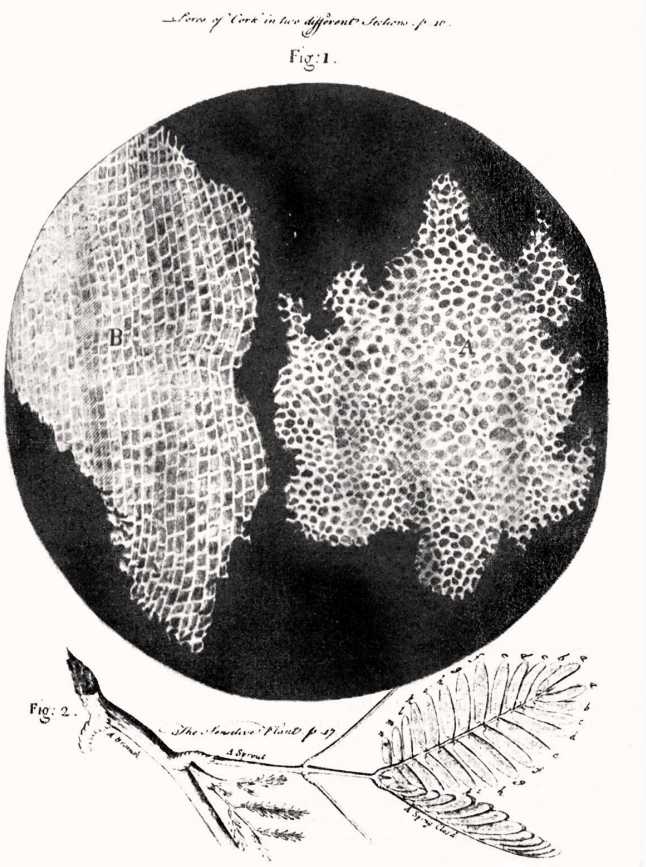
Single Cell Sequencing

Cells are Important

- Fundamental unit of life
- Autonomous and unique
- Interactive
- Dynamic - change over time
- Evolution occurs on the cellular level



Robert Hooke's drawing of cork cells, 1665

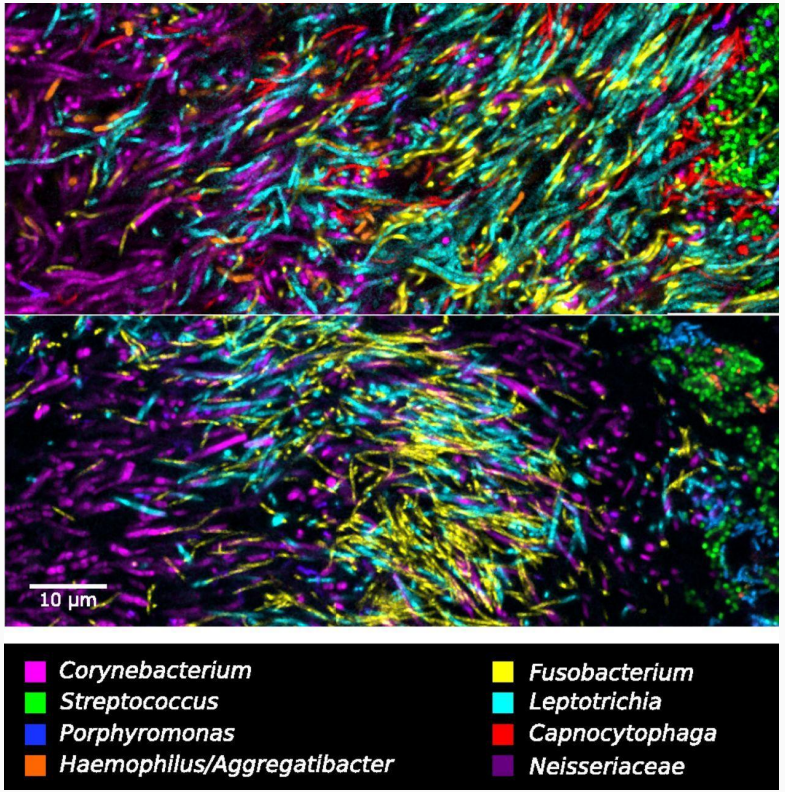
Cells are Diverse

Type	Prokaryotes	Eukaryotes
Typical size	~ 1-5 μm	~ 10-100 μm
DNA form	Circular	Linear
DNA location	Cytoplasm	Nucleus
DNA amount	~ .3-16 fg	~3-300,000 fg
RNA amount	~ 5-26,000 fg	~ 1,000-350,000 fg

