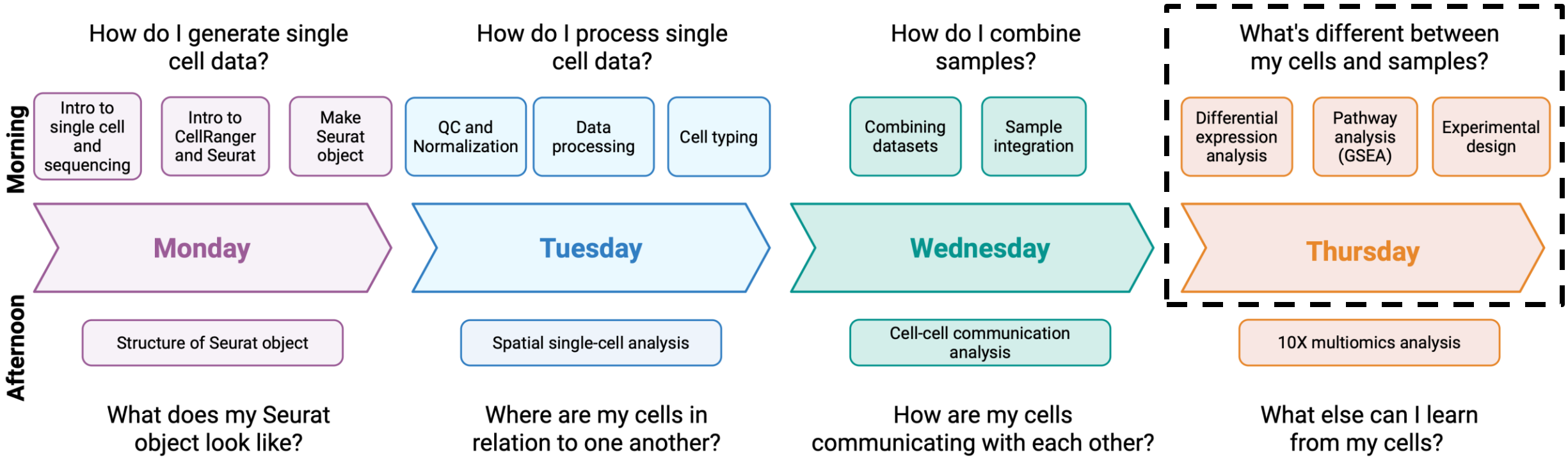
Thursday Beginner Session: Differential Expression and Gene Set Enrichment Analysis

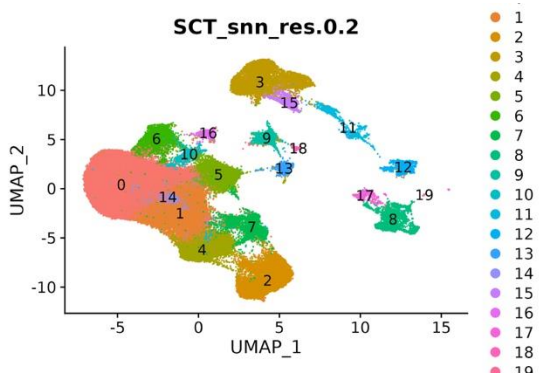
Corinne Strawser, PhD
Matt Gust

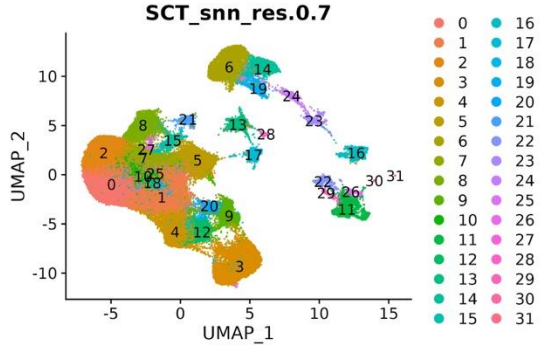
Nationwide Children's®
When your child needs a hospital, everything matters.

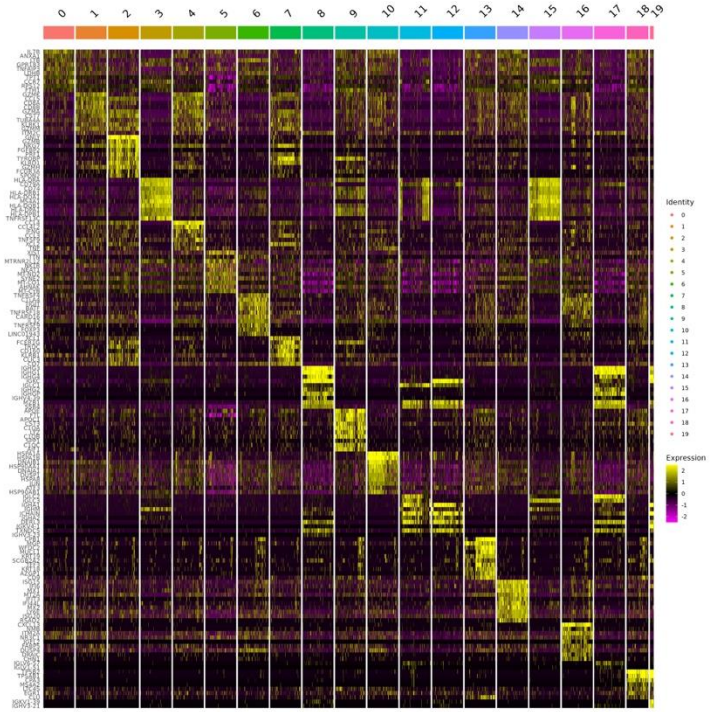

How do I generate single cell data?
How do I process single cell data?
How do I combine samples?
What's different between my cells and samples?
Morning
Intro to single cell and sequencing
Intro to CellRanger and Seurat
Make Seurat object
QC and Normalization
Data processing
Cell typing
Combining datasets
Sample integration
Differential expression analysis
Pathway analysis (GSEA)
Experimental design
Monday
Tuesday
Wednesday
Thursday
Afternoon
Structure of Seurat object
Spatial single-cell analysis
Cell-cell communication analysis
10X multiomics analysis
What does my Seurat object look like?
Where are my cells in relation to one another?
How are my cells communicating with each other?
What else can I learn from my cells?

