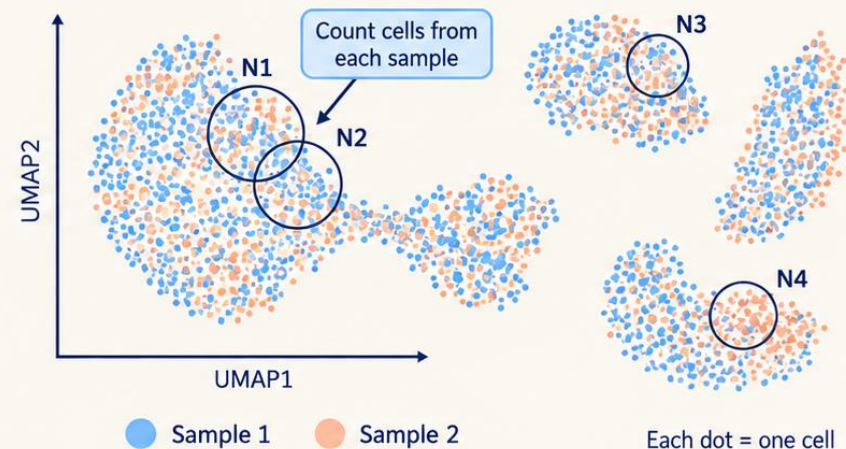
scCODA

Cell-type composition per sample



Milo

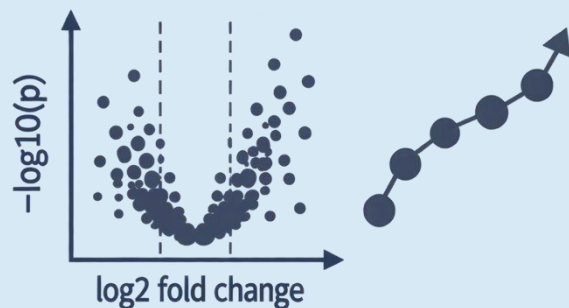
Overlapping kNN neighborhoods



	scCODA	Milo
Target	Annotated cell types	Overlapping kNN neighborhoods
Input	Cell counts by sample × cell type	Cell counts by sample × neighborhood
Replicate unit	Biological sample	Biological sample

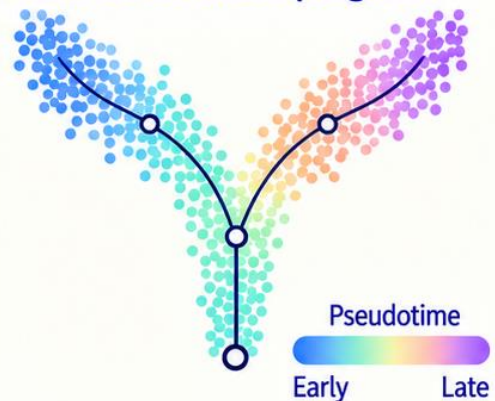
→ 4 Biological Inference

What changes, and what does it mean?

