Single-cell RNA-seq analysis

Xinlei Gao

BaRC Hot Topics – March, 14th 2024

Bioinformatics and Research Computing
Whitehead Institute

http://barc.wi.mit.edu/hot_topics/

Whitehead Institute

BaRC Hot Topics

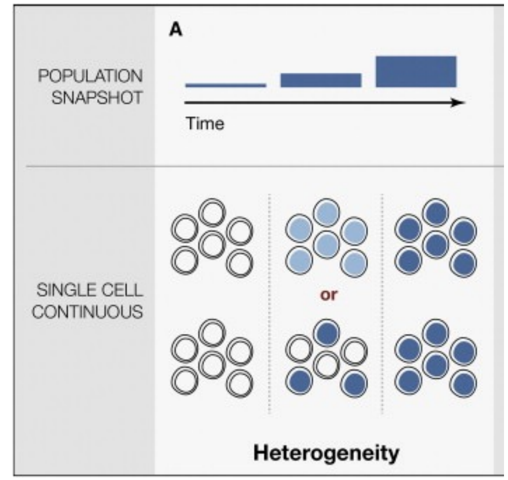
Outline

- Overview of scRNA-seq technology, cell barcoding, UMIs
- Experimental design
- Typical analysis pipeline
  - Preprocessing and quality control
  - Normalization
  - Dimensionality reduction
  - Clustering of cells
  - Labeling cells
  - Differential expression
  - Trajectory inference
- Integrating datasets
- Multimodal analysis

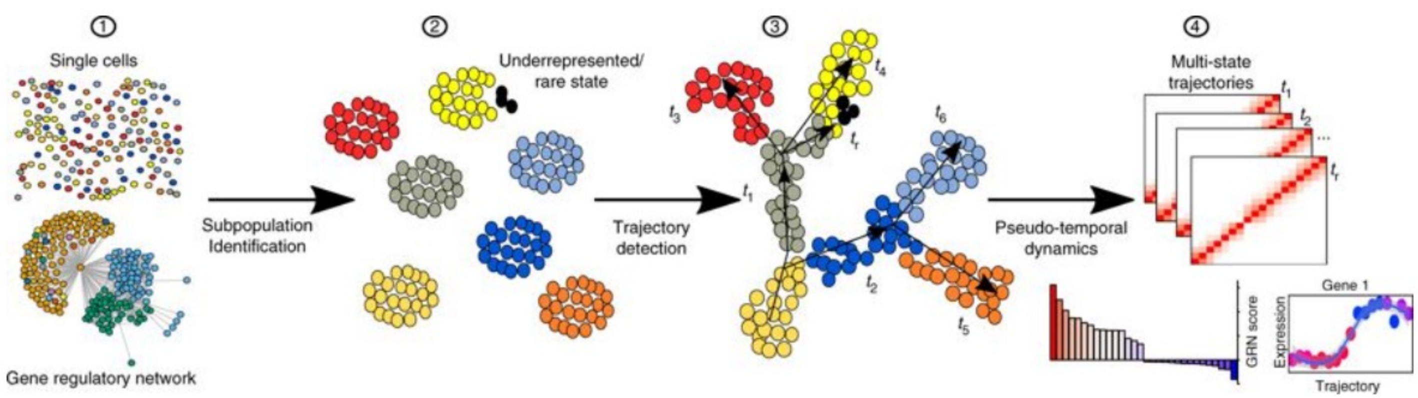
Why do single cell RNA-seq?

Access to expression profiles of individual cells allows us to:

- Learn about cellular heterogeneity
- Discover new cell populations
- Order cells within a developmental trajectory

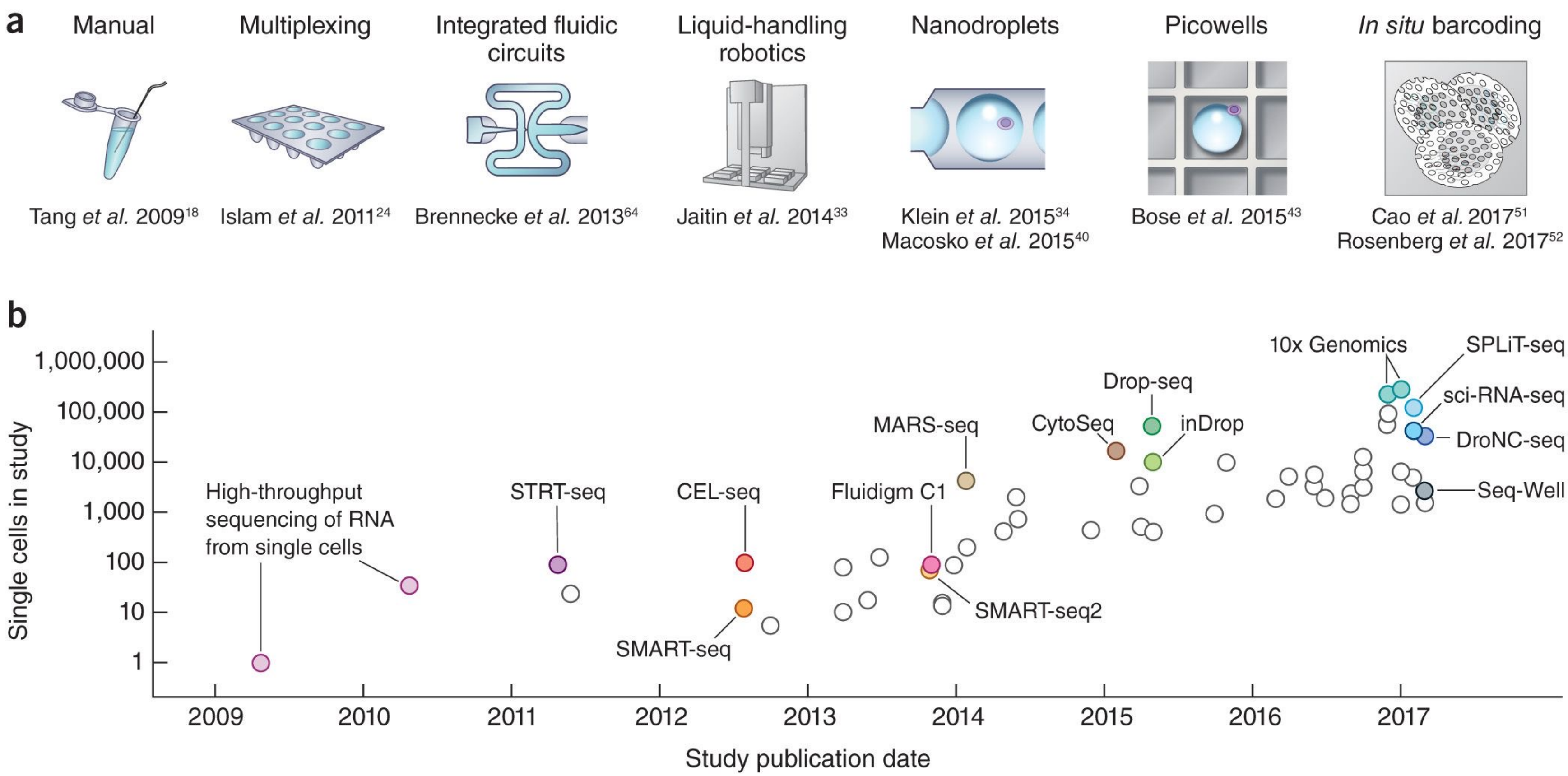


Etzrodt, *Cell Stem Cell* 2014



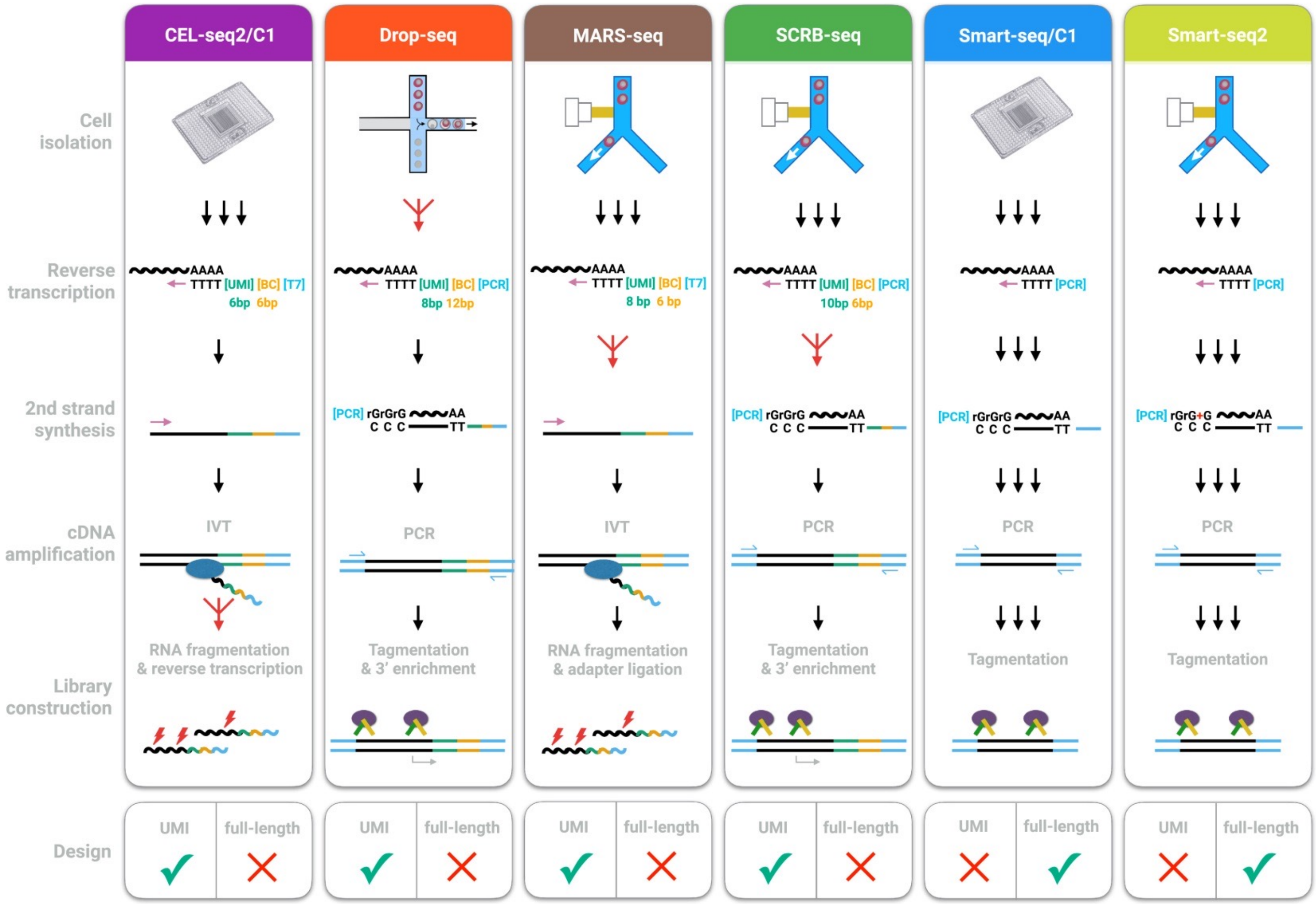
Lummertz da Rocha, *Nature Communications* 2018