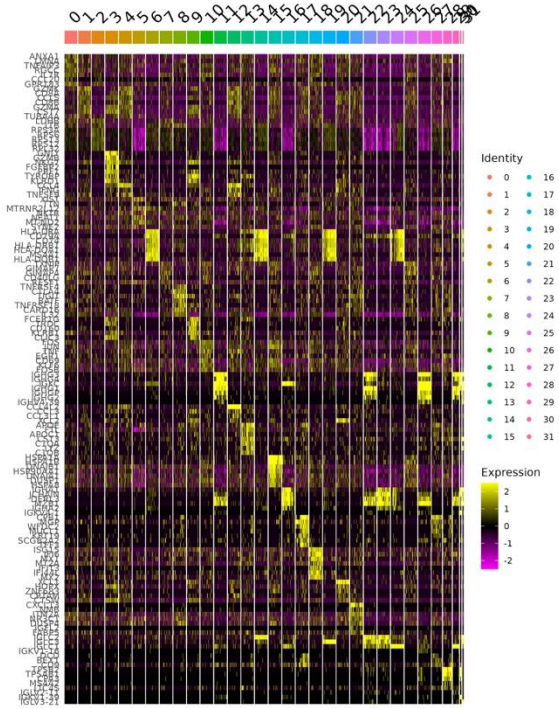
How do we use differential expression analysis in scRNA-seq?

1. Guiding clustering resolution



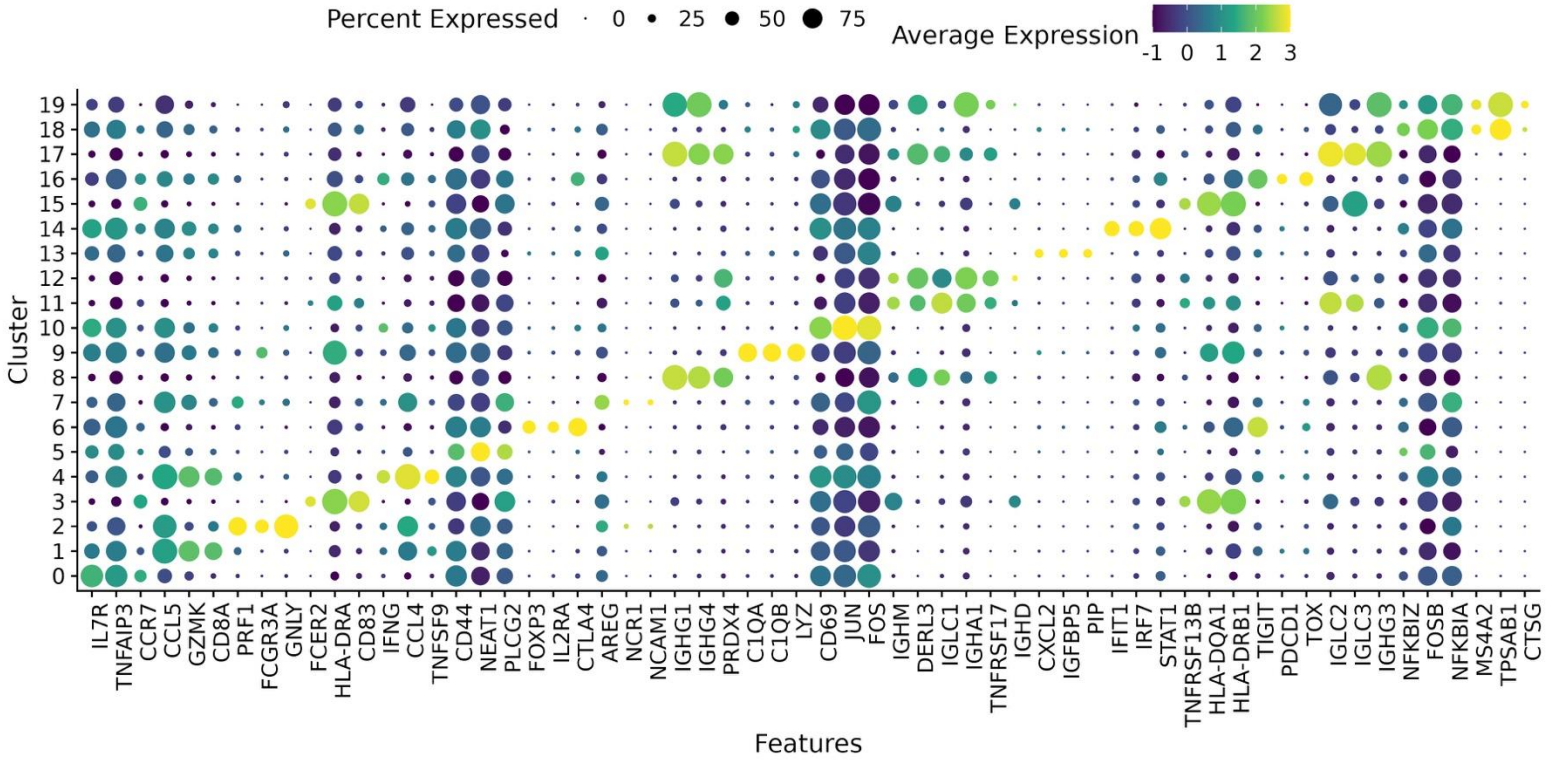






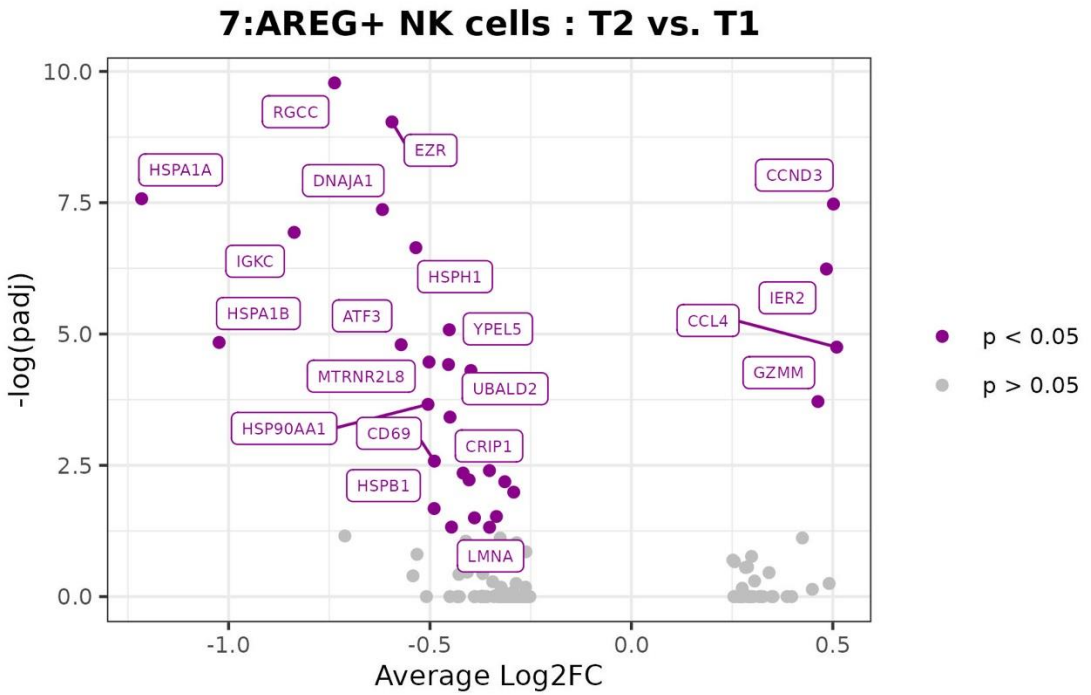
How do we use differential expression analysis in scRNA-seq?