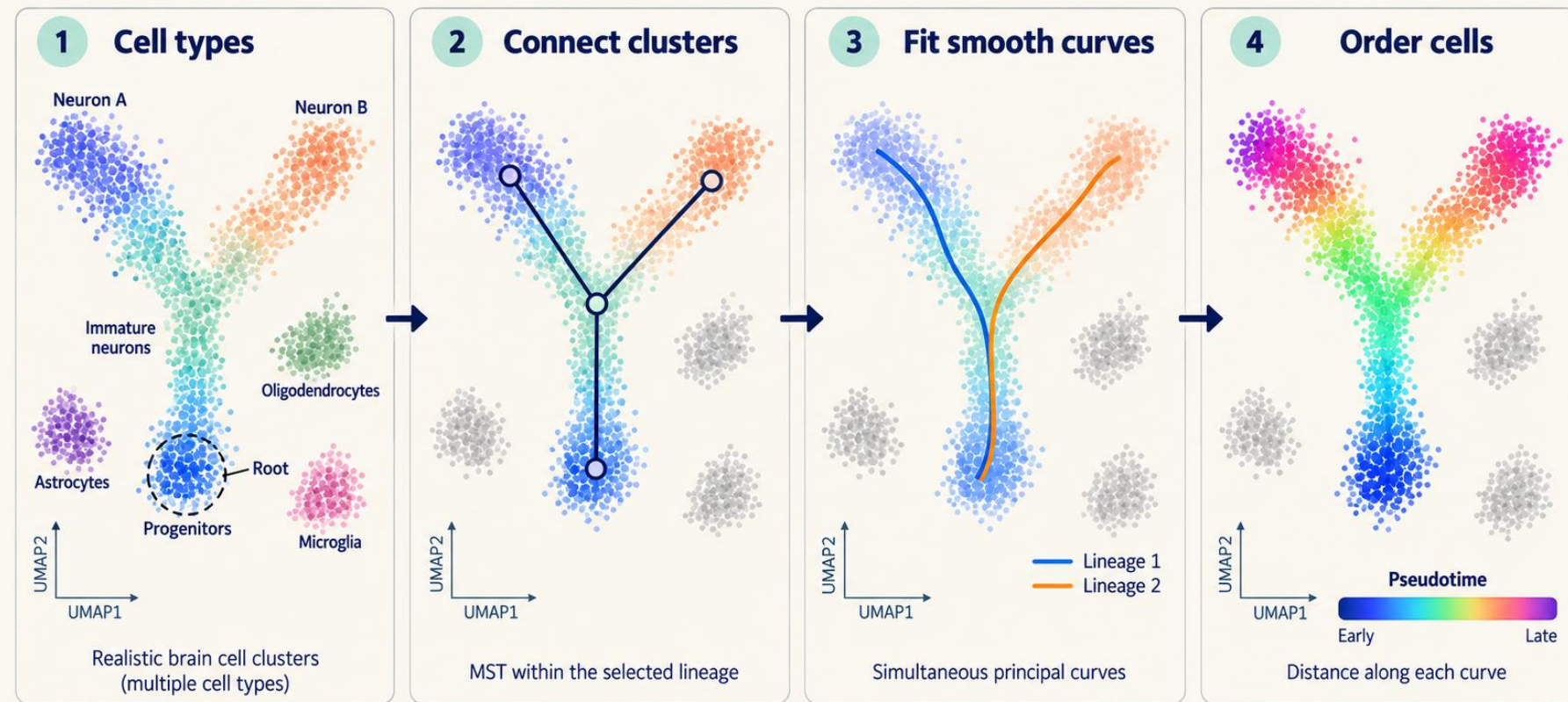


Trajectory Analysis

How do cell states progress?



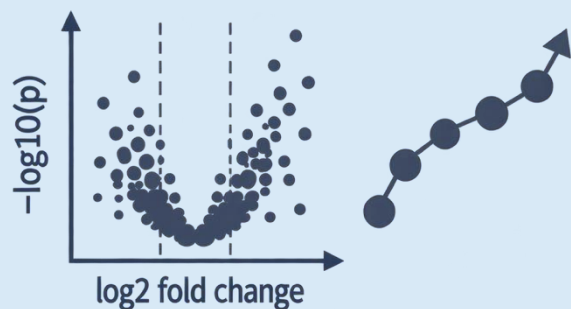
Slingshot: from cell clusters to pseudotime



Schematic brain UMAP • Trajectory fitted within a selected neuronal lineage

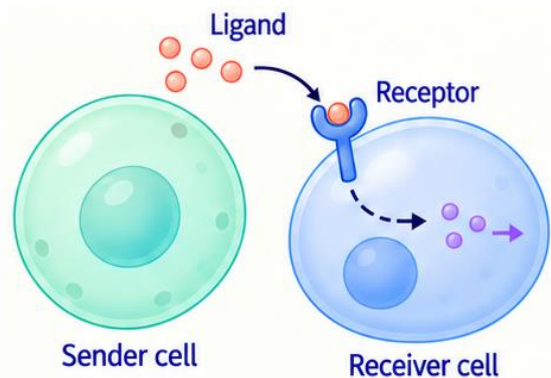
→ 4 Biological Inference

What changes, and what does it mean?