Exponential scaling of single-cell RNA-seq in the past decade



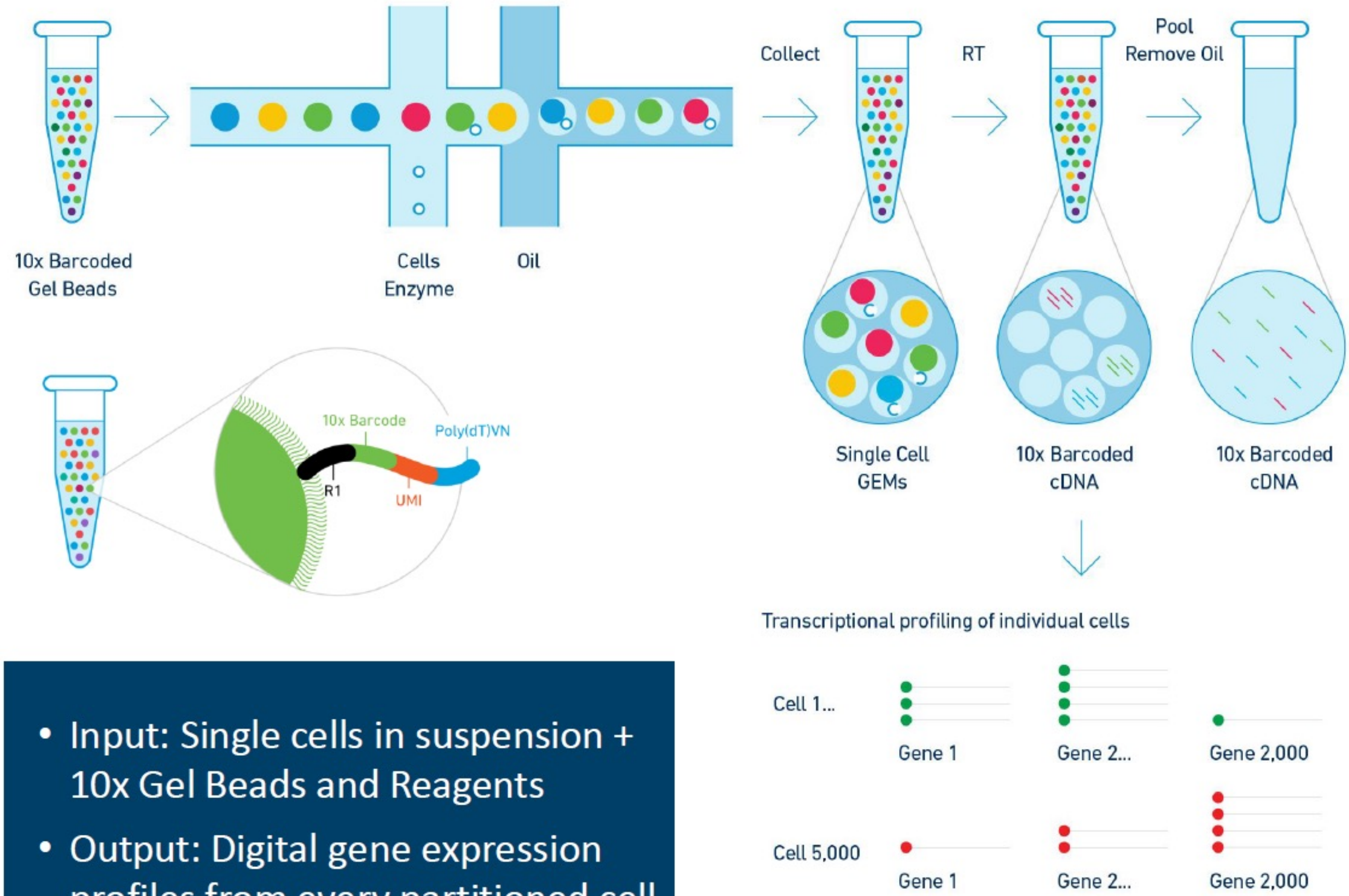
Svensson, V., Vento-Tormo, R. & Teichmann, S.
Nat Protoc **13,** 599–604 (2018). https://doi.org/10.1038/nprot.2017.149

Library preparation steps



Comparative Analysis of Single-Cell RNA Sequencing Methods
Ziegenhain et. al, Molecular Cell
Volume 65, Issue 4, 16 February 2017,

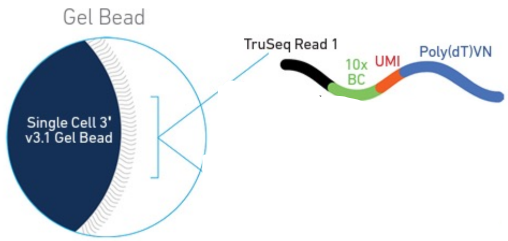
Single Cell Digital Gene Expression



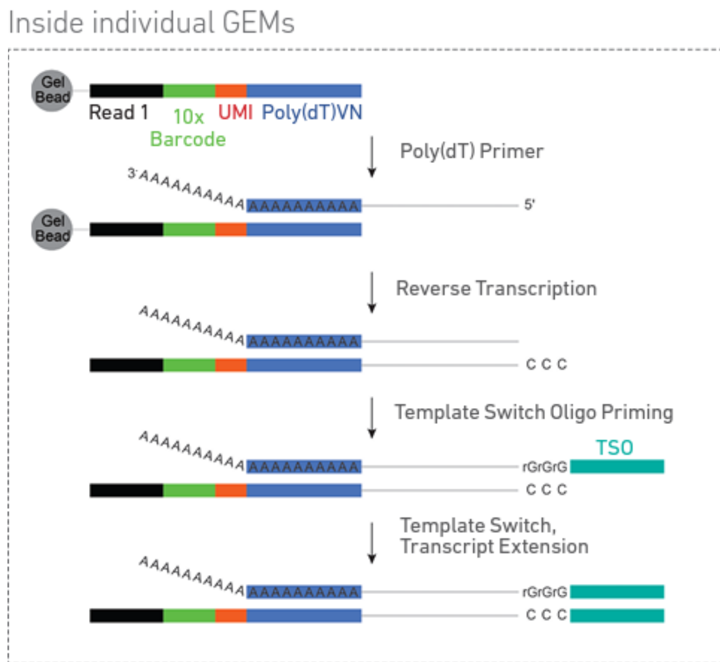
- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

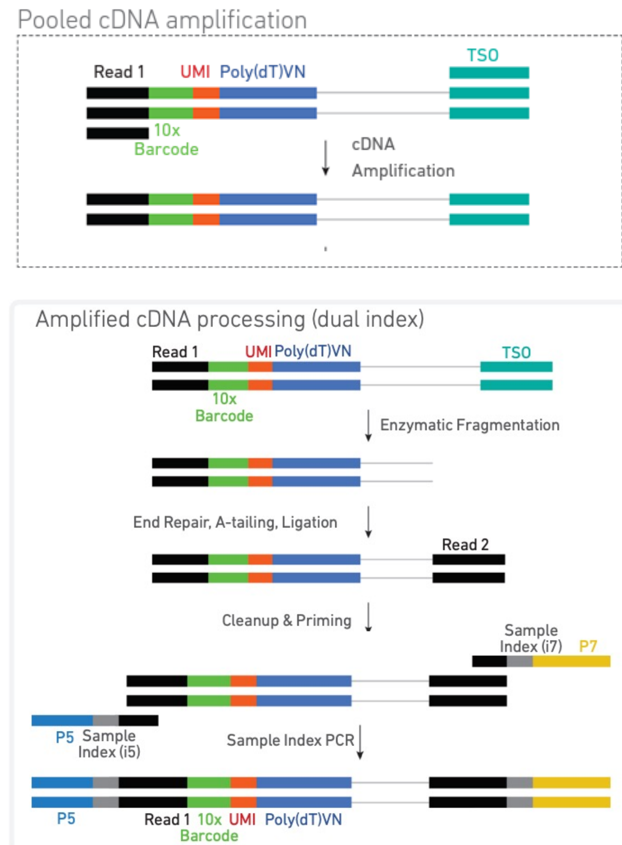
https://pages.10xgenomics.com/sup-training-single-cell-gene-expression-landing-page.html

Chromium Single Cell 3′ Reagent Kit



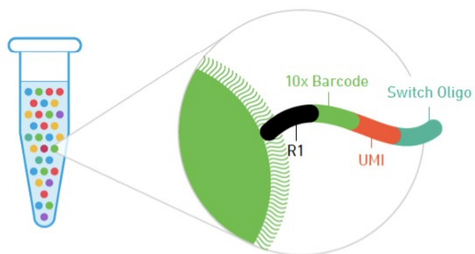
Single Cell 3′ Gel Beads





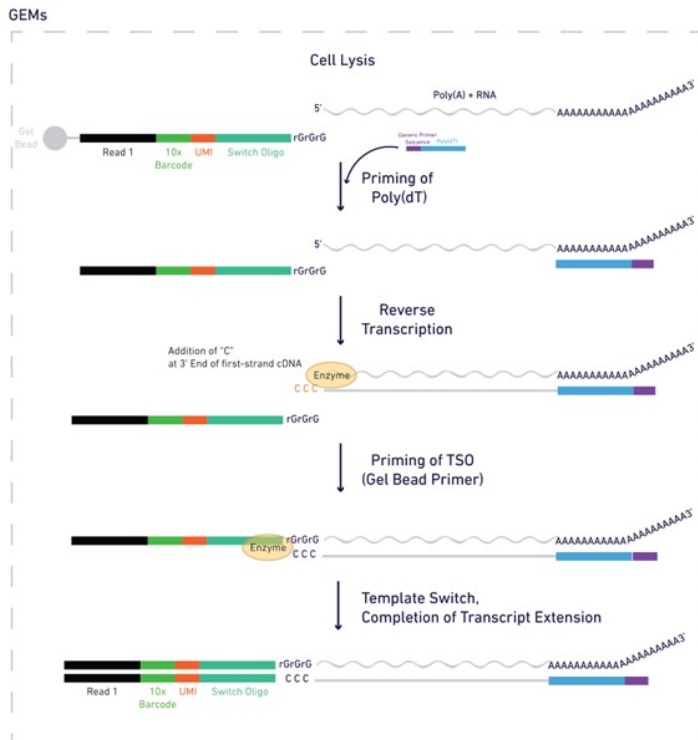
CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex__RevE.pdf

Chromium Single Cell 5’ Reagent Kit

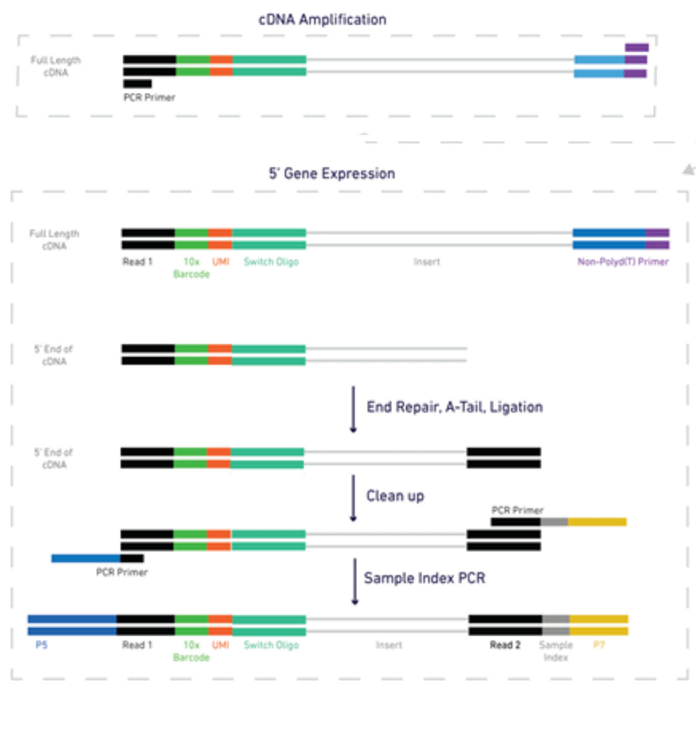


Single cell 5’ Gel Bead Oligo primer

Inside individual GEMs



Pooled cDNA processed in bulk



CG000331_ChromiumNextGEMSingleCell5-v2_UserGuide_RevE.pdf

Important differences between technologies

- Three prime bias
  - *i.e.* 3 prime versus 5 prime 10x genomics kits
- Gene coverage
  - *i.e.* Smart-seq versus 10x chromium 3’ kit
- High throughput or low throughput
  - *i.e.* 10x, Drop-seq versus Fluidigm C1
- Sensitivity (the number of detected genes)

Experimental design

- Process your samples in a way that the conditions can not be confounded with a batch effects, like processing date, facility, or reagents used.
  - *i.e.* If you have to process your cells in several batches, each batch should contain an equal number of samples and cells from each condition.
- Minimize processing time.
- For certain cell types, *i.e.* neurons, other techniques like single cell nuclei may be more appropriate.
- Number of reads required.
- Number of cells vs. coverage for each cell.