1. Guiding clustering resolution
2. Identifying cell types



How do we use differential expression analysis in scRNA-seq?

1. Guiding clustering resolution
2. Identifying cell types
3. Identifying transcriptional differences in samples

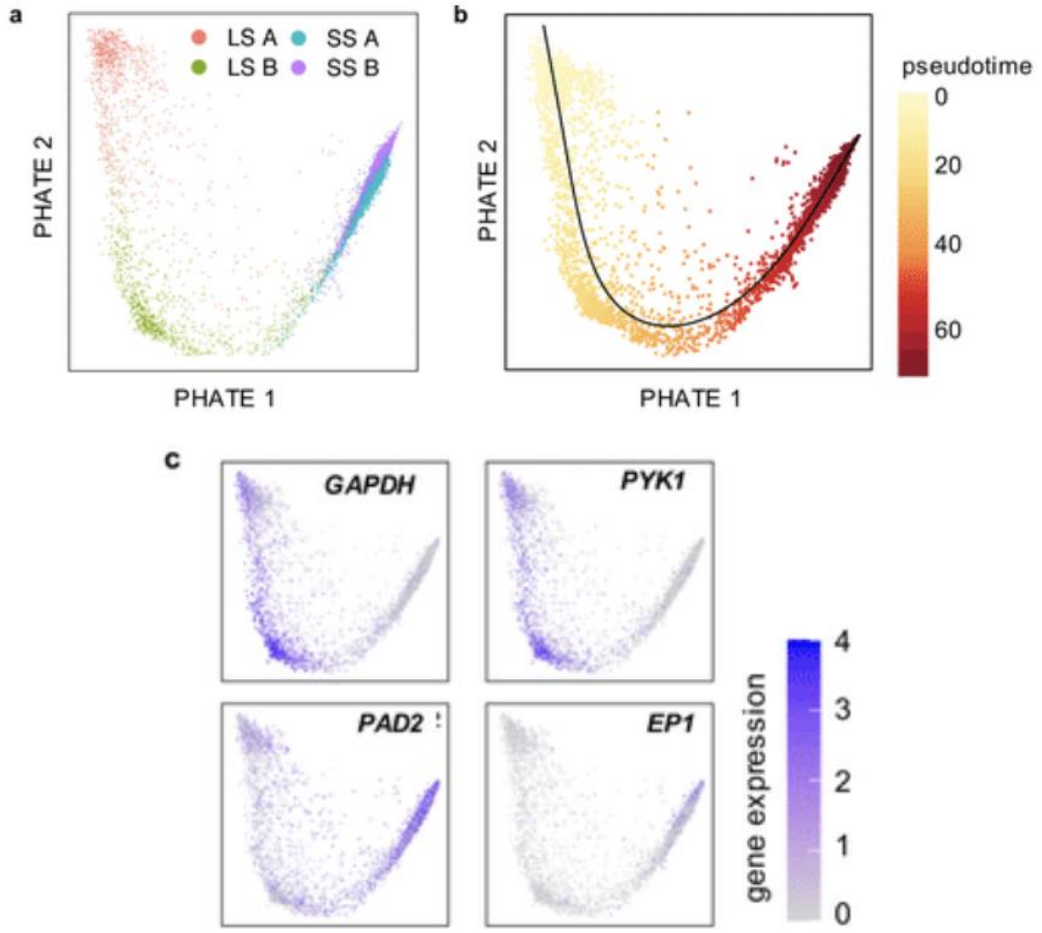


Lower in T2

Higher in T2

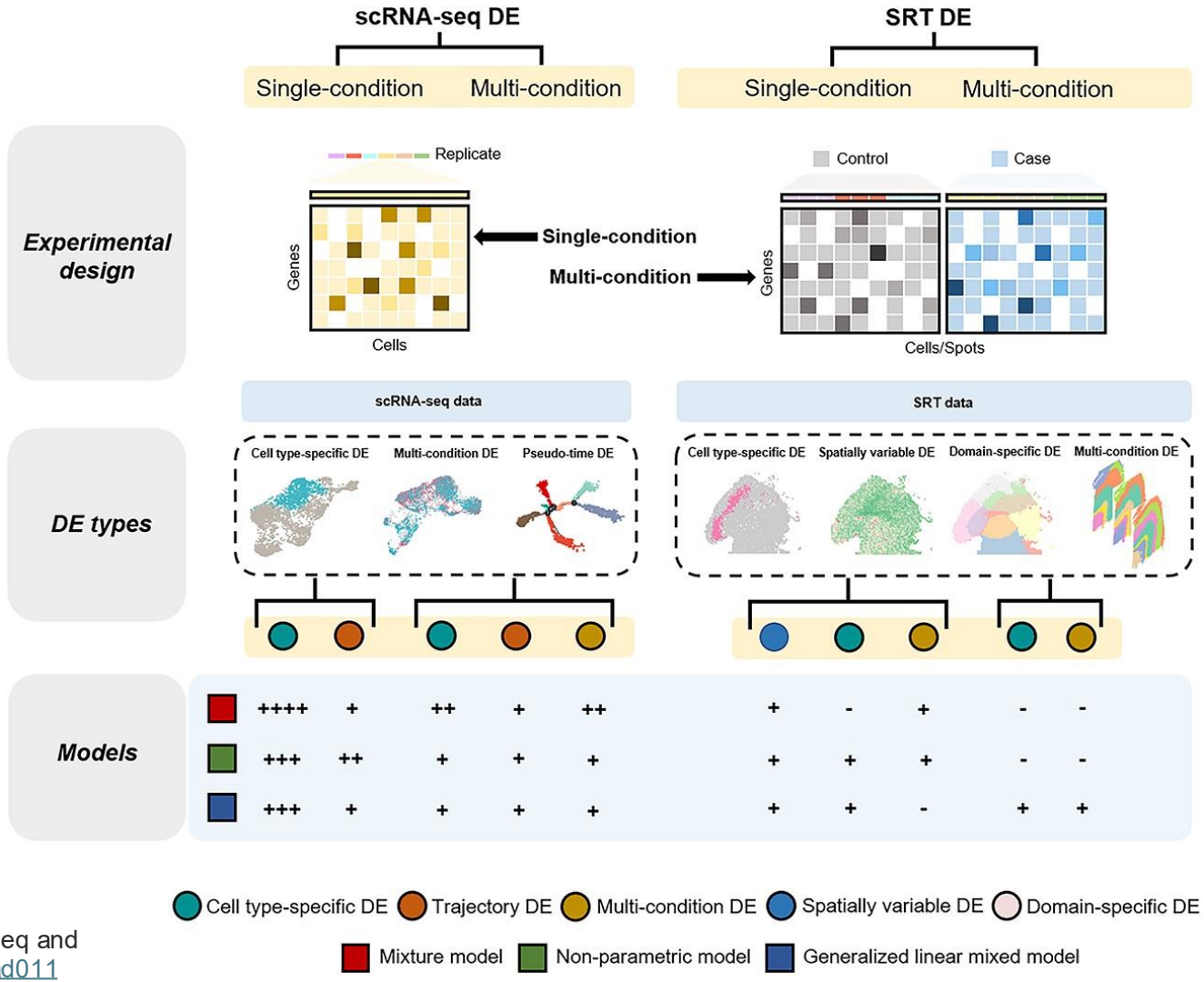
How do we use differential expression analysis in scRNA-seq?

