



BROTMAN BATY
INSTITUTE

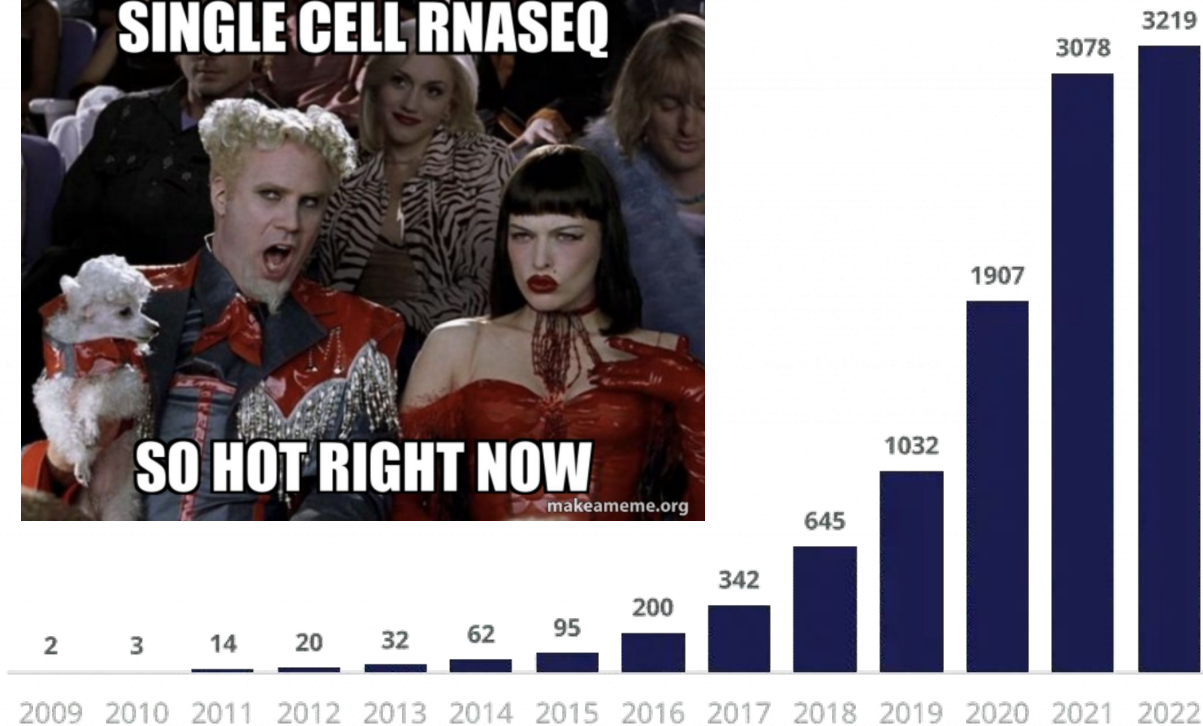
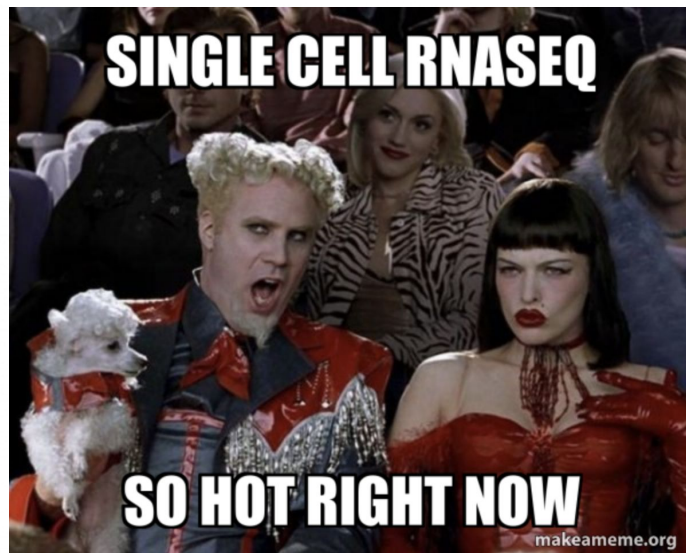
sc-RNA-seq analysis overview

