Transcriptómica II 2025 single-cell RNA-seq

PRÁCTICO I

Natalia Rego

nrego@fcien.edu.uy

bajar datos del DRIVE y arrancar script Colab

DRIVE

https://drive.google.com/drive/folders/18svEcgk20MucLScn4fU72US8Bxdvh_A0?usp=sharing

DEJAR CORRIENDO INSTALACIÓN

seminarios para 19 y 20 de noviembre

CattaPreta&2024 Leishmania

Hutchinson&2021 VSGs

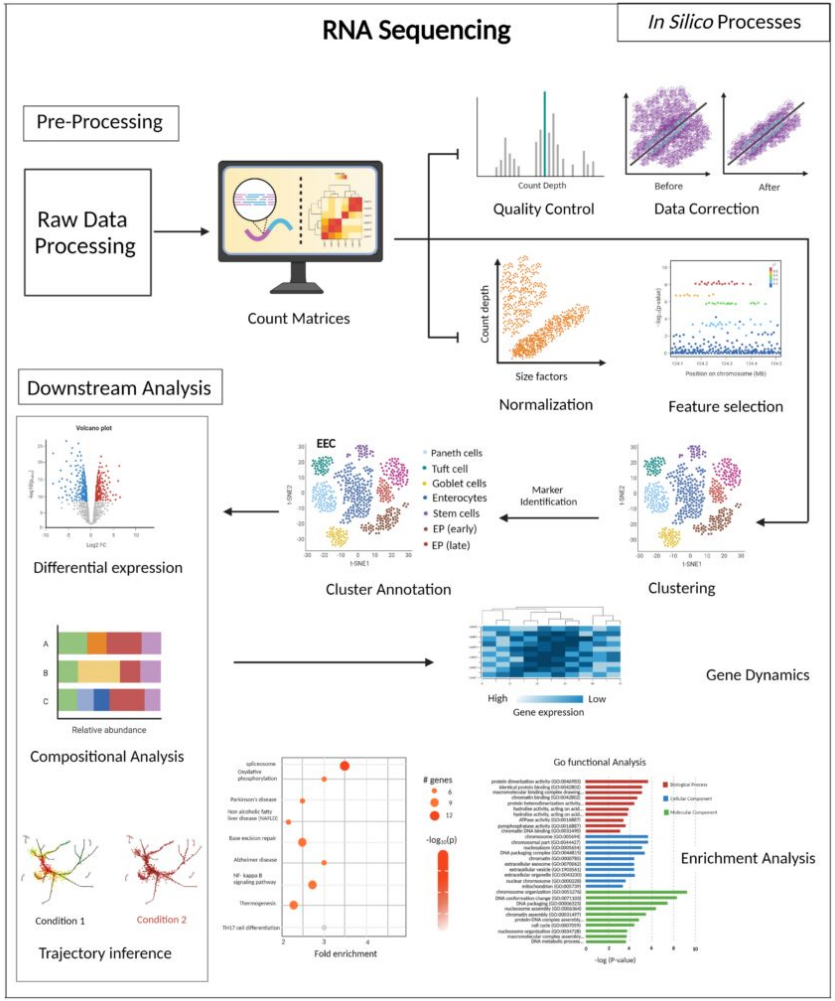
Hammond&2019 microglia

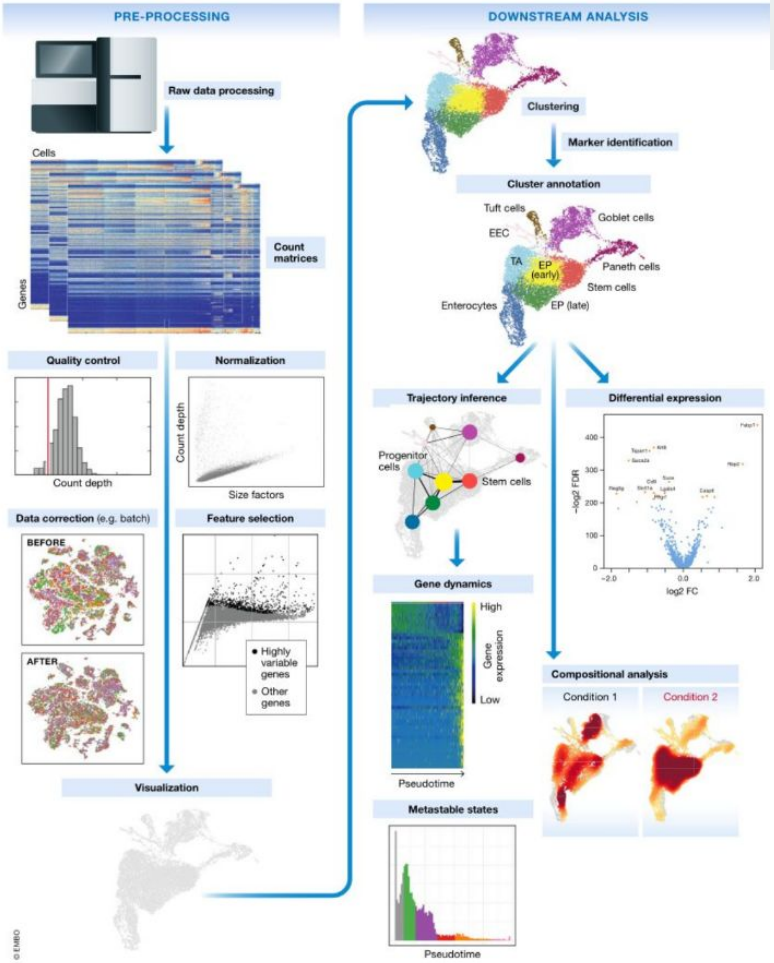
Randolph&2021 genetic ancestry

Roux&2023 C elegans

Segerstolpe&2016 type 2 diabetes

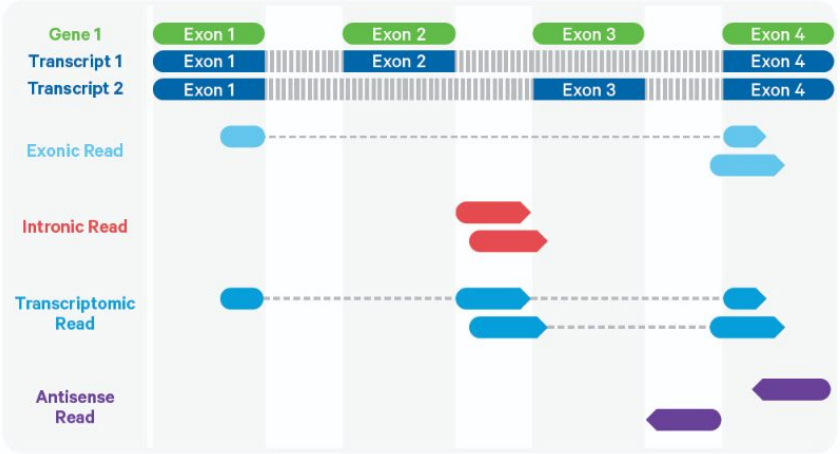
Tosches&2018 pallium evolution


RNA Sequencing
In Silico Processes
Pre-Processing
Raw Data Processing
Count Matrices
Count Depth
Quality Control
Before
After
Data Correction
Count depth
Size factors
Normalization
Feature selection
Downstream Analysis
Volcano plot
Differential expression
EEC
Paneth cells
Tuft cell
Goblet cells
Enterocytes
Stem cells
EP (early)
EP (late)
Cluster Annotation
Marker Identification
Clustering
Relative abundance
Compositional Analysis
Gene Dynamics
High
Low
Gene expression
Condition 1
Condition 2
Trajectory inference
Fold enrichment
genes
Go functional Analysis
Enrichment Analysis

