Biologia Molecular Computacional
IBI5035/QBQ2507 - 2023

Análise do transcritoma de células únicas

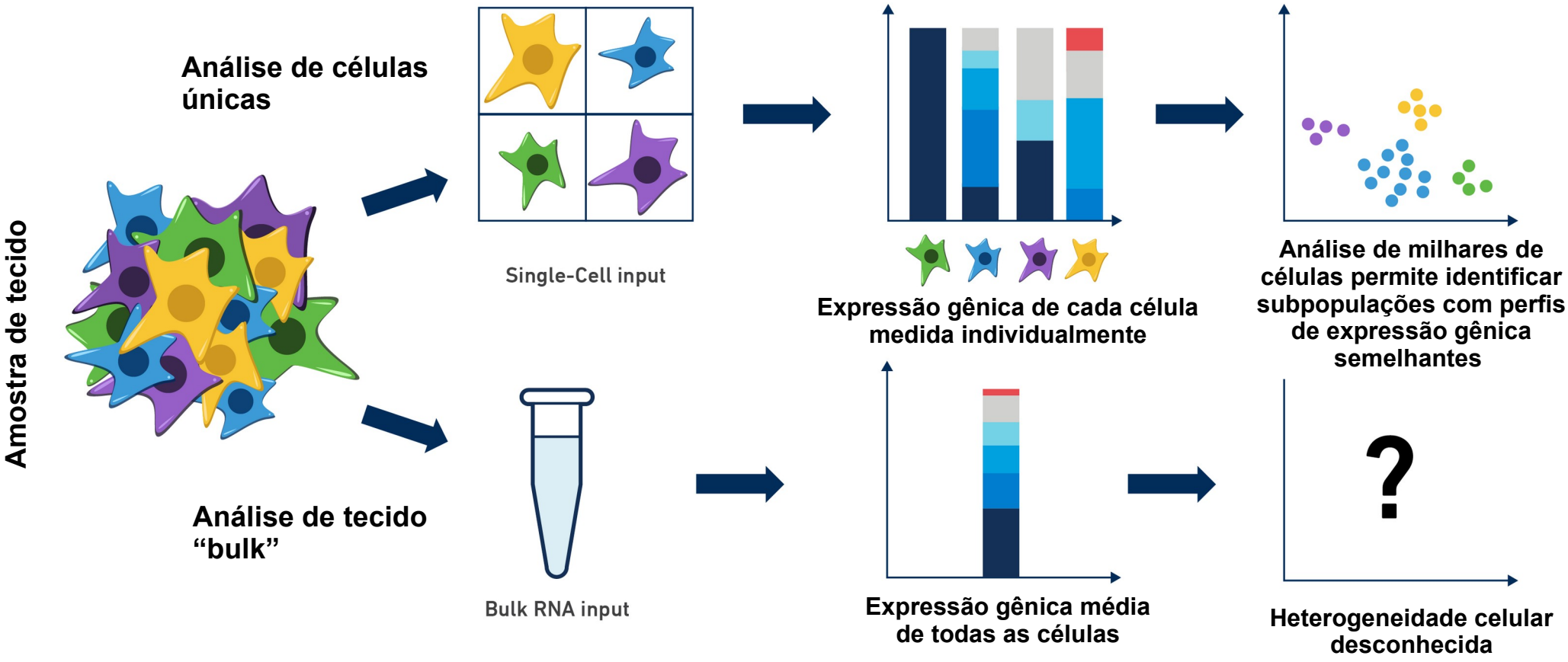
(tutorial)

Eduardo Moraes Rego Reis
Instituto de Química - USP

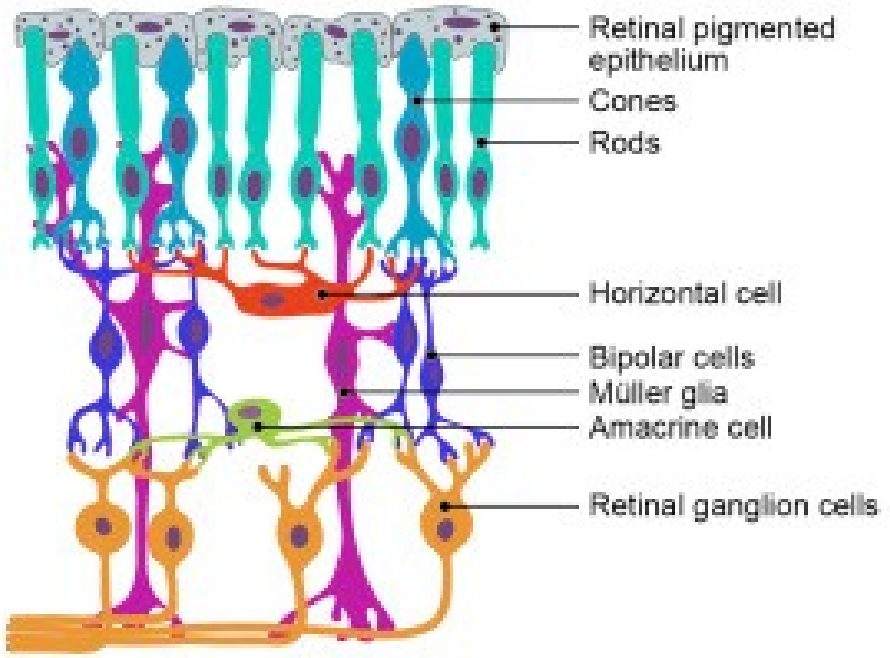
Estudos de transcritoma baseados em RNAseq de amostras de tecido (“bulk tissue”) tem a limitação de medirem a expressão média.

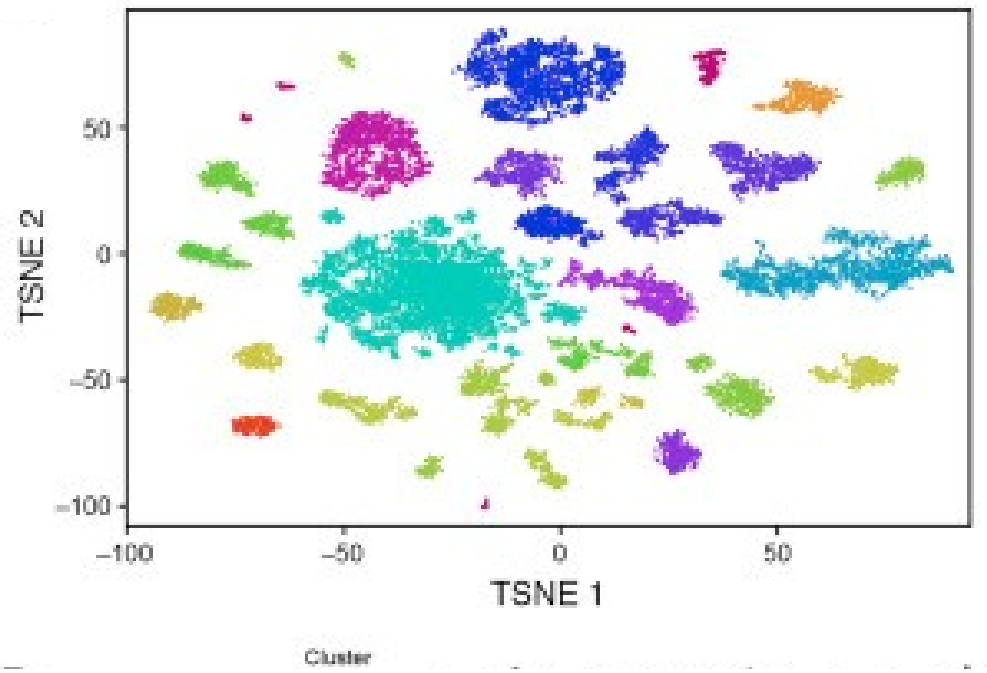
Não considera a existência de diferentes tipos celulares presentes no tecido

A abordagem de scRNA-seq revela a heterogeneidade celular que é mascarada em experimentos de RNAseq tradicional (“bulk tissue”)



Reconstrução do transcritoma de 39 tipos celulares da retina a partir de perfis de expressão gênica de 44,808 células por scRNA-seq

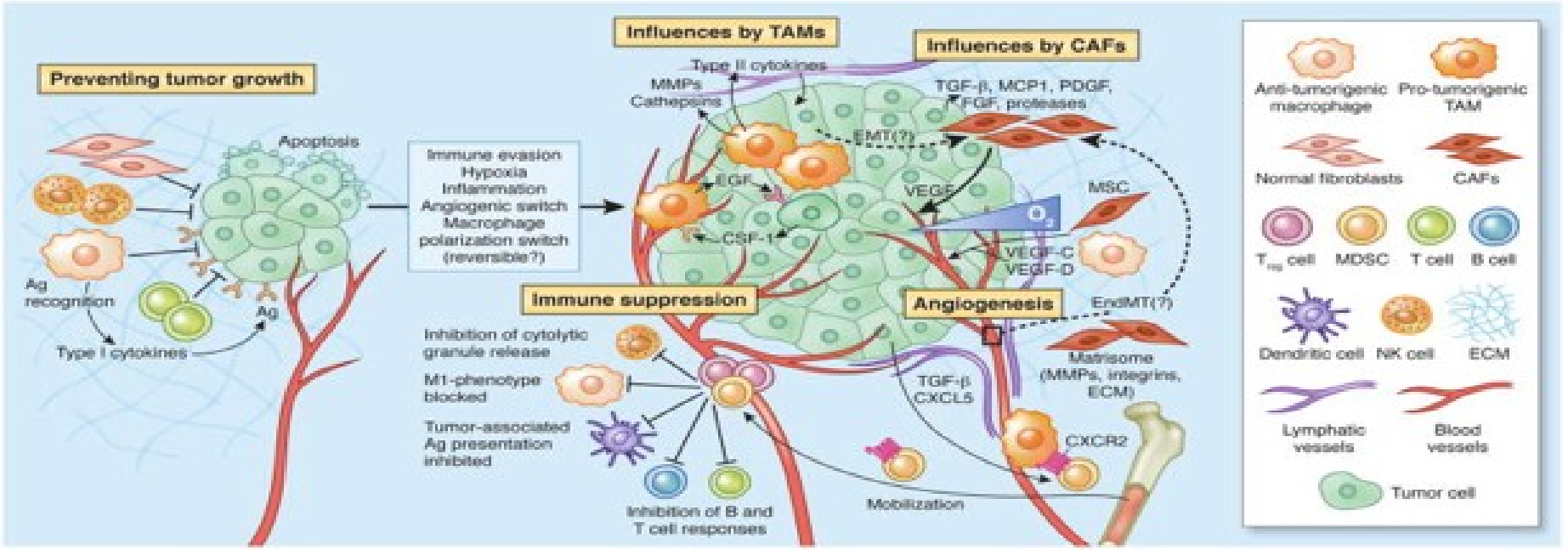




Cell, 2015 161:1202–1214

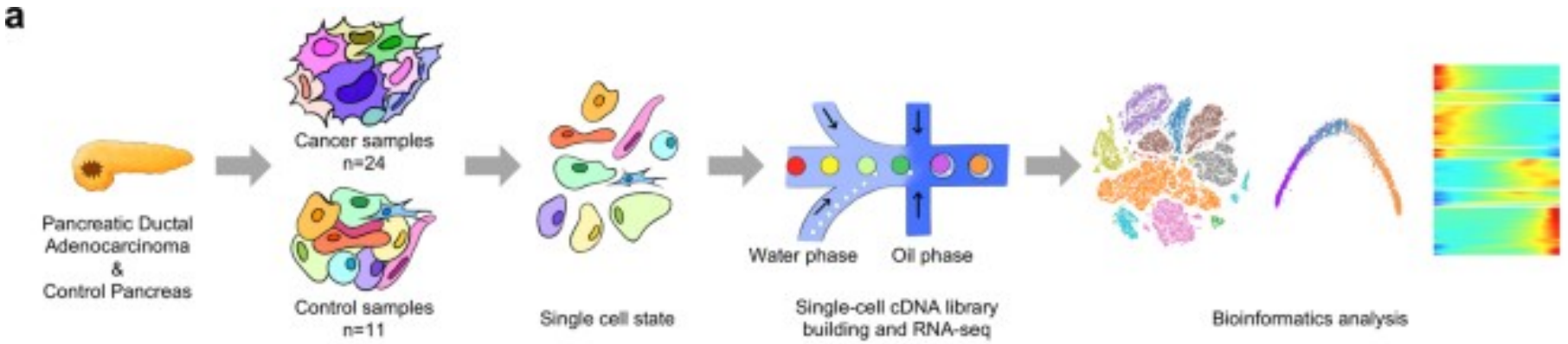
Diferenças transcricionais e funcionais entre células do mesmo tipo são muito relevantes na biologia

Diferentes tipos celulares presentes no estroma cooperam para estabelecer nichos que favorecem a tumorigênese e progressão do tumor



Quail and Joyce 2013 Nat Med. 19: 1423–1437

Subtipo celular exclusivo de PDAC identificado por análise transcriptômica de célula única.

