Transcriptómica II 2025 single-cell RNA-seq

PRÁCTICO 2

Natalia Rego

nrego@fcien.edu.uy

bajar datos del DRIVE y arrancar script Colab

DRIVE

https://drive.google.com/drive/folders/1ltESbSyw1VTO_irAXwPtrCajhAcZVrvz?usp=sharing

DEJAR CORRIENDO INSTALACIÓN

seminarios para 19 y 20 de noviembre

CattaPreta&2024 Leishmania: Agustín y Emilia (19)

Hutchinson&2021 VSGs

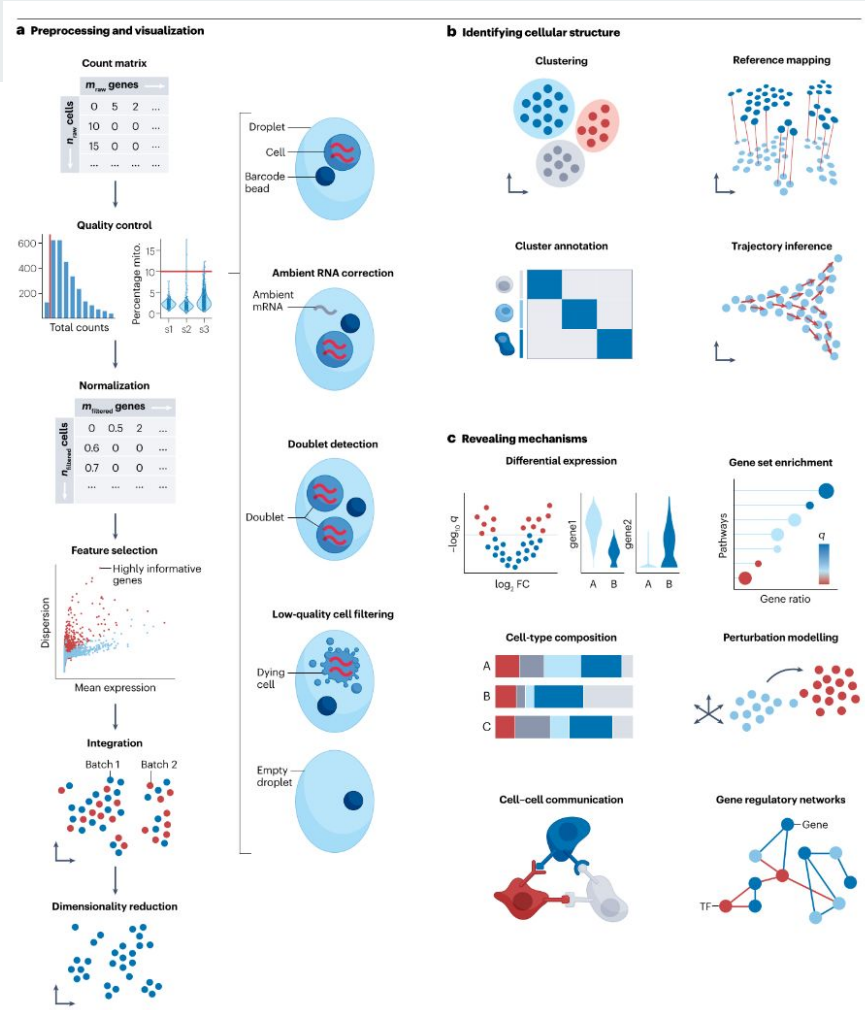
Hammond&2019 microglia

Randolph&2021 genetic ancestry: Sofía y Felipe

Roux&2023 C elegans

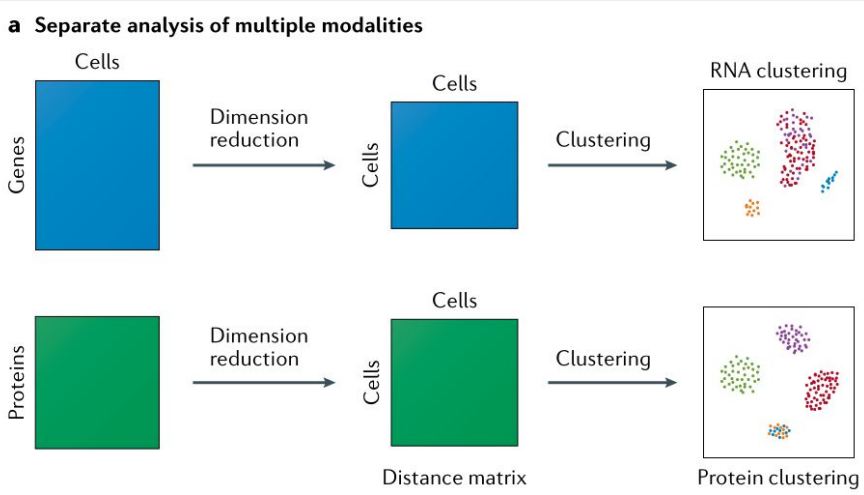
Segerstolpe&2016 type 2 diabetes; Laura y Antonella

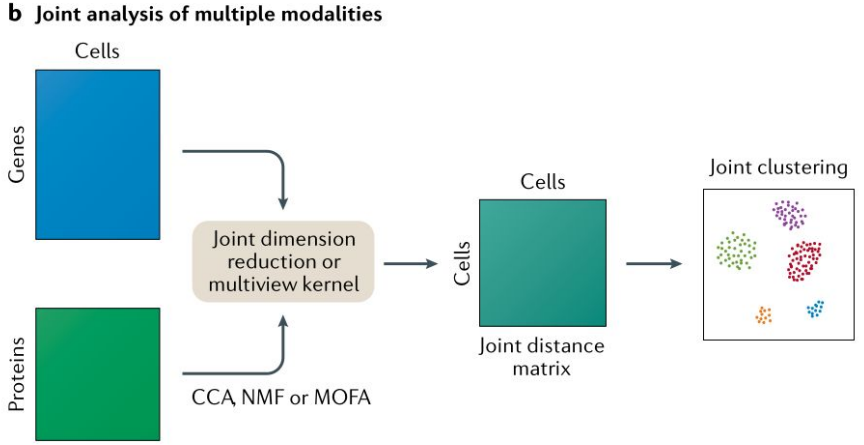
Tosches&2018 pallium evolution


a Preprocessing and visualization
Count matrix
Quality control
Total counts
Percentage mito
Normalization
Feature selection
Highly informative genes
Dispersion
Mean expression
Integration
Batch 1
Batch 2
Dimensionality reduction
Droplet
Cell
Barcode bead
Ambient RNA correction
Ambient mRNA
Doublet detection
Doublet
Low-quality cell filtering
Dying cell
Empty droplet
b Identifying cellular structure
Clustering
Reference mapping
Cluster annotation
Trajectory inference
c Revealing mechanisms
Differential expression
Gene set enrichment
Pathways
Gene ratio
Cell-type composition
Perturbation modelling
Cell–cell communication
Gene regulatory networks
Gene
TF