Mary O'Neill, Ph.D.
Director, Single Cell Genomics
ONeillMB@uw.edu
@ONeillMB1

Anh Vo, M.S.
Bioinformatician
athuyvo@uw.edu



Single cell RNA-seq has become ubiquitous

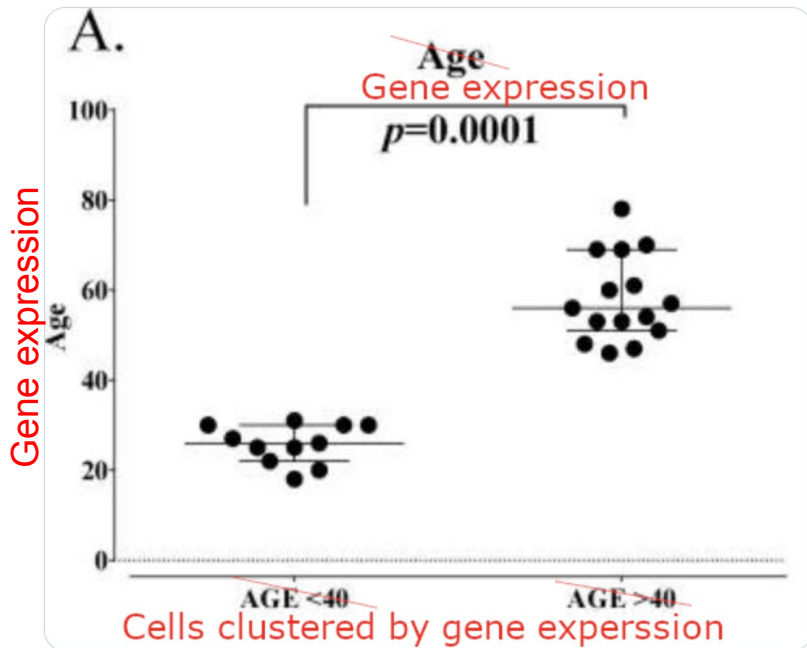


And the data is immense, often overwhelming



Alexis Vandenbon
@alexisvdb

Meanwhile, the single cell field

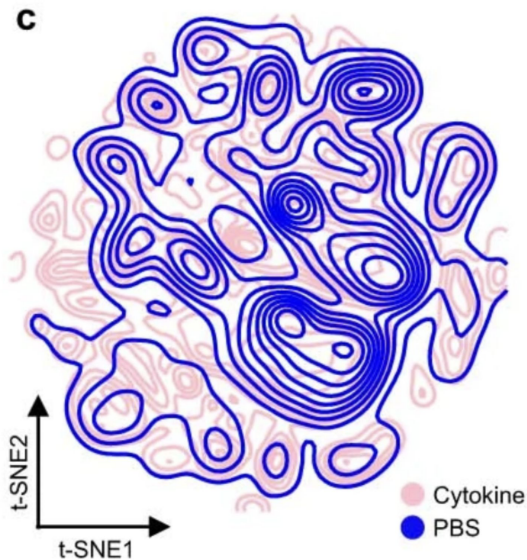


5:50 PM · Jun 20, 2022

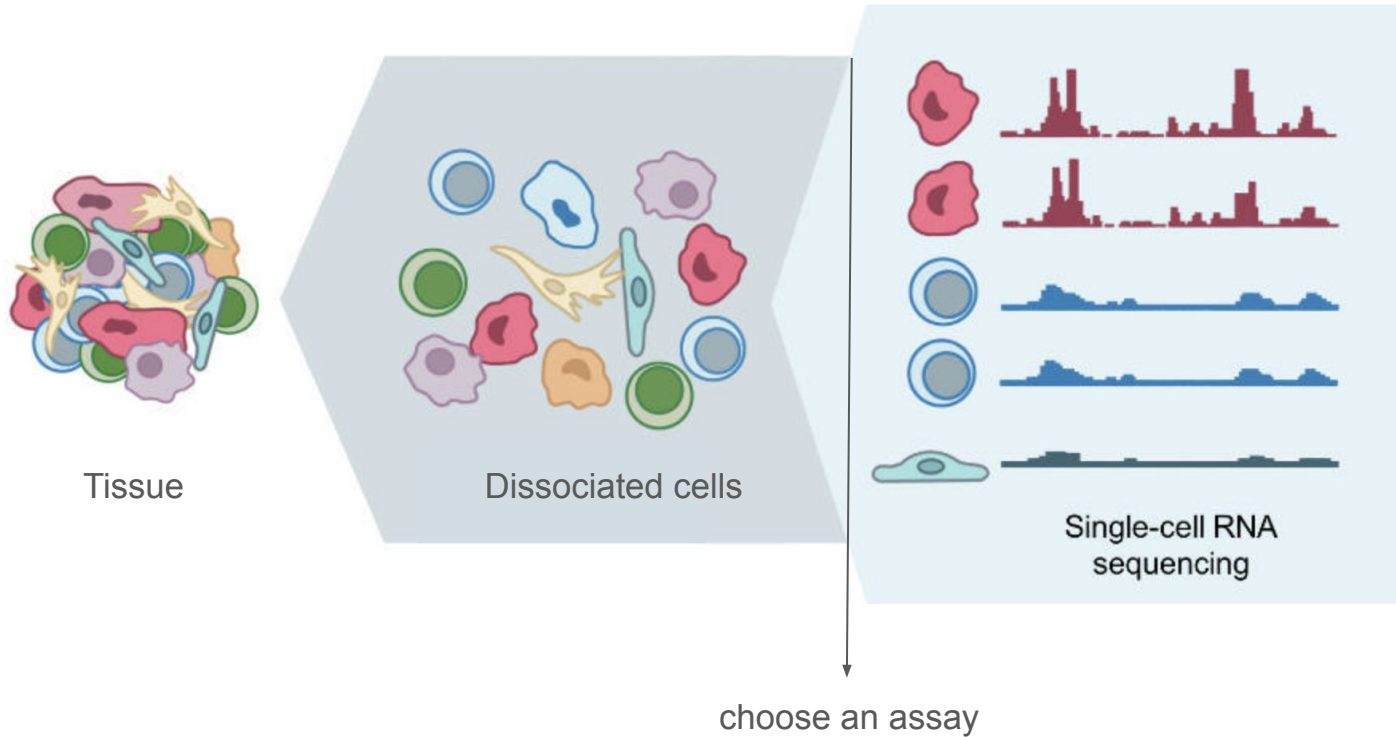


Lior Pachter
@lpachter

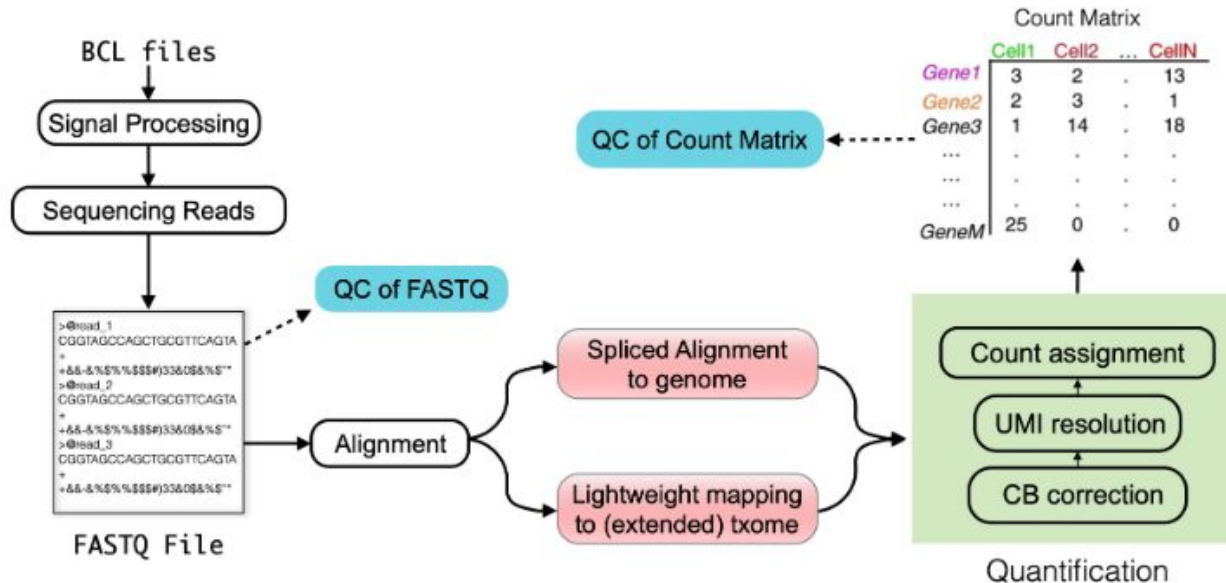