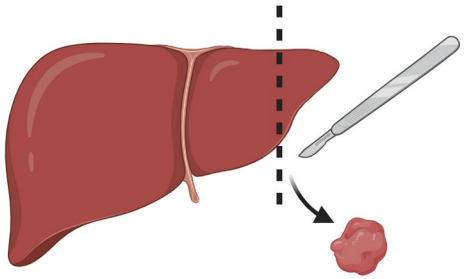


Cell Type	Number of Cells	Percentage
Ductal cell 1	2,646	6.30%
Ductal cell 2	11,315	26.95%
Acinar cell	512	1.22%
Endocrine cell	459	1.09%
Endothelial cell	5,134	12.23%
Fibroblast	5,802	13.82%
Stellate cell	5,284	12.59%
Macrophage	4,802	11.44%
T cell	3,616	8.61%
B cell	2,416	5.75%

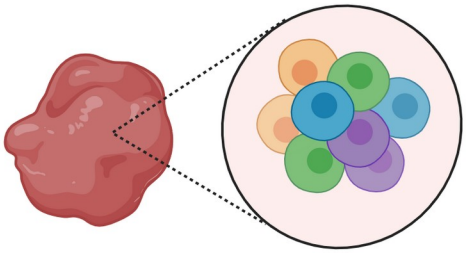
Peng at al., Cell Research volume 29, pages725–738 (2019)

Fluxograma de geração e análise de dados de scRNAseq

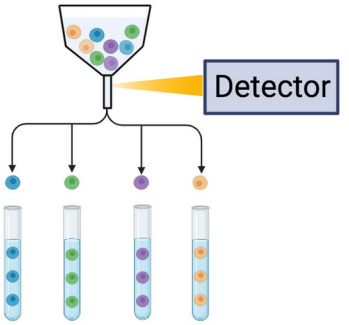
1 Fragmento de tecido



2 Método de dissociação das células presentes no tecido

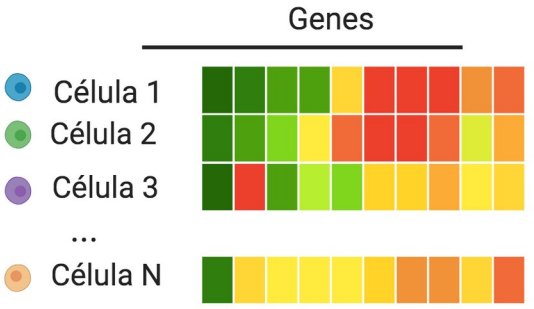


3 Separação das células únicas (microfluidica, citometria de fluxo

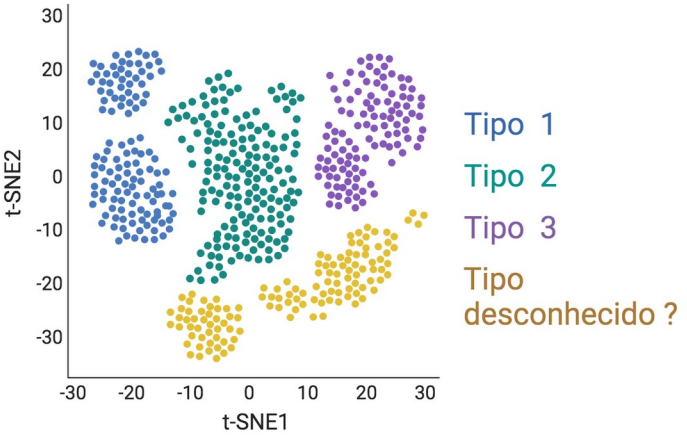


4 NGS

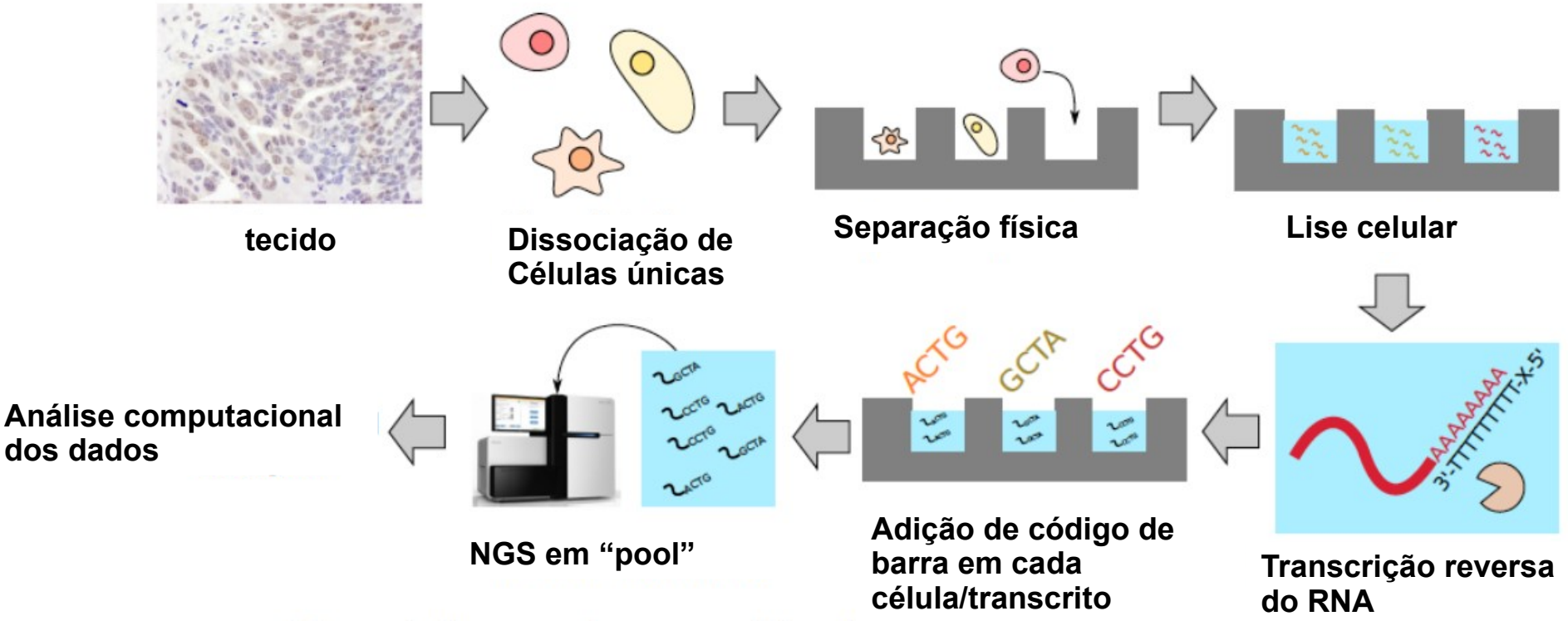
5 expressão gênica



6 Agrupamento ("clustering") e identificação de tipos celulares

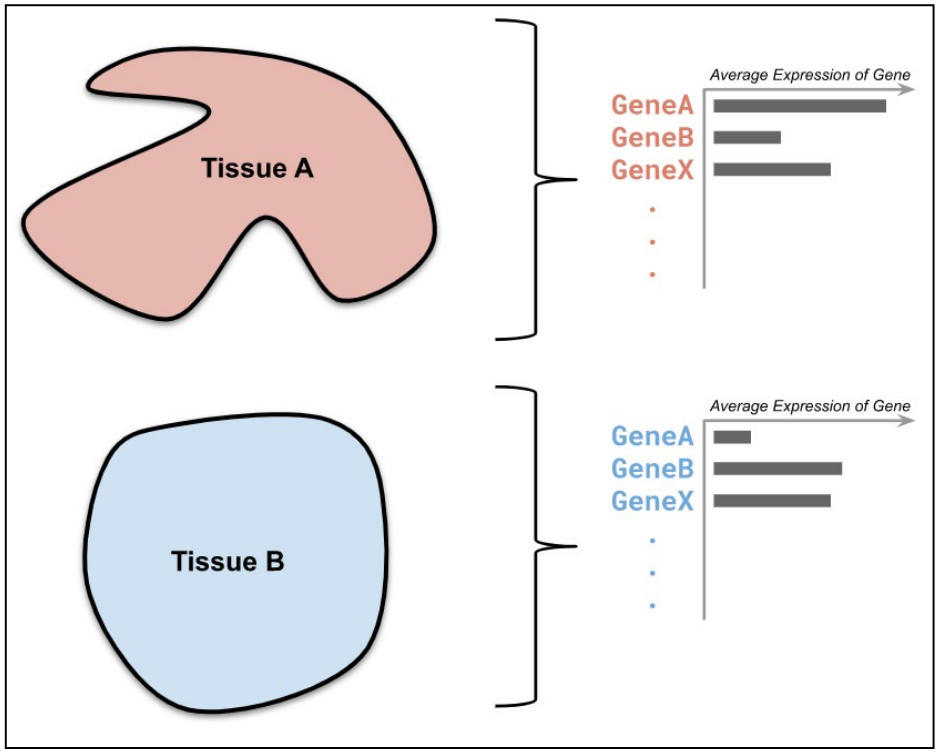


Geração de dados em um experimento de scRNAseq

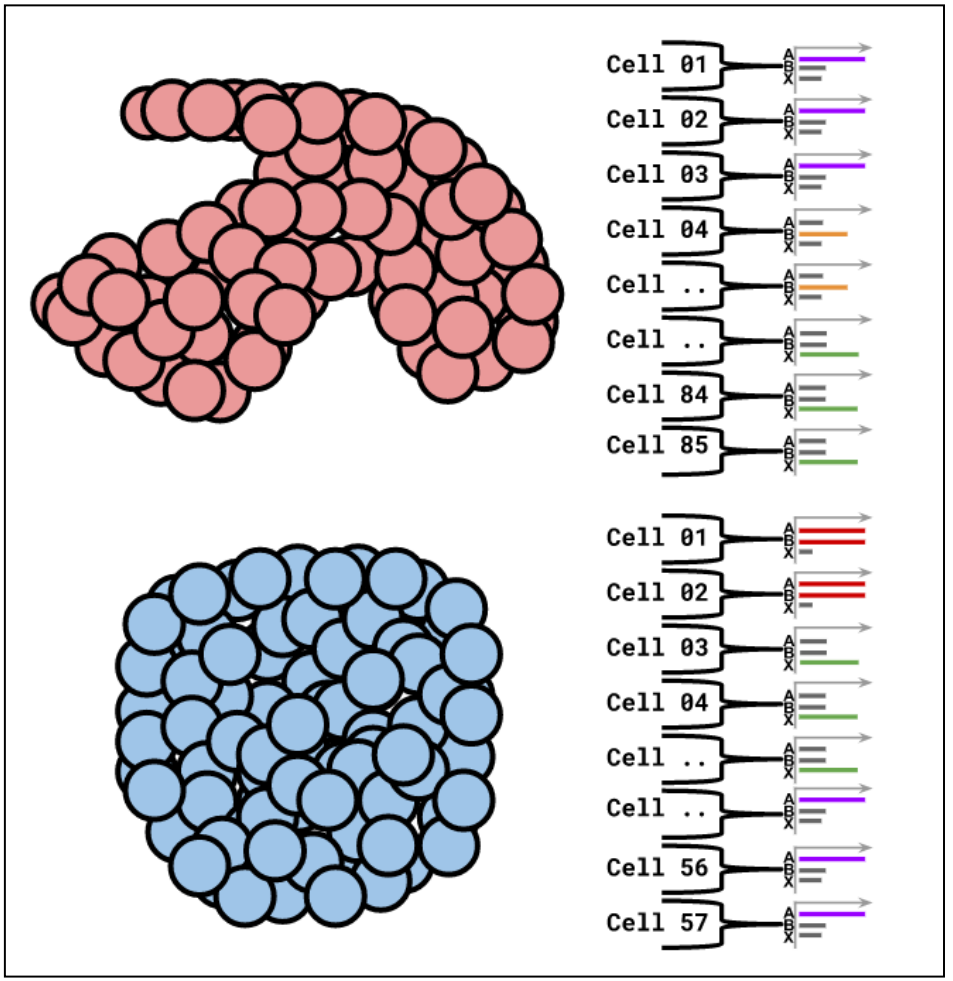


Análise do transcritoma “bulk” vs. célula única

RNA-seq “bulk”
média da expressão gênica nos diferentes tipos celulares presentes na amostra