Typical analysis pipeline

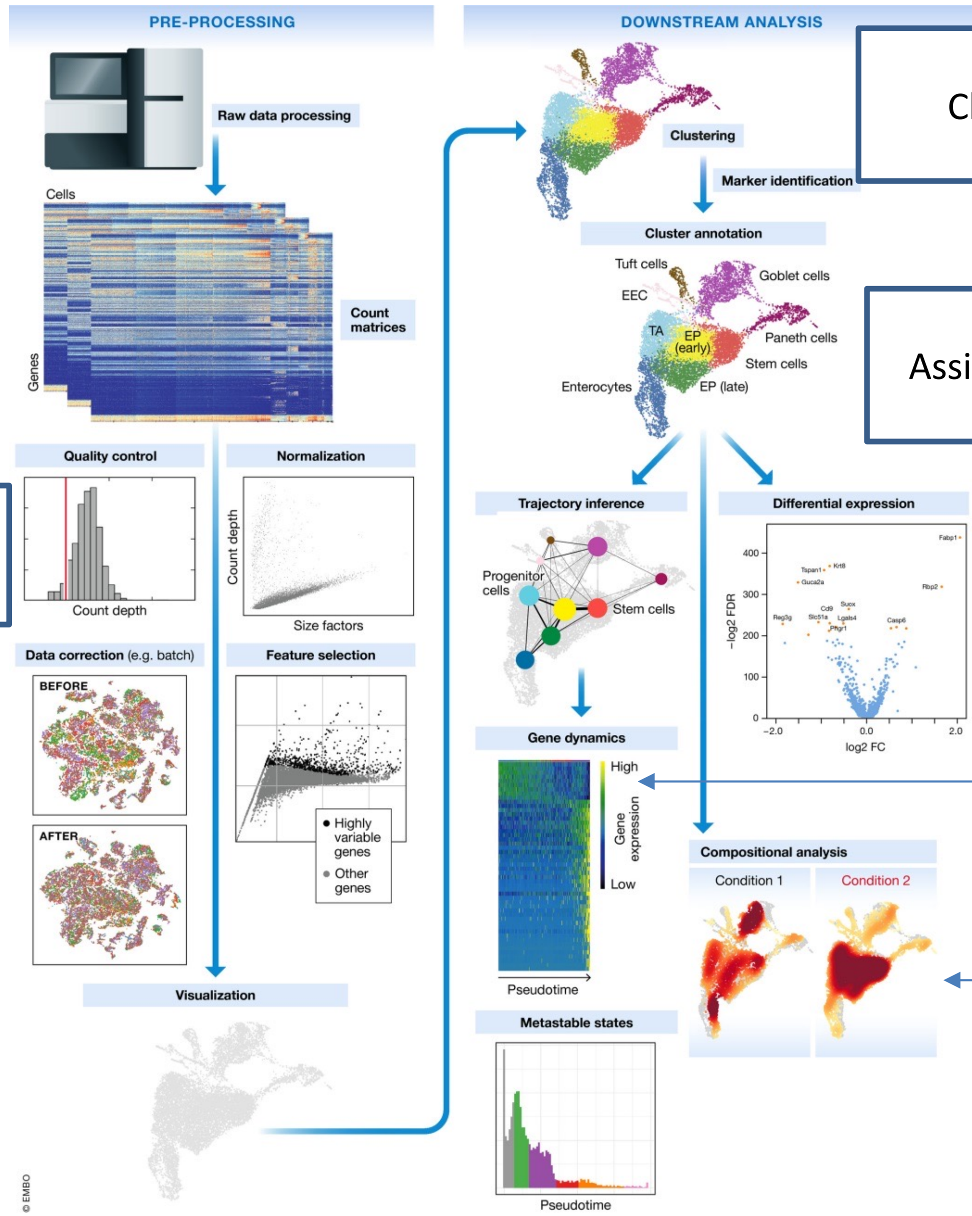
Expression Matrix (GENES x CELLS)

Filter Cells/Quality Control

Normalization

Identify Variable Genes

Dimensionality Reduction



Clustering

Exploring Known Marker Genes

Assign Cell Type

Differential Expression Analysis

Trajectory modeling and Gene expression dynamics

Compositional analysis

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

Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019

Technical challenges

- Data is noisy due to
  - cDNA amplification bias
  - mRNA capture efficiency
  - Large number of genes with 0 counts due to limiting mRNA. Zero expression doesn't mean the gene isn't on.
- Cells can change or die during isolation.

Review | Open Access | Published: 07 February 2020

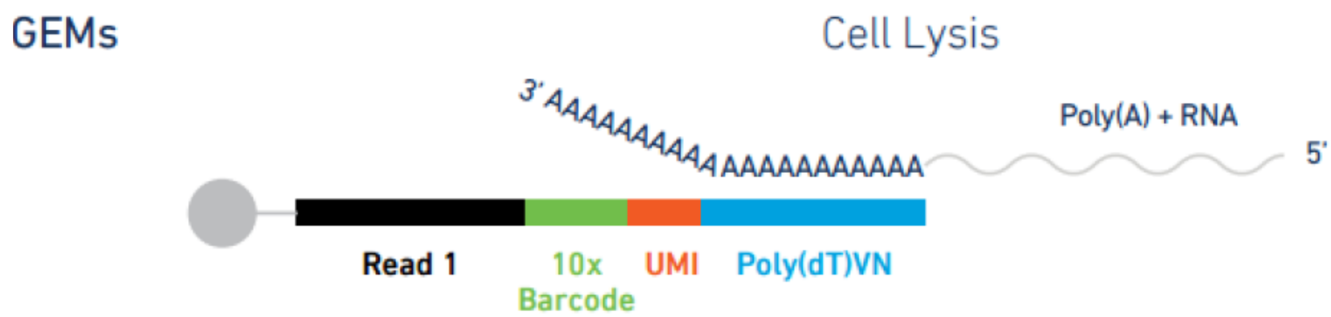
Eleven grand challenges in single-cell data science

David Lähnemann, Johannes Köster, … Alexander Schönhuth + Show authors

Genome Biology **21**, Article number: 31 (2020) | Cite this article

75k Accesses | **227** Citations | **286** Altmetric | Metrics

Preprocessing for technologies using Unique Molecular Identifiers (UMIs)



- Demultiplexing: assign all the reads with the same cell barcode to the same cell.
- Remove PCR duplicates: if several reads have the same UMI and map to the same location in the genome, keep only one.
  - Cell ranger software for 10x data (run by the genome technology core)
  - Drop-seq tools for drop-seq and seq-well data

https://www.10xgenomics.com/videos/training-modules/

Demultiplexing and counting 10x data

Cell Ranger™ Pipelines

10x GENOMICS™

Pipeline	Functionality
cellranger mkfastq	Barcode-aware demultiplexing from BCL to FASTQ

cellranger count	• Read-level analysis of a single library
	    • Transcriptome alignment with STAR
	    • Barcode processing
	    • Gene counting
	• Produces gene/cell matrix
	• Produces expression analysis and static visualizations
	• Produces .cloupe file for Loupe™ Cell Browser

https://www.10xgenomics.com/videos/training-modules/

CellRanger web summary

Cell Ranger

SUMMARY ANALYSIS

Estimated Number of Cells

2,580

Mean Reads per Cell	Median Genes per Cell
97,620	2,559

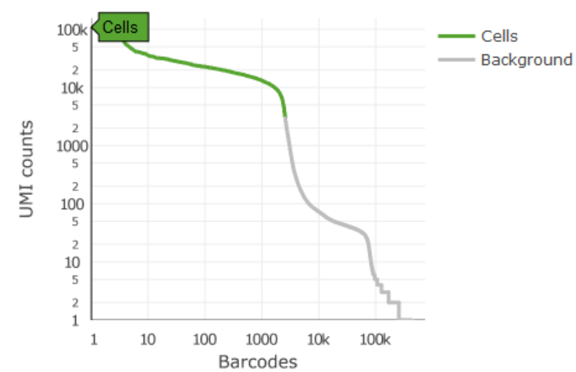
Sequencing

Number of Reads	251,861,835
Valid Barcodes	96.1%
Sequencing Saturation	78.1%
Q30 Bases in Barcode	94.7%
Q30 Bases in RNA Read	66.6%
Q30 Bases in Sample Index	87.9%
Q30 Bases in UMI	94.5%

Mapping

Reads Mapped to Genome	86.4%
Reads Mapped Confidently to Genome	80.5%
Reads Mapped Confidently to Intergenic Regions	2.7%
Reads Mapped Confidently to Intronic Regions	6.7%
Reads Mapped Confidently to Exonic Regions	71.0%
Reads Mapped Confidently to Transcriptome	69.2%
Reads Mapped Antisense to Gene	0.8%

Cells



Estimated Number of Cells	2,580
Fraction Reads in Cells	88.1%
Mean Reads per Cell	97,620
Median Genes per Cell	2,559
Total Genes Detected	15,848
Median UMI Counts per Cell	11,729

Sample

Name	L21_314
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.1

Lots of software available to analyze single-cell RNA-seq data

- **Seurat (R)**
- Scanpy (python)
- Monocle, Slingshot
- Destiny, scvelo, CellRank, Dynamo, MultiVelo
- See https://github.com/seandavi/awesome-single-cell