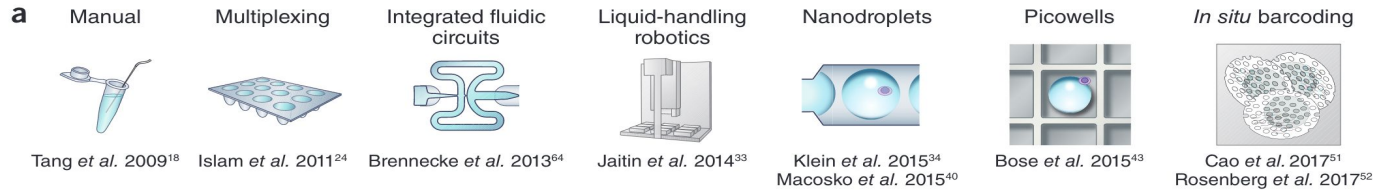
A reminder that contour plots on a t-SNE or UMAP just don't make any sense at all. Distances are distorted (a fact that even the authors of the methods concede), so there is no meaning to density of points. The contours are therefore misleading. See [journals.plos.org/ploscompbiol/a... 2/](https://journals.plos.org/ploscompbiol/article/doi/10.1371/journal.pcbi.1005481)



sc-RNA-seq Overview



No matter the technology...