RNAseq célula única (scRNA-seq)
expressão gênica em cada célula individual

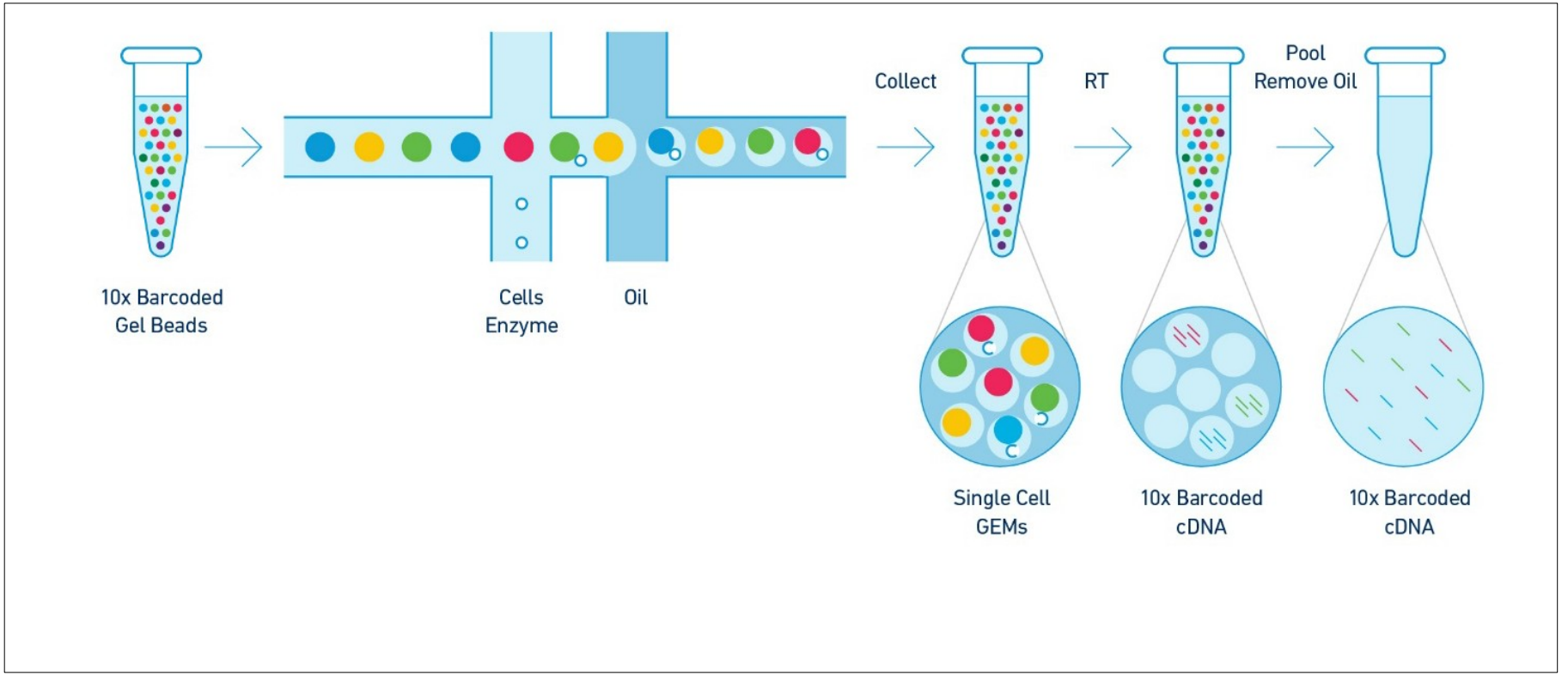


https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/scrna-intro/slides-plain.html

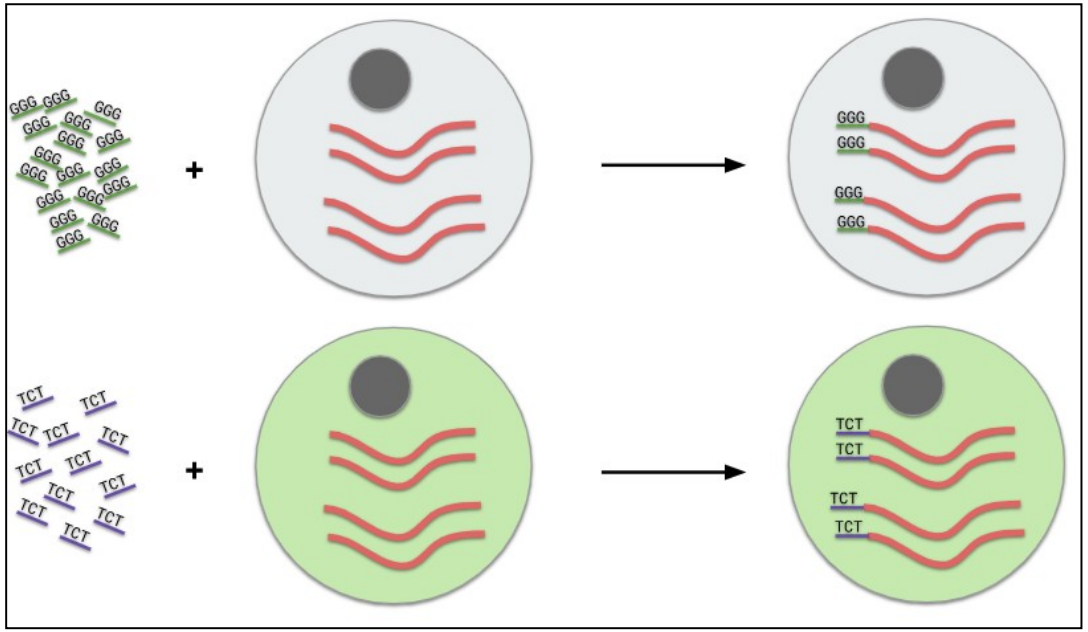
Plataformas para geração de dados de células únicas

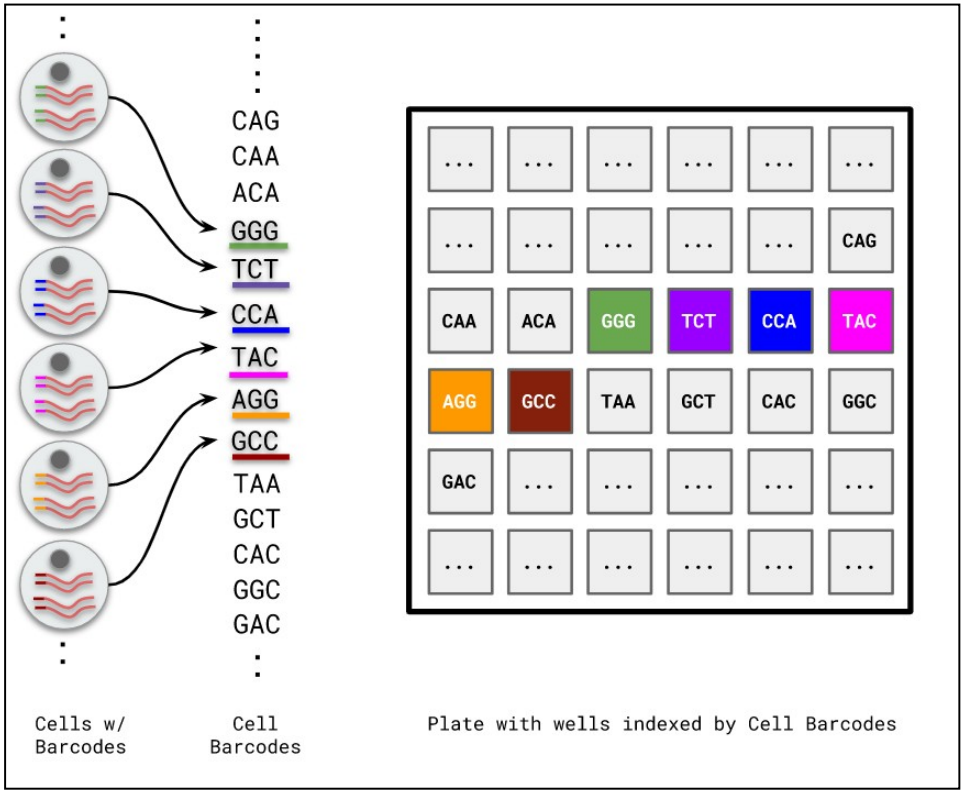
10x Chromium System

https://www.youtube.com/watch?v=4NAS1qTJmYA



Durante a preparação da biblioteca de cDNA, os RNAs de cada célula são marcados com códigos de barra moleculares únicos (8-12 bases).



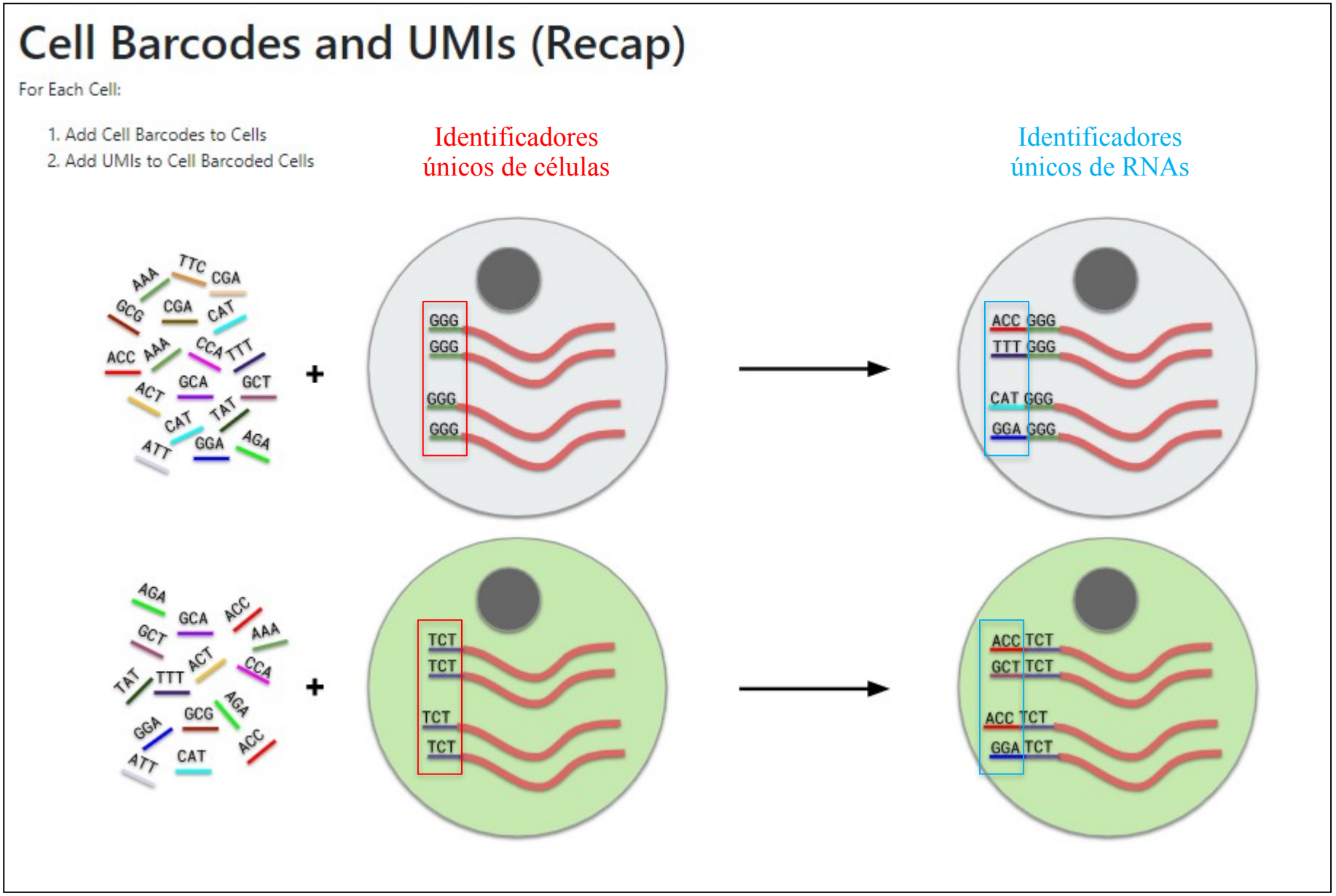


Permite o sequenciamento NGS em “pool” e em seguida assinalar os reads a cada uma das células analisadas

https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/scrna-intro/slides-plain.html

Também são adicionados identificadores moleculares únicos (UMIs) a cada transcrito