1. Guiding clustering resolution
2. Identifying cell types
3. Identifying transcriptional differences in samples
4. Identifying transcriptional differences in pseudotime



Overview of DE methods in single cell transcriptomics

- > 130 DE tools for single cell data
- Problem #1: So many 0s!
  - Are they biological or technical?
  - Drop-out refers to a gene that is expressed but due to shallow sequencing associated with scRNA-seq, the gene is not detect resulting in a 0 read count
  - Is this missing data? should it be imputed?
- Problem #2: Some integration/ batch correction processes modify the data.
  - Is this biological or technical variation?



Xiya Guo, et al. (2024) Recent advances in differential expression analysis for single-cell RNA-seq and spatially resolved transcriptomic studies, *Briefings in Functional.* https://doi.org/10.1093/bfgp/elad011

Interaction of sparsity, sequencing depth, and batch effects

Problem #1 Problem #2

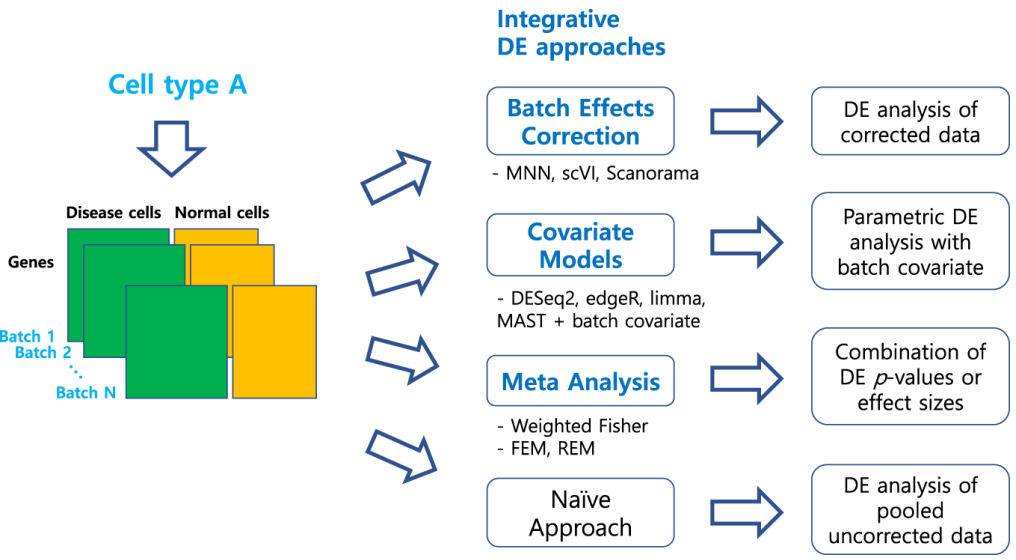


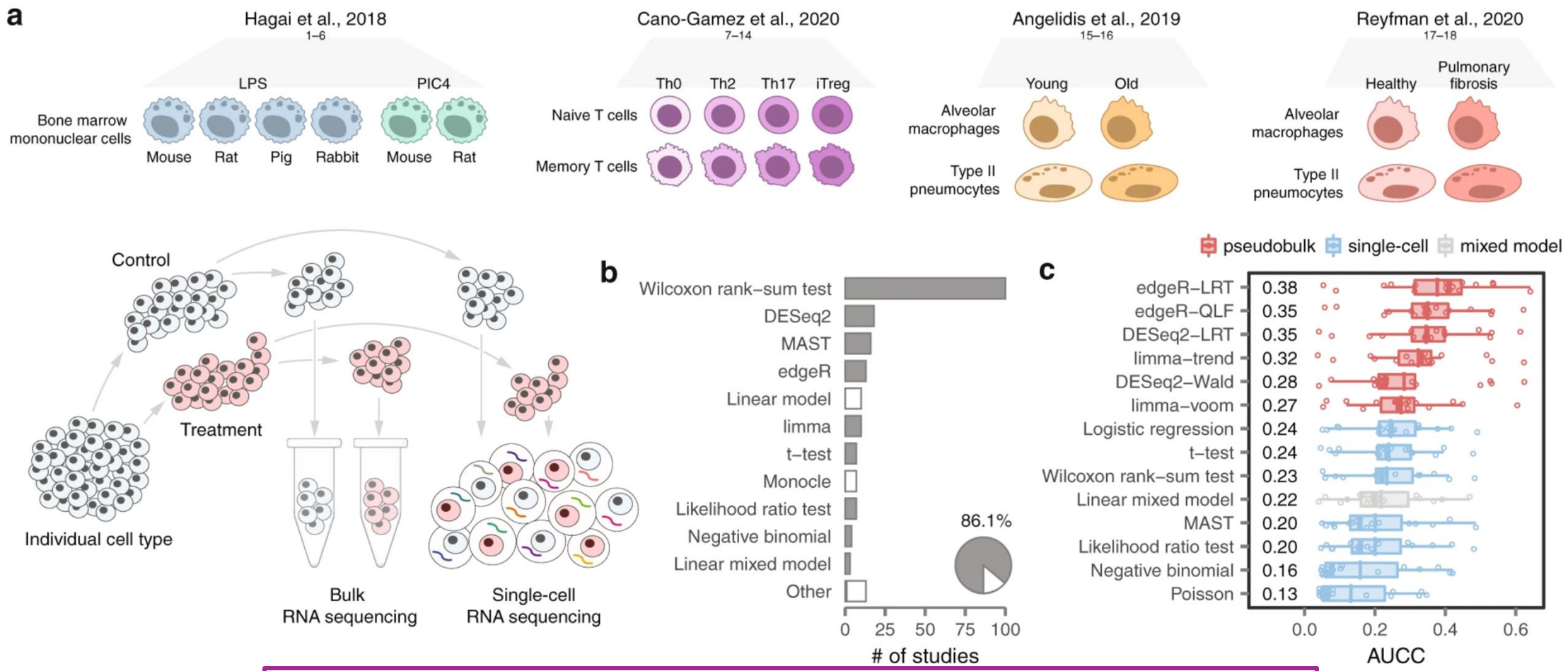
Table 1 | Recommended differential expression (DE) workflows for different experimental conditions