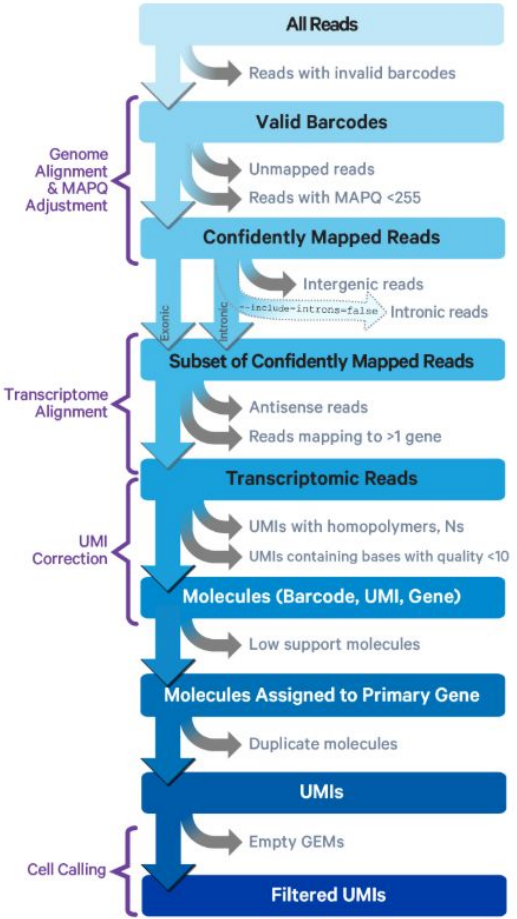

PRE-PROCESSING
DOWNSTREAM ANALYSIS
Raw data processing
Cells
Genes
Count matrices
Quality control
Count depth
Normalization
Count depth
Size factors
Data correction (e.g. batch)
BEFORE
AFTER
Feature selection
Highly variable genes
Other genes
Visualization
Clustering
Marker identification
Cluster annotation
Tuft cells
Goblet cells
EEC
TA
EP (early)
Paneth cells
Stem cells
Enterocytes
EP (late)
Trajectory inference
Progenitor cells
Stem cells
Differential expression
−log2 FDR
log2 FC
Gene dynamics
High
Low
Gene expression
Pseudotime
Compositional analysis
Condition 1
Condition 2
Metastable states
Pseudotime
© EMBO

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger (10X Genomics)

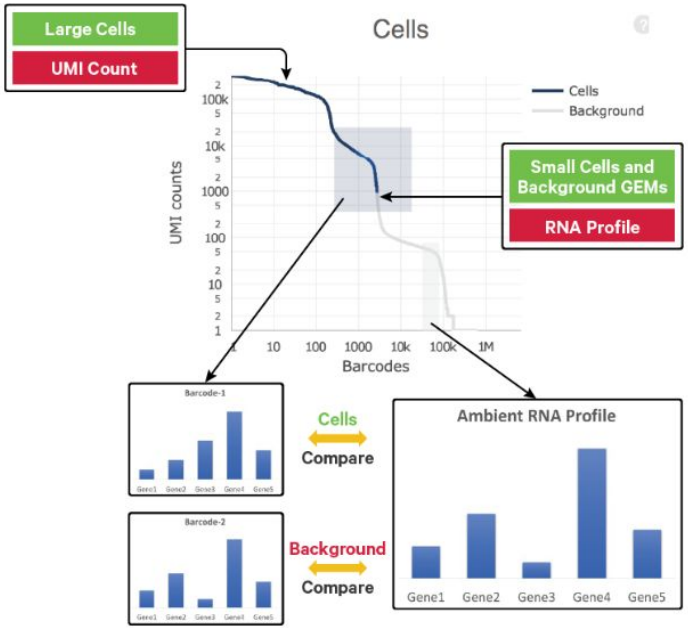


*A read is classified as exonic if at least 50% of it overlaps with an exon



https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger (10X Genomics)



The EmptyDrops algorithm uses RNA profile information of the cells (built based on a background model) along with total UMI counts to distinguish between background RNA and cellular RNA UMIs.