Review | Open Access | Published: 29 October 2021

Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape

Luke Zappia & Fabian J. Theis

Genome Biology **22**, Article number: 301 (2021) | Cite this article

6925 Accesses | **86** Altmetric | Metrics

Seurat

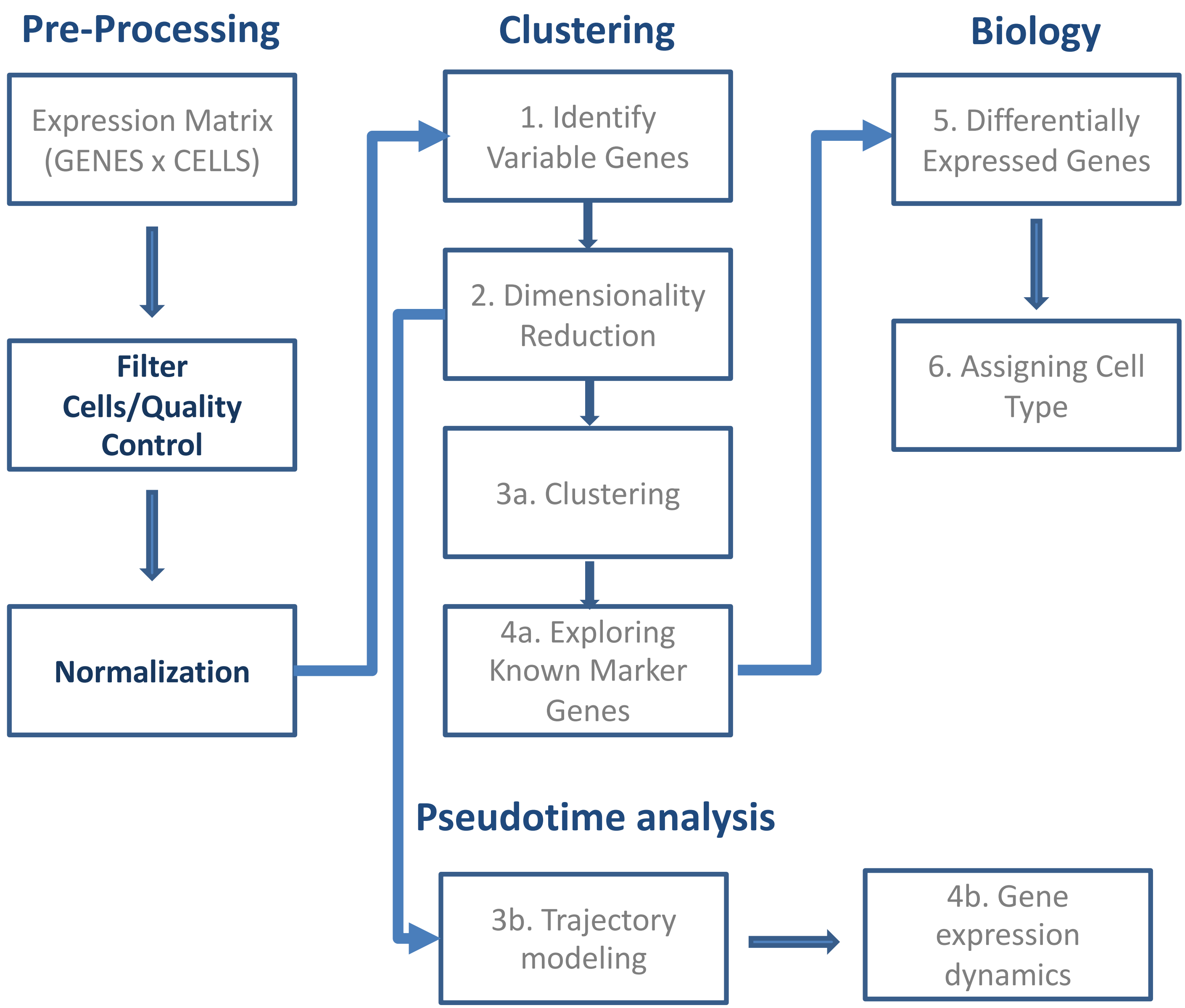
https://satijalab.org/seurat/

- Seurat is an R package designed for QC, analysis, and exploration of single cell RNA-seq data.
- Developed and by the Satija Lab at the New York Genome Center.
- It is well maintained and well documented.
- It has a built in function to read 10x Genomics data. It can de-multiplex hash tag data.
- It has implemented most of the steps needed in common analyses.

Major new features of Seurat in current version 5

- Integrative multimodal analysis
  - ‘bridge integration’
- Flexible, interactive, and highly scalable analysis
  - In-memory and on-disk storage
- Support for diverse spatial datasets
- Seurat object structure
  - Store data in layers (previously ‘slots’)
- Pseudobulk analysis
  - Reduce noise, improve quantification for lowly expressed genes

Typical analysis pipeline



Quality control and filtering

- Quality control
  - Number of reads per cell
  - Number of genes detected per cell
  - Proportion of transcript counts deriving from the mitochondria
- Remove cells with poor quality
  - Filter out cells with percentage of transcript counts deriving from the mitochondria higher than a cut off
  - Filter out cells with less than a lower threshold on the number of genes or counts per cell
- Remove doublets (two cells captured with one bead in the droplet)
  - Filter out cells with more than an upper threshold on the number of genes or counts per cell in your data
  - More sophisticated way of removing doublets
    - https://github.com/JonathanShor/DoubletDetection
    - https://github.com/AllonKleinLab/scrublet
    - DoubletFinder https://www.sciencedirect.com/science/article/pii/S2405471219300730?via%3Dihub

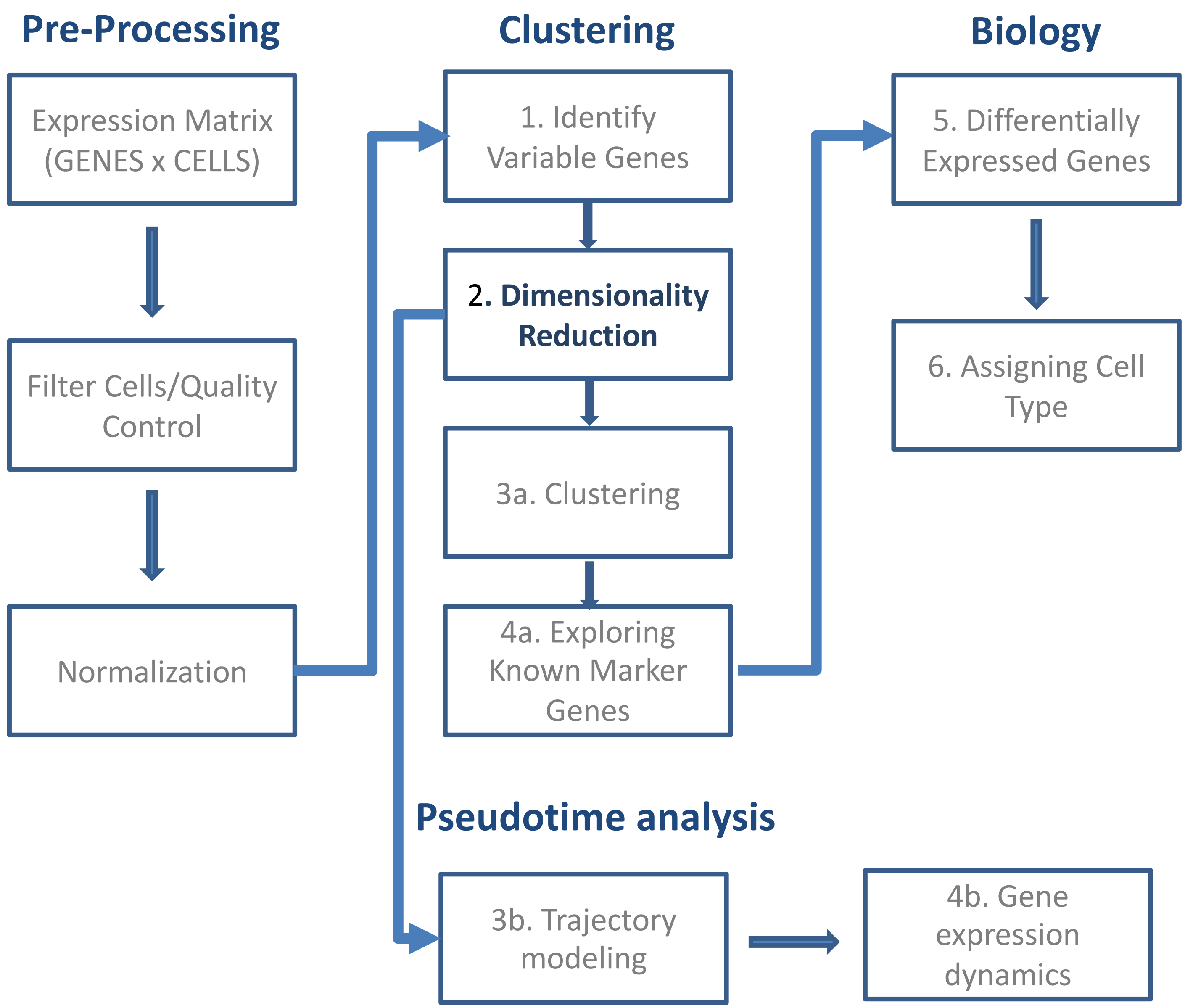
Normalization

Correct for sequencing depth (i.e. library size) of each cell so we can compare across cells

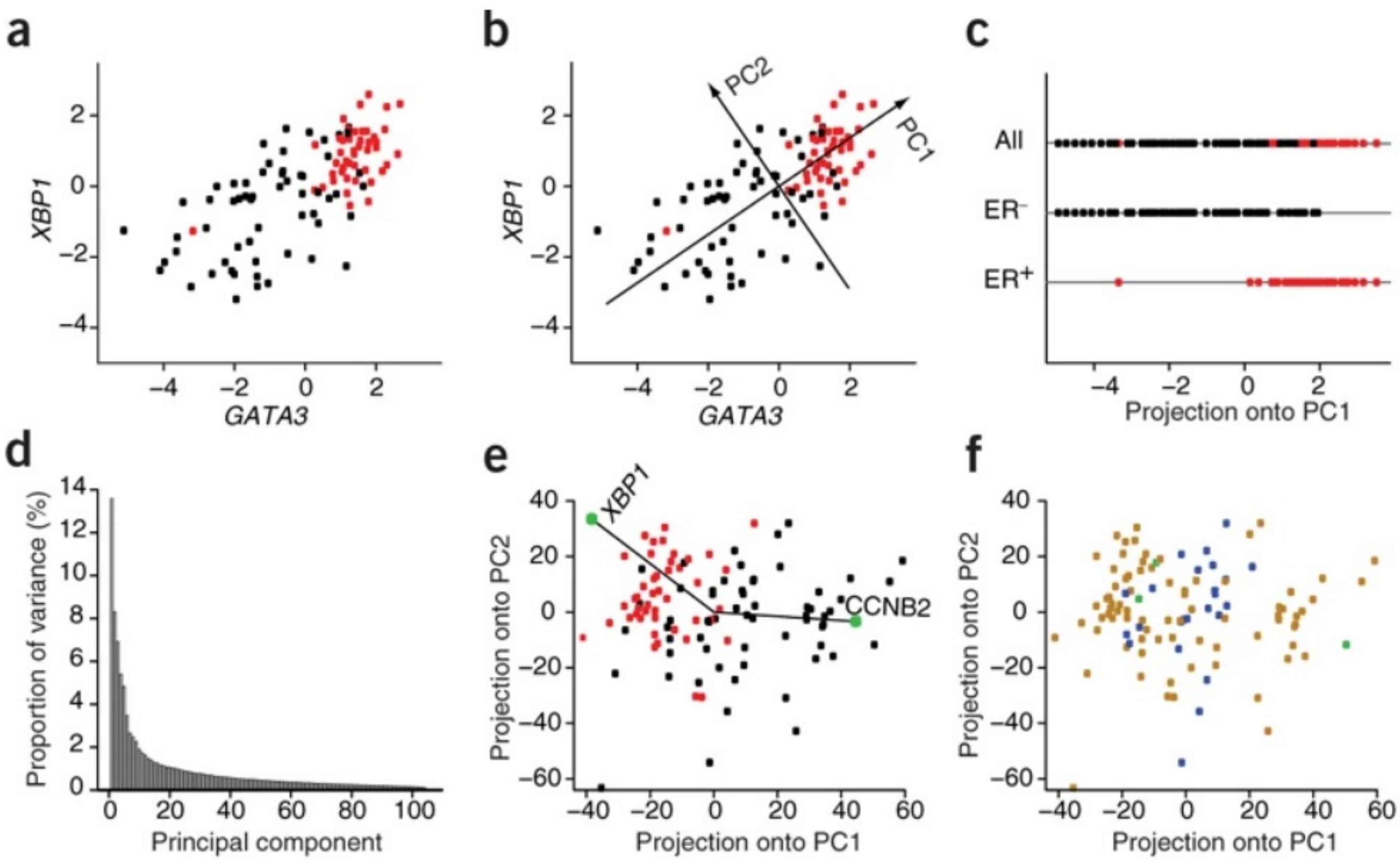
1. Normalize gene levels for each cell by total expression
2. Multiply by a scale factor (*i.e.* 10,000).
3. Log transform the scaled counts