Data Integration

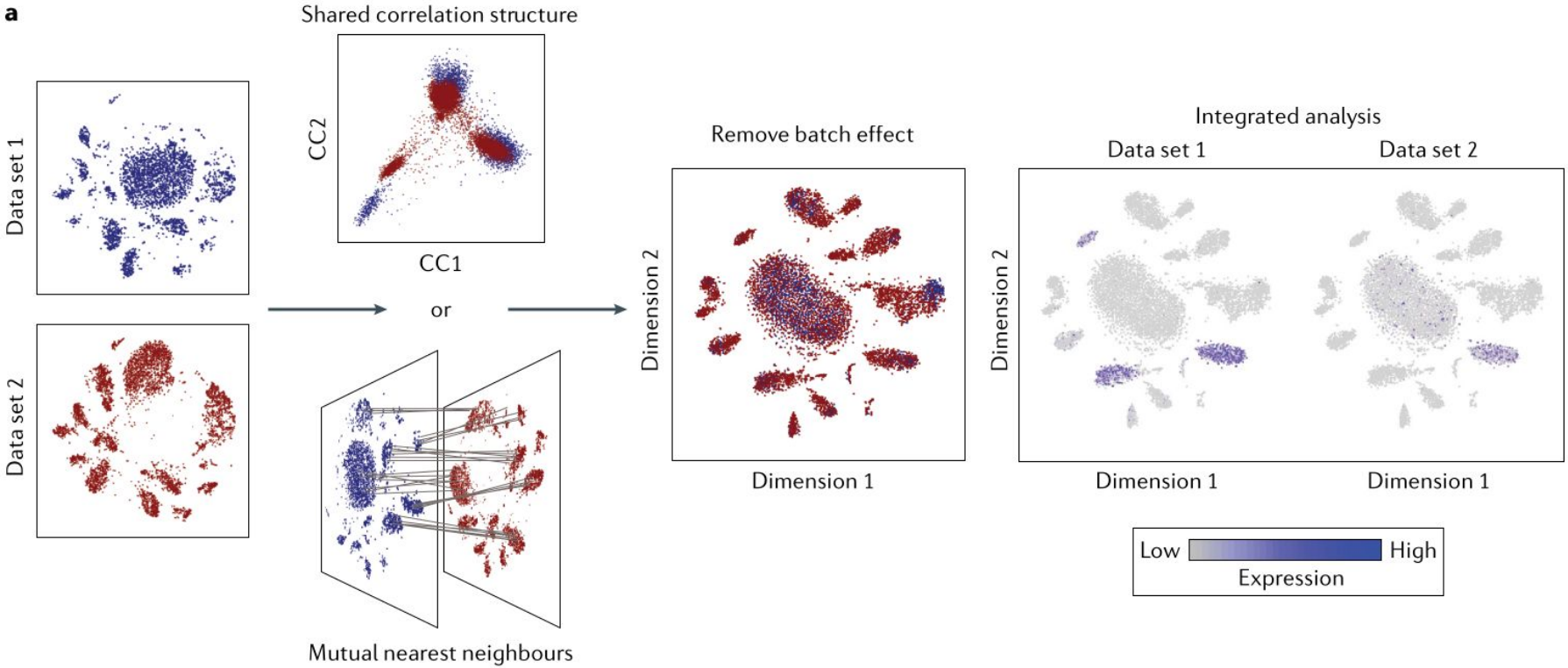
- multi-modal
  - CITE-seq
  - scRNA-seq & scATAC-seq
  - scRNA-seq & spatial
  - etc
- multiple samples
  - batch effects
  - different patients
  - different conditions & treatments
  - different experiments / labs
  - different technologies (e.g. smart-Seq2 and 10X)
  - different species



