Introduction to Single-cell RNA-seq

Colton McNinch, PhD.

February 14, 2024

Bioinformatics and Computational Biosciences Branch (BCBB)

National Institute of Allergy and Infectious Diseases (NIAID)

National Institute of Allergy and Infectious Diseases

BCBB

Bioinformatics and Computational Biosciences Branch

Centralized resource to provide:

- expert training
- consultation
- collaboration

Domain expertise provided in numerous areas:

- Clinical Genomics
- Metagenomics
- Microbial Genomics
- Data Science, Biostatistics, and Informatics
- Structural Biology
- 3D Printing and Biovisualization
- Imaging
- Software development

More information:

https://www.niaid.nih.gov/research/bioinformatics-computational-biosciences-branch

Feel free to email us!

bioinformatics@niaid.nih.gov

Outline

scRNA-Seq Overview

- Differences from bulk RNA-Seq
- Evolution of the technology used
- Major steps of a scRNA-Seq project
- Comparison of current protocols
- Cell isolation strategies
- Transcript quantification strategies

scRNA-Seq experimental design

- Choosing the appropriate protocol
- Avoiding batch effects

Data analysis

- Raw data processing
- Major steps following raw data processing
    - Quality control
    - Normalization
    - Variable feature selection
    - Dimensionality reduction
    - Cell clustering
    - Further downstream analyses