This is the log normalization implemented in Seurat

Typical analysis pipeline



Visualization: Principal Component Analysis



Ringnér, M. What is principal component analysis?. *Nat Biotechnol 26, 303–304 (2008).* https://doi.org/10.1038/nbt0308-303

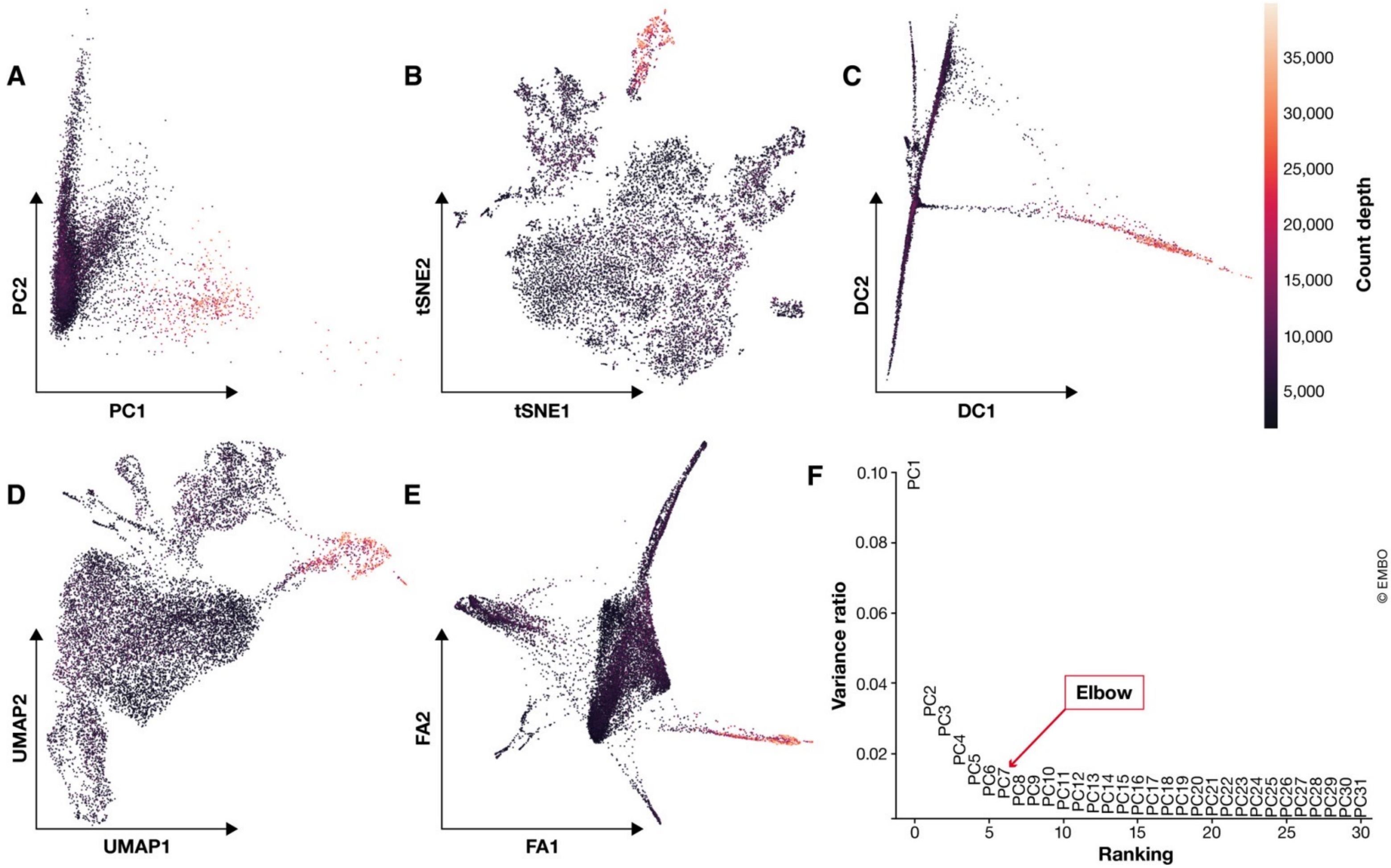
Other dimensionality reduction methods

Cells in 20000 (genes) dimensional space $\xrightarrow{\text{PCA}}$ Cells in 10-50 principal components space

How can we further summarize these multiple PCAs into just 2 dimensions?

Cells in 10-50 principal components space $\xrightarrow{\text{tSNE, UMAP, other}}$ Cells in 2D space

Visualization: dimensionality reduction



Current best practices in single-cell RNA-seq analysis: a tutorial
Mol Syst Biol, *Volume: 15, Issue: 6, First published: 19 June 2019*

t-Distributed Stochastic Neighbor Embedding (tSNE)

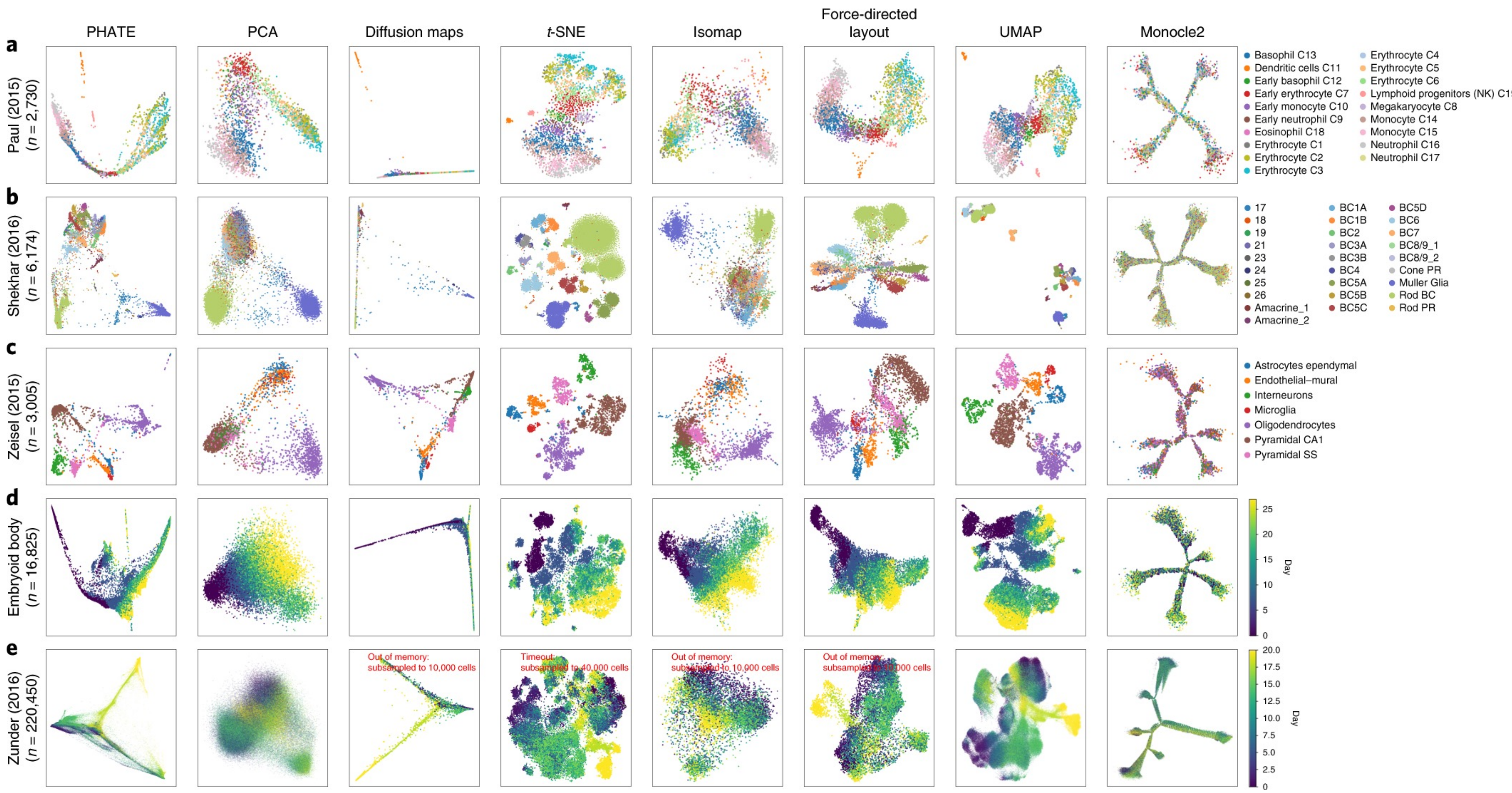
- Takes a set of points in a high-dimensional space and finds a faithful representation of those points in a lower-dimensional space, typically the 2D plane.
- The algorithm is **non-linear** and adapts to the underlying data, performing different transformations on different regions.
- The t-SNE algorithm adapts its notion of “distance” to regional density variations in the data set. As a result, it naturally expands dense clusters, and contracts sparse ones, evening out cluster sizes.
- Distances between clusters might not be biologically meaningful.