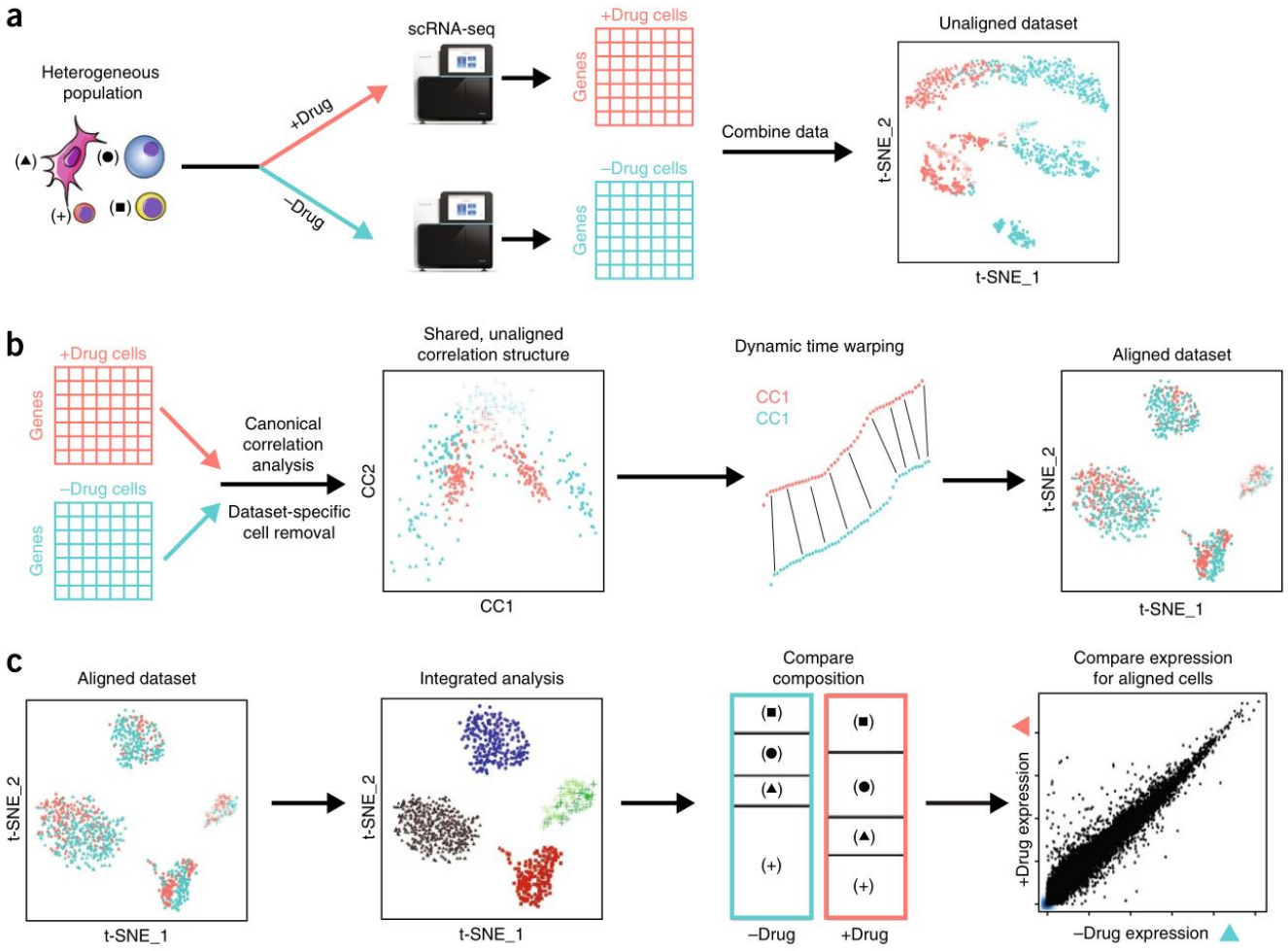


multimodal data sets are likely to reveal subtle differences in cell state that cannot be captured by a single modality alone

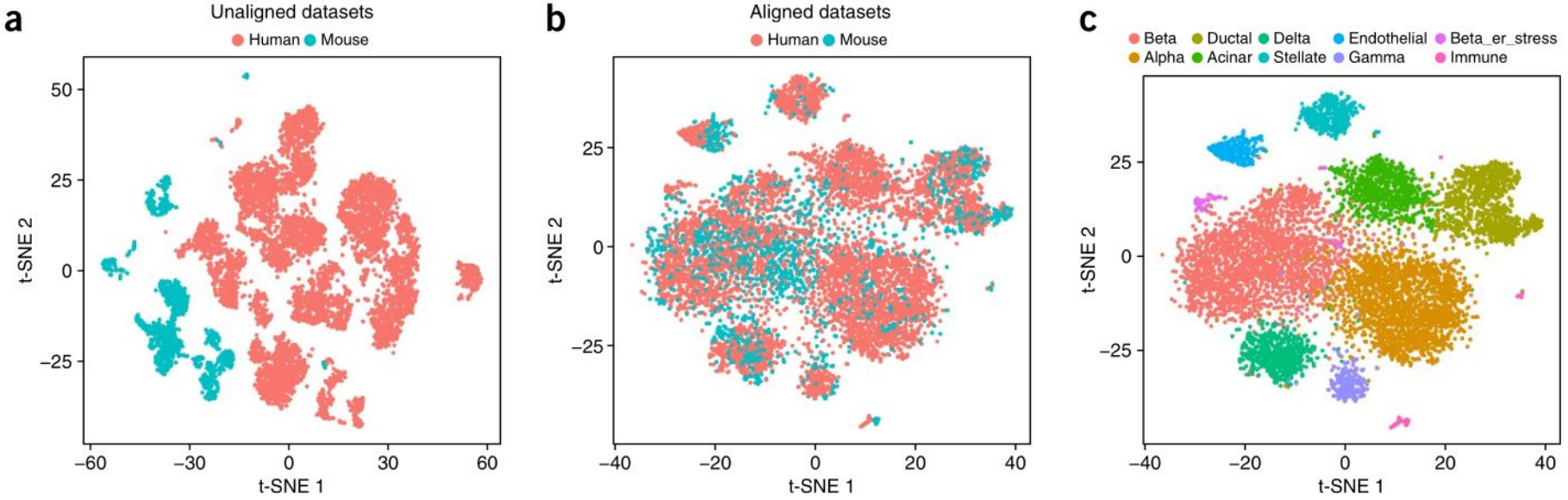
Q: e.g. CITE-seq: RNA & surface protein??
e.g. RNA-seq + ATAC-seq?