Raw Data Processing

10x Genomics Support /

Cell Ranger



bbi-lab

[Overview](#) [Repositories 36](#) [Projects](#) [Packages](#) [Teams 6](#) [People 12](#)

Popular repositories

[bbi-sci](#)

Public

Nextflow ☆ 5 🍴 1

[bbi-dmux](#)

Public

Python ☆ 3 🍴 3

[STAR](#) / [docs](#) / [STARsolo.md](#) 



[alexdobin](#) Fixed issues with documentation.

fb84afe · 2 years ago [History](#)

Preview

Code

Blame

458 lines (387 loc) · 30.4 KB

Raw



**STARsolo: mapping, demultiplexing and quantification
for single cell RNA-seq**



Inputs for Cell Ranger (10X)



Inputs for count

- **FASTQ files:** *Required.* The FASTQ format is a text-based file format that contains nucleotide sequence information along with quality scores for each sequenced nucleotide. Typically, FASTQ files are provided by your sequencing core. If not available, you must convert BCL files to FASTQ using Illumina's [BCL Convert](#) or [bcl2fastq](#).
- **Reference transcriptome:** *Required.* A reference transcriptome is a collection of all known transcript sequences from a given organism. 10x Genomics provides [downloadable pre-built references transcriptomes](#) for human and mouse for your Cell Ranger run. The [cellranger mkref](#) pipeline also enables custom reference creation for non-human, non-mouse species with a reference genome sequence (FASTA file) and gene annotations (GTF file). Reference transcriptome is optional for standalone Antibody Capture libraries (without Gene Expression).



Outputs for Cell Ranger (10X)



For primary analysis, the `cellranger count`, `cellranger vdj`, and `cellranger multi` pipelines will output the following types of files:

- web summary (HTML) (count, multi, vdj)
- metrics summary CSV
- BAM
- raw and filtered feature-barcode matrices (MEX, H5)
- secondary analysis files (CSV)
- molecule info (H5)
- Loupe files (cloupe and vloupe)



Common Sources for References



1. Ensembl

- **Website:** [Ensembl Genome Browser](#)
- Provides reference genomes and annotations for many species.
- Look for species-specific FASTA files and corresponding GTF annotations.

2. GENCODE

- **Website:** [GENCODE Project](#)
- High-quality annotations for human and mouse genomes.
- Often used for scRNA-seq studies due to its comprehensive transcript models.

3. UCSC Genome Browser

- **Website:** UCSC Downloads
- Offers FASTA and GTF files for various assemblies (e.g., GRCh38, mm10).

4. NCBI

- **Website:** [NCBI RefSeq](#)
- Contains genome sequences and RefSeq annotations, which are reliable and widely used.

*10X Genomics
has pre-built
references!*



An alternative...

STARsolo

STARsolo is a turnkey solution for analyzing droplet single cell RNA sequencing data (e.g. 10X Genomics Chromium System) built directly into STAR code. STARsolo inputs the raw FASTQ reads files, and performs the following operations

- error correction and demultiplexing of cell barcodes using user-input whitelist
- mapping the reads to the reference genome using the standard STAR spliced read alignment algorithm
- error correction and collapsing (deduplication) of Unique Molecular Identifiers (UMI)
- quantification of per-cell gene expression by counting the number of reads per gene
- quantification of other transcriptomic features: splice junctions; pre-mRNA; spliced/unspliced reads similar to Velocyto

STARsolo output is designed to be a drop-in replacement for 10X CellRanger gene quantification output. It follows CellRanger logic for cell barcode whitelisting and UMI deduplication, and produces nearly identical gene counts in the same format. At the same time STARsolo is ~10 times faster than the CellRanger.



Downstream data analysis (our focus)

- Cell Detection
- Quality Control
- Doublet Detection
- *Ambient RNA*
- *Batch Correction*
- Normalization
- Feature Selection
- Dimensionality Reduction
- Clustering
- Annotation
- Differential Gene Expression
- *Gene Set / Pathway Enrichment*
- *Trajectory Analysis*
- *Gene Regulatory Network Analysis*
- *Deconvolution*
- (. . .)

Count Matrix

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...
...
...
GeneM	25	0	.	0

Cell Detection

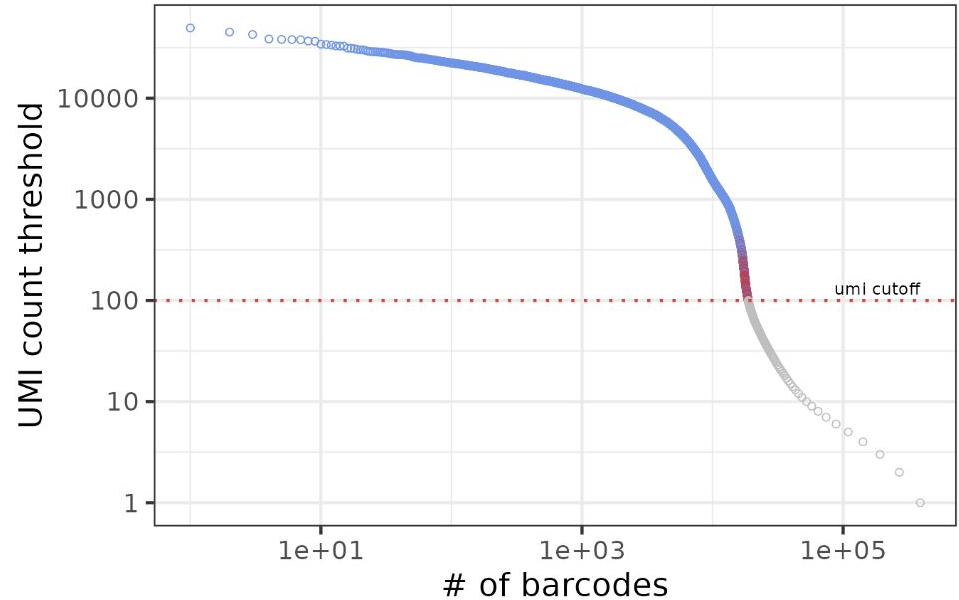
Purpose:

Identifying true cells versus empty droplets or debris in single-cell RNA sequencing data.