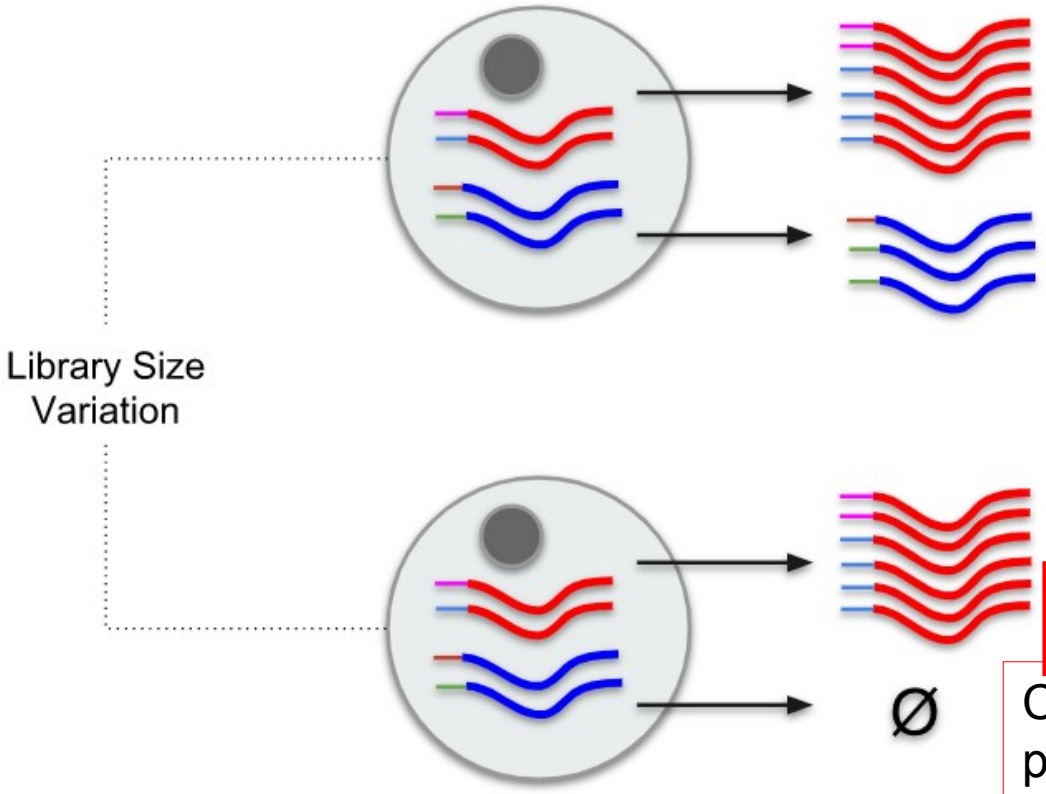
- Utilizados para a normalização dos dados (corrigir viés de amplificação durante o processo)
- Permitem inferir a abundância relativa dos diferentes genes expressos de forma mais precisa



https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/scrna-intro/slides-plain.html

Problemas a serem superados durante a normalização de dados de scRNAseq

Variabilidade técnica



Corrigida considerando o numero de RNAs com UMIs únicos

Corrigida aumentando a profundidade de sequencianento, eliminando células com expressão nula

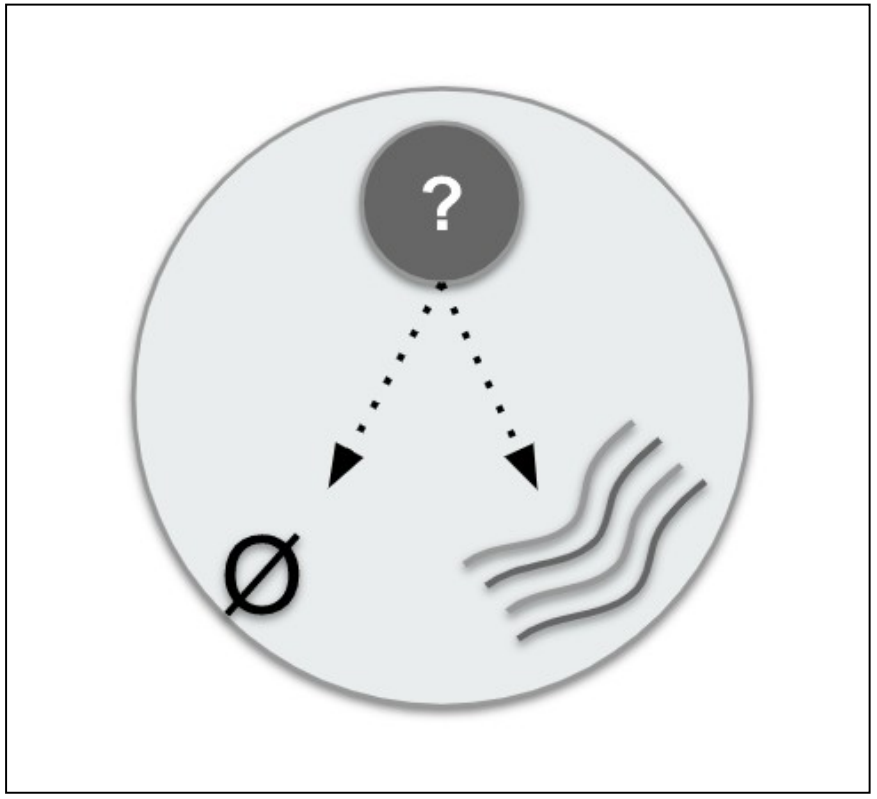
Células com poucos ou nenhum read

Problemas a serem superados durante a normalização de dados de scRNAseq

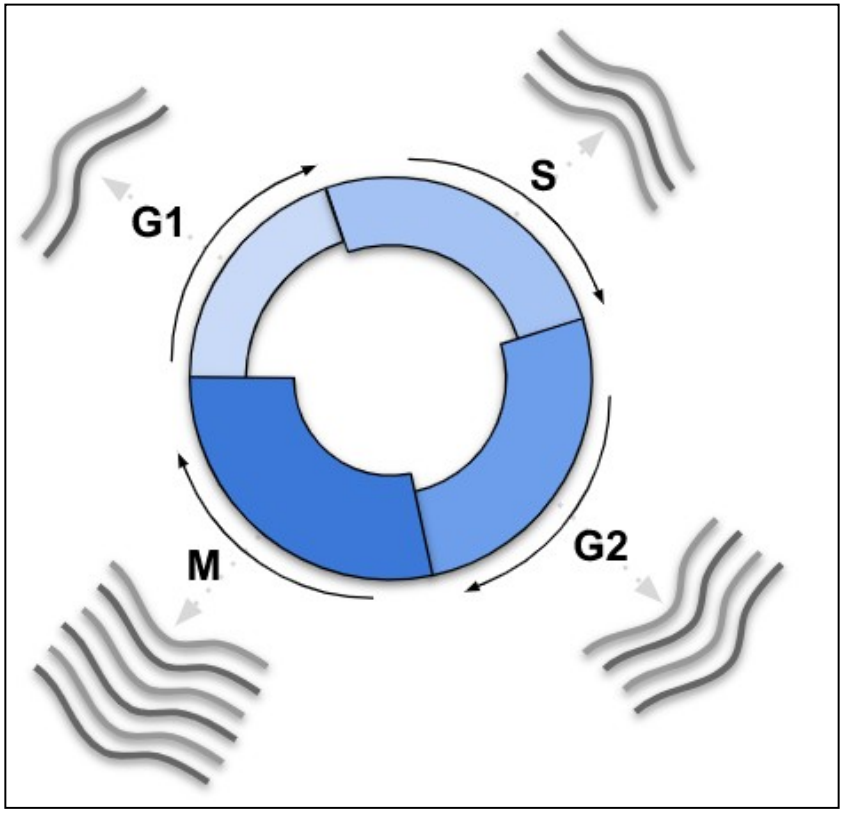
Variabilidade biológica

Células do mesmo tipo podem ter níveis de expressão diferentes em um dado instante

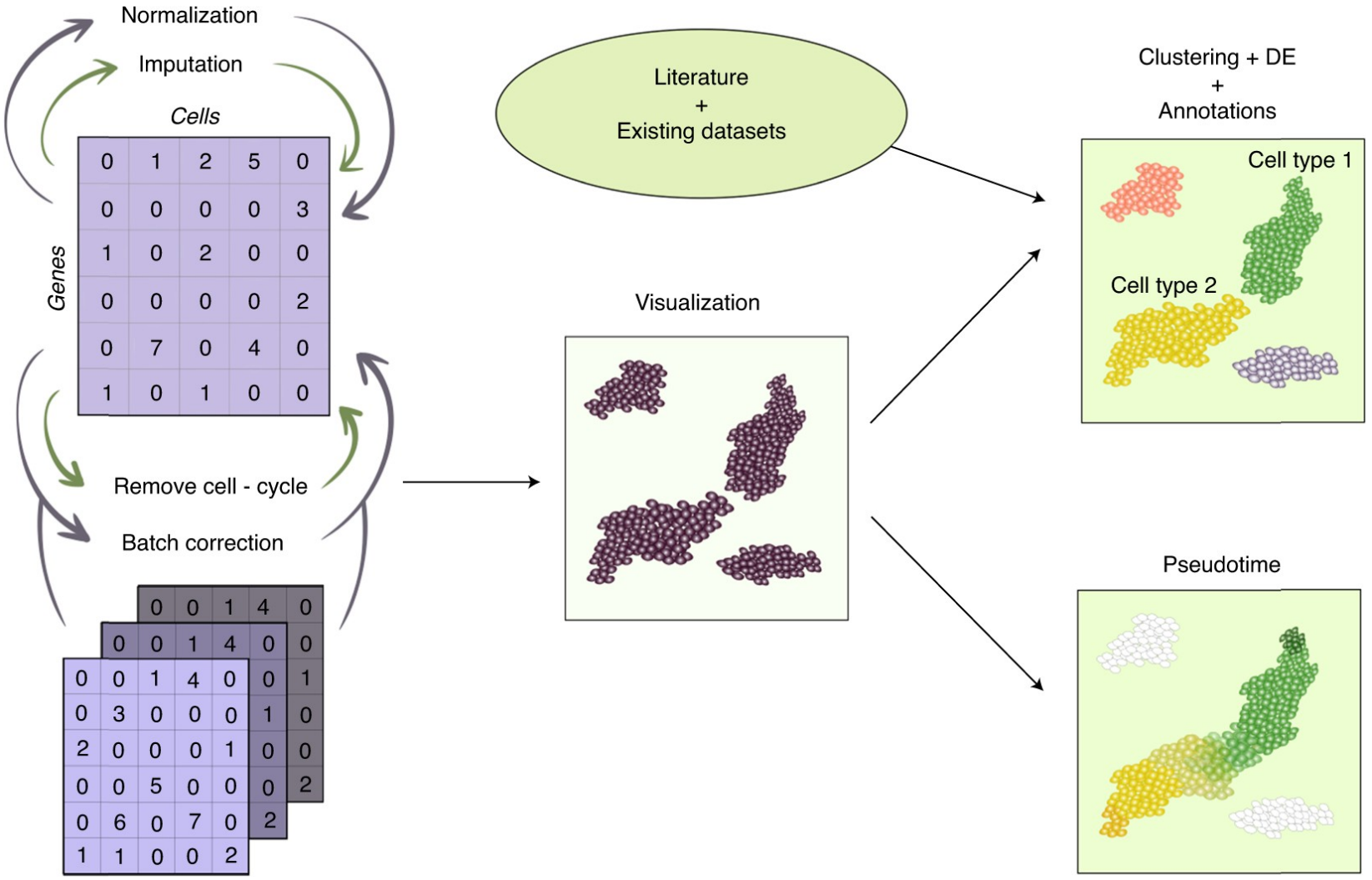
Expressão gênica acontece em "pulsos"



Expressão gênica varia ao longo do ciclo celular



Etapas na análise de dados de scRNAseq



Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

Alguns pipelines computacionais para análise de dados de scRNAseq:

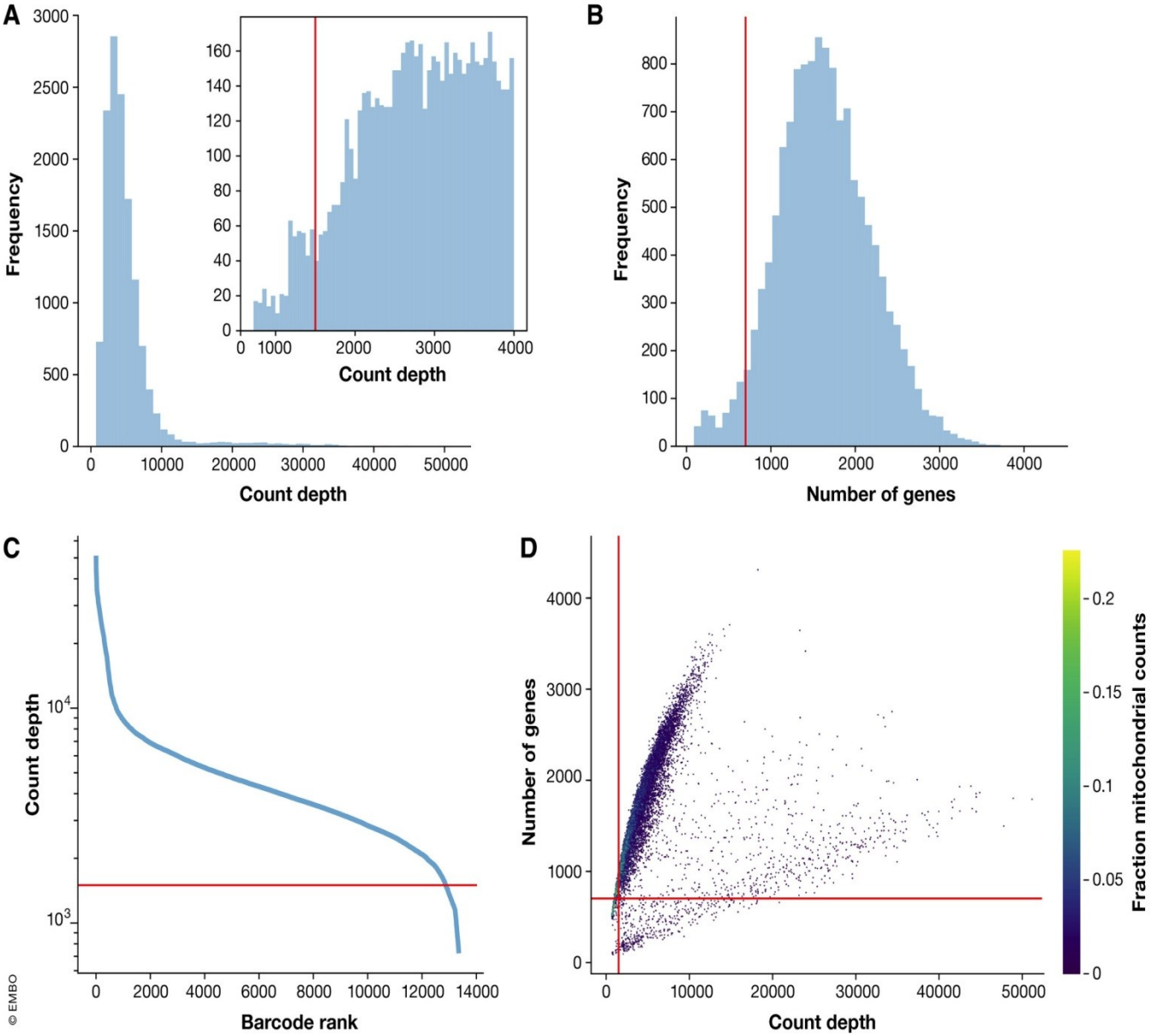
R and bioconductor tools:
https://github.com/drisso/bioc2016singlecell
https://hemberg-lab.github.io/scRNA.seq.course/

Seurat:
https://satijalab.org/seurat/get_started.html;

Scanpy:
https://scanpy.readthedocs.io/en/stable/tutorials.html)

Etapas de controle de qualidade e filtragem de dados com baixa qualidade

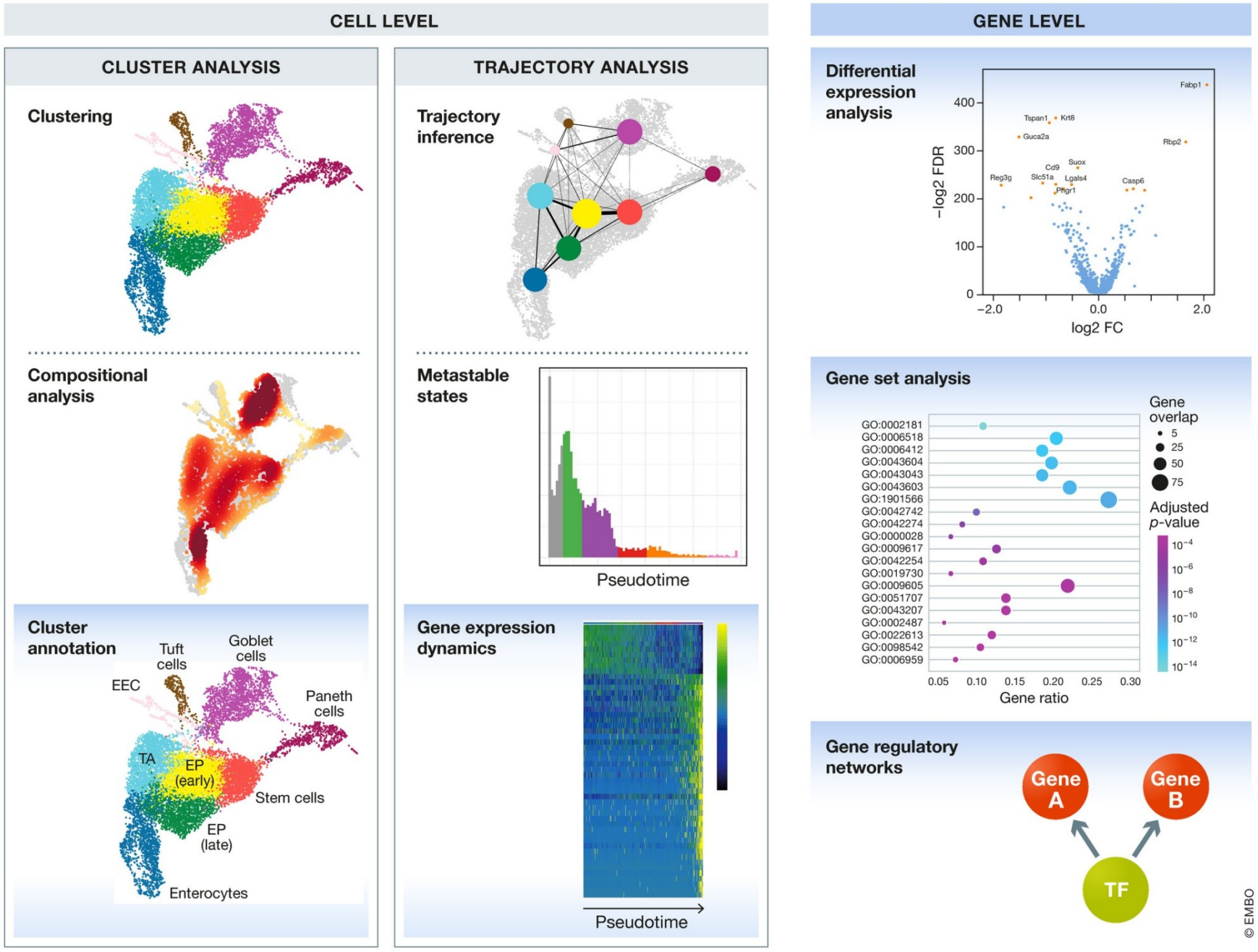
Eliminação de células com baixo número de reads, baixo número de genes/UMIs detectados, alta presença de genes mitocondriais (células danificadas)



Current best practices in single−cell RNA−seq analysis: a tutorial

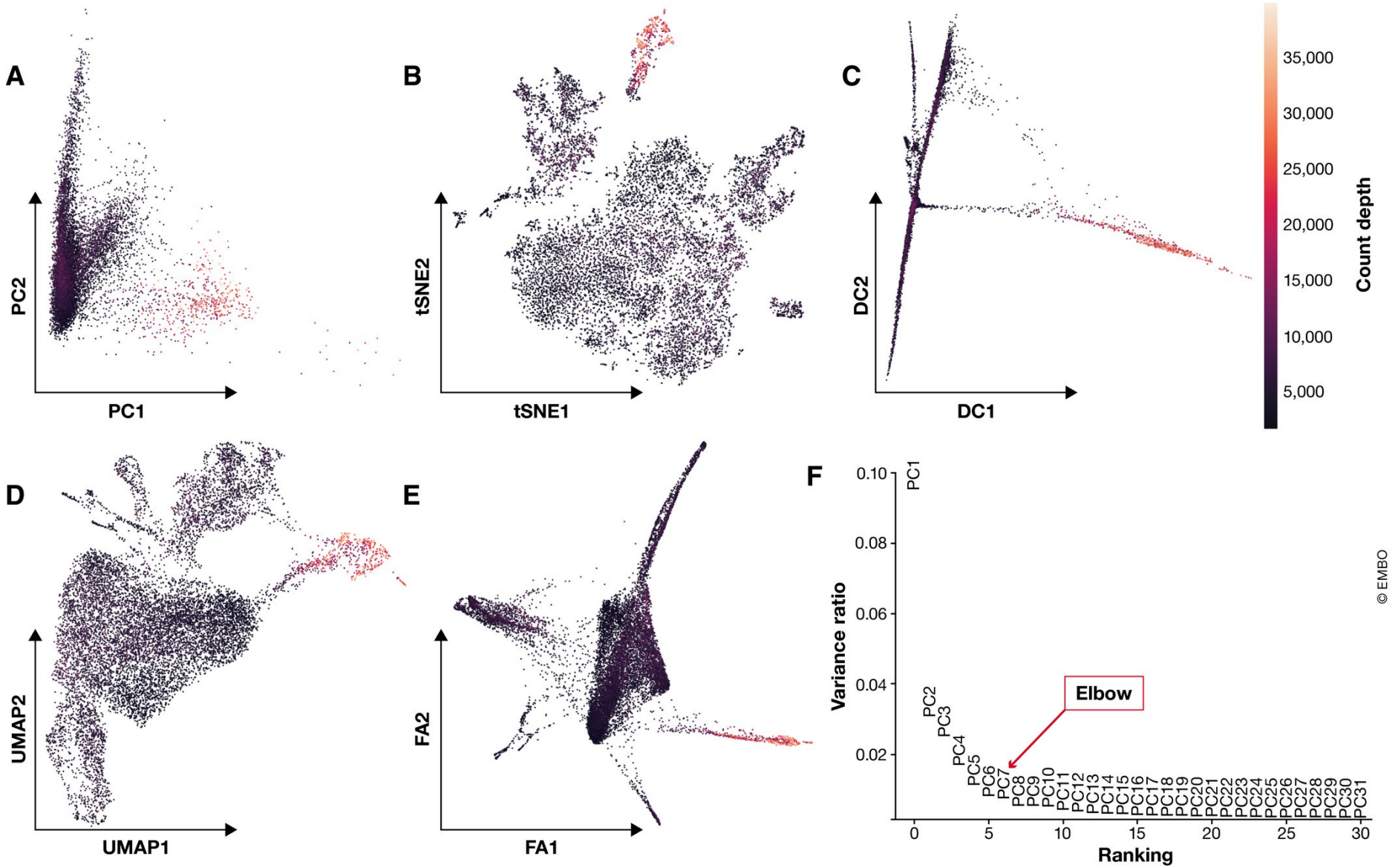
Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