Landenmark HKE, Forgan DH, Cockell CS (2015). An Estimate of the Total DNA in the Biosphere. PLoS Biol 13(6): e1002168. https://doi.org/10.1371/journal.pbio.1002168

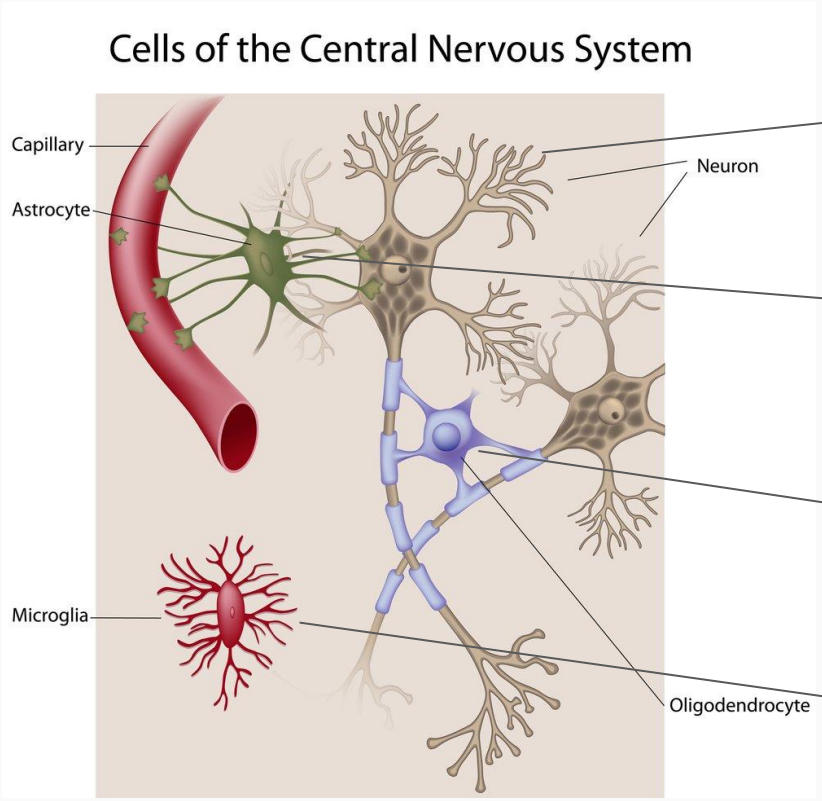
Cells are Diverse: Microbial Ecology

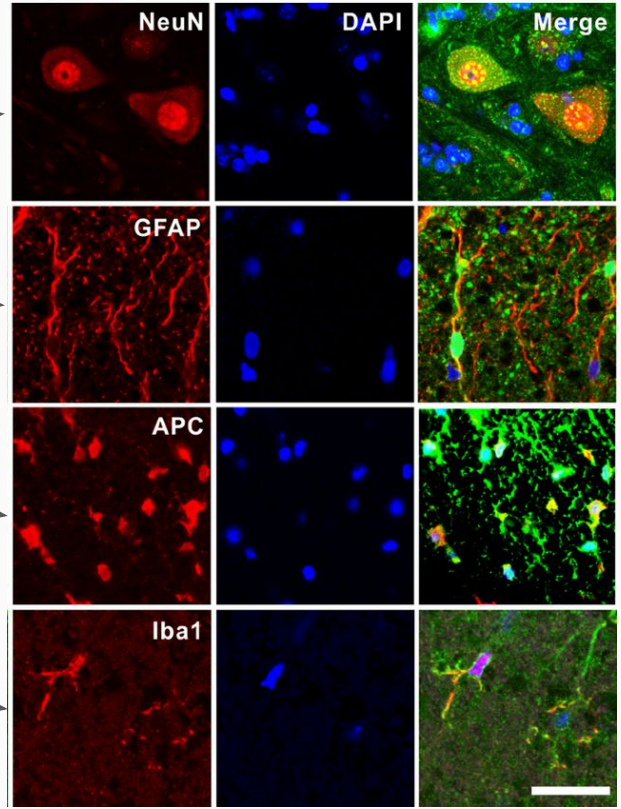
- Nearly every environment on Earth supports microbial life
- Coresident microbes usually work together in a balance
- Imbalances (or invaders) can disrupt the function of overall ecology
- Which specific microbes out of millions cause a particular effect?