Common approaches:

- **Knee Plot:** Identify a "knee" where UMI counts sharply drop.
- **Inflection Point:** Separates cells from ambient RNA.
- **EmptyDrops:** Statistical method

Most pipelines have cell detection incorporated into them and will output both 'filtered' and 'unfiltered' cell-by-gene matrices.



emptyDrops_FDR_0.01 FALSE NA TRUE



Quality Control



Motivation for QC:

- Low-quality libraries arise from technical issues (e.g., cell damage, inefficient library prep).
- They can bias results by:
 - Forming distinct, misleading clusters.
 - Inflating variance estimates in dimensionality reduction.
 - Artificially upregulating ambient or contaminant genes.

Key QC Metrics:

1. **Library size:** Total RNA counts per cell.
 - Low counts indicate RNA loss or prep inefficiencies.
2. **Expressed genes:** Number of genes with non-zero expression.
 - Low values suggest incomplete transcript capture.
3. **Mitochondrial percentage:** Proportion of reads from mitochondrial genes.
 - High values point to cytoplasmic RNA loss during cell damage.
4. **Spike-in percentage:** Proportion of reads mapping to spike-ins.
 - Elevated levels indicate endogenous RNA loss.



Identifying Low-Quality Cells



Fixed Thresholds:

- Predefined cutoffs (e.g., $<5,000$ genes, $>10\%$ mitochondrial reads) exclude low-quality cells.
- Limitations: Thresholds vary by protocol and dataset.

Adaptive Thresholds:

- Use median absolute deviation (MAD) for dynamic thresholding.
- Advantages: Adjusts to dataset-specific variations.

Diagnostic Plots:

- Inspect QC metric distributions to avoid misclassification of biologically relevant cells.
- Example: Scatter plots of mitochondrial % vs. total counts to ensure high-quality cells aren't removed.



Normalization in scRNA-seq



Motivation:

- Variability in sequencing coverage arises from technical biases like differences in cDNA capture or PCR efficiency.
- Normalization removes these biases to enable meaningful biological comparisons.

Scaling Normalization:

- Divides counts by a **size factor** (e.g., library size or deconvolution-based).
- Assumes biases scale equally across all genes within a cell.



Normalization methods



Library Size Normalization:

- Uses total counts per cell to compute size factors.
- Assumes balanced differential expression (DE) between cells, but composition biases can introduce inaccuracies.
- Effective for clustering but less so for DE analyses.

Deconvolution Normalization (scrn):

- Accounts for composition biases by pooling cells for size factor estimation.
- Adjusts for cell type-specific deviations.
- Useful for accurate downstream analyses like DE testing.

Spike-in Normalization:

- Scales based on added spike-in RNAs, preserving total RNA content differences.
- Ideal when RNA content changes are biologically relevant (e.g., T cell activation).



Feature Selection in scRNA-seq



Motivation:

- Identify genes that contribute meaningful biological variation.
- Eliminate genes dominated by noise or "uninteresting" variation.
- Improve computational efficiency for clustering, dimensionality reduction, and other downstream steps.

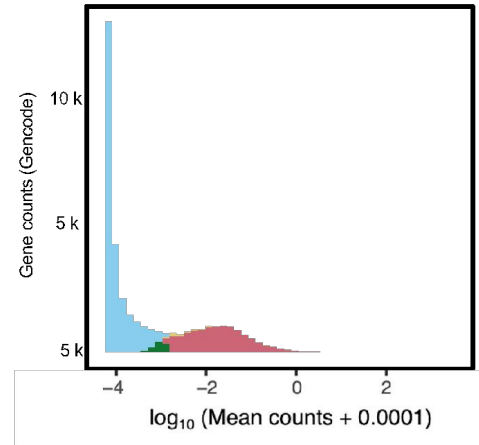
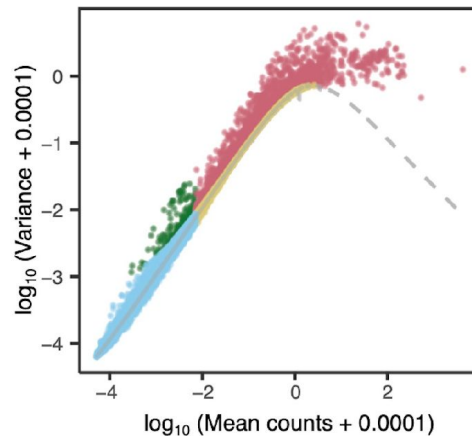
Key Concept:

- Use **highly variable genes (HVGs)**, identified as genes with a biological variance component larger than expected given their abundance.
- $\text{Biological variance} = \text{Total variance} - \text{Technical noise}.$

Methods for Quantifying Gene Variance

Trend Fitting:

- Model the **mean-variance relationship** using functions like `modelGeneVar()`.
- Account for **non-linear trends** between gene expression levels and variance.