Análises computacionais ao nível de células ou genes



Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

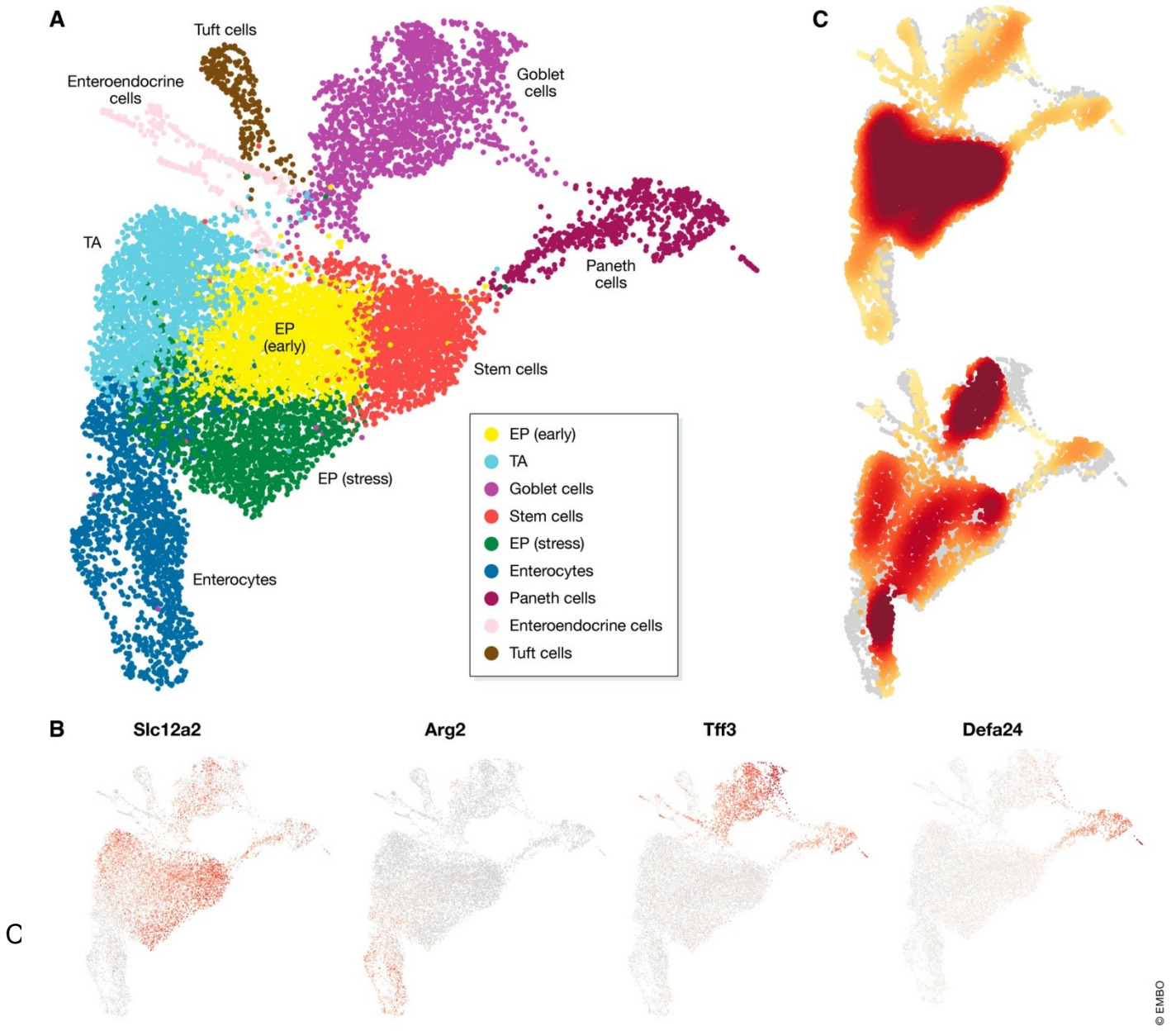
Utilizam-se diferentes métodos não supervisionados de agrupamento e redução de dimensionalidade (ex. PCA, tSNE, UMAP) para visualização de células com padrão de expressão gênica semelhante



Current best practices in single−cell RNA−seq analysis: a tutorial

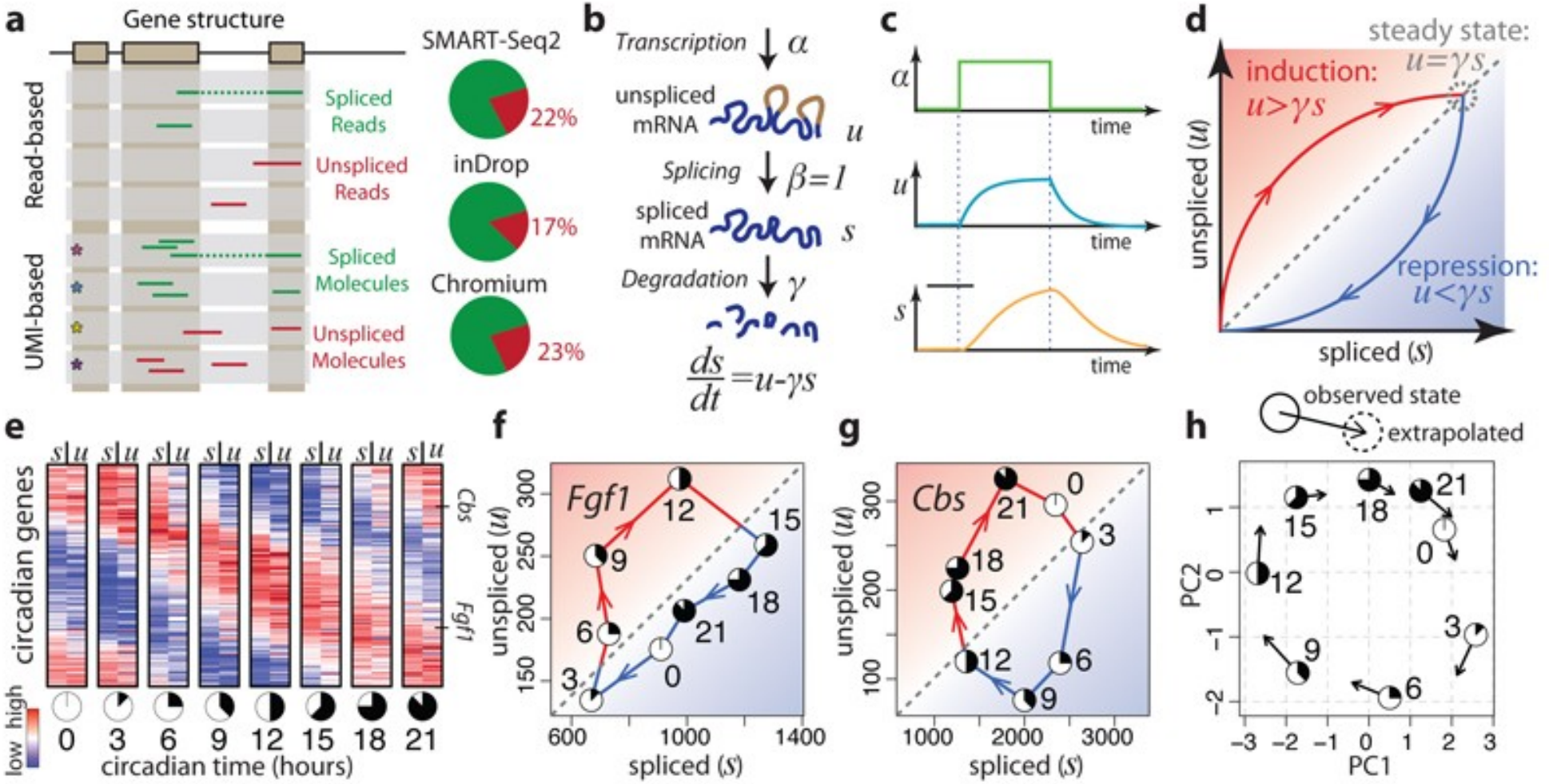
Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

Genes expressos em subtipos específicos são usados para anotar os agrupamentos de células representados em duas dimensões.



Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

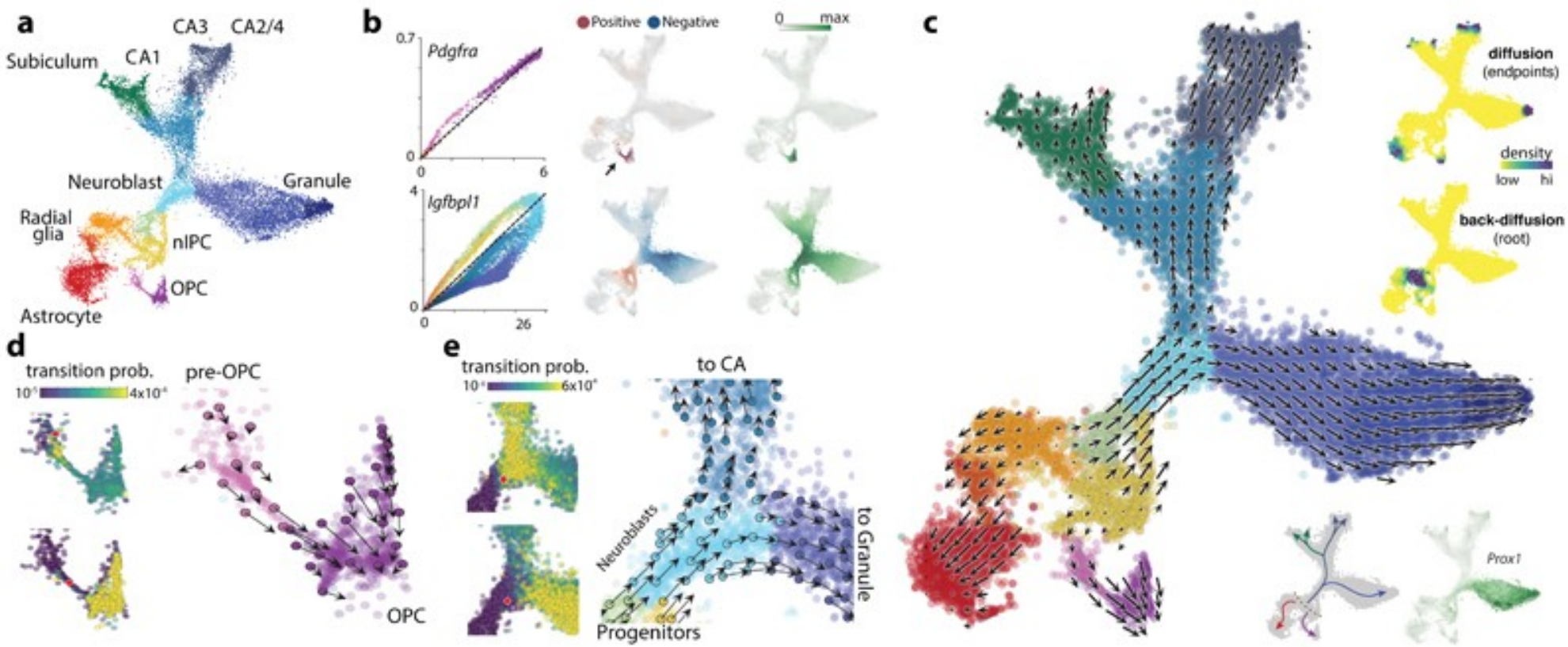
O balanço entre a fração de mRNAs nãoprocessados e processados (spliced) é preditivo da progressão do estado celular



RNA velocity of single cells, La Manno et al., Nature. 2018 Aug; 560(7719): 494–498.

A análise da velocidade de processamento de RNAs em células únicas permite inferir a dinâmica da diferenciação celular

Trajetória de diferenciação das principais linhagens neurais no hipocampo.



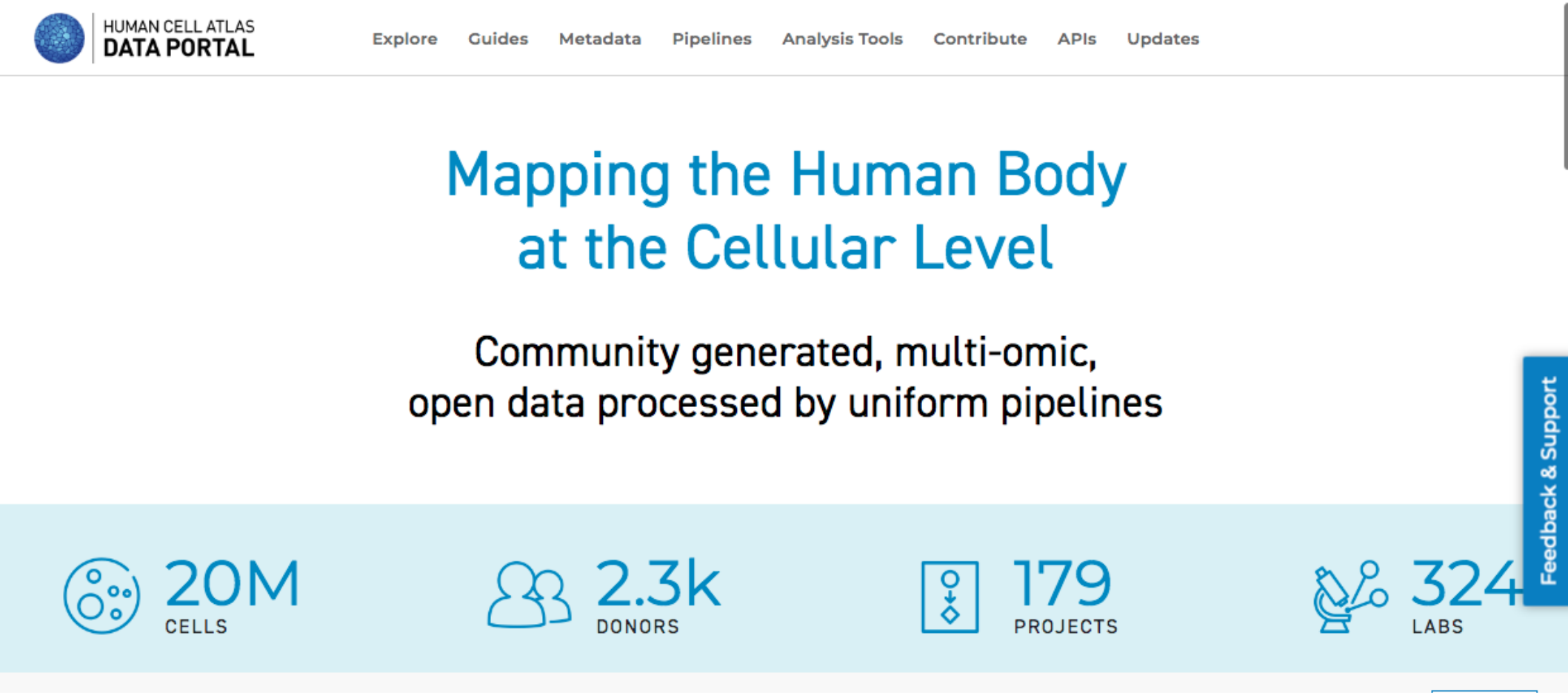
RNA velocity of single cells, La Manno et al., Nature. 2018 Aug; 560(7719): 494–498.