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger (10X Genomics)

```
pwd
#/mnt/TORTUGAS/TORTUGA_GENOMA
cellranger mkref --nthreads 6 --genome=cellranger_TSEv1
--fasta=/mnt/cuatroteras1/TORTUGAS/TORTUGA_GENOMA/NCBI_20may24/ncbi_dataset/data/GCF_013100865.1/GCF_0131008
65.1_CAS_Tse_1.0_genomic.fna \
--genes=/mnt/cuatroteras1/TORTUGAS/0_PALLIUM/2_star_se_MACE/turtle_mace_utr-override_2500Will_GTF.gtf


pwd
#/mnt/singlecell/natalia/TORTUGASsc_MAR24
cellranger count --localcores=8 --localmem=64 \
      --id=TSEmarch24 --sample=T05_CKDL240015316-1A_H3TNNDSXC \
      --transcriptome=/mnt/cuatroteras1/NATA/TORTUGAS/TORTUGA_GENOMA/cellranger_TSEv1 \
      --fastqs=/mnt/cuatroraid/NATA/TORTUGASsc/TORTUGASsc_march24_exp2 \
      --create-bam=true
```

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger (10X Genomics)

```
natalia@tlaloc:~$ cd /mnt/singlecell/natalia/TORTUGASsc_MAR24/TSEmarch24/outs/
natalia@tlaloc:/mnt/singlecell/natalia/TORTUGASsc_MAR24/TSEmarch24/outs$ ll
total 20070812
drwxr-sr-x. 7 natalia rstudio          98 jun 10  2024 analysis
-rw-r--r--. 1 natalia rstudio    10155483 jun 10  2024 cloupe.cloupe
-rw-r--r--. 1 natalia rstudio   522839662 jul  3  2024 filtered.bam
drwxr-sr-x. 2 natalia rstudio          89 jun 10  2024 filtered_feature_bc_matrix
-rw-r--r--. 1 natalia rstudio     1833024 jun 10  2024 filtered_feature_bc_matrix.h5
-rw-r--r--. 1 natalia rstudio        5287 jul  3  2024 header
-rw-r--r--. 1 natalia rstudio         643 jun 10  2024 metrics_summary.csv
-rw-r--r--. 1 natalia rstudio    31311367 jun 10  2024 molecule_info.h5
-rw-r--r--. 1 natalia rstudio      652615 jul  3  2024 my_filtered_fastqc.html
-rw-r--r--. 1 natalia rstudio      998949 jul  3  2024 my_filtered_fastqc.zip
-rw-r--r--. 1 natalia rstudio  3593359054 jul  3  2024 my_filtered.fq
-rw-r--r--. 1 natalia rstudio 12954512155 jun 10  2024 possorted_genome_bam.bam
-rw-r--r--. 1 natalia rstudio     4133088 jun 10  2024 possorted_genome_bam.bam.bai
drwxr-sr-x. 2 natalia rstudio          89 jun 10  2024 raw_feature_bc_matrix
-rw-r--r--. 1 natalia rstudio    12003179 jun 10  2024 raw_feature_bc_matrix.h5
-rw-r--r--. 1 natalia rstudio  3415962015 jul  3  2024 sam.tmp
-rw-r--r--. 1 natalia rstudio     4712090 jun 10  2024 web_summary.html
natalia@tlaloc:/mnt/singlecell/natalia/TORTUGASsc_MAR24/TSEmarch24/outs$
```

```
t/singlecell/natalia/TORTUGASsc_MAR24/TSEmarch24/outs$ ll filtered_feature_bc_matrix

lia rstudio   50968 jun 10  2024 barcodes.tsv.gz
lia rstudio  174733 jun 10  2024 features.tsv.gz
lia rstudio 3828973 jun 10  2024 matrix.mtx.gz
t/singlecell/natalia/TORTUGASsc_MAR24/TSEmarch24/outs$
```

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger EXP#1

Alerts

The analysis detected 1 error and 1 warning.