Approaches to Address Noise:

- **No spike-ins:** Model variance using assumptions (e.g., Poisson noise for UMI data).
- **With spike-ins:** Fit variance trend to spike-ins (e.g., ERCC), providing a robust estimate of technical noise.



Dimensionality Reduction



Purpose:

- Compress high-dimensional gene expression data into fewer dimensions.
- Reduce noise, computational complexity, and enable visualization.

Why It Works:

- Genes are correlated through shared biological processes, allowing data compression.
- Dominant patterns in data captured by principal components or other derived dimensions.

Benefits:

- Simplifies downstream analyses (e.g., clustering).
- Improves interpretability and visualization (e.g., PCA plots).



Methods for Dimensionality Reduction



Principal Components Analysis (PCA):

- Captures variation along orthogonal axes (principal components).
- First PCs typically represent biological signals; later PCs capture noise.
- Widely used for denoising, clustering, and input for advanced methods.

Non-Linear Methods:

- **t-SNE:**
 - Preserves local structure for complex population visualization.
 - Computationally intensive, sensitive to parameters (e.g., perplexity).
- **UMAP:**
 - Faster than t-SNE, retains more global structure.
 - Increasingly favored for large datasets.



Clustering



What is Clustering in scRNA-seq?

- **Clustering** groups cells based on similarity in gene expression profiles, identifying distinct populations or cell types.
- The goal is to identify **biologically meaningful groups** in high-dimensional scRNA-seq data (e.g., cell types, states, or developmental stages).

Types of Clustering

- **Unsupervised Clustering:** No prior labels; groups cells based on intrinsic similarities.
- **Supervised Clustering:** Uses known labels (e.g., cell types) to guide the clustering process.



Common Clustering Approaches



- **Graph-based Clustering:** Treats cells as nodes in a graph, using algorithms like Louvain or Leiden for community detection.
- **K-means:** Partitions cells into K clusters.
- **Hierarchical Clustering:** Builds a dendrogram to group cells based on similarity.
- **Density-based Clustering:** Identifies clusters as regions of high density in the data (e.g., DBSCAN).



Annotation

What is Annotation in scRNA-seq?

- **Annotation** is the process of assigning biological meaning to clusters of cells, usually based on their gene expression profiles.
- The goal is to map clusters to known **cell types**, **states**, or **lineages** based on expression patterns and marker genes.

Why is Annotation Important?

- Provides biological context for identified clusters.
- Helps **interpret cellular diversity** in tissues or developmental stages.
- Facilitates **comparative analysis** across conditions or diseases.



Annotation

What is Annotation in scRNA-seq?

- **Annotation** is the process of assigning biological meaning to clusters of cells, usually based on their gene expression profiles.
- The goal is to map clusters to known **cell types**, **states**, or **lineages** based on expression patterns and marker genes.

Why is Annotation Important?

- Provides biological context for identified clusters.
- Helps **interpret cellular diversity** in tissues or developmental stages.
- Facilitates **comparative analysis** across conditions or diseases.



Differential Gene Expression



What is Differential Gene Expression?

- **DGE** refers to the identification of genes whose expression levels significantly differ between two or more conditions, clusters, or cell types.
- In **scRNA-seq**, DGE helps identify:
 - Genes that distinguish different cell types or subpopulations.
 - Genes that respond to treatment or disease conditions.

Why is DGE Important in scRNA-seq?

- Unveils **biological insights** into cell-type-specific responses.
- Helps in understanding **disease mechanisms**, **developmental processes**, and **treatment effects**.
- Essential for **identifying biomarkers** or therapeutic targets.



Approaches to DGE



1. Common Approaches for DGE

- **DeSeq2, EdgeR:** Popular tools for RNA-seq, adapted for single-cell data.
- **Mast:** A method designed specifically for single-cell RNA-seq that accounts for the zero-inflated nature of scRNA-seq data.
- **Wilcoxon Rank-Sum Test:** Non-parametric test to compare gene expression across conditions.

2. Challenges in scRNA-seq DGE

- **Zero Inflation:** Many genes have zero counts in many cells.
 - Approaches like **Mast** and **DESeq2** handle this well.
- **Dropout Events:** Technical noise where genes are expressed but not detected in all cells.
- **Cell-to-Cell Variability:** High biological variation that requires large sample sizes for reliable results.

3. Statistical Considerations

- **Multiple Testing Correction:** To control the False Discovery Rate (FDR) due to the large number of genes being tested.
 - Common corrections: **Benjamini-Hochberg**, **Bonferroni**.
- **Fold Change vs. p-value:** Important to balance both metrics for interpretation.
- **Normalization:** Correct for differences in sequencing depth and cell-specific biases (e.g., library size normalization).