Referencias

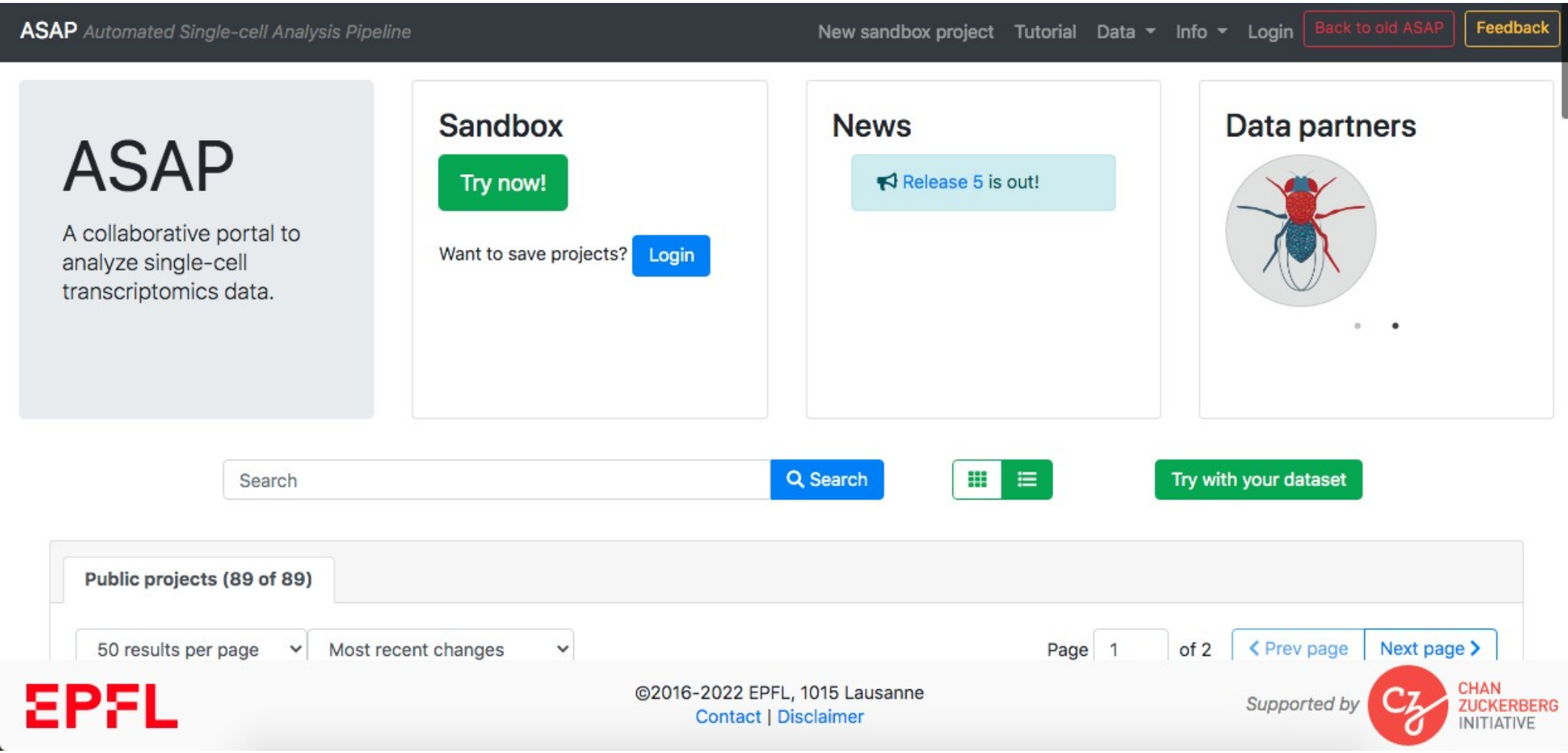
- Andrews, T.S., Kiselev, V.Y., McCarthy, D. *et al.* Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc* **16,** 1–9 (2021). https://doi.org/10.1038/s41596-020-00409-w

- Malte D Luecken & Fabian J Theis; Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology15:e8746 (2019) https://doi.org/10.15252/msb.20188746

https://data.humancellatlas.org/

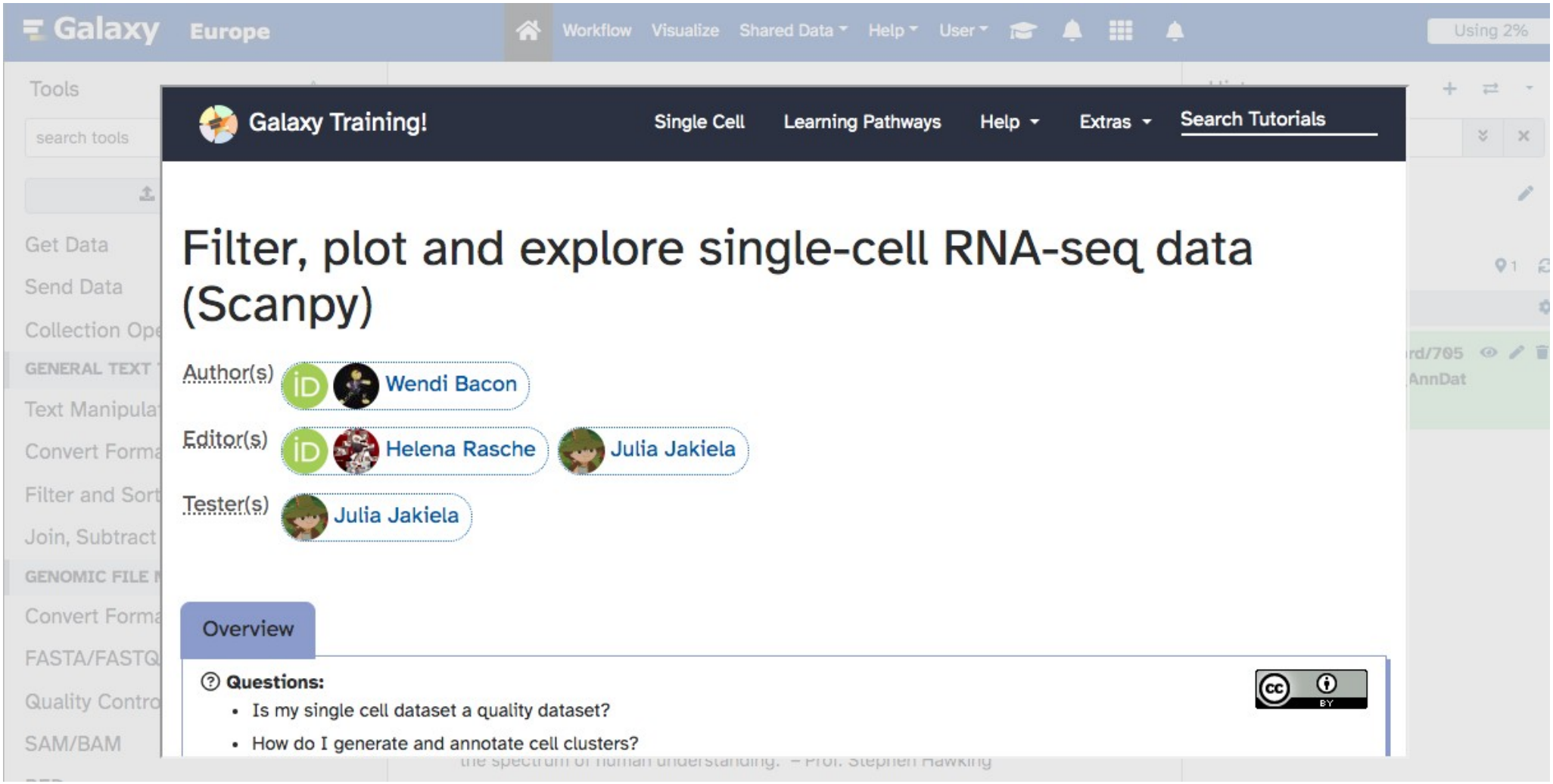


https://asap.epfl.ch/



Tutorial scRNAseq

Entre no Galaxy Europe
Criar uma nova sessão de trabalho (“history”): scRNAseq
Na aba “Galaxy Training Materials”, localize o tutorial abaixo



ORIGINAL RESEARCH article

Front. Immunol., 01 November 2018
Sec. T Cell Biology
Volume 9 - 2018 | https://doi.org/10.3389/fimmu.2018.02523

This article is part of the Research Topic
Single Cell Approaches To Study The Immune System
View all 13 Articles >

Single-Cell Analysis Identifies Thymic Maturation Delay in Growth-Restricted Neonatal Mice

 Wendi A. Bacon[1,2,3*†]  Russell S. Hamilton[2,3†] Ziyi Yu[4] Jens Kieckbusch[1,2]  Delia Hawkes[1]

Ada M. Krzak[1] Chris Abell[4]  Francesco Colucci[1,2]  D. Stephen Charnock-Jones[1,2*]

[1] Department of Obstetrics and Gynaecology, University of Cambridge, Cambridge, United Kingdom
[2] Centre for Trophoblast Research, University of Cambridge, Cambridge, United Kingdom
[3] Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom
[4] Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Fetal growth restriction (FGR) causes a wide variety of defects in the neonate which can lead to increased risk of heart disease, diabetes, anxiety and other disorders later in life. However, the effect of FGR on the immune system, is poorly understood. We used a well-characterized mouse model of FGR in which placental Igf-2 production is lost due to deletion of the placental specific *Igf-2* P_0 promotor. The thymi in such animals were reduced in mass with a ~70% reduction in cellularity. We used single cell RNA sequencing (Drop-Seq) to analyze 7,264 thymus cells collected at postnatal day 6. We identified considerable heterogeneity among the Cd8/Cd4 double positive cells with one subcluster showing marked upregulation of transcripts encoding a sub-set of proteins that contribute to the surface of the ribosome. The cells from the FGR animals were underrepresented in this cluster. Furthermore, the distribution of cells from the FGR animals was skewed with a higher proportion of immature double negative cells and fewer mature T-cells. Cell cycle regulator transcripts also varied across clusters. The T-cell deficit in FGR mice persisted into adulthood, even when body and organ weights approached normal levels due to catch-up growth. This finding complements the altered immunity found in growth restricted human infants. This reduction in T-cellularity may have implications for adult immunity, adding to the list of adult conditions in which the *in utero* environment is a contributory factor.