identification of either a shared space or equivalent cells across groups (e.g. using CCA or MNNs) can then be used to eliminate batch-specific variation, enabling direct comparison between the groups



Butler et al 2018


a
Unaligned datasets
Human Mouse
t-SNE 2
t-SNE 1
b
Aligned datasets
Human Mouse
t-SNE 2
t-SNE 1
c
Beta Ductal Delta Endothelial Beta_er_stress
Alpha Acinar Stellate Gamma Immune
t-SNE 2
t-SNE 1

Fast, sensitive and accurate integration of single-cell data with Harmony

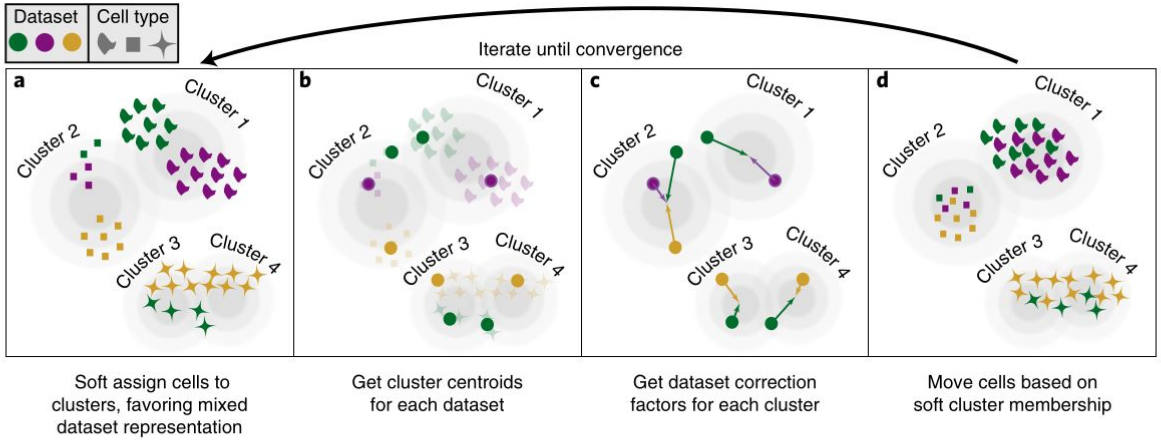


Fig. 1 | Overview of Harmony algorithm. PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects. **a**, Harmony uses fuzzy clustering to assign each cell to multiple clusters, while a penalty term ensures that the diversity of datasets within each cluster is maximized. **b**, Harmony calculates a global centroid for each cluster, as well as dataset-specific centroids for each cluster. **c**, Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. **d**, Finally, Harmony corrects each cell with a cell-specific factor: a linear combination of dataset correction factors weighted by the cell's soft cluster assignments made in step **a**. Harmony repeats steps **a** to **d** until convergence. The dependence between cluster assignment and dataset diminishes with each round. Datasets are represented with colors, cell types with different shapes.

Efficient integration of heterogeneous single-cell transcriptomes using Scanorama