Favorite online resources



<https://www.singlecellcourse.org/index.html>

<https://bioconductor.org/books/release/OSCA/>

Orchestrating Single-Cell Analysis with Bioconductor



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

Single-cell best practices



Awesome Single Cell

Community-curated list of software packages and data resources for single-cell, including RNA-seq, ATAC-seq, etc.

<https://github.com/seandavi/awesome-single-cell?tab=readme-ov-file>




eg. Doublet detection methods in Awesome Single Cell



Doublet Identification

- [AMULET](#) - [shell, Python, R] - A count based method for detecting multiplets from single nucleus ATAC-seq (snATAC-seq) data. [Genome Biology](#)
- [demuxlet](#) - [shell] - [Multiplexed droplet single-cell RNA-sequencing using natural genetic variation](#)
- [DoubletFinder](#) - [R] - Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. [BioRxiv](#)
- [DoubletDecon](#) - [R] - Cell-State Aware Removal of Single-Cell RNA-Seq Doublets. [BioRxiv](DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq Doublets)
- [DoubletDetection](#) - [R, Python] - A Python3 package to detect doublets (technical errors) in single-cell RNA-seq count matrices. An [R implementation](#) is in development.
- [Scrublet](#) - [Python] - Computational identification of cell doublets in single-cell transcriptomic data. [BioRxiv](#)
- [solo](#) - [Python] - Doublet detection via semi-supervised deep learning.



Demuxafy

DOUBLET DETECTING METHODS

Overview of Doublet Detecting Softwares

- DoubletDecon
- DoubletDetection
- DoubletFinder
- ScDbfFinder
- Scds
- Scrublet
- Solo



General advice for analysis



- Someone just spend a bunch of time, money, and effort generating this data - don't skimp or rush the analysis
- Set realistic expectations for everyone involved. The analysis will take time. A lot of time. (Sometimes years.)
- The analysis will require some programming skills. Point and click software is severely limiting.
- Every dataset is unique. Default thresholds and parameters are often not appropriate. There is no one-size-fits-all approach.
- Explore your data. Become the expert of it. Ask yourself if results make sense. Just because a program gave you an answer does not make it true.
- Flag it rather than delete it (e.g. low quality cells, doublets, etc.)
- Try multiple methods and look for concordance
- Don't try to interpret non-linear dimensionality reductions of your data (e.g. tSNE, UMAPs)
- Don't limit yourself to one software program!

Don't limit yourself to one software ...