Sparsity (zero rate)	Depth*	Batch Effects	Recommended DE workflows
80%	77	Substantial	MAST_Cov, ZW_edgeR_Cov, ZW_DESeq2_Cov, scVI_limmatrend, DESeq2_FEM, limmatrend_Cov
80%	77	Small	MAST, ZW_edgeR, ZW_DESeq2, Pseudobulk_limma, DESeq2_FEM, limmatrend_Cov
80%	10	Substantial	DESeq2_Cov, limmatrend_Cov, DESeq2_wFisher, LogN_FEM, MAST_Cov, Raw_Wilcox
80%	10	Small	DESeq2, limmatrend_Cov, LogN_FEM, Pseudobulk_edgeR, Pseudobulk_limma, Raw_Wilcox
80%	4	Substantial	LogN_FEM, limmatrend, Raw_Wilcox, RISC_QP
80%	4	Small	LogN_FEM, limmatrend, Raw_Wilcox, RISC_QP
40%	77	Substantial	MAST_Cov, ZW_edgeR_Cov, ZW_DESeq2_Cov, limma_BEC_Wilcox, Scanorama_limmatrend, logN_FEM
40%	77	Small	Pseudobulk_limma, Raw_Wilcox, Pseudobulk_edgeR, DESeq2, LogN_FEM
40%	4	Substantial	limmavoom_Cov, limmatrend_Cov, ZW_edgeR_Cov, ZW_DESeq2_Cov, logN_FEM, limma_BEC_Wilcox
40%	4	Small	Pseudobulk_limma, limmvoom, limmatrend, Pseudobulk_edgeR, Raw_Wilcox, edgeR

In 40% sparsity cases, recommended methods were selected based on simulation results only.
*Average nonzero count in each cell after filtering sparsely expressed genes (zero rate > 0.95).

Nguyen, H.C.T., Baik, B., Yoon, S. *et al.* Benchmarking integration of single-cell differential expression. *Nat Commun* **14**, 1570 (2023). https://doi.org/10.1038/s41467-023-37126-3

Best practices for single-cell differential expression analysis are still being developed