Mark Welch, et al. 2016. "Biogeography of a Human Oral Microbiome at the Micron Scale." *Proceedings of the National Academy of Sciences of the United States of America* 113 (6): E791–800.

Cells are Diverse: Human Brain



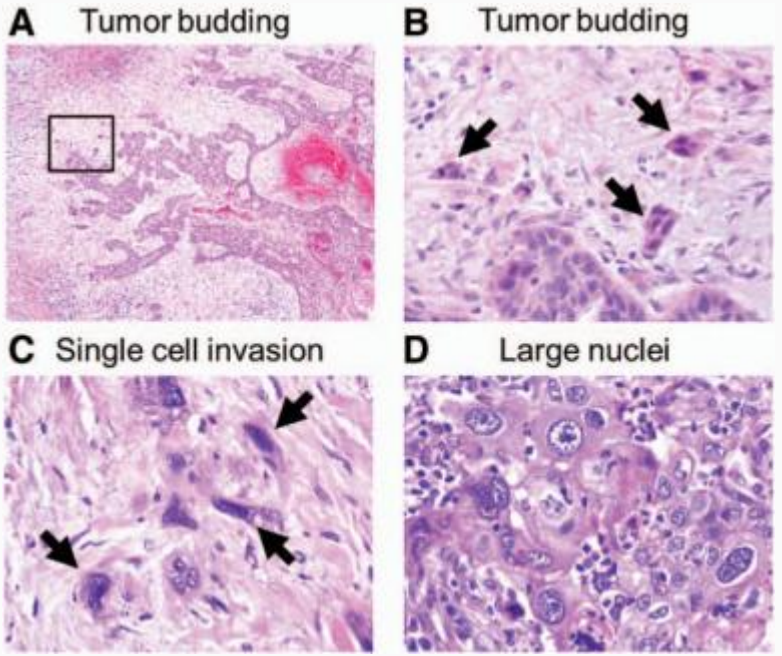


Glia

The vast majority of cells (10x-50x) in your brain are glia, not neurons

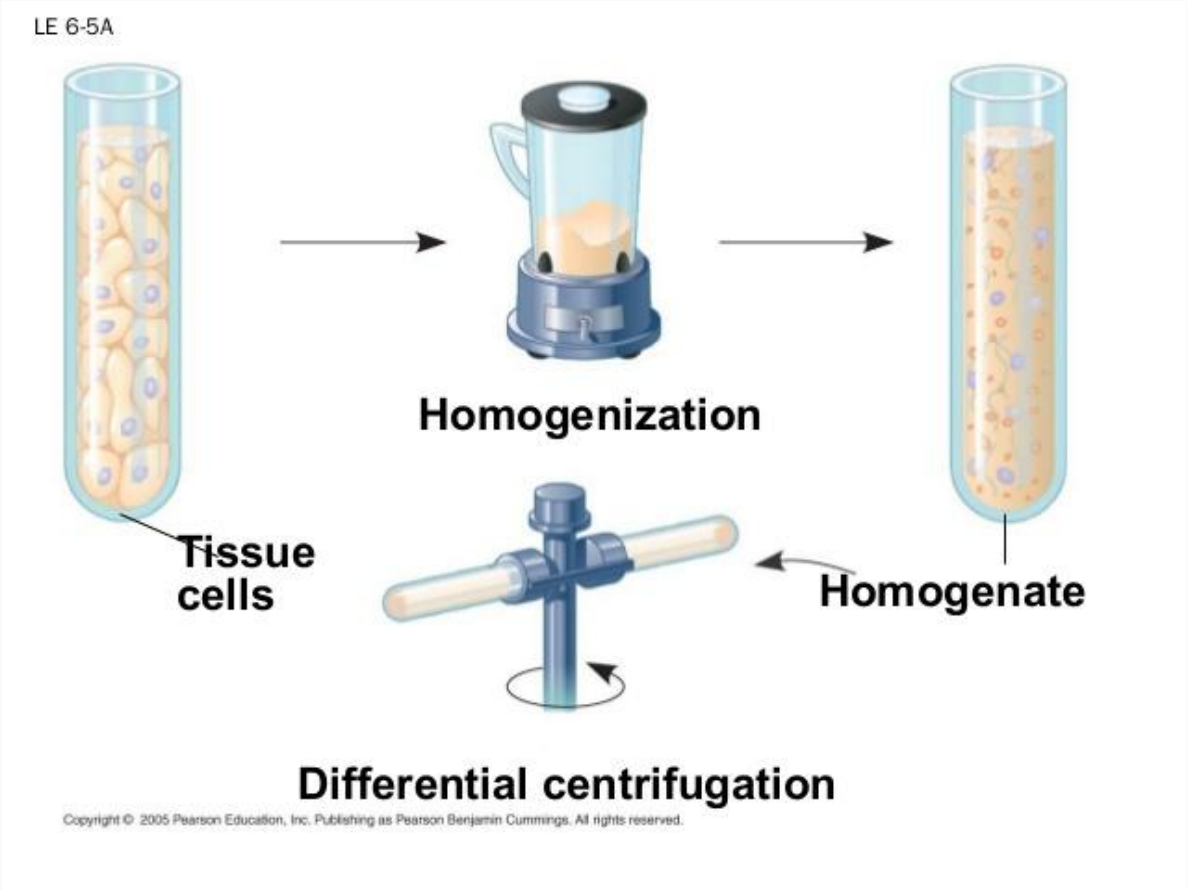
Cells are Diverse: Tumors

- Tumors are seeded by single mutated cells
- Founder cells divide and further mutate
- Large tumors undergo angiogenesis
- Selectively kill the cancerous cells: cure the cancer
- But which cells to target?