https://distill.pub/2016/misread-tsne/

UMAP
Uniform manifold approximation and projection

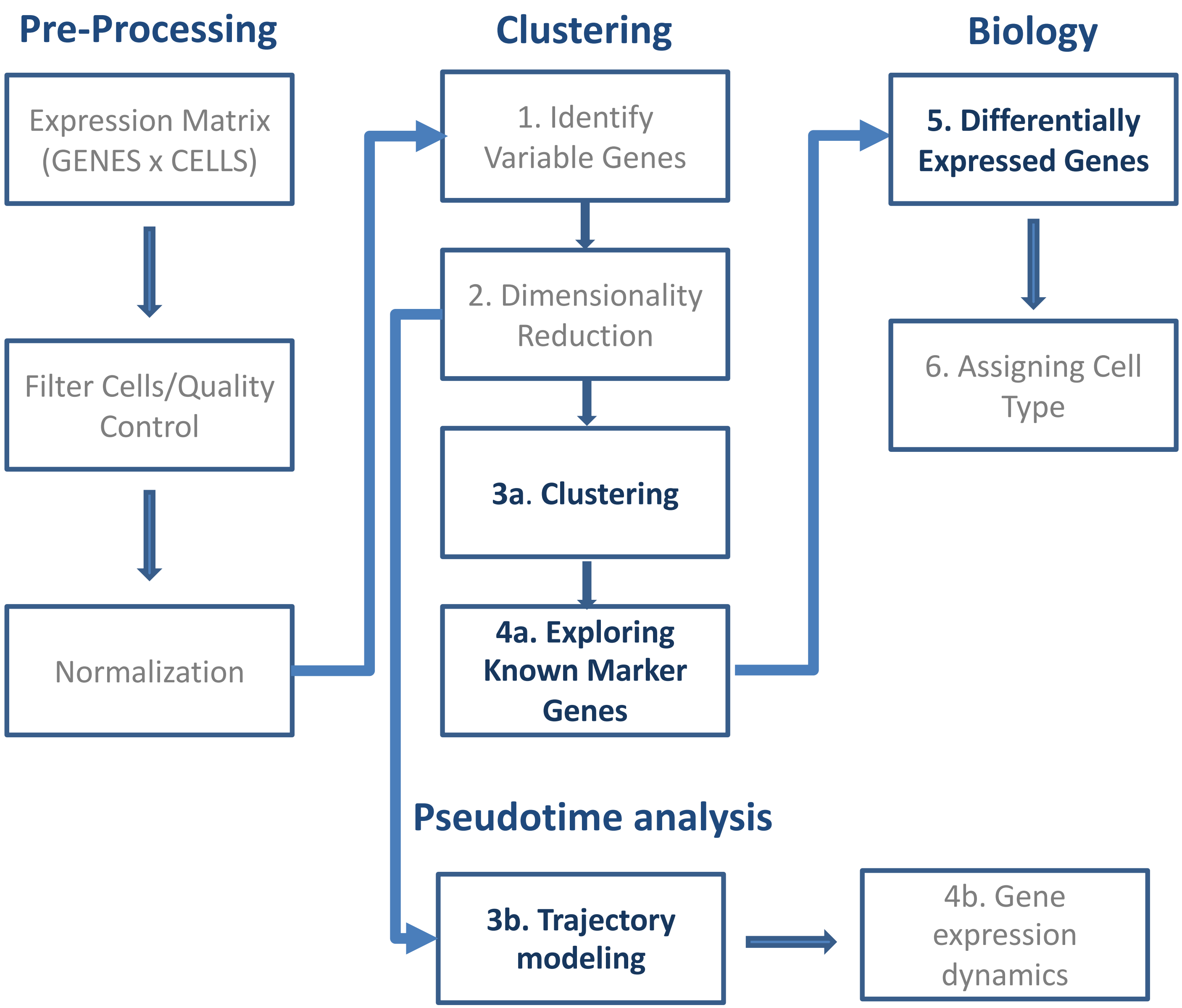
- It is a non linear dimensionality reduction algorithm.
- Preserves the local structure but also the global structure and the continuity of the cell subsets better.
- See PMID: 30531897 for comparison of tSNE and UMAP.

Comparison of visualization methods on biological datasets

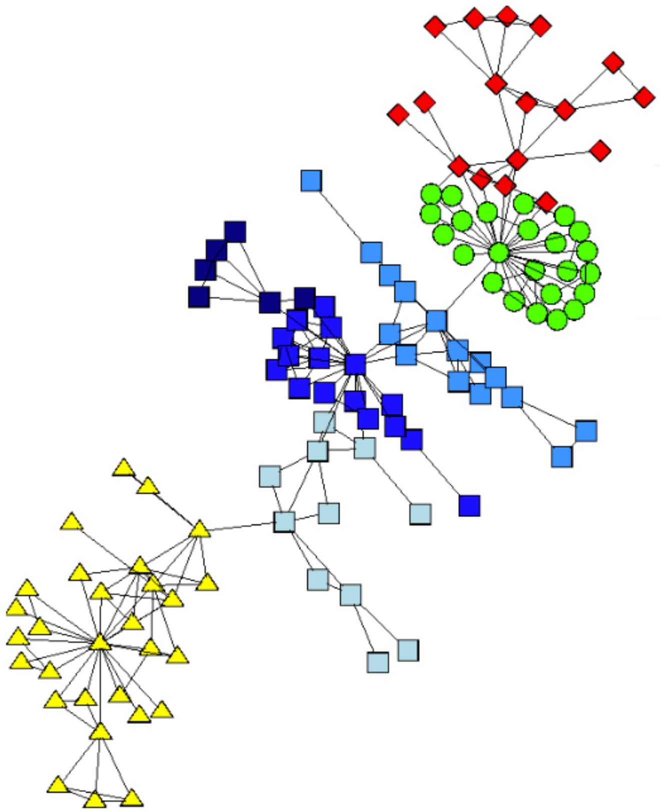


Moon, K.R., van Dijk, D., Wang, Z. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37,** 1482–1492 (2019).

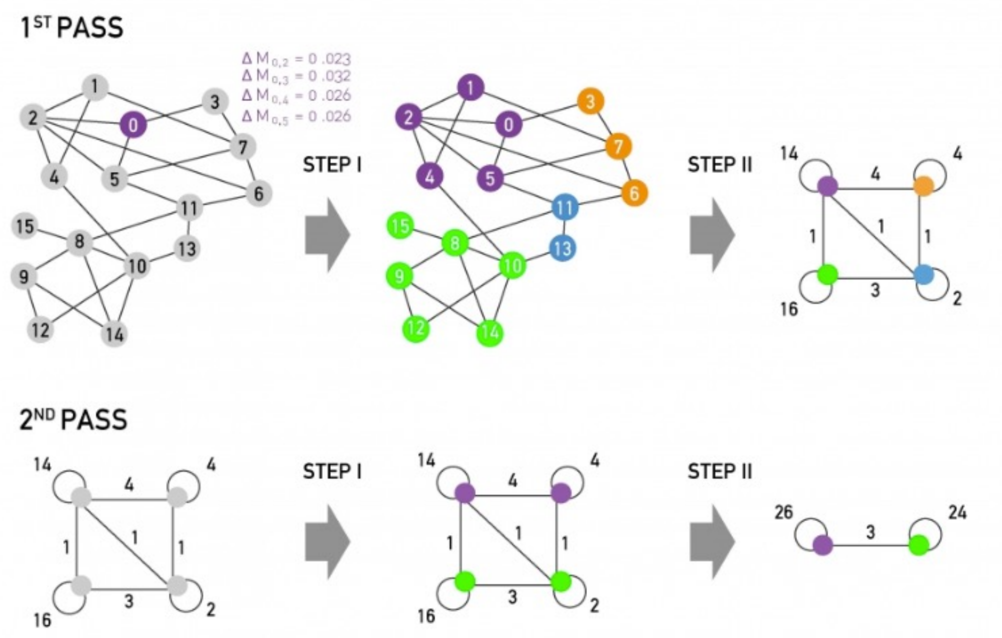
Typical analysis pipeline



Graph based clustering



Example of communities in a graph



Louvain algorithm Blondel et al.
Journal of Statistical Mechanics: Theory and Experiment 2008

https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/index.html

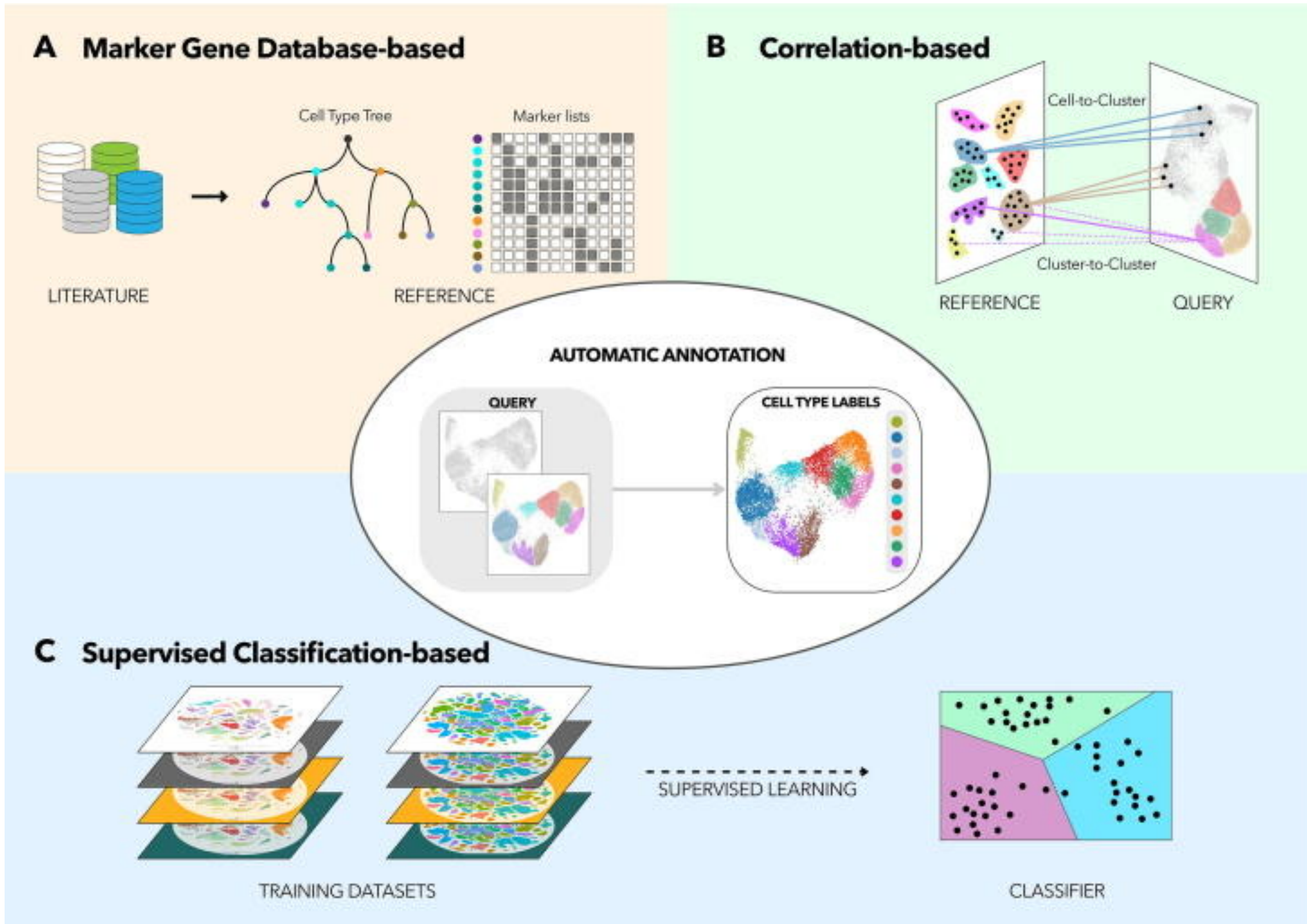
Clustering and Biology: What do you want to learn from the experiment?