NOTE: Pseudobulk methods should have at least 3 samples per group

Squair et al 2021

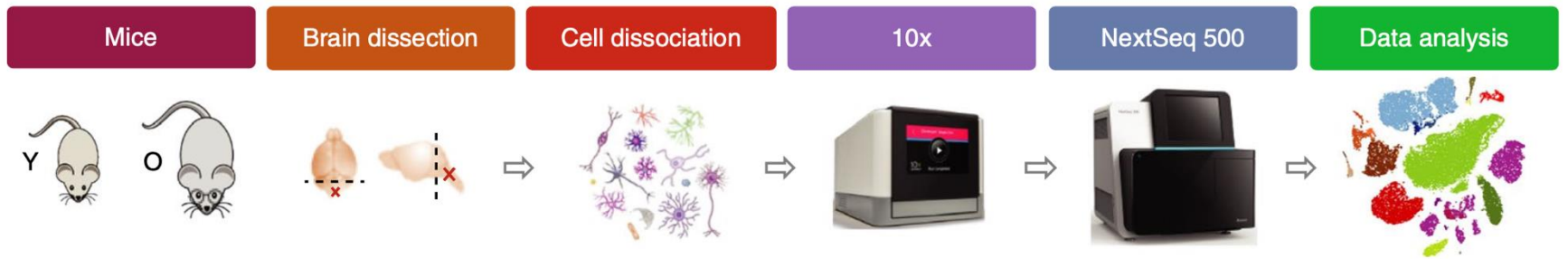
scRNA-seq demo data

RESOURCE
https://doi.org/10.1038/s41593-019-0491-3
nature neuroscience

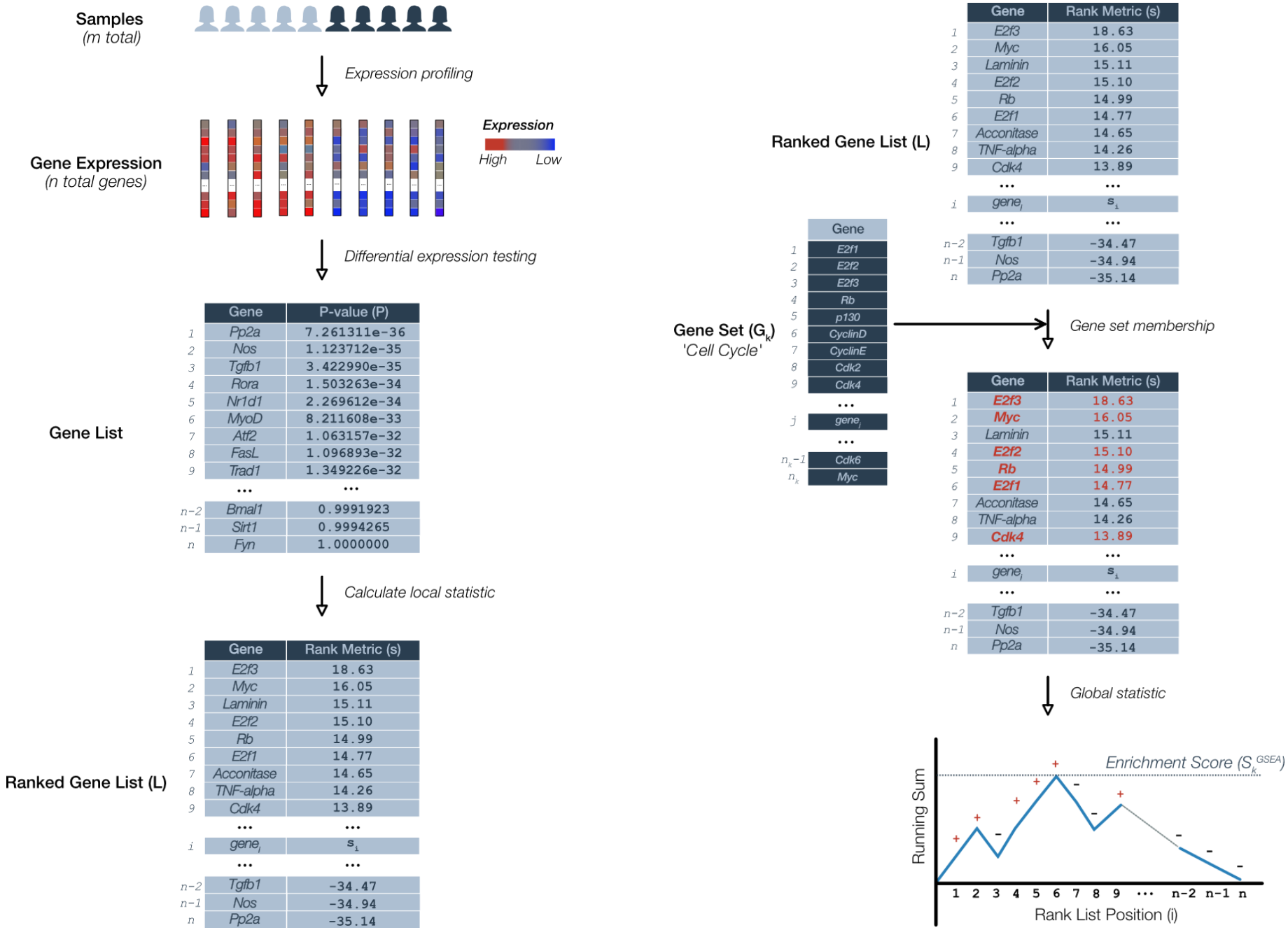
Single-cell transcriptomic profiling of the aging mouse brain

Methodios Ximerakis[1,2,3,8*], Scott L. Lipnick[1,2,3,4,8], Brendan T. Innes[5], Sean K. Simmons[3], Xian Adiconis[3], Danielle Dionne[3], Brittany A. Mayweather[1,2], Lan Nguyen[3], Zachary Niziolek[6], Ceren Ozek[1,2], Vincent L. Butty[7], Ruth Isserlin[5], Sean M. Buchanan[1,2], Stuart S. Levine[7], Aviv Regev[3], Gary D. Bader[5], Joshua Z. Levin[3] and Lee L. Rubin[1,2,3*]