Helpful Resources

scRNA-Seq Overview

Differences from bulk RNA-Seq

Bulk RNA-Seq

Measure **average gene expression** across a population of cells

More sensitive
lowly expressed genes often detected

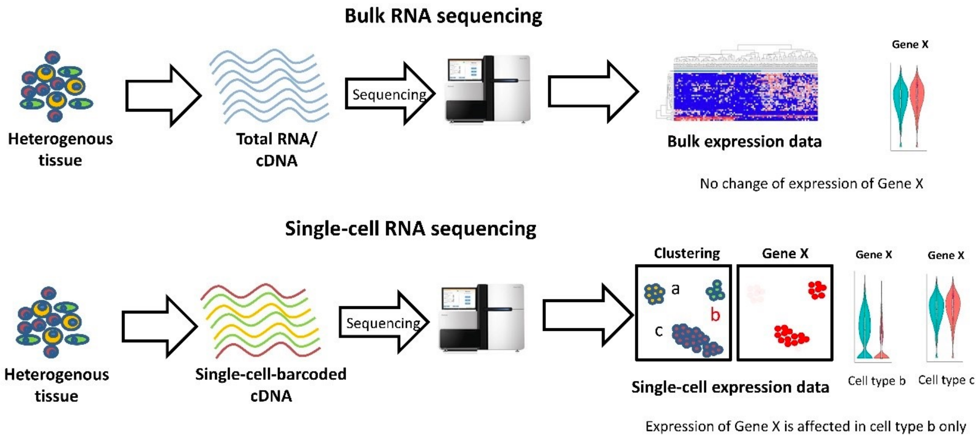
Examine all RNA types

scRNA-Seq

Measure **cell-specific expression** in each cell type

Less sensitive
lowly expressed genes often missed

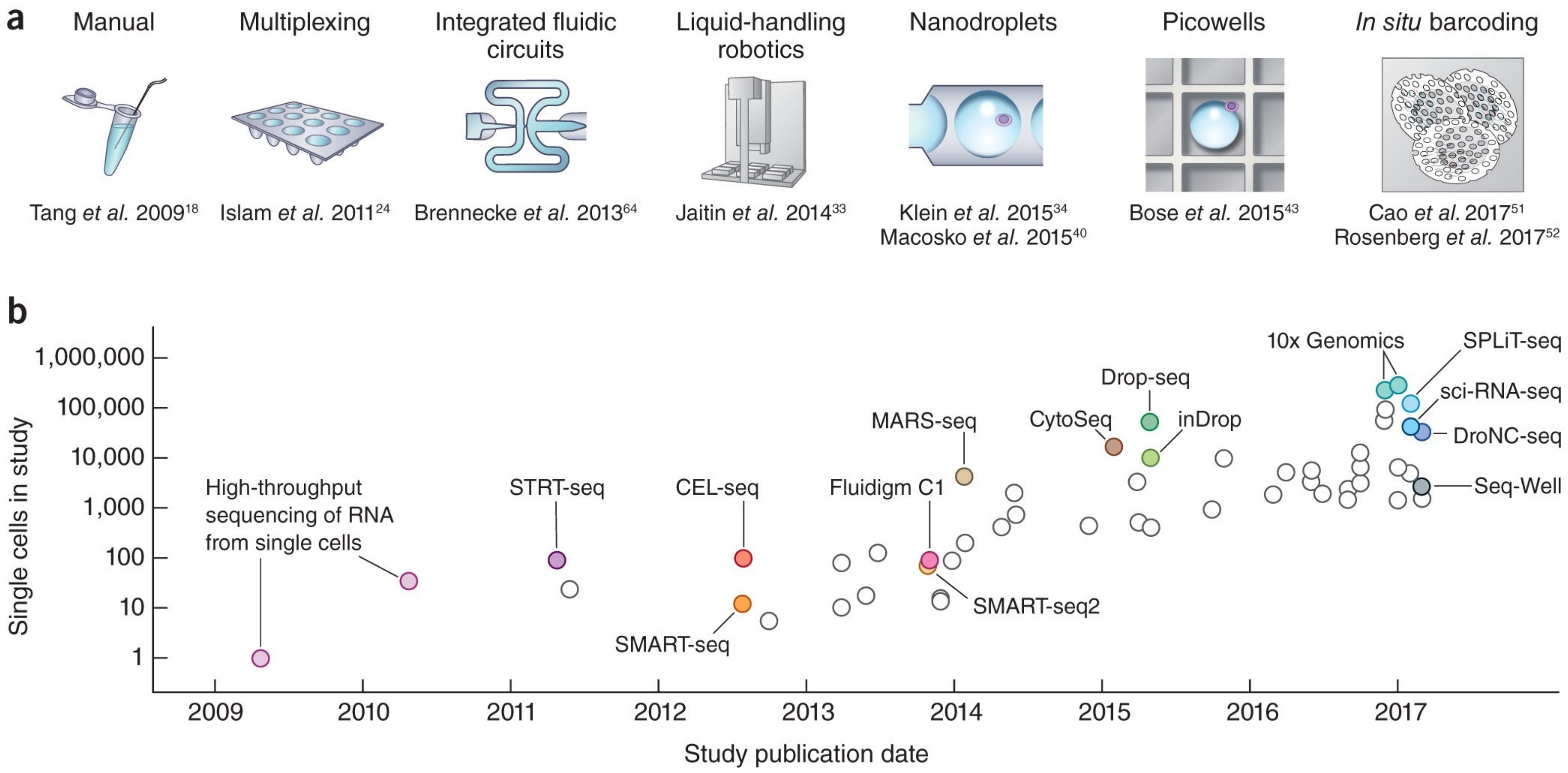
Examine only poly(A) mRNAs



https://cdn.technologynetworks.com/tn/images/body/11570107721805.jpg

scRNA-Seq Overview

Evolution of the technology used



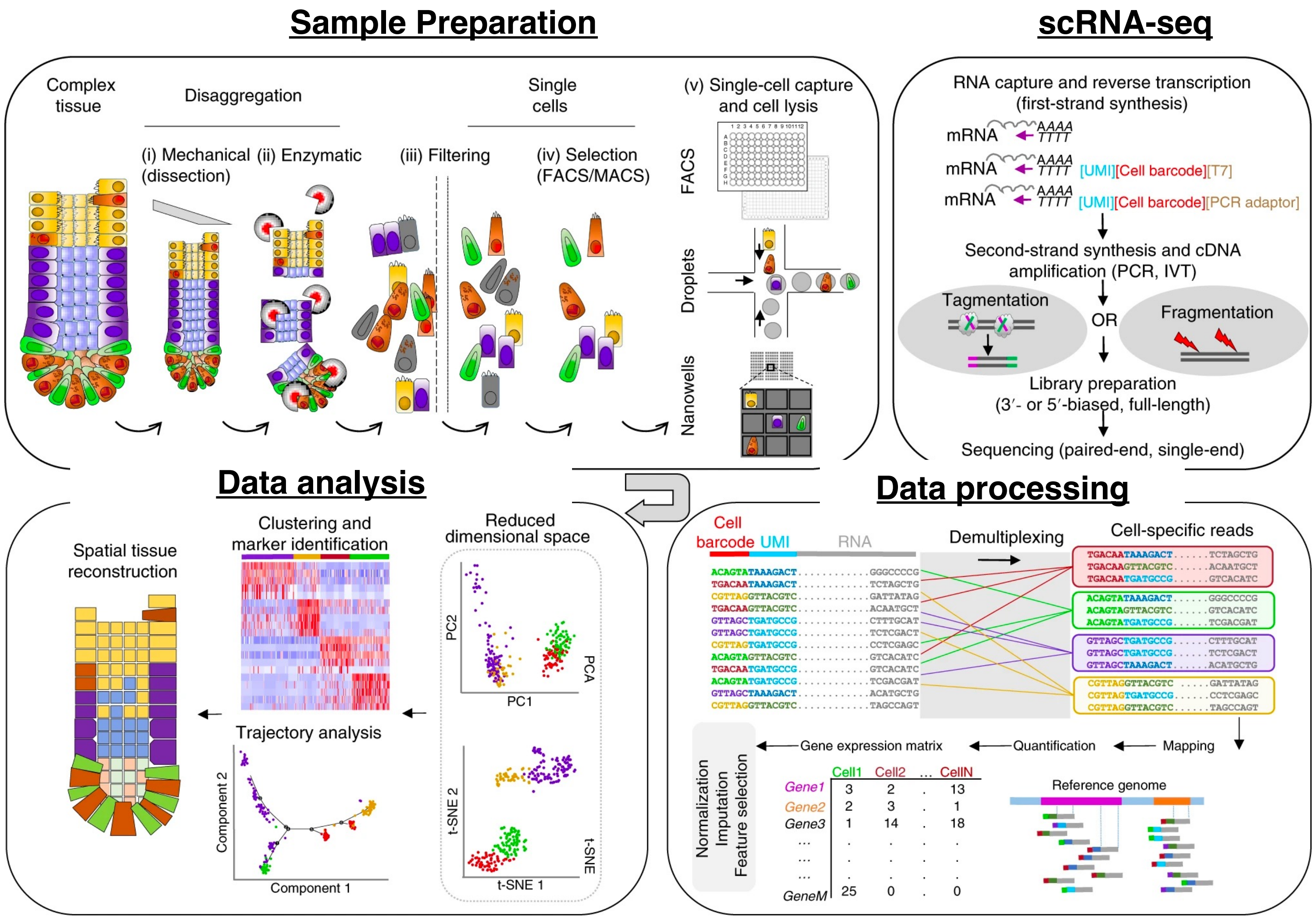
https://doi.org/10.1038/nprot.2017.149

scRNA-Seq Overview

Major steps of a scRNA-Seq project

National Institute of Allergy and Infectious Diseases

1. **Sample Preparation**
   - Isolate cells from complex tissue
   - Lyse cells
   - Add sequencing reagents
2. **scRNA-seq**
   - Prep RNA for sequencing
   - Sequence RNA libraries
3. **Data processing**
   - Separate (demultiplex) reads by cell barcode
   - Align reads to reference genome
   - Correct read errors
4. **Data analysis**
   - Quality control
   - Dimensionality reduction (PCA/UMAP/t-SNE)
   - Cell-type clustering
   - Differential expression analysis
   - Many other analyses