Kadota, Kyuichi, et al. 2014. "Comprehensive Pathological Analyses in Lung Squamous Cell Carcinoma: Single Cell Invasion, Nuclear Diameter, and Tumor Budding Are Independent Prognostic Factors for Worse Outcomes." *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 9 (8): 1126–39.

The Forest: Tissue Homogenate



The Trees: Cells

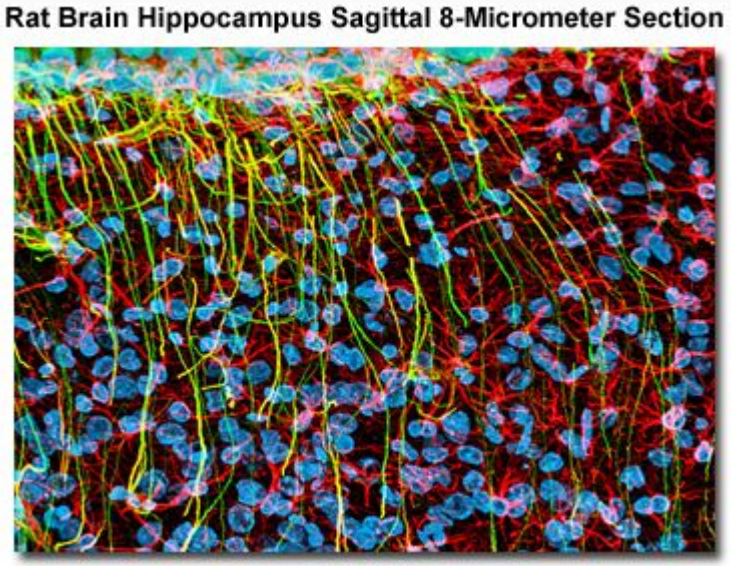
- What cell types are in a sample?
- What are their proportions?
- How does their transcription differ?
- Which/how do specific cells respond to stimulus?
- How do cells develop over time?
- What is the level of mosaicism in tissues?