	Alert	Value	Detail
!	Low Fraction Reads in Cells	21.7%	Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using --force-cells.
⚠	Low Fraction Reads Confidently Mapped To Transcriptome	25.7%	Ideal > 30%. This can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.

Summary | Gene Expression

10,071
Estimated Number of Cells

17,856
Mean Reads per Cell

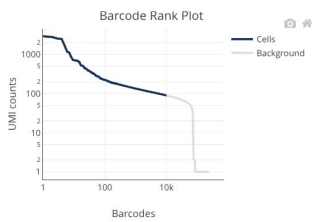
90
Median Genes per Cell

Run Summary

Sample ID	TSEmarch24
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Transcriptome	cellranger_TSEv1-
Pipeline Version	cellranger-8.0.1

Sequencing

Cells



Estimated Number of Cells	10,071
Fraction Reads in Cells	21.7%
Mean Reads per Cell	17,856
Median UMI Counts per Cell	101
Median Genes per Cell	90
Total Genes Detected	16,144

Sequencing

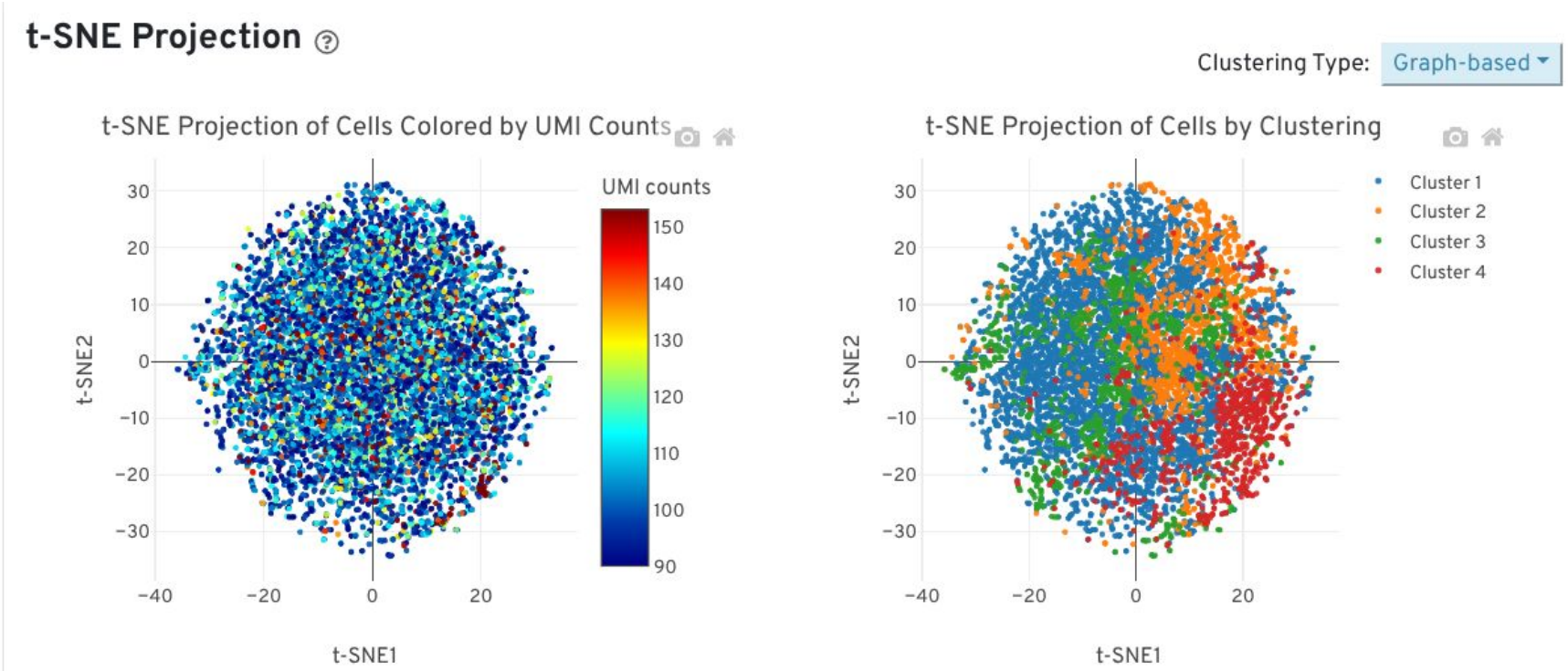
Number of Reads	179,830,397
Number of Short Reads Skipped	0
Valid Barcodes	95.5%
Valid UMIs	100.0%
Sequencing Saturation	88.4%
Q30 Bases in Barcode	95.6%
Q30 Bases in RNA Read	73.2%
Q30 Bases in UMI	91.1%