https://doi.org/10.1038/s41596-018-0073-y

scRNA-Seq protocols

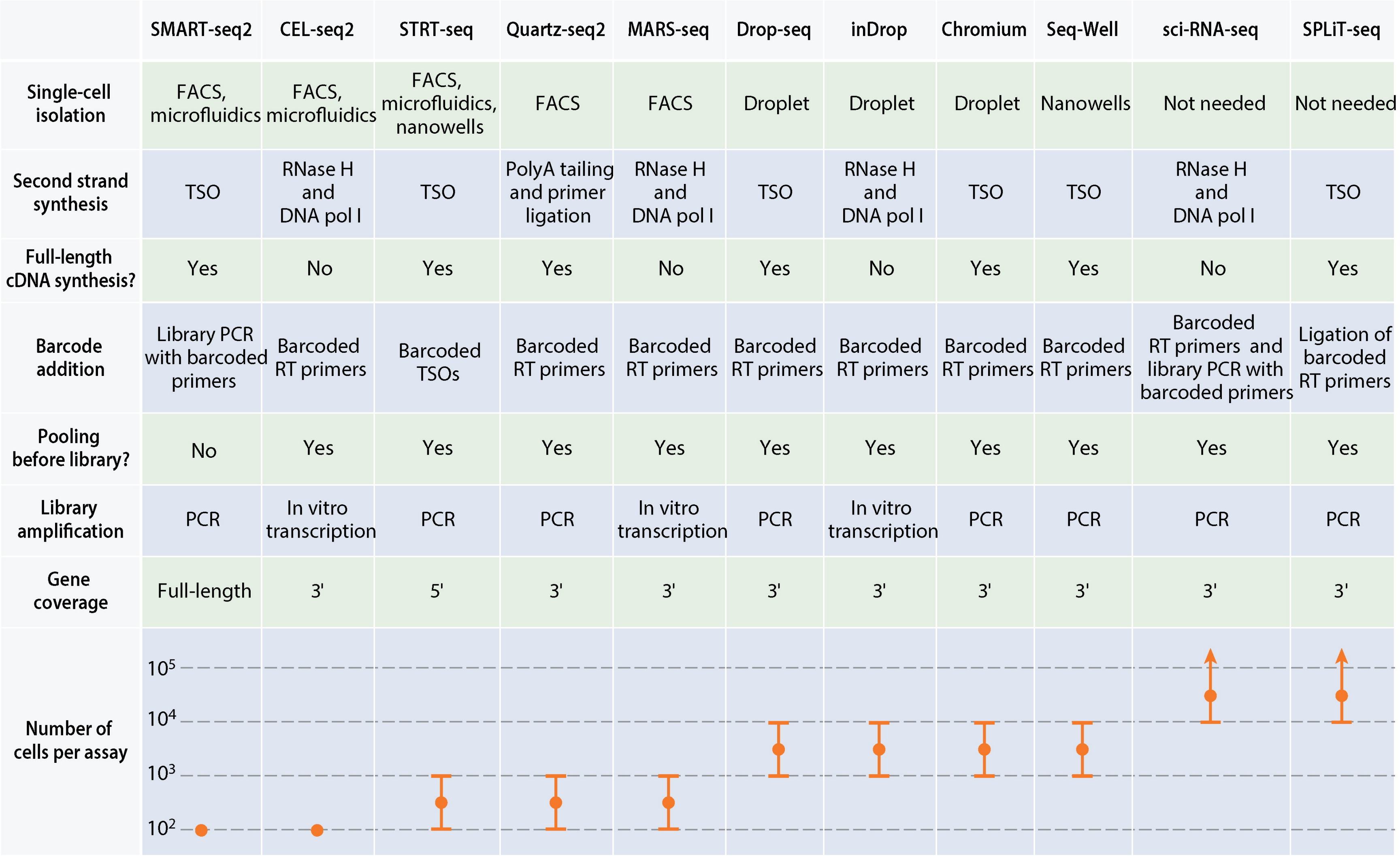
Comparison of current protocols

Numerous scRNA-Seq protocols

- Each have strength & weaknesses

Biggest differences

- How cells are isolated
- How transcripts are quantified



	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay											

Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

https://doi.org/10.1146/annurev-biodatasci-080917-013452

scRNA-Seq protocols

Cell isolation strategies

Microtitre-plate

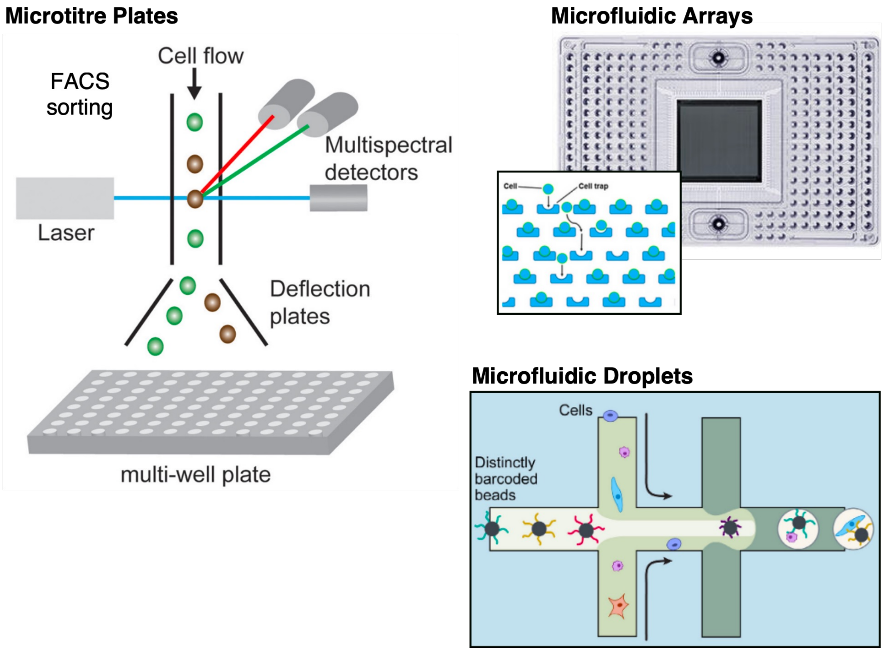
- Isolate cells into individual wells of a plate
  - Fluorescent Activated Cell Sorting (FACS)
- Low throughput
- High sensitivity

Microfluidic-array

- Isolate cells into individual wells of a microfluidic chip
  - Cells travel through microscopic channels and chambers and are sorted by size and other physical properties
- Medium throughput
- Medium sensitivity

Microfluidic-droplet (Droplet)

- Isolate cells into nanoliter-sized oil droplets
  - 10x Genomics
- High throughput
- Low sensitivity



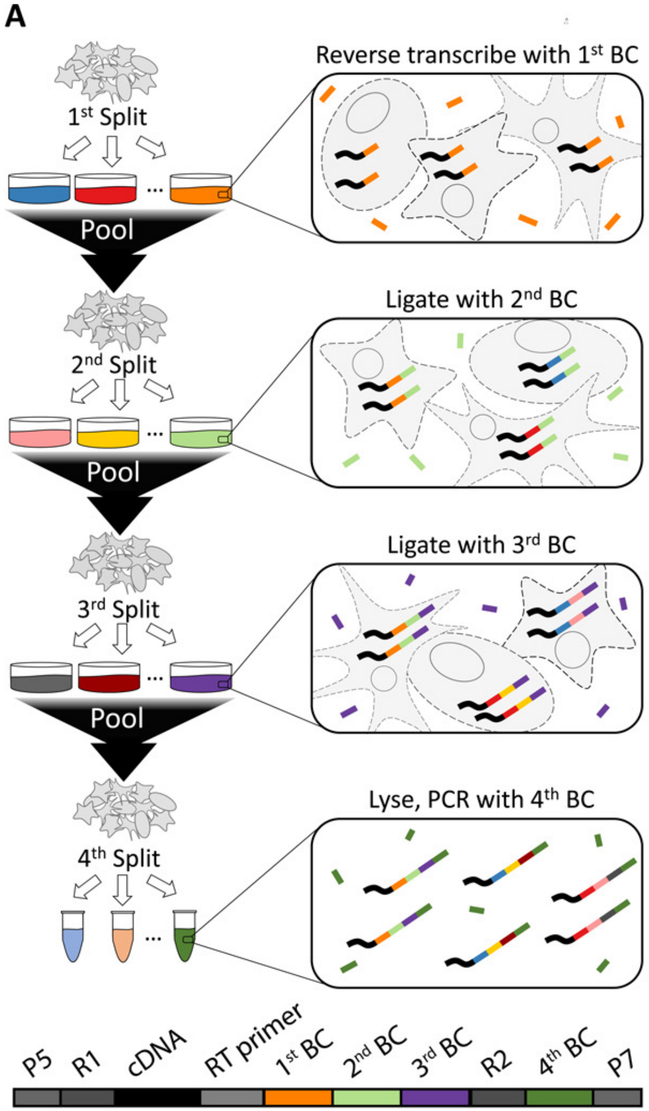
https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html#cell-capture

scRNA-Seq protocols

Cell isolation strategies

<u>Combinatorial barcoding</u>

- No need to isolate cells
  - Parse Biosciences
- High throughput
  - Up to a million cells
- Low sensitivity



https://doi.org/10.1126%2Fscience.aam8999

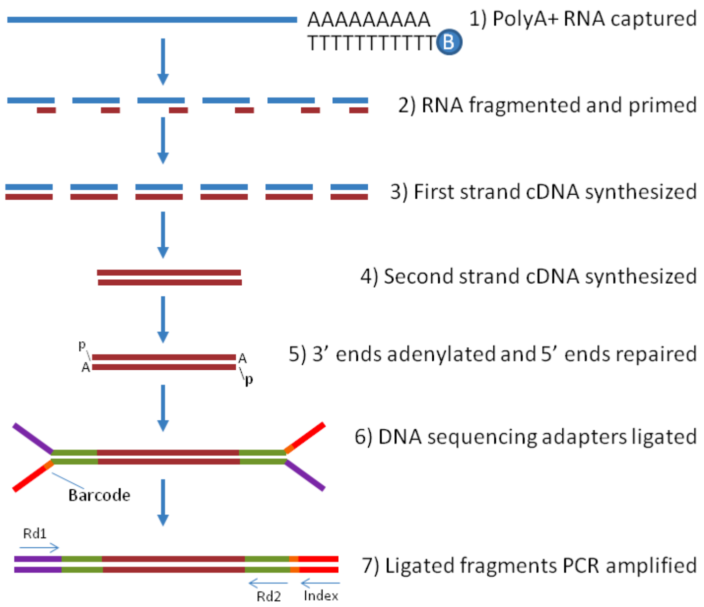
scRNA-Seq protocols

Transcript quantification strategies

Full-length

Uniform read coverage across transcript

Lower throughput



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

5) 3' ends adenylated and 5' ends repaired

6) DNA sequencing adapters ligated

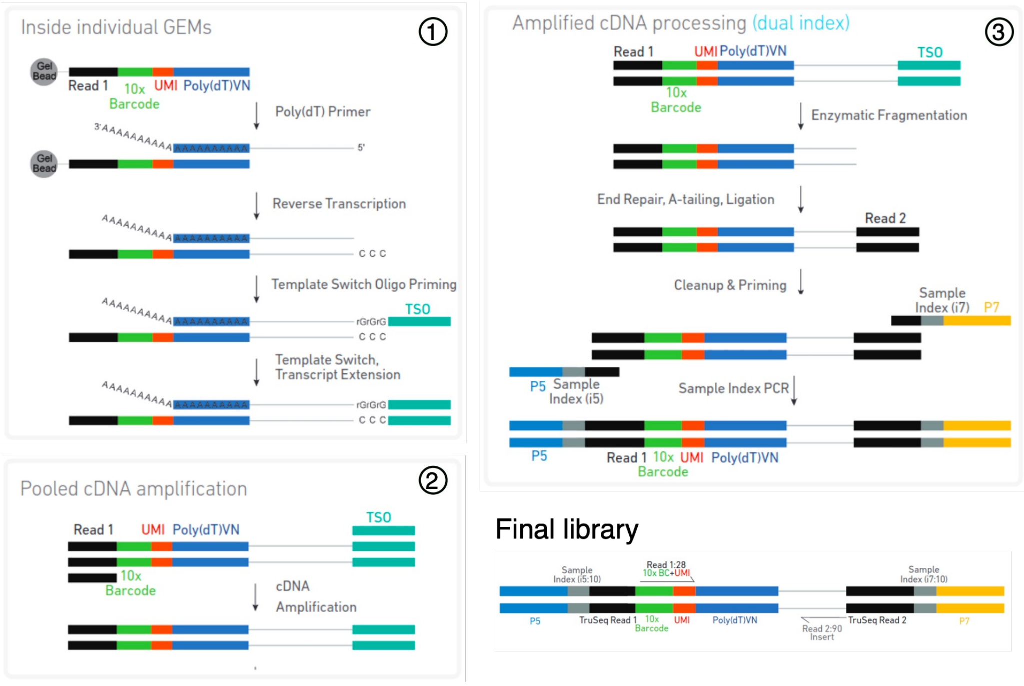
7) Ligated fragments PCR amplified

https://www.labome.com/method/RNA-seq.html

Tag-based

Capture only either the 5' or 3' ends

Higher throughput



https://assets.ctfassets.net/an68im79xiti/1C16trEdzy1Folq5xbOijE/7e6fb1f504e130bd561d898384da99d9/CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex__RevB.pdf

scRNA-Seq experimental design

Choosing the appropriate protocol

What is the research goal?