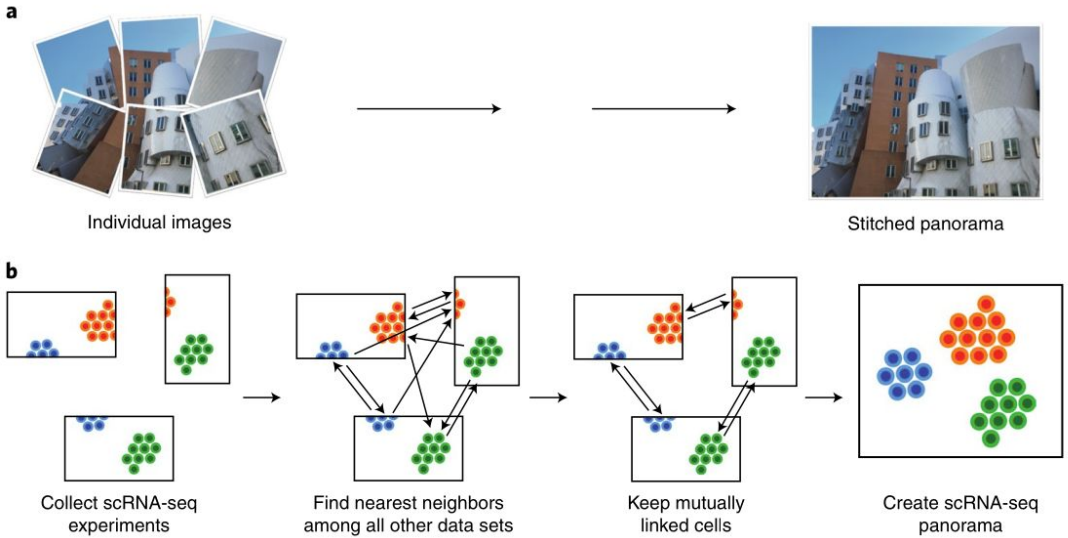


Fig. 1 | Illustration of 'panoramic' dataset integration. a, A panorama stitching algorithm finds and merges overlapping images to create a larger, combined image. **b**, A similar strategy can also be used to merge heterogeneous scRNA-seq datasets. Scanorama searches nearest neighbors to identify shared cell types among all pairs of datasets. Dimensionality reduction techniques and an approximate nearest-neighbors algorithm based on hyperplane locality sensitive hashing and random projection trees greatly accelerates the search step. Mutually linked cells form matches that can be leveraged to correct