Mapping

Reads Mapped to Genome	70.1%
Reads Mapped Confidently to Genome	51.2%
Reads Mapped Confidently to Intergenic Regions	8.8%
Reads Mapped Confidently to Intronic Regions	32.7%
Reads Mapped Confidently to Exonic Regions	9.8%
Reads Mapped Confidently to Transcriptome	25.7%
Reads Mapped Antisense to Gene	16.7%

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger EXP#1



https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger EXP#2

Alerts

The analysis detected 1 error and 1 warning.

	Alert	Value	Detail
!	Low Fraction Reads in Cells	34.8%	Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using --force-cells.
⚠	High Fraction of Reads Mapped Antisense to Genes	27.7%	Ideal < 20%. Rates of up to 40% are common for single nuclei samples. Higher fraction of antisense reads may indicate use of an incorrect chemistry type, or an issue with the reference transcriptome.

Summary | Gene Expression

5,797
Estimated Number of Cells

43,242
Mean Reads per Cell

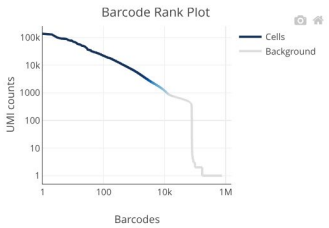
1,636
Median Genes per Cell

Run Summary

Sample ID	TSEaugust24
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Transcriptome	cellranger_TSEv1-
Pipeline Version	cellranger-8.0.1

Sequencing

Cells



Estimated Number of Cells	5,797
Fraction Reads in Cells	34.8%
Mean Reads per Cell	43,242
Median UMI Counts per Cell	2,822
Median Genes per Cell	1,636
Total Genes Detected	19,424

Sequencing

Number of Reads	250,675,601
Number of Short Reads Skipped	0
Valid Barcodes	94.4%
Valid UMIs	99.5%
Sequencing Saturation	32.9%
Q30 Bases in Barcode	94.6%
Q30 Bases in RNA Read	94.7%
Q30 Bases in UMI	94.5%