Characterize isoform expression

- Full-length transcript quantification protocol

Characterize cell expression in heterogeneous tissue

- Droplet-based cell isolation protocol

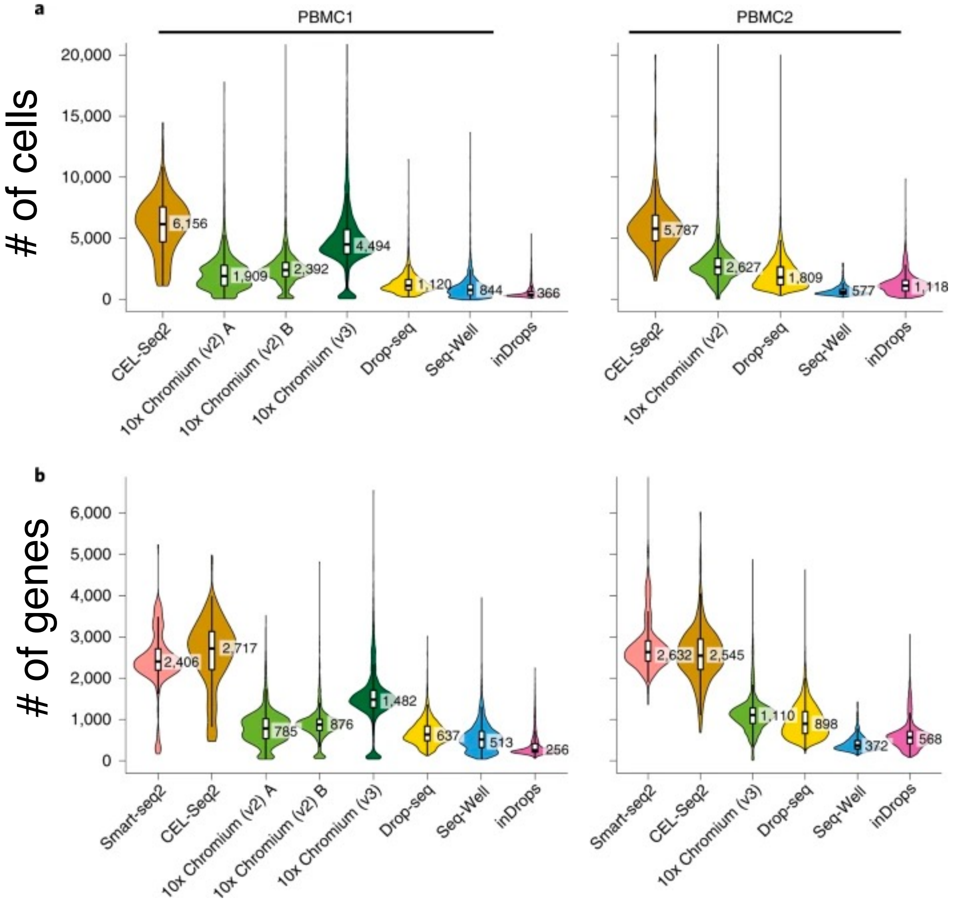
Popular methods

SMART-seq2

- best for small number of cells at great detail
- low-throughput
- full-length transcript quantification

10x Chromium

- best for large number of cells from heterogeneous tissue
- high-throughput
- 3’ or 5’ ends of transcripts