for batch effects and merge experiments together (Methods), whereby the datasets forming connected components on the basis of these matches become a scRNA-seq 'panorama'.

Cell type annotation

- It can be performed with **manual** or **automatic** approaches
  a three-step approach is recommended:
  automated annotation –> expert manual annotation –> verification step

- Manual approach using gene markers (issue with P values, known markers, time consuming)

- Automated approaches:

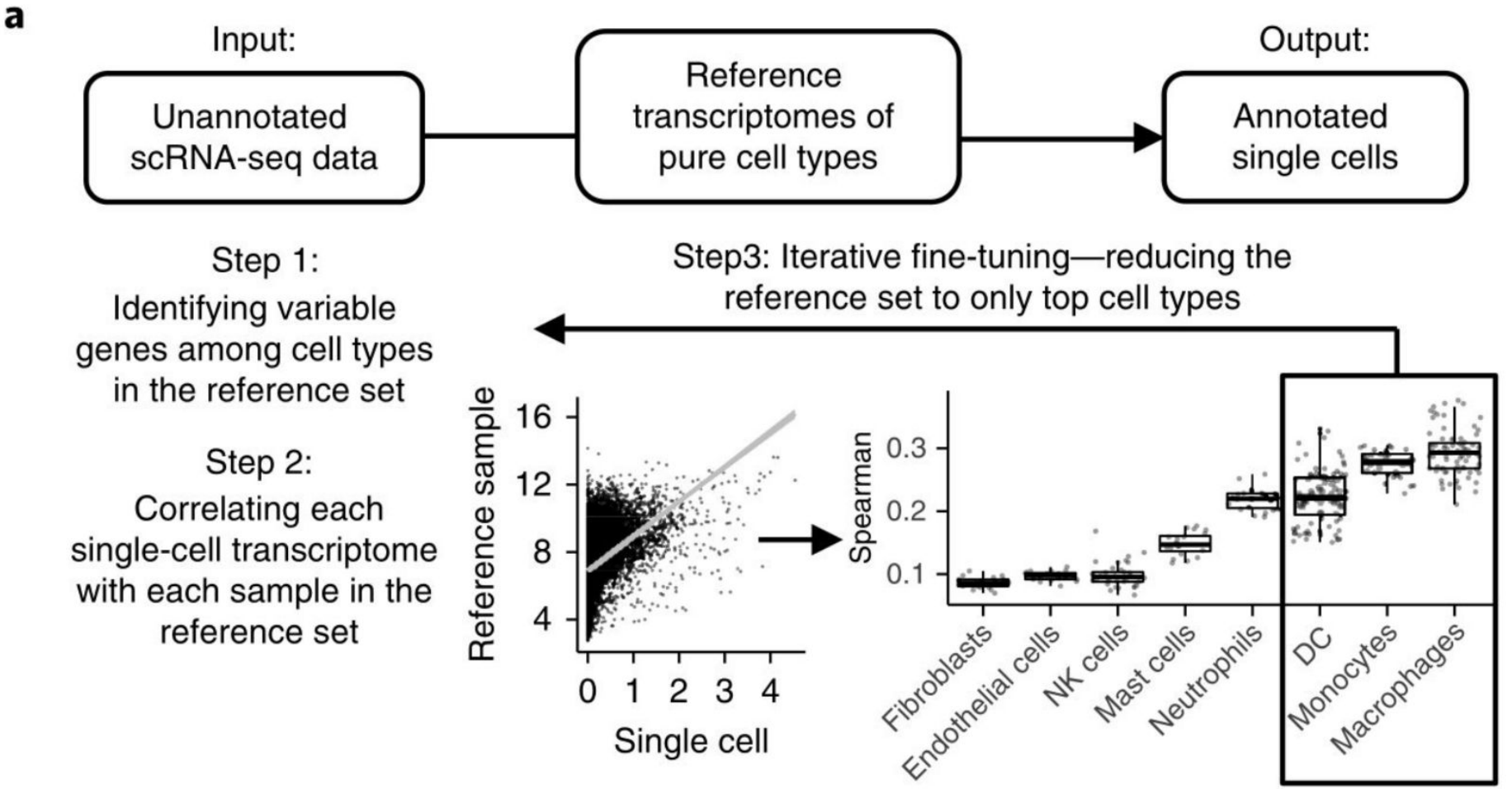
  classifier-based methods: e.g. CellTypist
  pre-trained classifiers; results depends on method and reference data