- Classify cells and discover new cell populations (*i.e.* Louvain algorithm)
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Differential expression analysis between clusters

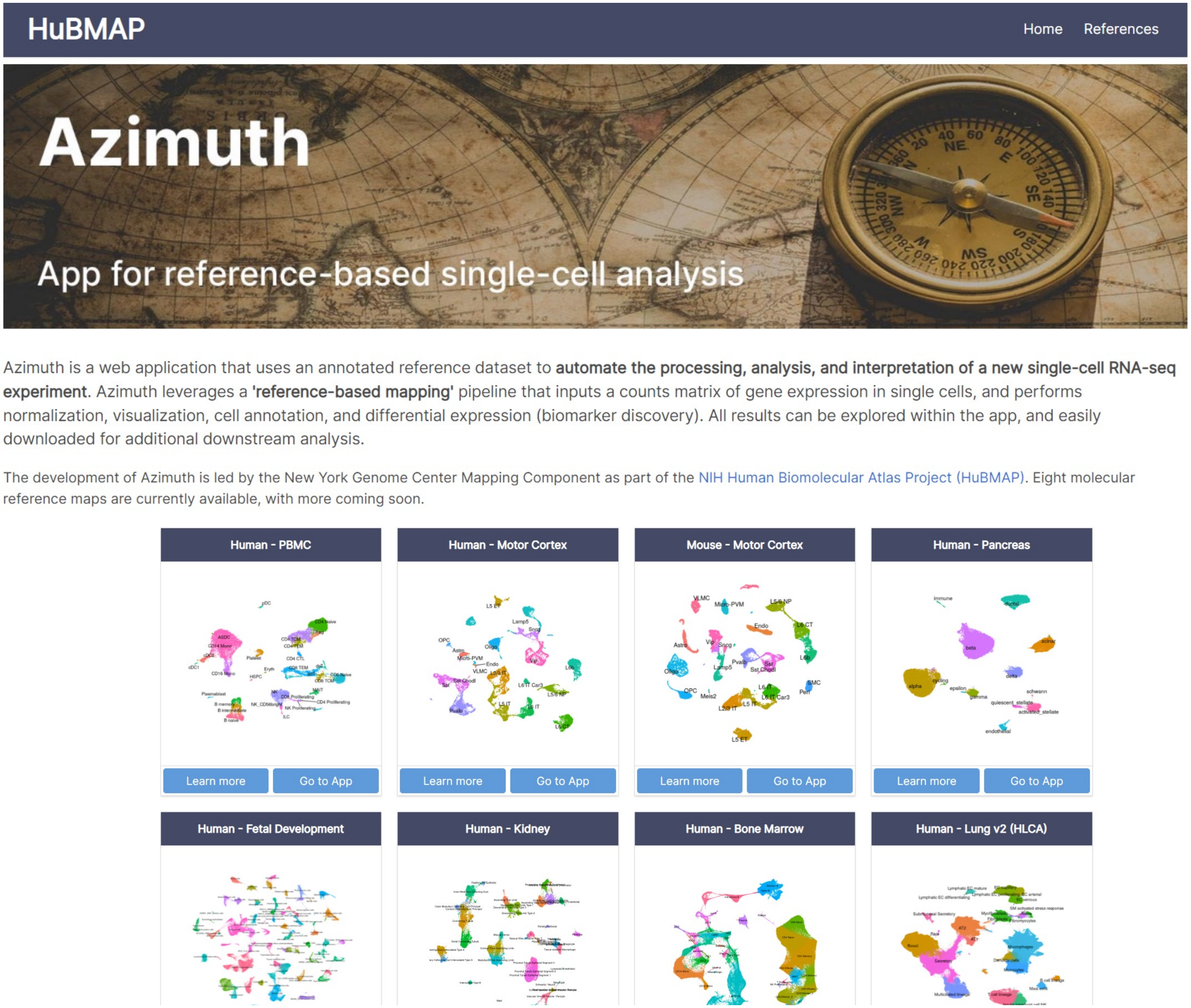
- Finds marker genes that will help determine the identity of the clusters.
- Since the expression data used to find the clusters and the markers is the same, the P-values are inflated and can lead to an overestimation of marker genes.
- The ranking of genes based on P-values is unaffected and it is a better way of selecting marker genes.

Cell type annotation

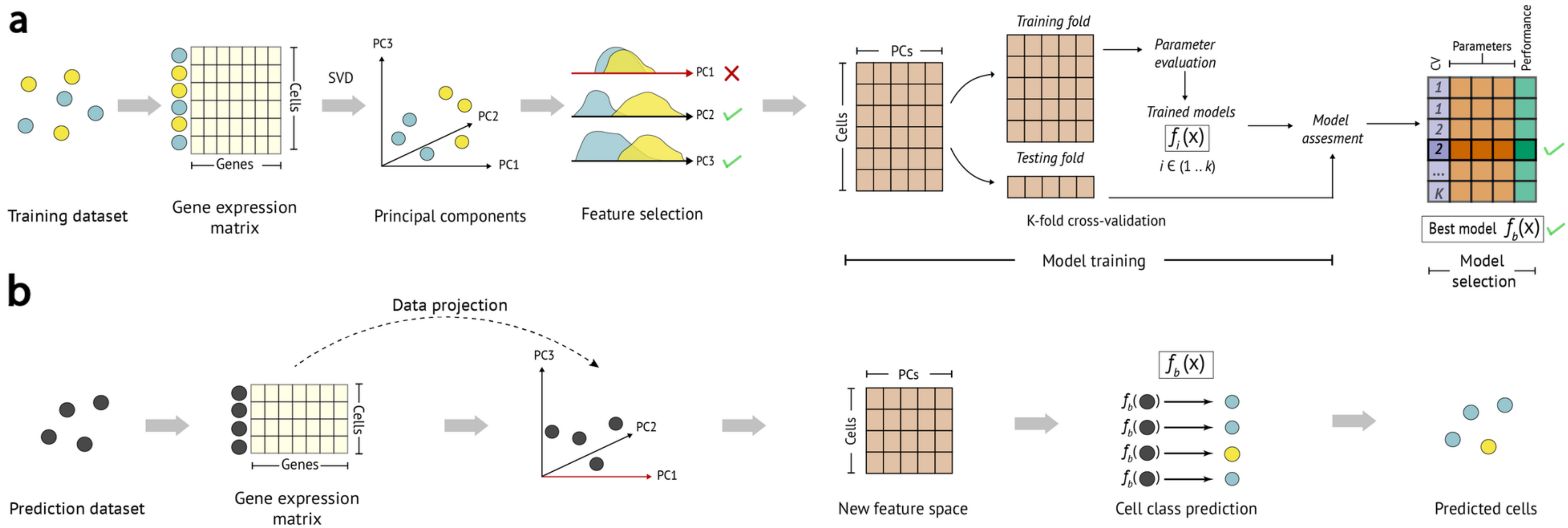


Pasquini *et al.* Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J. 2021 Jan 19*.

Cell type annotation: Azimuth



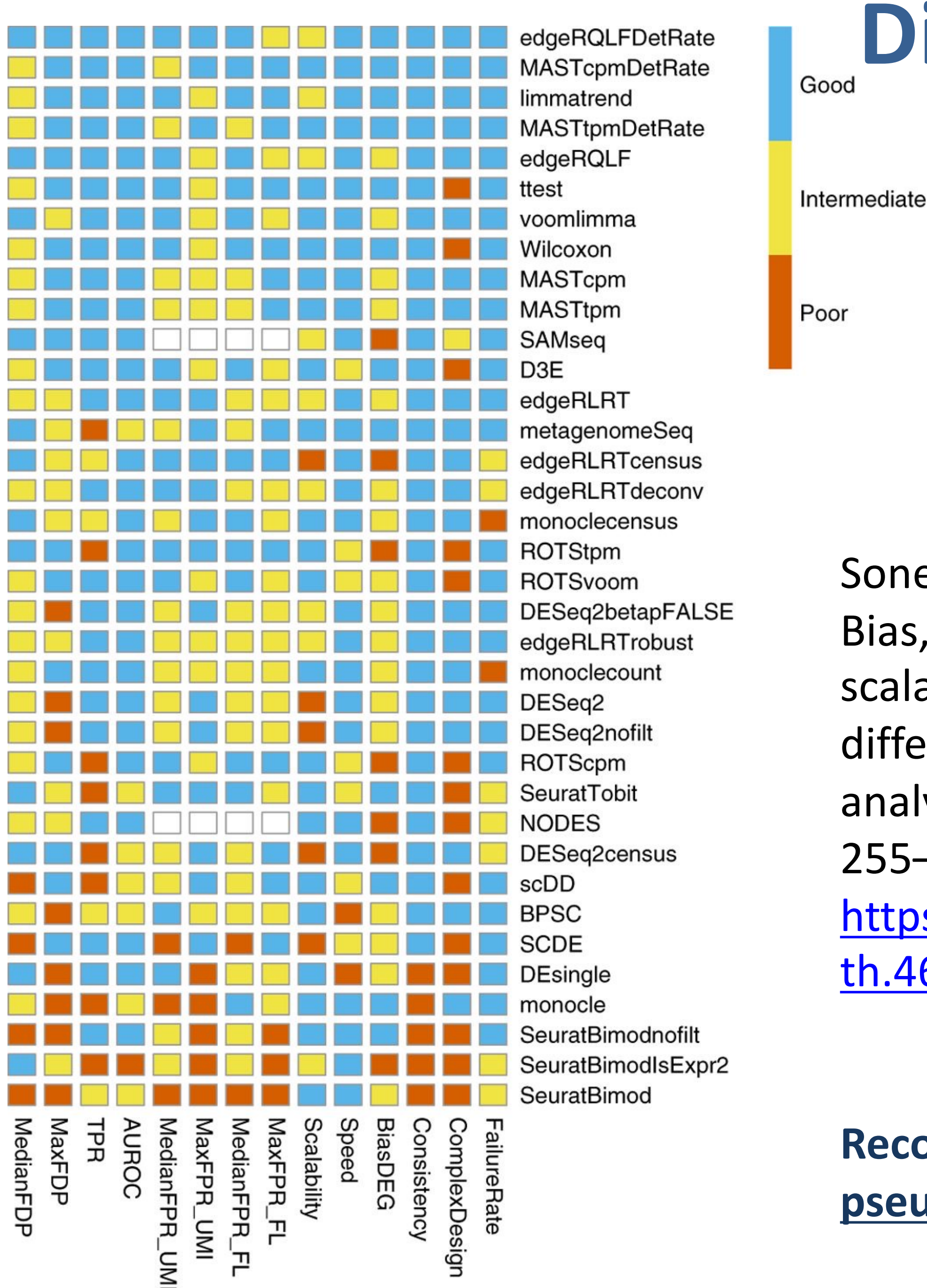
Cell type annotation: scPred



Alquicira-Hernandez *et al. scPred*: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* **20**, 264 (2019).

We have a command line tool to execute the 'scPred' method and pre-trained models using Tabula Sapiens data on the slurm cluster: /nfs/BaRC_Public/BaRC_code/R/scRNAseq_cell_type_classification/

Differential expression analysis between conditions



Soneson, C., Robinson, M. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* 15, 255–261 (2018). https://doi.org/10.1038/nmeth.4612

Recommended:
pseudo bulk methods

Clustering and Biology:

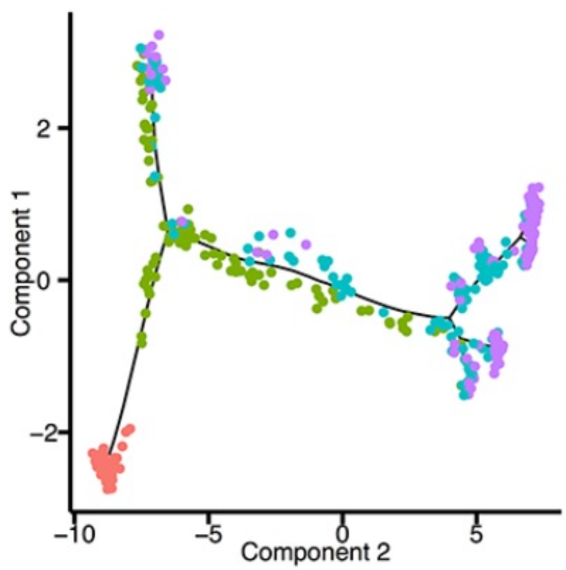
What do you want to learn from the experiment?