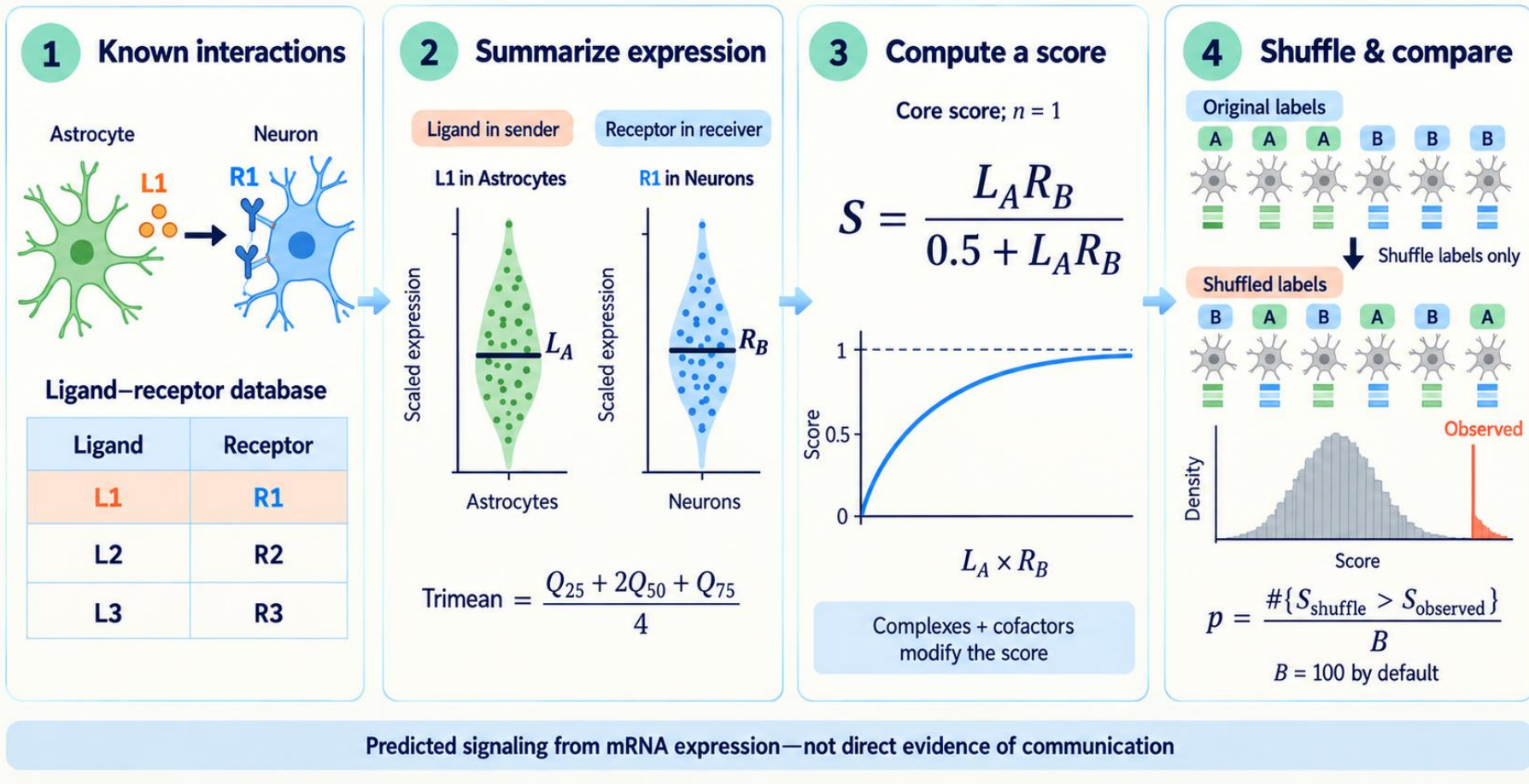


Cell-Cell Communication

Which signaling interactions are predicted?



CellChat: inferring cell-cell communication

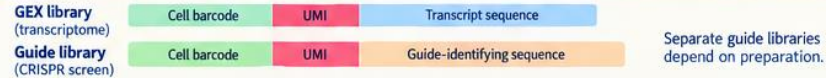


Single-cell RNA-seq Analysis – Complete Cheat Sheet

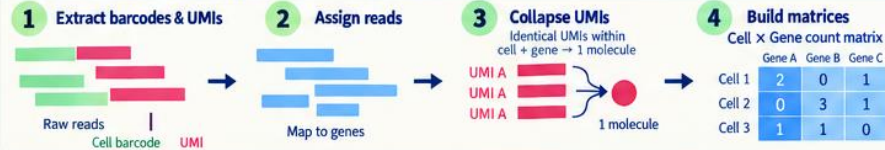
From raw reads to biological insight

1 PRIMARY PROCESSING

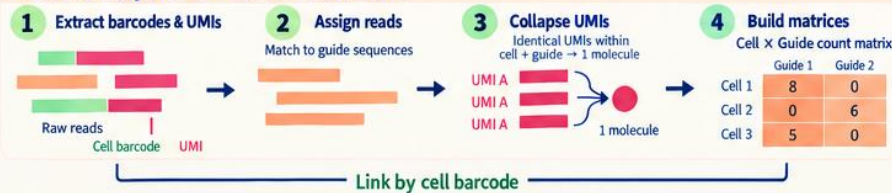
How do reads become counts?