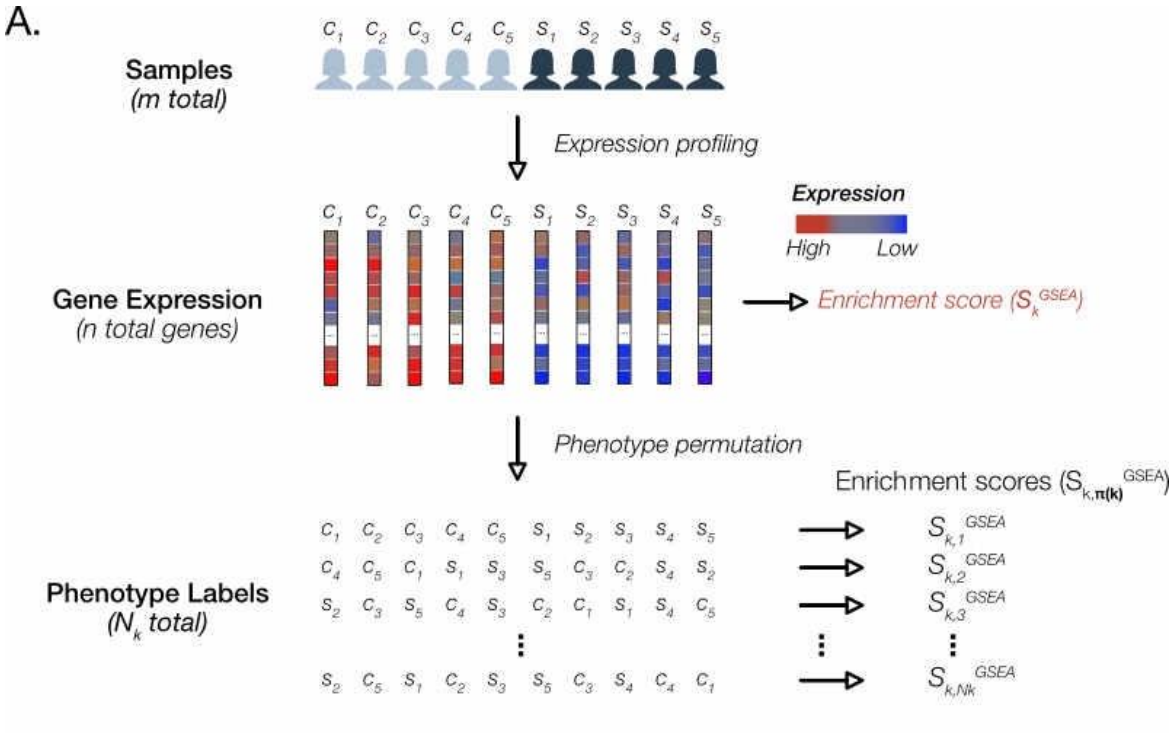
a

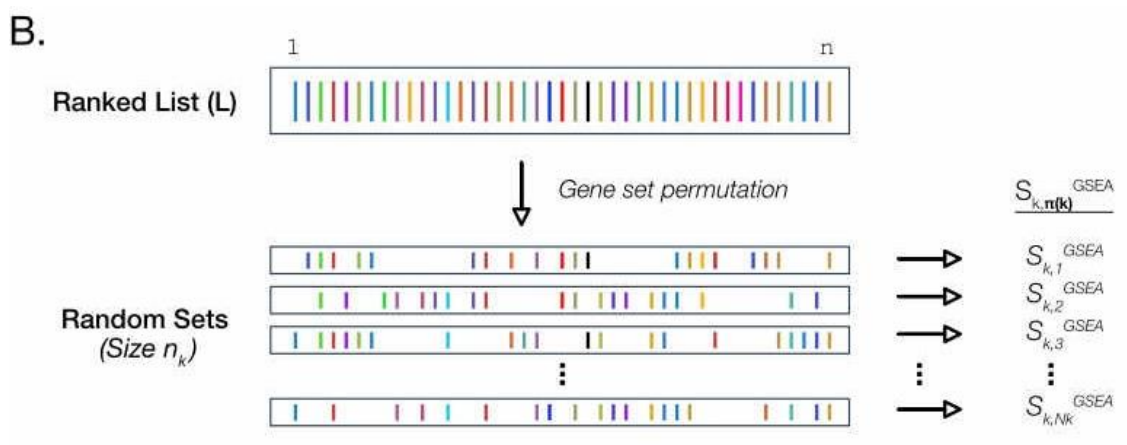


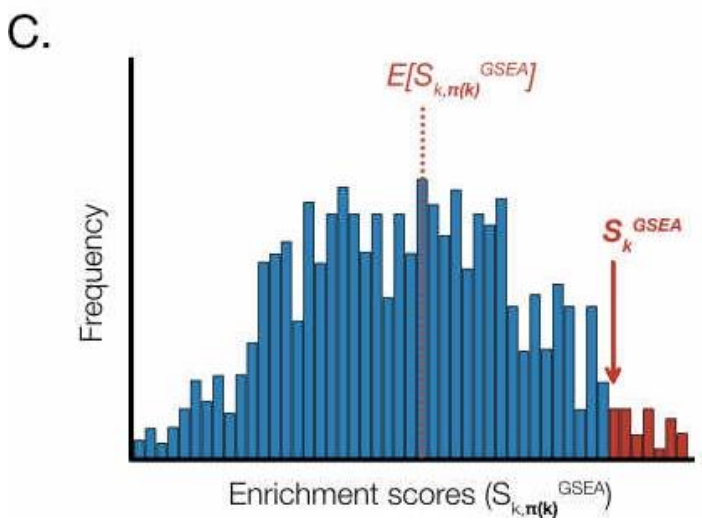
NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.



https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/





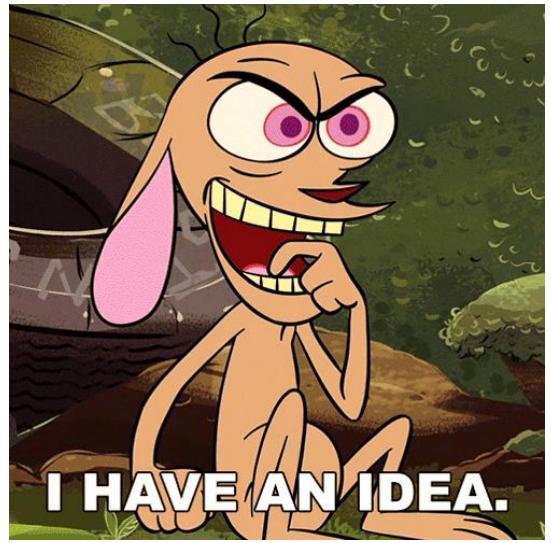


+ Normalize Enrichment Scores (NES) to Gene Set Size
+ Multiple Test Corrections

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

Experimental Design Considerations

1. Do I really need to do scRNA-seq?
2. What type of protocol do I need to use?
3. How much money do I have?
4. How do I mitigate batch effects?



Do I really need to do scRNA-seq?

- single cell type of interest - probably not
- landscape of cell types in a sample – probably
- Make sure you think through your specific single cell or spatial question
- TIP: Data is often hypothesis generating - should be validated with secondary assay (ex: qPCR)

What type of protocol do I need to use?

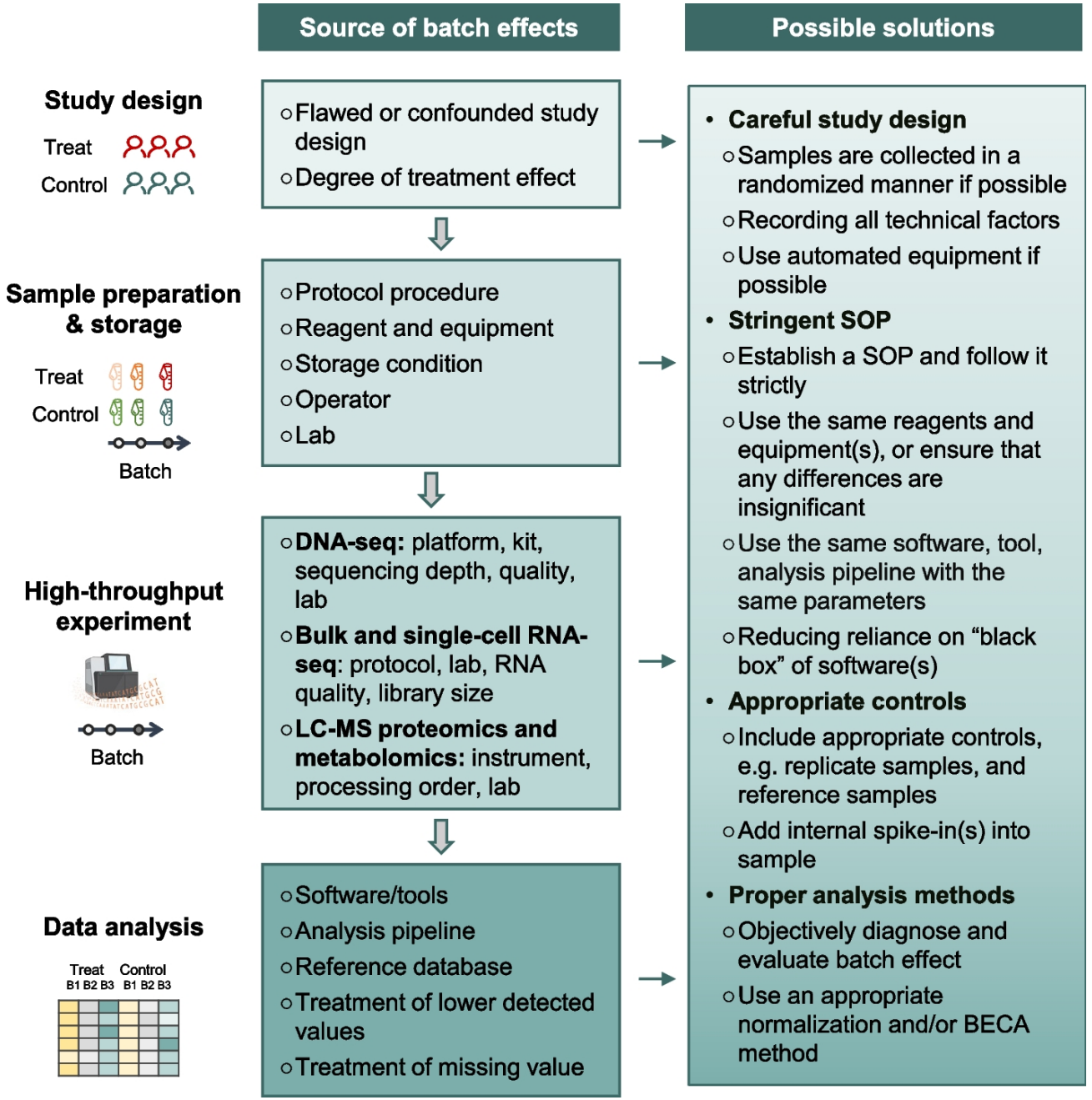
- species of interest (reference species vs. unusual organism
- transcript capture (poly-A 3' GEX, poly-A 5' GEX, probe-based)
- number of samples (multiplex vs. single plex)
- frequency of cell types of interest (FACS sorting, plate-based)
- frequency of transcripts of interest (droplet-based, plate-based)
- ability to identify cell types of interest (protein + GEX assays)
- input cell type/quality (fresh vs frozen vs fixed vs FFPE vs OCT)
- “Multiomics” – GEX paired with ATAC/TCR+BCR/Spatial/etc.