In this tutorial, we will cover:

1. Introduction
    1. Get data
2. Filtering
    1. Generate QC Plots
    2. Analysing the plots
    3. Apply the thresholds
3. Processing
4. Preparing coordinates
    1. Principal components
    2. Neighborhood graph
    3. Dimensionality reduction for visualisation
5. Cell clusters & gene markers
    1. FindMarkers
6. Plotting!
7. Insights into the beyond
    1. Biological Interpretation
    2. Technical Assessment
8. Interactive visualisations
9. Conclusion

Remover dados de baixa qualidade
Células com baixa contagem de read, baixo numero de genes, alta % de genes mitocondriais

Identificar e visualizar tipos celulares

AnnData

The most common format, called `AnnData`, stores the matrix as well as gene and cell annotations in a concise, compressed and extremely readable manner:

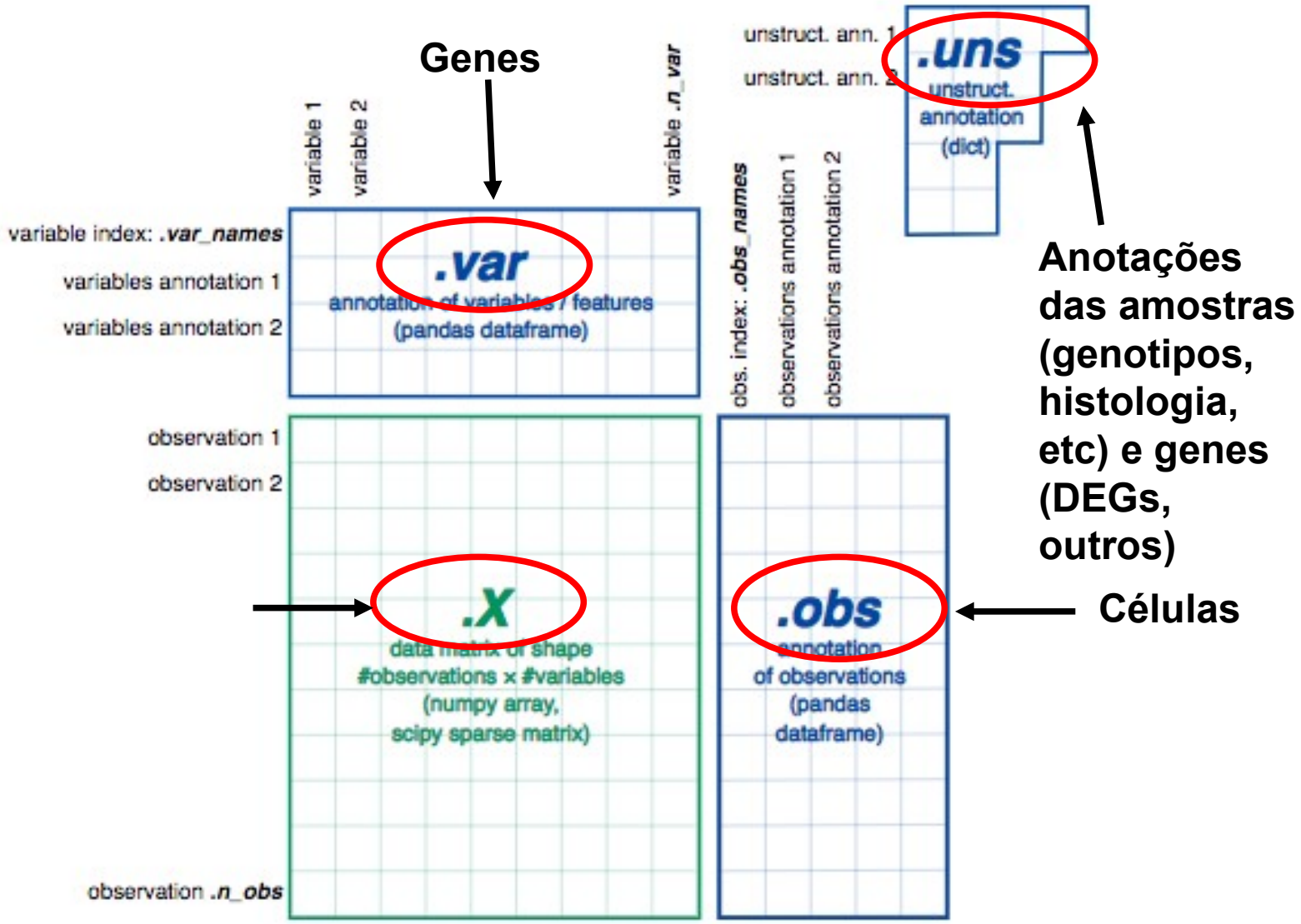
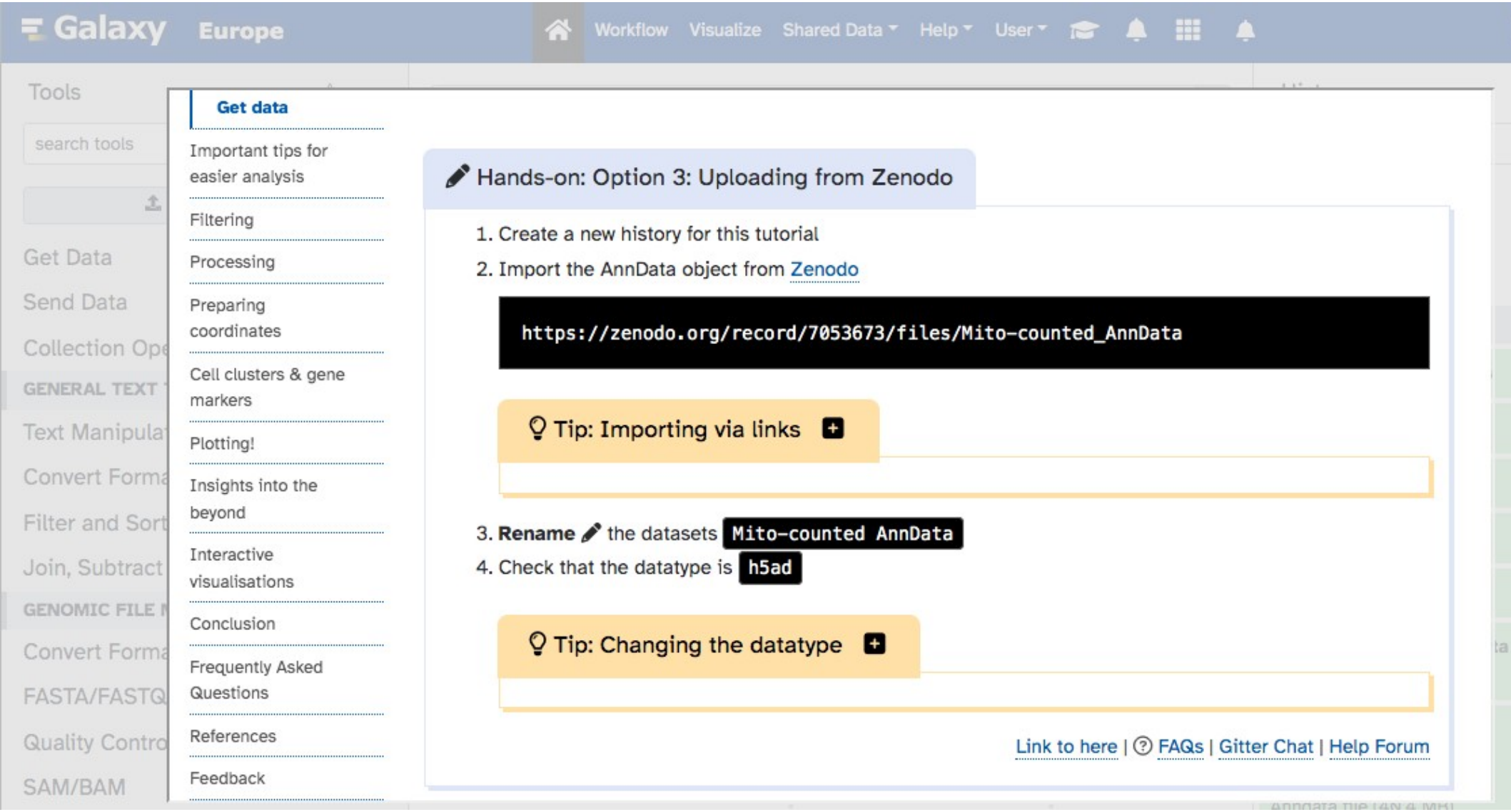


Figure 1: `AnnData` format stores a count matrix `X` together with annotations of observations (i.e. cells) `obs`, variables (i.e. genes) `var` and unstructured annotations `uns`.

Como obter os dados



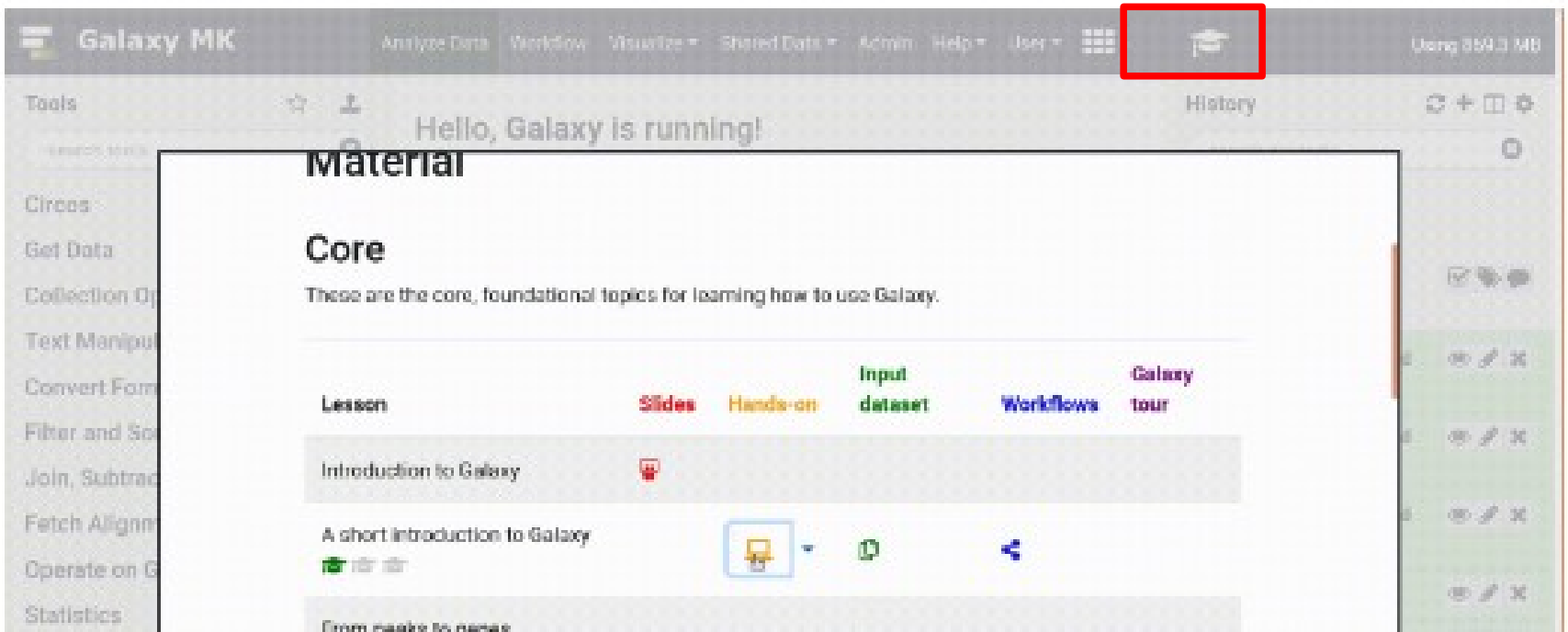
Muito importante!

Executar os comandos dentro do "tutorial mode" do Galaxy, como indicado no protocolo

Tip: Using tutorial mode

Tools are frequently updated to new versions. Your Galaxy may have multiple versions of the same tool available. By default, you will be shown the latest version of the tool. This may NOT be the same tool used in the tutorial you are accessing. Furthermore, if you use a newer tool in one step, and try using an older tool in the next step... this may fail! To ensure you use the same tool versions of a given tutorial, use the **Tutorial mode** feature.

- Open your Galaxy server
- Click on the icon on the top menu, this will open the GTN inside Galaxy.
- Navigate to your tutorial
- Tool names in tutorials will be blue buttons that open the correct tool for you
- **Note:** this does not work for all tutorials (yet)



Instruções para confecção do relatório:

- Responder as perguntas abaixo
- O relatório deve ser entregue através do e-disciplinas.

1. Qual o dado experimental que foi utilizado na análise do tutorial?
2. Quais etapas de filtragem e transformação (normalização) de dados foram realizadas? Qual o objetivo dessas etapas?
3. Qual o significado de células “drop out”?
4. Qual a razão de filtrar células que possuem uma alta porcentagem de contagens em genes ribosomais?
5. Quantas células e quantos genes foram detectados no início e após cada etapa de filtragem?
6. A análise de componentes principais (PCs) permite reduzir a dimensionalidade dos dados e identificar subconjuntos de genes que contribuem para a maior parte da variação entre as células. Quantos PCs explicam 90% ou mais da variancia da expressão gênica?
7. Quantos clusters diferentes de células foram identificados?
8. Após a clusterização, os diferentes grupos de células são anotados com os genes marcadores identificados no estudo original (Il2ra,Cd8b1,Cd8a,Cd4,Itm2a,Aif1,Hba-a1). Um dos passos da análise identificou marcadores dos clusters e gerou um arquivo de saida com os 50 melhores marcadores de cada cluster ranqueados. Os marcadores usados estão nessa lista? Em que posição no ranking em cada cluster?