Transcriptome (GEX) reads → Cell × Gene matrix



Guide library reads → Cell × Guide matrix



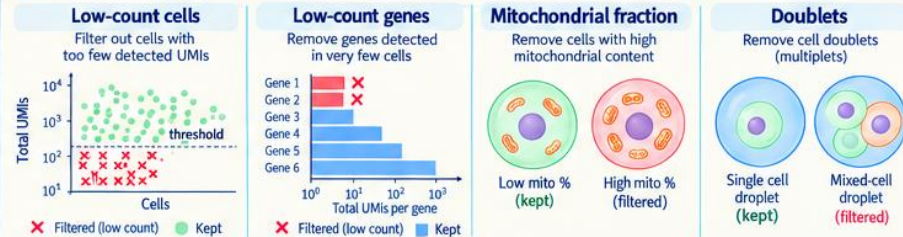
Output: Count matrices

2 QC & PREPROCESSING

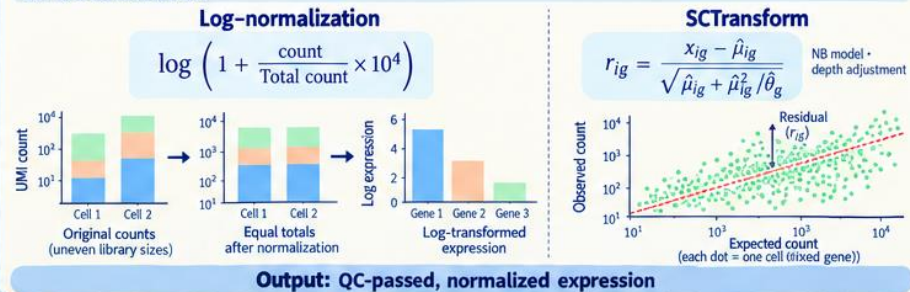
Which cells and counts can we trust?

a. QC Filtering

Context-dependent QC thresholds

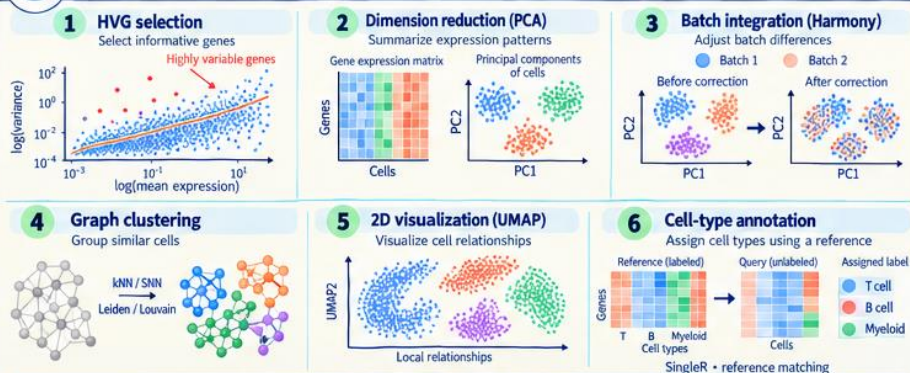


b. Normalization



3 CELL LANDSCAPE

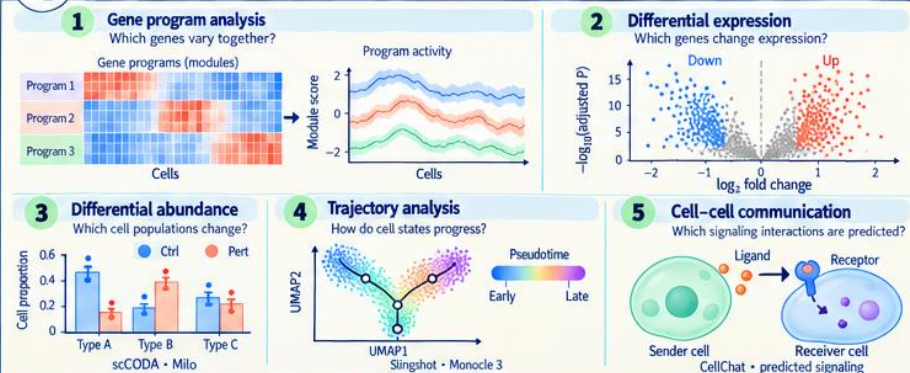
What cell types and states are present?



Output: Cell types & states

4 BIOLOGICAL INFERENCE

What changes, and what does it mean?



Output: Biological insight

Thanks!