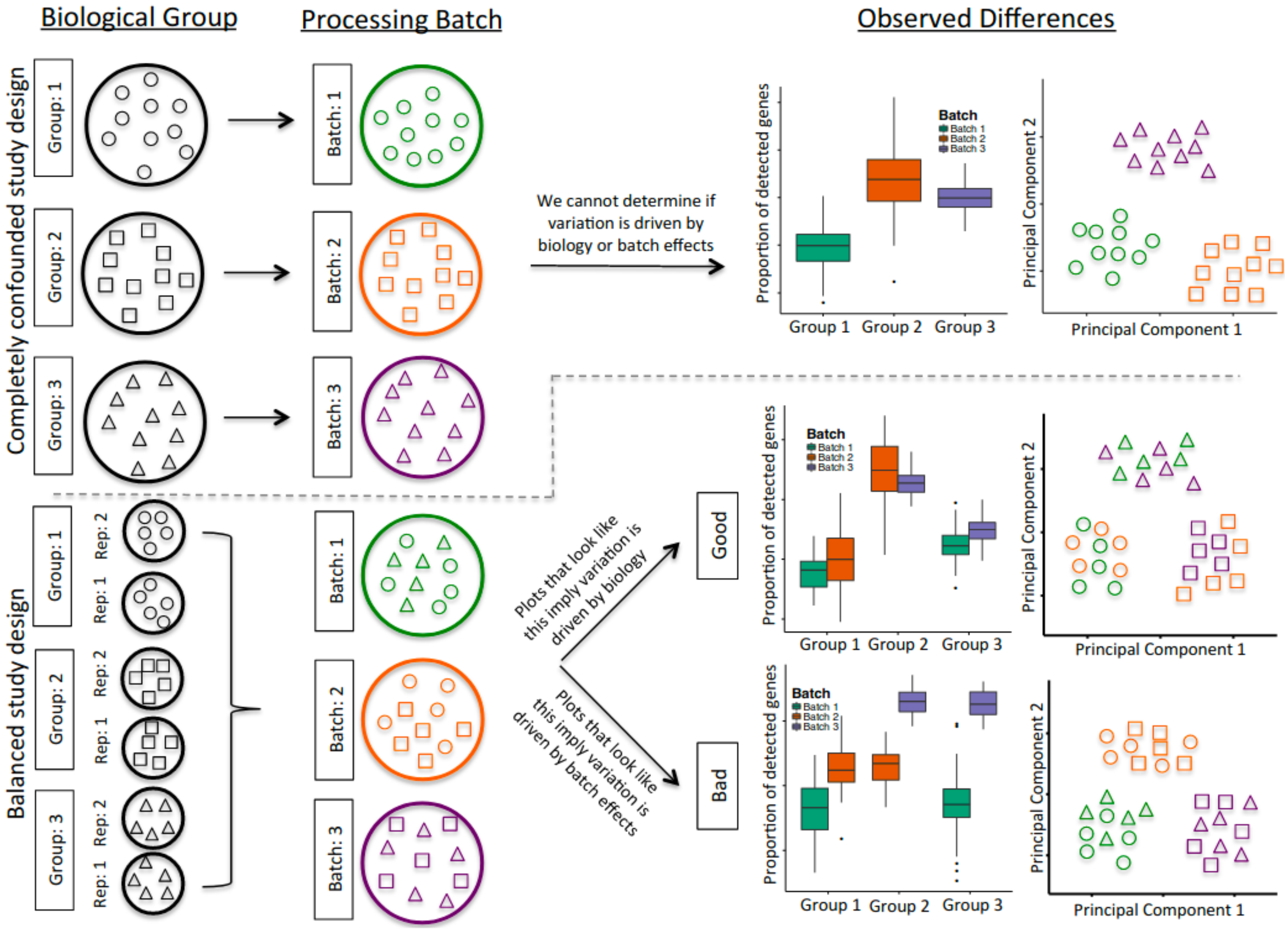


https://doi.org/10.1038/s41587-020-0465-8

scRNA-Seq experimental design

NIH National Institute of Allergy and Infectious Diseases

Avoiding batch effects



We cannot determine if variation is driven by biology or batch effects

Plots that look like this imply variation is driven by biology

Plots that look like this imply variation is driven by batch effects

https://doi.org/10.1093/biostatistics/kxx053

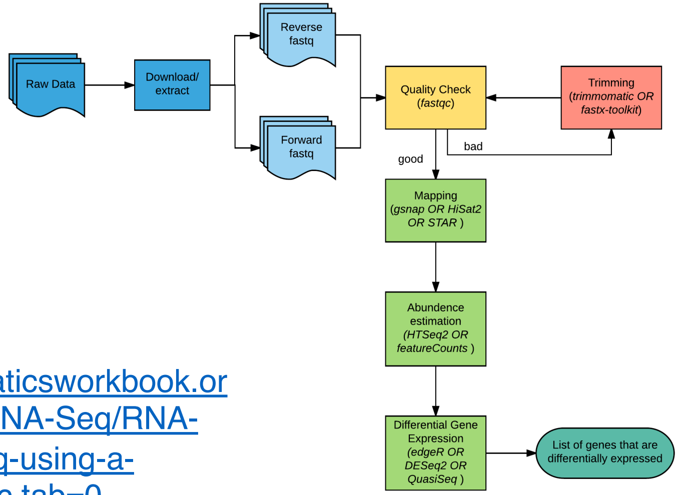
Data analysis

Processing raw full-length scRNA-Seq data

- Input = raw RNA-Seq reads (.fastq.gz files)
- Output = gene count matrix

Using a reference genome

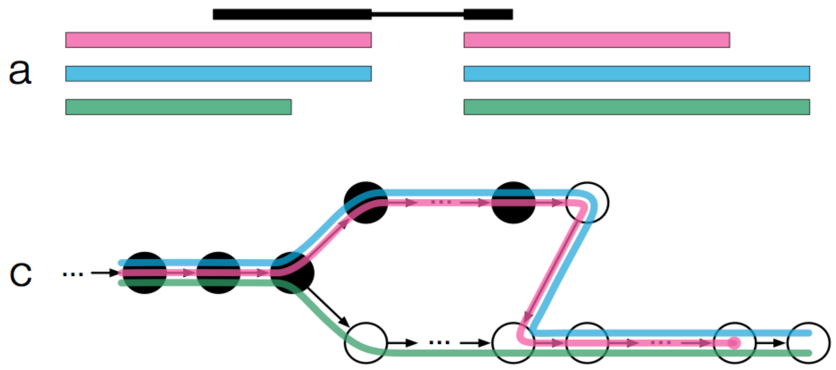
- Align reads with splice-aware aligner
  - STAR or hisat2
- Quantify gene/transcript read counts
  - HTSeq2 or featureCounts


Raw Data
Download/ extract
Reverse fastq
Forward fastq
Quality Check (*fastqc*)
Trimming (*trimmomatic OR fastx-toolkit*)
good
bad
Mapping (*gsnap OR HiSat2 OR STAR*)
Abundence estimation (*HTSeq2 OR featureCounts*)
Differential Gene Expression (*edgeR OR DESeq2 OR QuasiSeq*)
List of genes that are differentially expressed

https://bioinformaticsworkbook.org/dataAnalysis/RNA-Seq/RNA-SeqIntro/RNAseq-using-a-genome.html#gsc.tab=0

Using a reference transcriptome

- Using a pseudo-alignment tool
  - kallisto
  - Salmon
- Alignment & quantification taken care of by the same tool



https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

Data analysis

Processing raw tag-based scRNA-Seq data

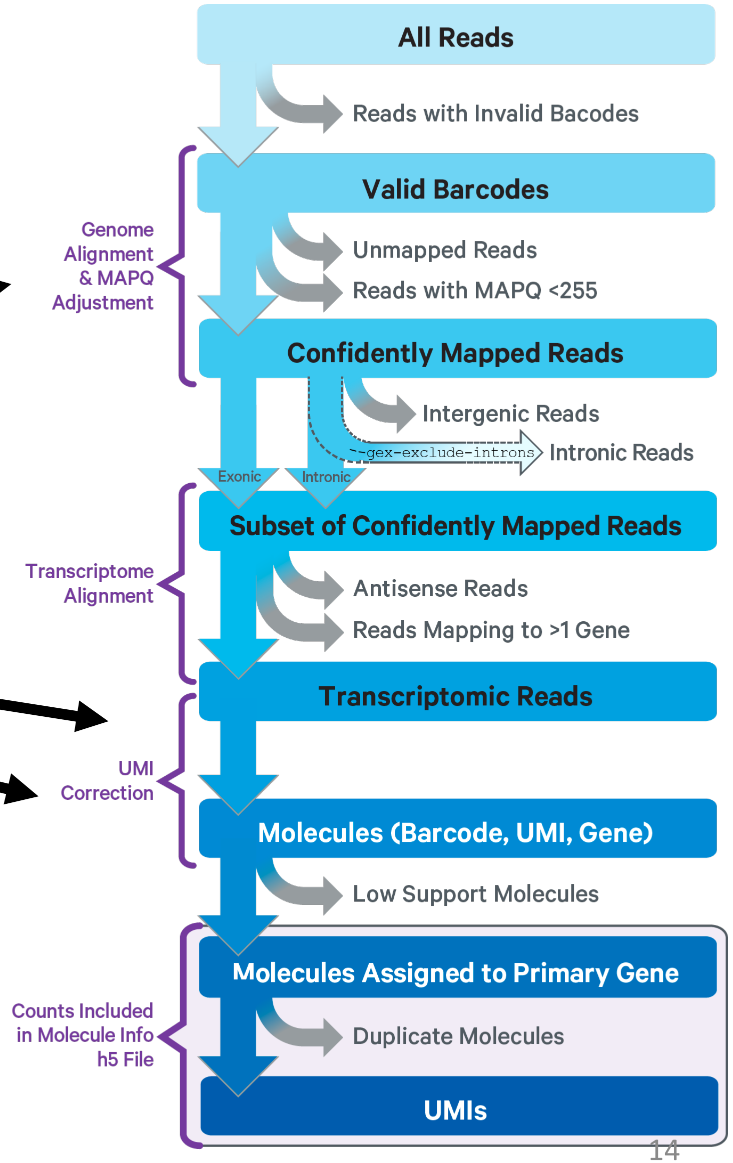
Input = raw RNA-Seq reads (.fastq.gz files)
Output = gene count matrix

Five main steps:

1) Mapping reads to reference genome (transcriptome)
2) Assigning reads to genes
3) Assigning reads to cells (cell barcode demultiplexing)
4) Counting the number of unique RNA molecules (UMI deduplication)
5) Cell filtering

Cell Ranger performs all five steps

- Default tool for 10x Genomics Chromium scRNA-Seq data
- Easy to use and very thorough



All Reads

Reads with Invalid Bacodes

Genome Alignment & MAPQ Adjustment

Valid Barcodes

Unmapped Reads

Reads with MAPQ <255

Confidently Mapped Reads

Intergenic Reads

--gex-exclude-introns

Intronic Reads

Exonic

Intronic

Transcriptome Alignment

Subset of Confidently Mapped Reads

Antisense Reads

Reads Mapping to >1 Gene

Transcriptomic Reads

UMI Correction

Molecules (Barcode, UMI, Gene)

Low Support Molecules

Counts Included in Molecule Info h5 File

Molecules Assigned to Primary Gene

Duplicate Molecules

UMIs

Data analysis

Processing raw tag-based scRNA-Seq data

Input = raw RNA-Seq reads (.fastq.gz files)
Output = gene count matrix

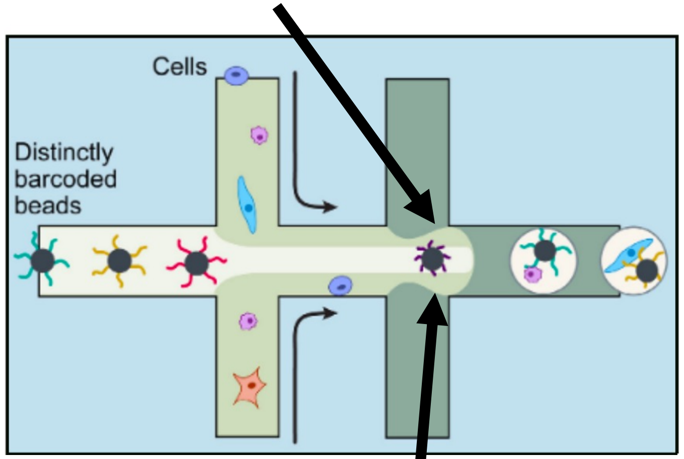
Five main steps:

1) Mapping reads to reference genome (transcriptome)
2) Assigning reads to genes
3) Assigning reads to cells (cell barcode demultiplexing)
4) Counting the number of unique RNA molecules (UMI deduplication)
5) Cell filtering

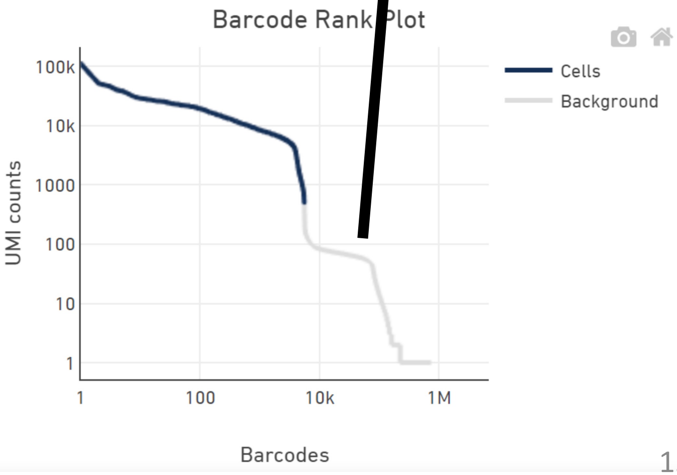
Cell Ranger performs all five steps

- Default tool for 10x Genomics Chromium scRNA-Seq data
- Easy to use and very thorough

"Empty Droplets"
No cell



UMI counts



Cells ordered by # of UMIs

https://assets.ctfassets.net/an68im79xiti/163qWiQBTVi2YLbskJphQX/e90bb82151b1cdab6d7e9b6c845e6130/CG000329_TechnicalNote_InterpretingCellRangerWebSummaryFiles_RevA.pdf

Data analysis

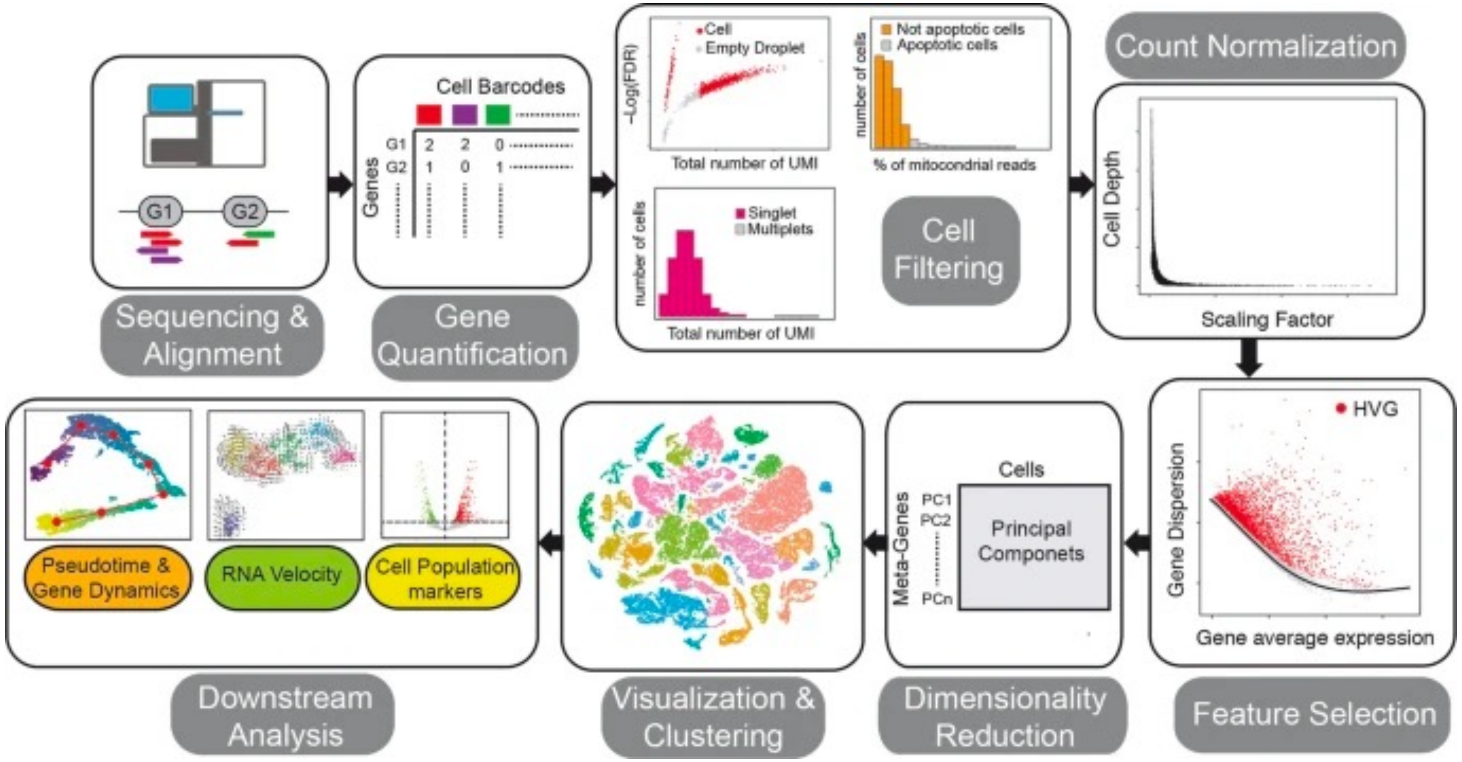
Major steps following raw data processing

Pre-processing
- Quality control
- Normalization
- Variable feature selection

Dimensionality reduction
- Linear
- Non-linear

Cell clustering

Downstream analysis
- Identifying cell population marker
- Differential expression analysis
- Trajectory analysis

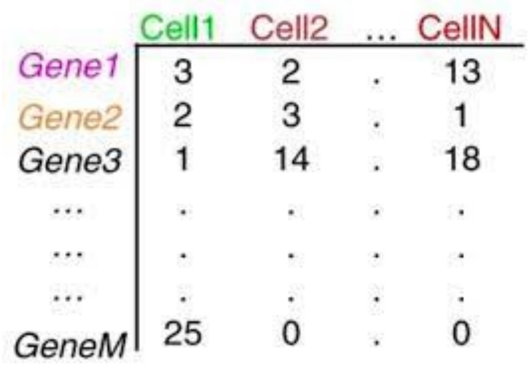


https://doi.org/10.1007/978-1-0716-1307-8_19

Data analysis

Quality control

Main goal:
Remove poor quality cells



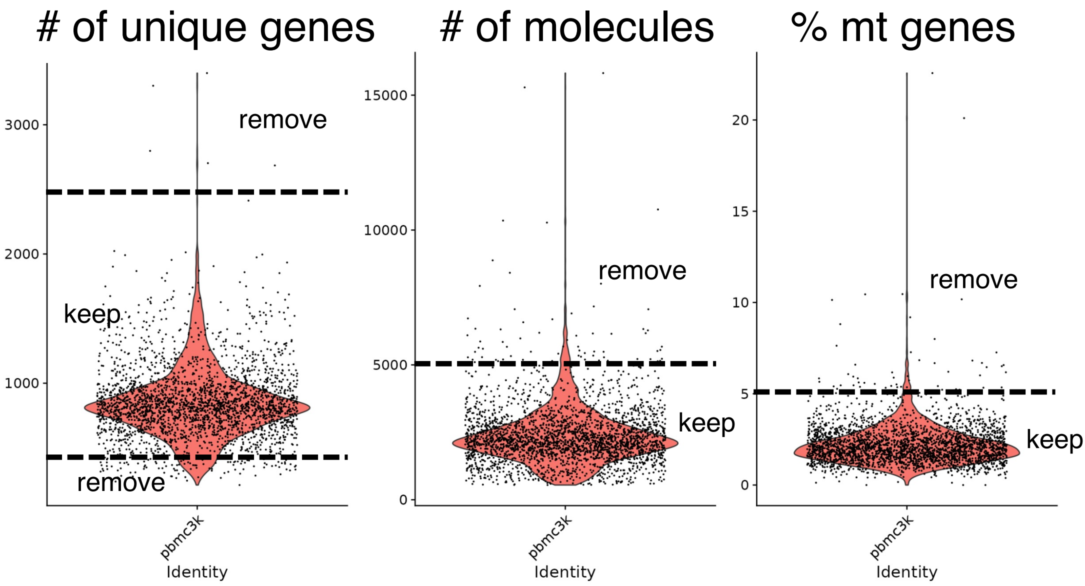
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Common criteria used:

1) **# of unique genes detected in a cell**
   - Low-quality cells or empty droplets have fewer genes detected
   - Cell doublets or multiplets have many genes detected

2) **# of molecules detected within a cell (correlates with unique genes)**

3) **% of reads mapping to the mitochondrial genome**
   - Low-quality / dying cells exhibit mitochondrial contamination



# of unique genes

# of molecules

% mt genes

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Data analysis

Normalization

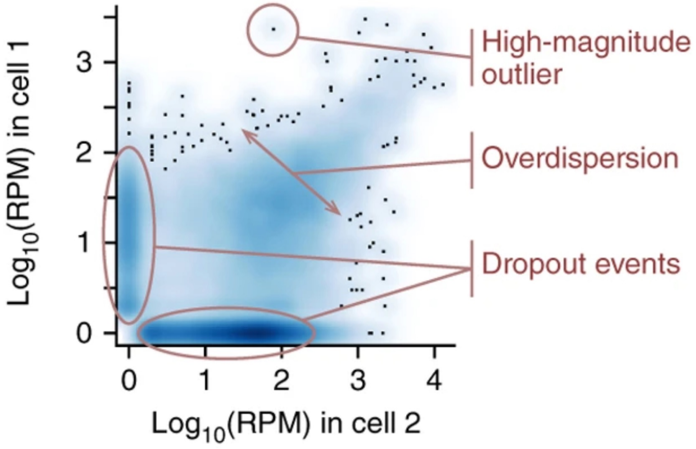
Main goals:

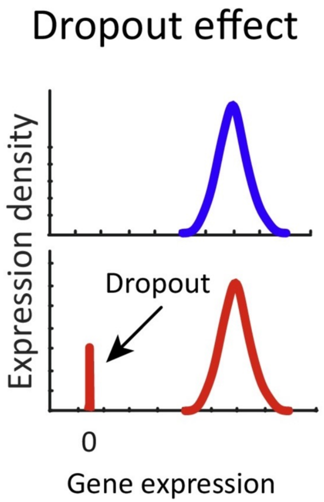
1) Remove technical bias from gene expression data
2) Ensure downstream analyses aren't dominated by only the most highly expressed genes

scRNA-seq uses small amount of input RNA = more inaccurate/variable measurements

Major challenge of scRNA-seq data:

- Transcripts often 'missed' (not detected) during sequencing though they are actually expressed
- Known as dropouts
- Requires different transformation methods than Bulk RNA-Seq





https://doi.org/10.1038/nmeth.2967

https://doi.org/10.1016/j.it.2016.12.001

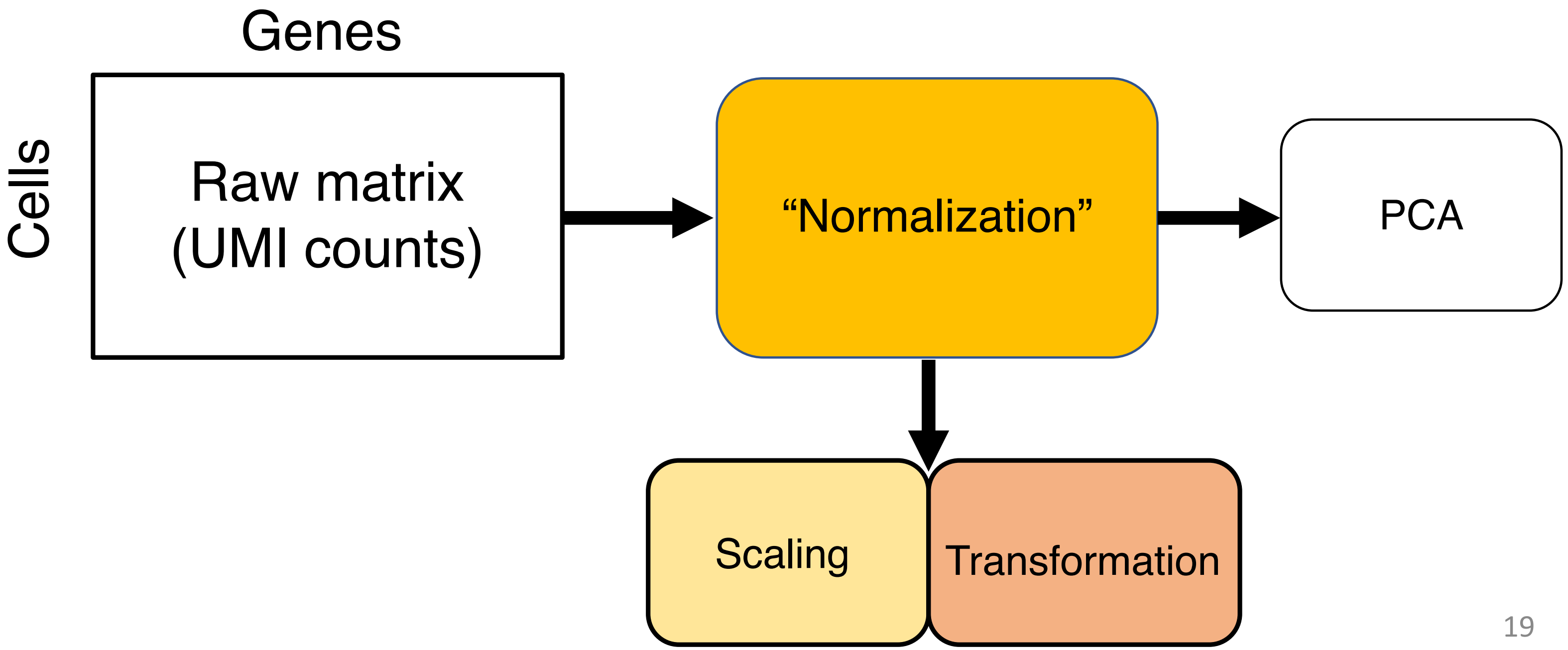
Data analysis

Normalization

Two steps:

1) Scaling
   - Accounts for cells not having same sequencing depth or same amount of input RNA

2) Transformation
   - Accounts for genes being expressed at different levels and with different variation
   - Accounts for "dropouts" (moderate/high expression in one cell but not detected in another)
   - Numerous methods:
     - Log-normalization
     - Square-root
     - sctransform



https://doi.org/10.1186/s13059-019-1874-1

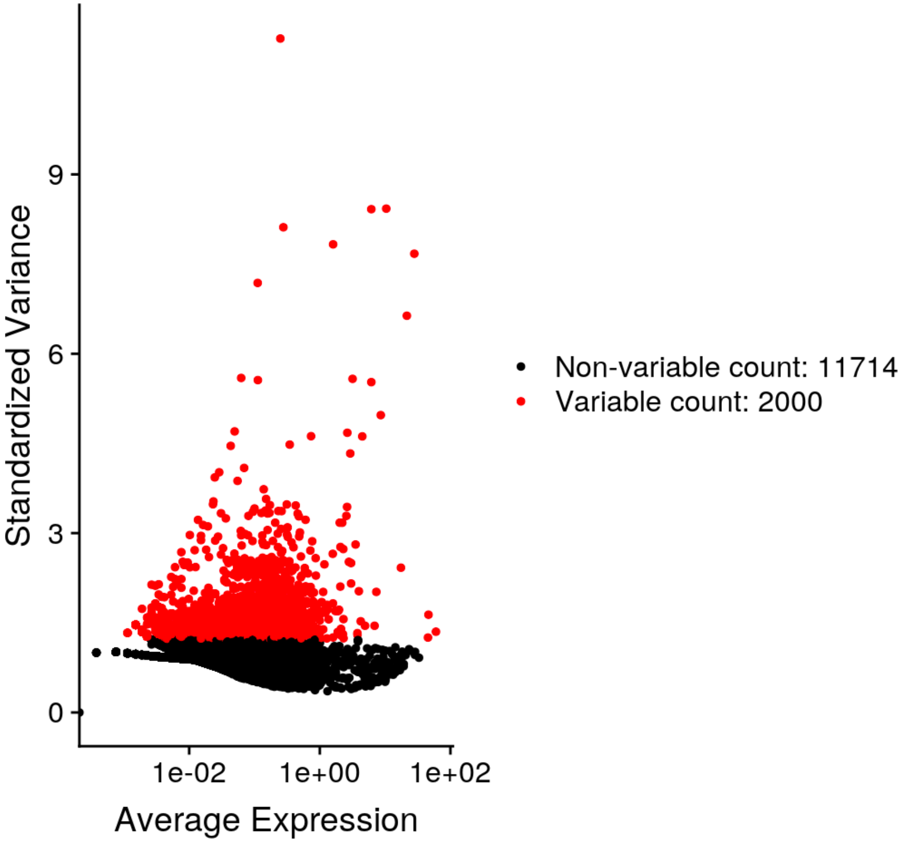
Data analysis

Variable feature selection

Main goal:

Keep genes with relevant biological information, while excluding uninformative genes

- Reduces dimensionality of data
- Enhances the ability to detect biological signal in dataset
- Typically aims to keep 500-2000 genes with most cell-to-cell variability



https://satijalab.org/seurat/archive/v3.0/pbmc3k_tutorial.html

Data analysis

Dimensionality reduction

Main goal:
Condense complex (multi-dimensional) data into simpler (lower-dimensional) representations while keeping the most important properties of the data

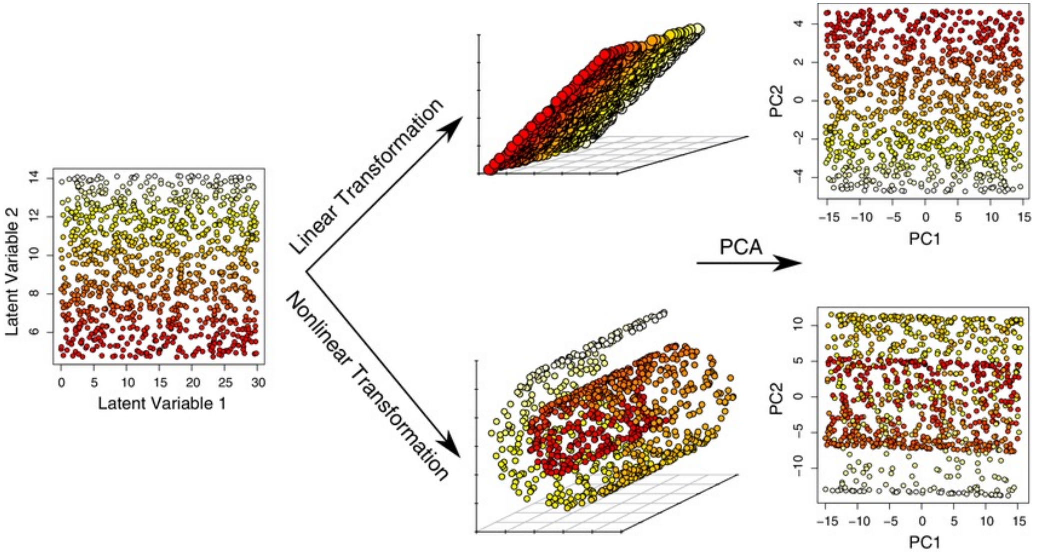
Required to perform important downstream analyses (e.g. clustering and visualization)

Linear

- PCA (Principal Component Analysis)

Non-linear

- t-sne (t-distributed stochastic neighbor embedding)
- UMAP (Uniform Manifold Approximation and Projection)



https://doi.org/10.1007/s11692-018-9464-9

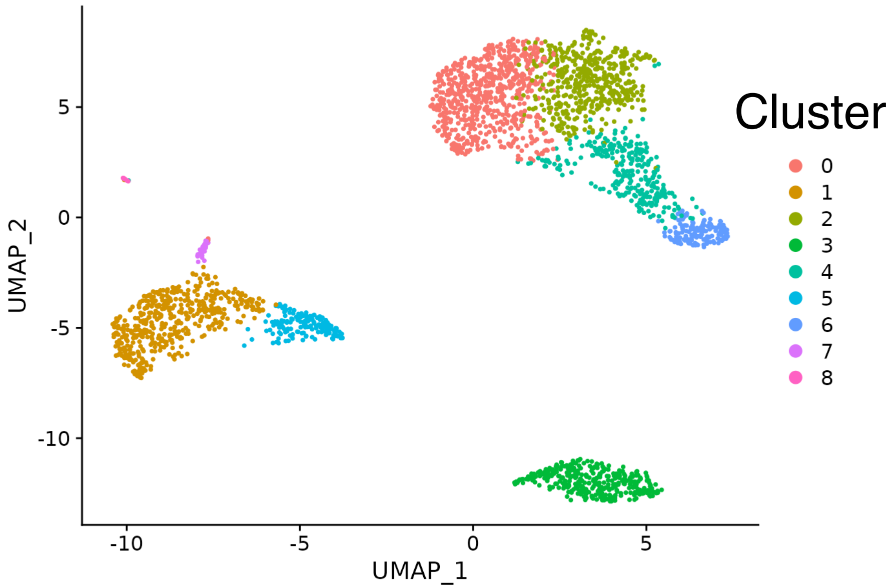
Data analysis

Cell clustering

Main goal:
Separate a population of cells into transcriptionally distinct sub-populations (clusters)

Main steps:

1) Calculate how similar each of the cells are to each other (a similarity score metric)
2) Partition/group (cluster) cells based on those scores



https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

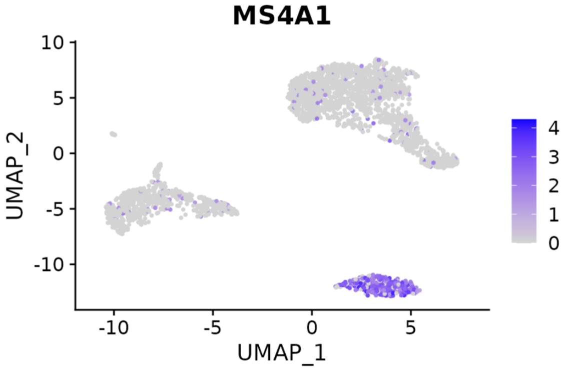
Data analysis

Further downstream analyses

Identifying cell population markers

Goal: Determine the genes most differentially expressed between cell clusters

Helps to determine cell identities of different cell clusters



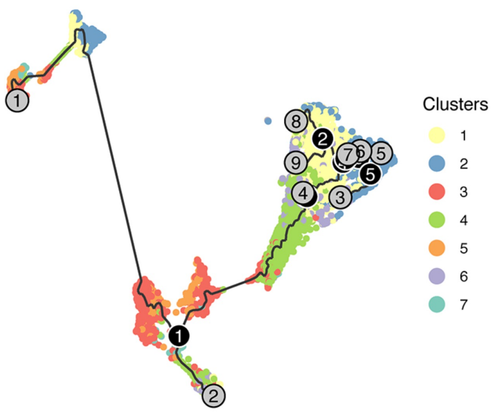
Differential expression analysis

Goal: Determine the genes most differentially expressed between cell clusters or conditions

Trajectory analysis

Goal: Determine the differentiation trajectory of a set of cells

Monocle 3



Murine cortex differentiation

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

https://doi.org/10.1371/journal.pcbi.1008205

Helpful resources

Guided courses and vignettes

- Wellcome Sanger Institute: https://www.singlecellcourse.org/
- Broad Institute: https://broadinstitute.github.io/2020_scWorkshop/
- Seurat vignettes: https://satijalab.org/seurat/articles/get_started.html

Review articles

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

"Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data"

Preparing for the Hands-on Tutorial

Request an HPC server account

- Biowulf HPC instructions:
  - https://hpc.nih.gov/docs/accounts.html
  - All NIH researchers in the Enterprise directory can request access

- Skyline HPC instructions:
  - NIAID researchers are automatically provided account access
  - Can test access here: https://skyline.niaid.nih.gov/access/

- **Install R and R studio**
  - **https://posit.co/download/rstudio-desktop/**
  - R version 4.0 or greater