Single Cell Sequencing Workflow

1. Dissociation of tissue, isolation of cells
2. FAC sorting (optional)
3. Nucleic acid extraction and processing
4. Sequencing library prep + sequencing
5. Analysis

Dissociation of Tissue

- Cells in complex tissue are highly intermingled
- Must separate cells from each other *without destroying them or breaking membranes*
- Complex cellular morphology (e.g. neurons) makes dissociation challenging
- Can isolate nuclei instead:
  - Contain DNA/some RNA
  - Much more input material needed

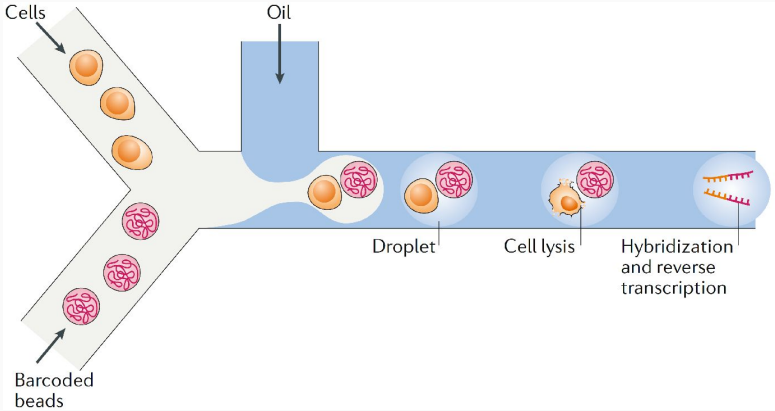


Cell Isolation Techniques

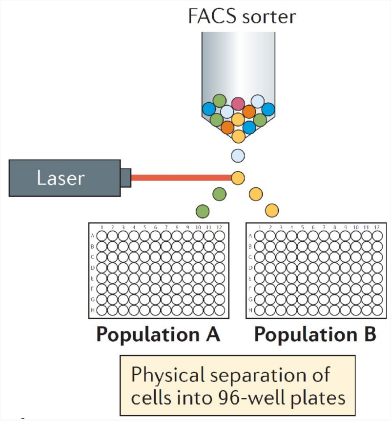
Microfluidics

Fluidigm C1
Integrated Fluidic Circuit (IFC)

Droplet Based



Fluorescence Activated Cell Sorting (FACS)

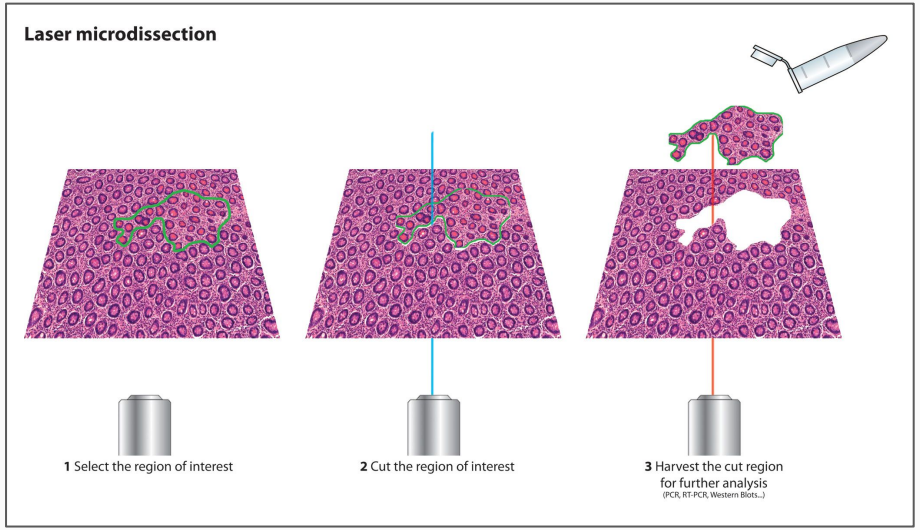


Some technologies use all three!