How much money do I have?

- capital investment (10X Chromium vs. PIP-seq vs. plate-based)
- number of samples
- n=1 per condition still acceptable, but n=3 or more allows for pseudobulking and DESeq2 differential expression
- sequencing requirements

How do I mitigate batch effects?

TIP: don't separate samples into different lanes/sequencing runs if you want to compare them!



Yu, Y., Mai, Y., Zheng, Y. *et al.* Assessing and mitigating batch effects in large-scale omics studies. *Genome Biol* **25**, 254 (2024). https://doi.org/10.1186/s13059-024-03401-9

PRO TIP: NCH Genomic Services Laboratory (GSL) has single cell core to help with planning and executing your experiments. Please reach out if you're interested!

Contact: Dr. Katie Miller, GSL Director (Katherine.Miller@nationwidechildrens.org)

- **EXPERT TIP**: The scRGOT group meets monthly and is here to help and share ideas! Come join us!

Get on the mailing list!
Contact: Dr. Katie Miller
(Katherine.Miller@nationwidechildrens.org)

- We also have a weekly bioinformatics meeting on Thursdays at 1PM!

Contact: Dr. Matthew Cannon

(Matthew.Cannon@nationwidechildrens.org)

nature reviews genetics

https://doi.org/10.1038/s41576-023-00586-w

Expert recommendation

Check for updates

Best practices for single-cell analysis across modalities

Lukas Heumos[1,2,3,28], Anna C. Schaar[1,4,5,28], Christopher Lance[1,6], Anastasia Litinetskaya[1,4], Felix Drost[1,3], Luke Zappia[1,4], Malte D. Lücken[1,7], Daniel C. Strobl[1,3,8,9], Juan Henao[1], Fabiola Curion[1,4], Single-cell Best Practices Consortium*, Herbert B. Schiller[2] & Fabian J. Theis[1,3,4,5]

https://www.nature.com/articles/s41576-023-00586-w

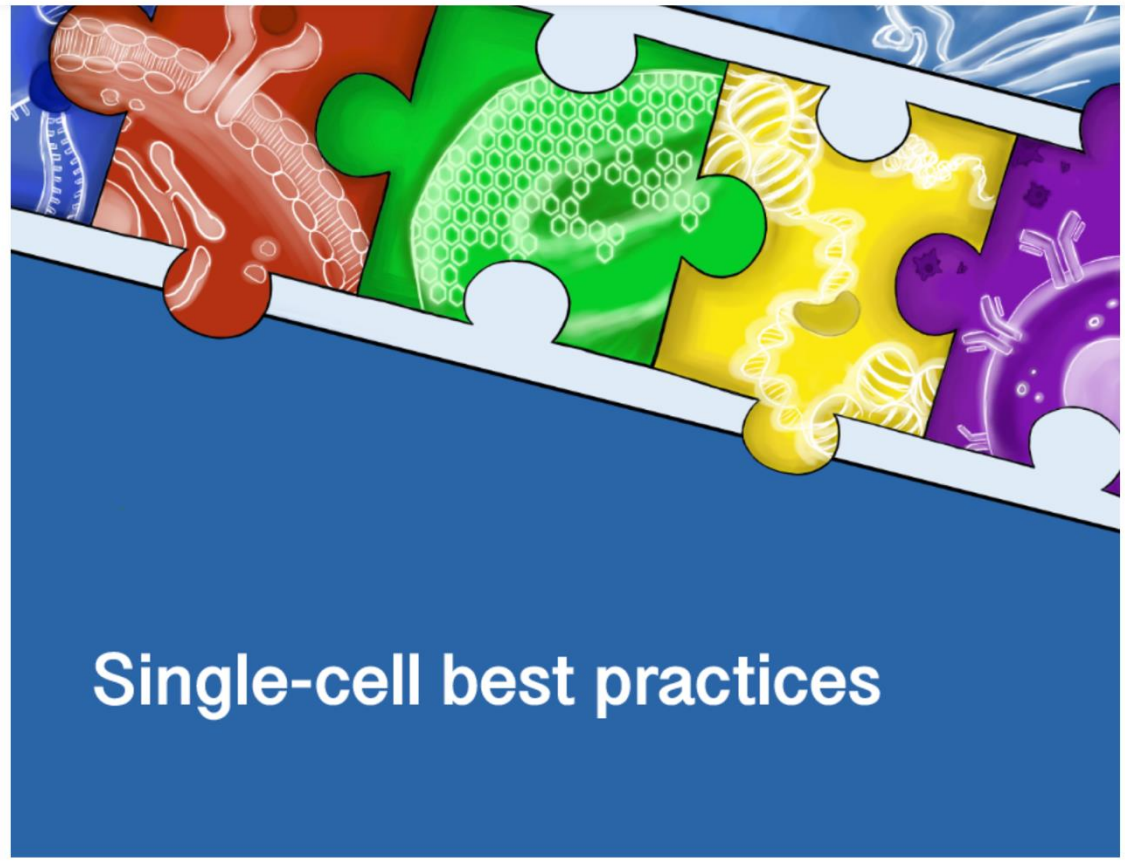
Search ⌘ + K

Introduction

1. Prior art
2. Single-cell RNA sequencing
3. Raw data processing
4. Analysis frameworks and tools
5. Interoperability

Preprocessing and visualization

6. Quality Control
7. Normalization
8. Feature selection
9. Dimensionality Reduction



Single-cell best practices

Contents

https://www.sc-best-practices.org/preamble.html