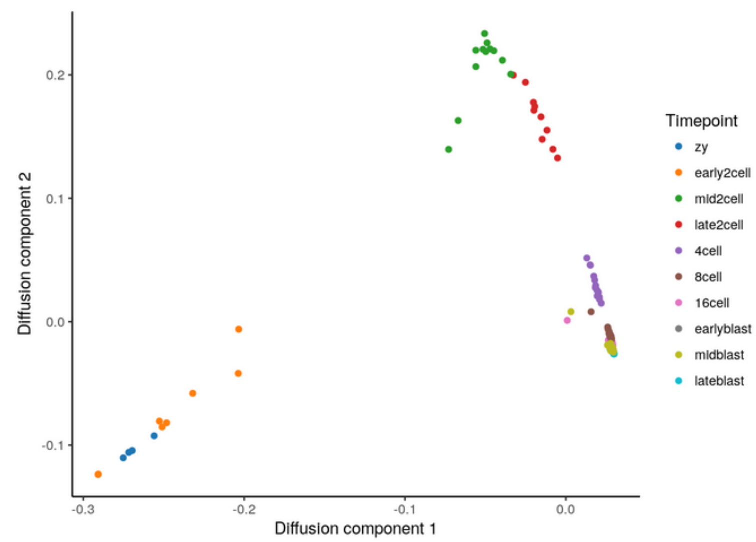
- Classify cells and discover new cell populations
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Reconstructing 'trajectories' Pseudotime analysis

Applicable when studying a process where cells change continuously. For example cell differentiation during development, or cell response to a stimulus.

- **Monocle**
- TSCAN
- Slicer
- **Slingshot**
- Diffusion maps
  - ✓ Scanpy (python)
  - ✓ **destiny** (R)
- PHATE



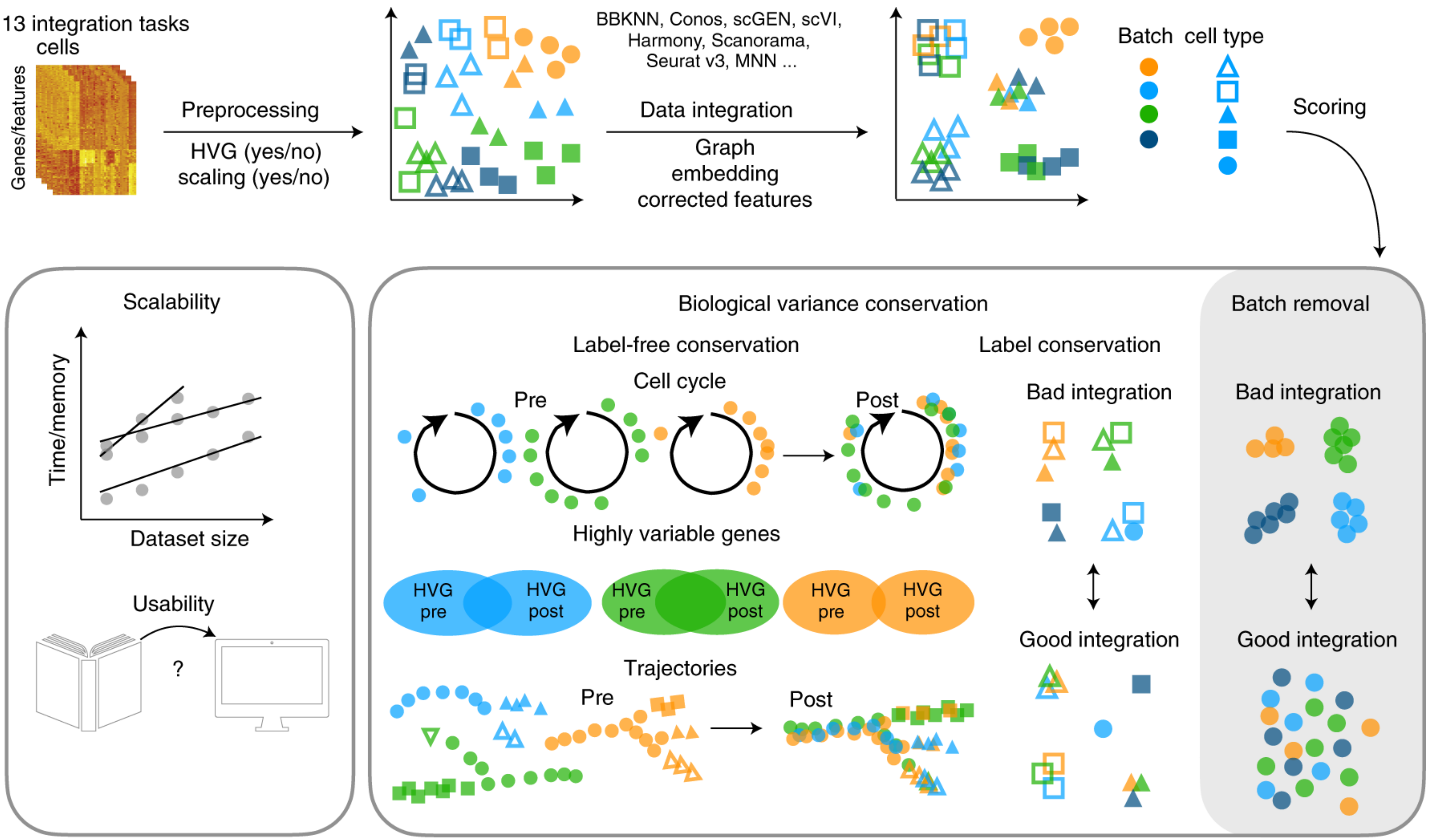


Integrating datasets

Dataset integration: removing batch effects

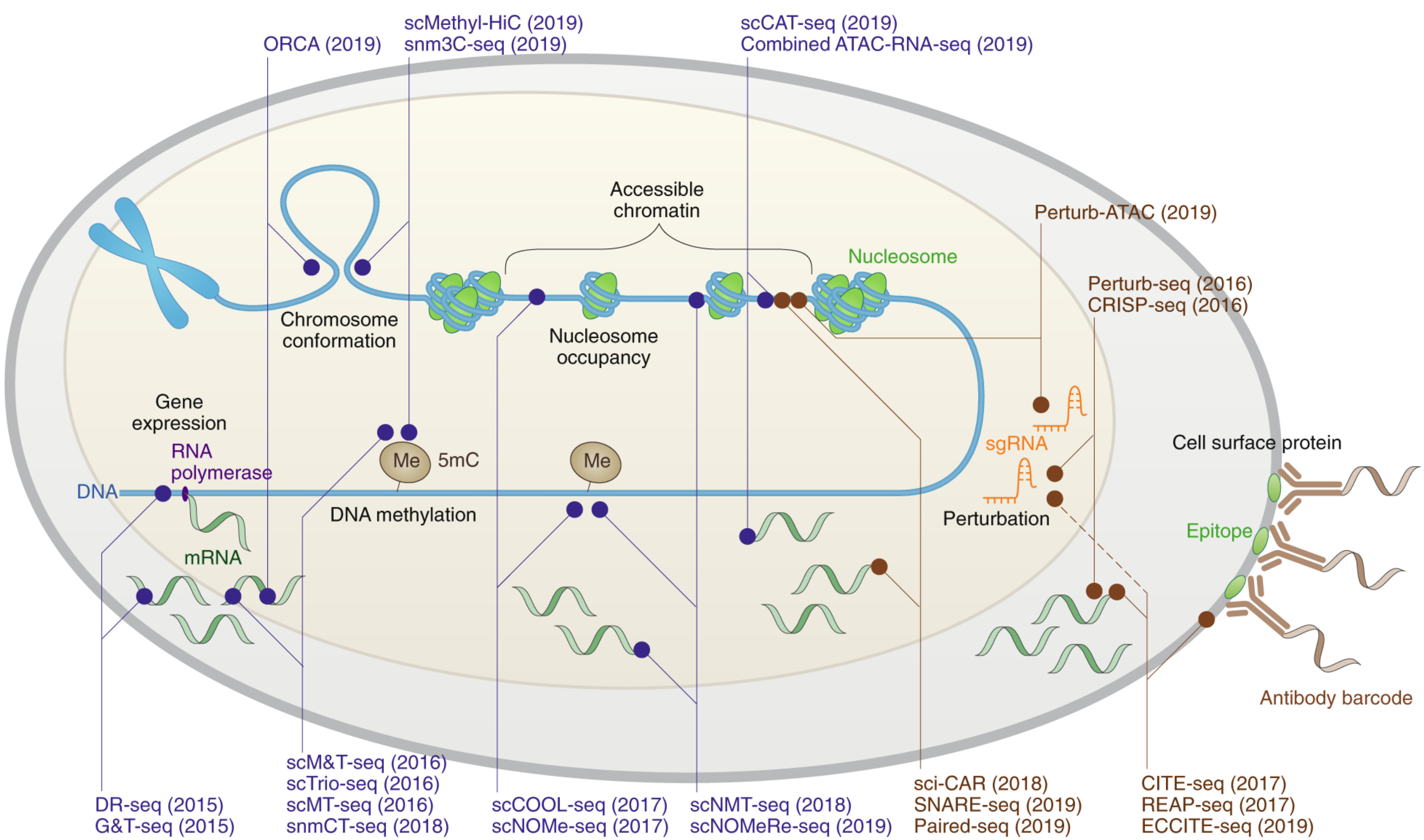
- R packages like **Combat** can be used for this (https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat)
- **CCA** in Seurat. *Cell* **177**, 1888-1902 (2019) Link to SOP
- **Harmony**. *Nature Metshods* **16**, 1289-1296 (2019) Link to SOP
- **LIGER**. *Nature Biotechnology* **37**, 1873–1887 (2019)
- **CSS**: cluster similarity spectrum integration. *Genome Biology* **21**, (2020)
- See "Dealing with confounders" section of the "Analysis of single cell RNA-seq data" course (Hemberg Group).
- Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21,** 12 (2020).
- Deep learning methods: scVAEIT, scVI, totalVI, MultiVI, scVI

Benchmarking of integration methods



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022).

Multimodal analysis



Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat Methods* **17,** 11–14 (2020).

Analysis Demo

- Goal:
  - To walk you through an example analysis of scRNA-seq data.
    - Exploring the data
    - Performing quality control
    - Identifying cell type subsets
  - To introduce you to scRNA-seq analysis using the Seurat package.
- We will be analyzing the a dataset of Non-Small Cell Lung Cancer Cells (NSCLC) freely available from 10X Genomics (https://support.10xgenomics.com/single-cell-vdj/datasets/2.2.0/vdj_v1_hs_nsclc_5gex)

Helpful links

- Single cell day: https://satijalab.org/scgd23/
- Seurat vignettes:

  Version 5: https://satijalab.org/seurat/articles/get_started_v5_new

  Version 4: https://satijalab.org/seurat/vignettes.html
- https://scrnaseq-course.cog.sanger.ac.uk/website/seurat-chapter.html
- Analysis, visualization, and integration of spatial datasets with Seurat
- https://icb-scanpy.readthedocs-hosted.com/en/stable/tutorials.html
- https://github.com/theislab/single-cell-tutorial/blob/master/supplementary_scripts/Splatter-marker-genes-random-data.ipynb
- https://github.com/theislab/single-cell-tutorial/blob/master/latest_notebook/Case-study_Mouse-intestinal-epithelium_1906.ipynb

References and resources

- A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. PMID: 28821273
- Current best practices in single-cell RNA-seq analysis: a tutorial. PMID: 31217225
- "Analysis of single cell RNA-seq data" course (Hemberg Group).
- 2017/2018 Single Cell RNA Sequencing Analysis Workshop (UCD,UCB,UCSF)
- seandavi/awesome-single-cell
- Broad Institute single cell portal https://singlecell.broadinstitute.org/single_cell
- Tabula Muris https://tabula-muris.ds.czbiohub.org/
- UCSC Cell Browser https://cells.ucsc.edu

Upcoming Hot Topics

April 4th

ATAC-seq analysis

April 25th

CUT & Tag analysis