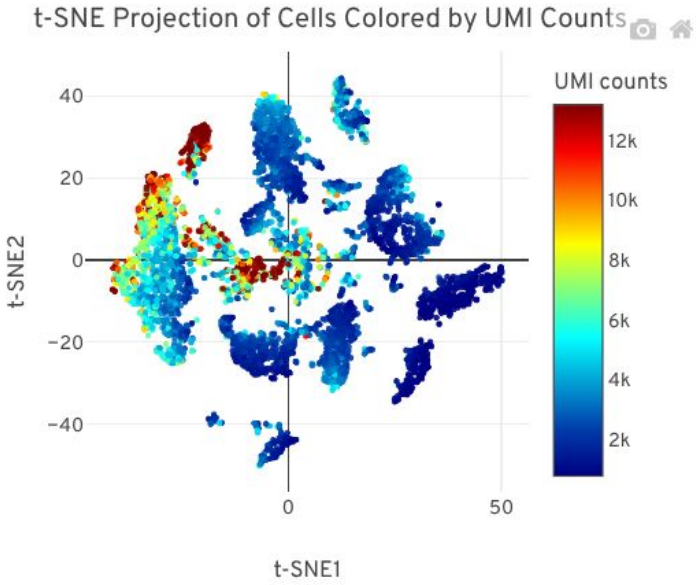
Mapping

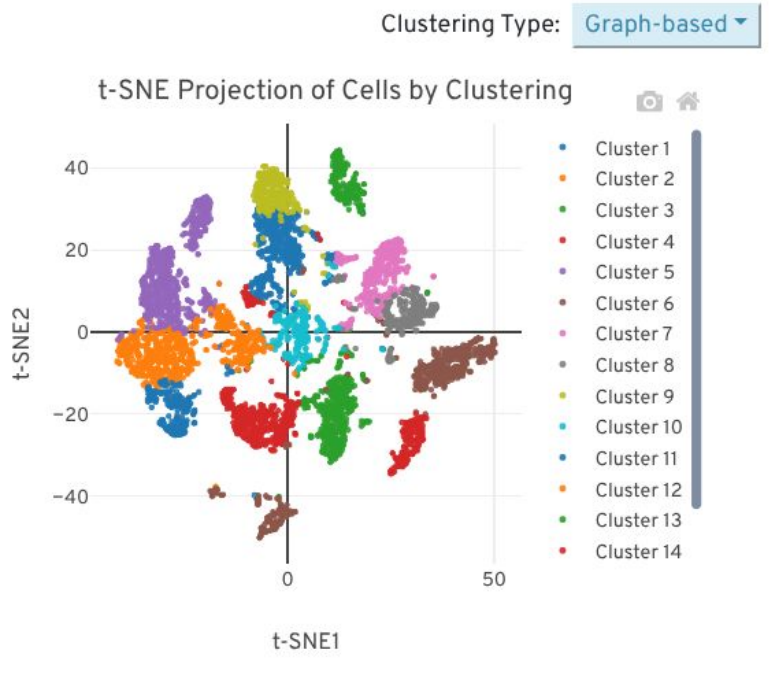
Reads Mapped to Genome	92.1%
Reads Mapped Confidently to Genome	89.4%
Reads Mapped Confidently to Intergenic Regions	14.7%
Reads Mapped Confidently to Intronic Regions	60.4%
Reads Mapped Confidently to Exonic Regions	14.2%
Reads Mapped Confidently to Transcriptome	46.8%
Reads Mapped Antisense to Gene	27.7%

https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

CellRanger EXP#2

t-SNE Projection





Otras herramientas

PROS: mayor flexibilidad, otras tecnologías, mayor control de parámetros, especies no modelo, reproducibilidad e integración con Snakemake/Nextflow.

- STARsolo
  - mapeo y conteo integrado
  - compatible con Seurat / Scanpy
- Kallisto | bustools (kb-python)
  - pseudoalineamiento y procesamiento de barcodes/UMIs
  - matrices de genes y transcriptos
- Salmon | Alevin-fry
  - pseudomapeo con modelos avanzados de errores de UMIs y barcodes
  - matrices de genes, transcriptos y clases de equivalencia