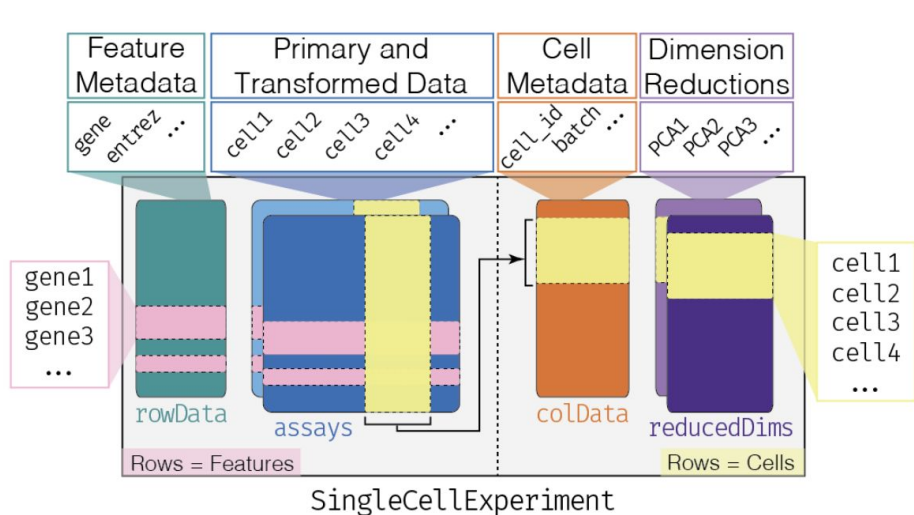


Figure 4.1: Overview of the structure of the `SingleCellExperiment` class. |

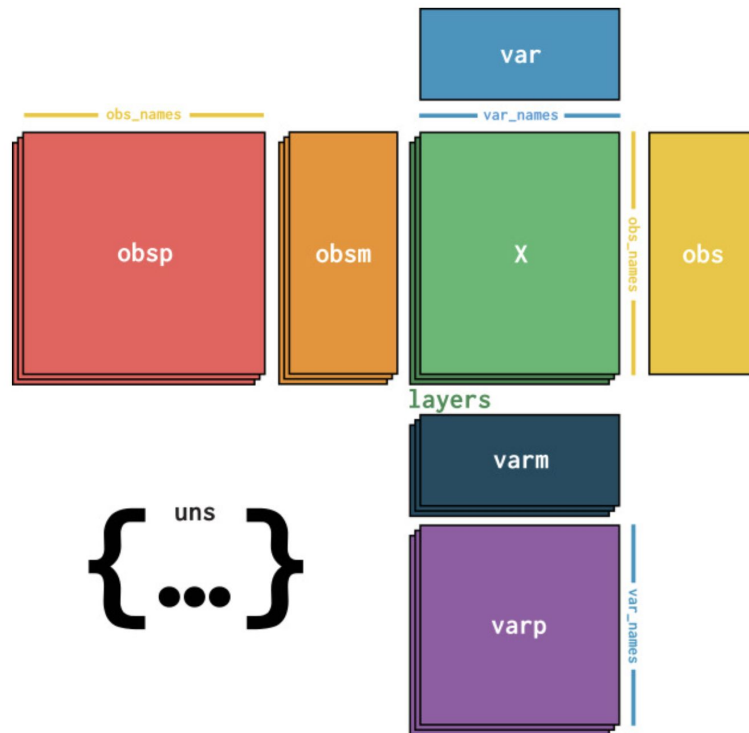


Fig. 4.1 AnnData overview. Image obtained from [Virshup et al., 2021].

Tutorial

Background on the datasets



10X Genomics, PBMCs



Cell Ranger · [pbmc_1k_v3](#) · Peripheral blood mononuclear cells (PBMCs) from a healthy donor

[SUMMARY](#) [ANALYSIS](#)

Estimated Number of Cells

1,222

Mean Reads per Cell

54,502

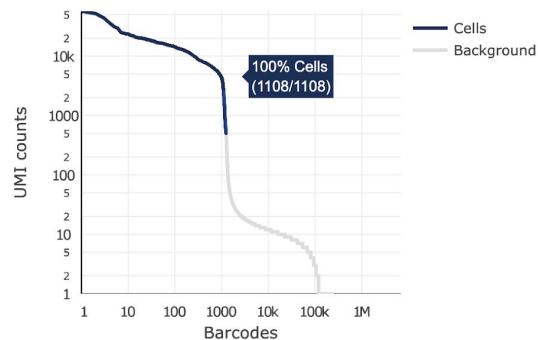
Median Genes per Cell

1,919

Sequencing

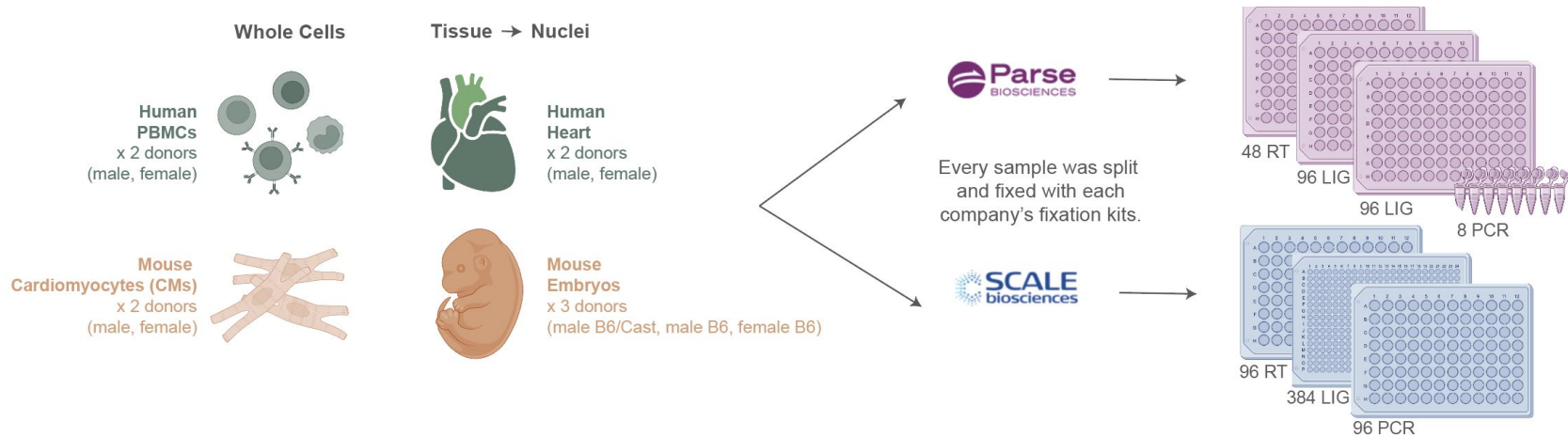
Number of Reads	66,601,887
Valid Barcodes	97.4%
Sequencing Saturation	70.8%
Q30 Bases in Barcode	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in Sample Index	91.1%
Q30 Bases in UMI	92.7%

Cells



Estimated Number of Cells	1,222
Fraction Reads in Cells	94.9%
Mean Reads per Cell	54,502
Median Genes per Cell	1,919
Total Genes Detected	18,391
Median UMI Counts per Cell	6,628

Combinatorial-indexing benchmarking (generated *in-house*)



Tutorial: summary statistics for a random sample of 50K barcodes