  reference mapping: e.g Azimuth (Seurat)
  mapping to existing, annotated single-cell reference and performing label transfer on the resulting join embedding

SingleR

SingleR is an automatic annotation method for single-cell RNA sequencing (scRNAseq) data (Aran et al. 2019). Given a reference dataset of samples (single-cell or bulk) with known labels, it labels new cells from a test dataset based on similarity to the reference. Thus, the burden of manually interpreting clusters and defining marker genes only has to be done once, for the reference dataset, and this biological knowledge can be propagated to new datasets in an automated manner.

`https://bioconductor.org/packages/release/bioc/html/SingleR.html`

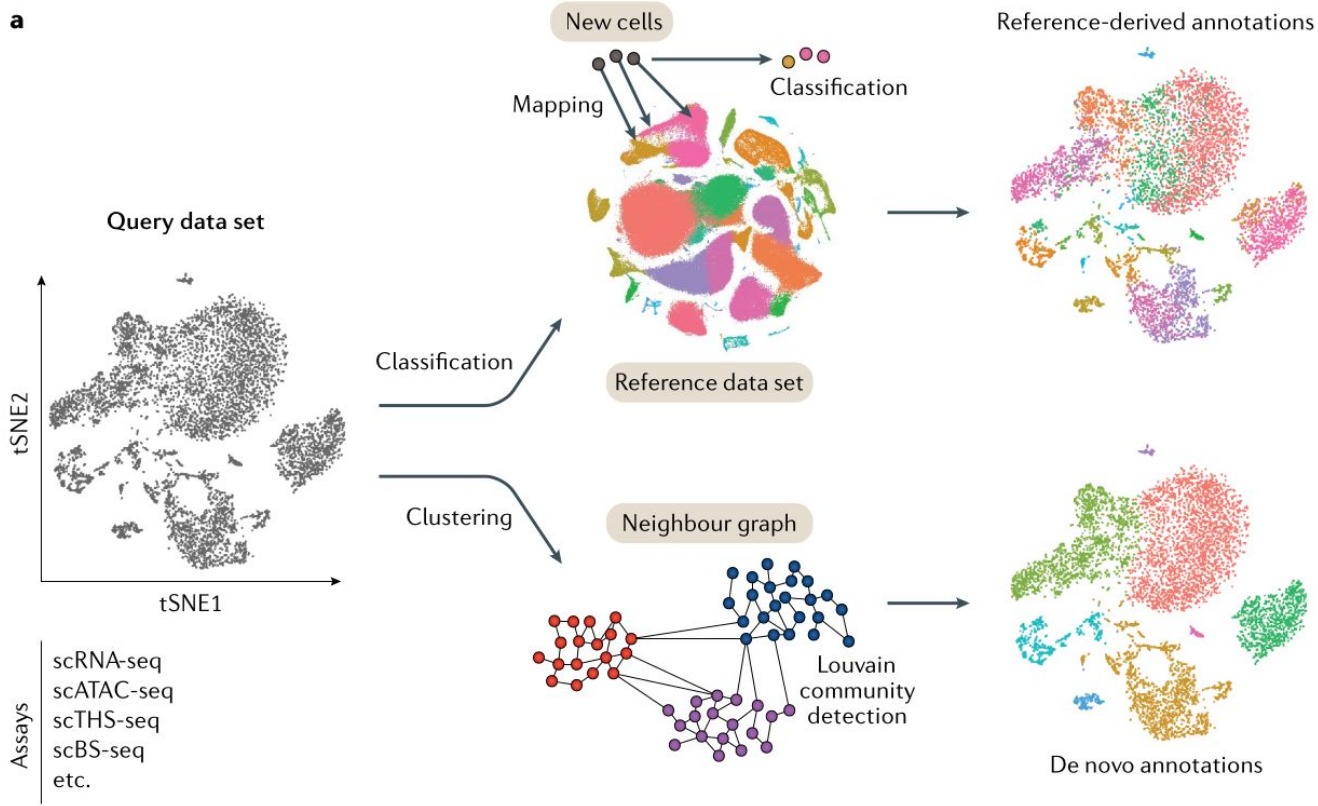

a
Input:
Unannotated scRNA-seq data
Reference transcriptomes of pure cell types
Output:
Annotated single cells
Step 1:
Identifying variable genes among cell types in the reference set
Step 2:
Correlating each single-cell transcriptome with each sample in the reference set
Step3: Iterative fine-tuning—reducing the reference set to only top cell types
Reference sample
16
12
8
4
0 1 2 3 4
Single cell
Spearman
0.3
0.2
0.1
Fibroblasts
Endothelial cells
NK cells
Mast cells
Neutrophils
DC
Monocytes
Macrophages

SingleR

the `SingleR()` function: identifies marker genes from the reference and uses them to compute assignment scores (based on the Spearman correlation across markers) for each cell in the test dataset against each label in the reference. The label with the highest score is the assigned to the test cell, possibly with further fine-tuning to resolve closely related labels.

`plotScoreHeatmap()` displays the scores for all cells across all reference labels, which allows users to inspect the confidence of the predicted labels across the dataset. Ideally, each cell (i.e., column of the heatmap) should have one score that is obviously larger than the rest, indicating that it is unambiguously assigned to a single label. A spread of similar scores for a given cell indicates that the assignment is uncertain, though this may be acceptable if the uncertainty is distributed across similar cell types that cannot be easily resolved.

Using cell atlases as a reference for cell assignment: projecting cells onto an existing dataset to facilitate the transfer of cell labels



3 Teóricos, 4 Prácticos, 2 jornadas de seminarios, 3 hs c/u

TRANSCRIPTÓMICA II, SINGLE-CELL RNA-seq código B0058				
TEÓRICO 03/11 - 06/11/25	lunes	16:00 - 19:00	salón 201/203	
	miércoles	16:00 - 19:00	salón 209	
	jueves	16:00 - 19:00	salón 209	
PRÁCTICO 10/11 - 20/11	lunes	16:00 - 19:00	salón 107	
	miércoles	16:00 - 19:00	salón 109	
	jueves	16:00 - 19:00	salón 107	

Linux:

https://swcarpentry.github.io/shell-novice/key-points.html

R:

https://www.rforbiologists.org/

https://melbournebioinformatics.github.io/r-intro-biologists/intro_r_biologists.html

Google Colab:

https://colab.research.google.com/

https://www.youtube.com/watch?v=inN8seMm7UI

https://www.youtube.com/watch?v=FXKMmilL70w