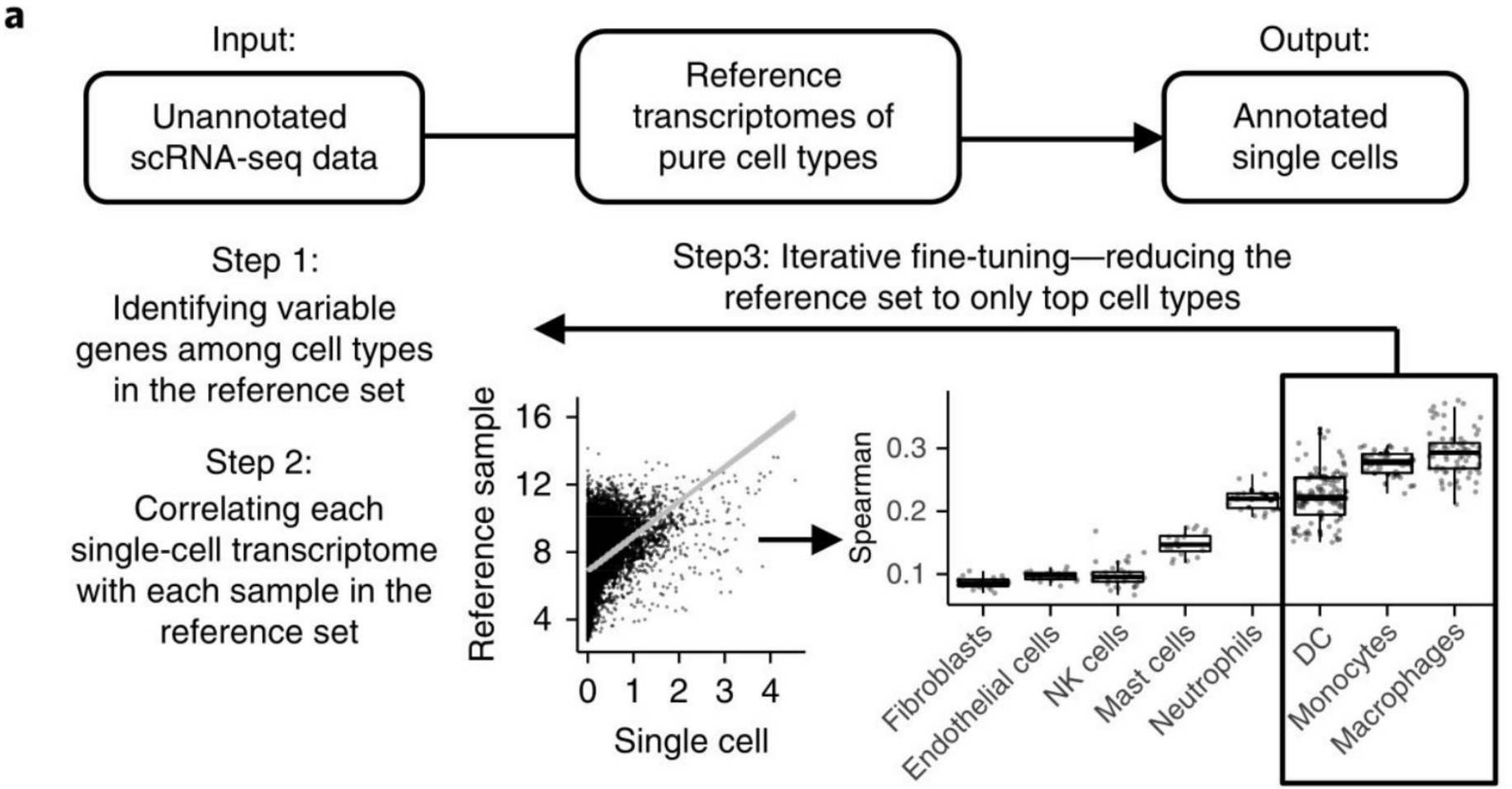
recomendable: probar STARsolo o/y Alevin-fry para organismos no modelos

CONS: no hacen cell calling con modelo probabilístico (knee-plots + heuristicas o `dropletUtils::emptyDrops`)

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.

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