Potter, S. S. 2018.
Papalexi, E. Satija, R. 2018.

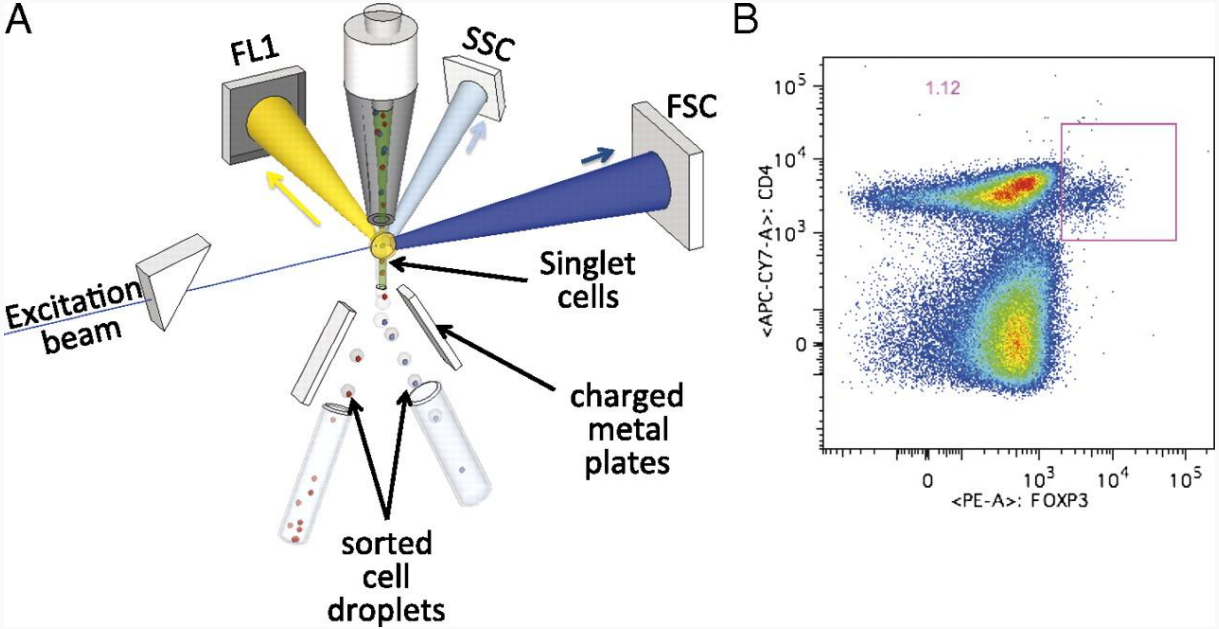
Laser Capture Microdissection

- Technique for isolating groups of cells *in situ*
- Low throughput, requires expensive equipment
- Laser causes damage to tissue and degrades RNA
- **Not generally suitable for single cell sequencing**



Fluorescence-Activated Cell Sorting (FACS)

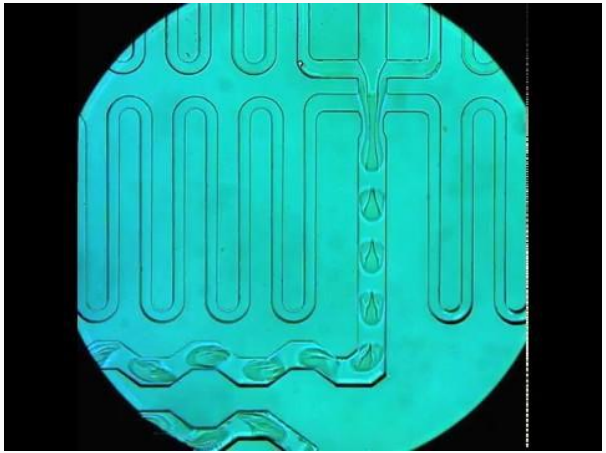
- Cells with known surface markers are tagged with fluorescent antibodies
- Tagged cells excited by lasers during flow cytometry
- Excited and non-excited cells separated and collected
- Cell type specific populations can be sequenced and studied

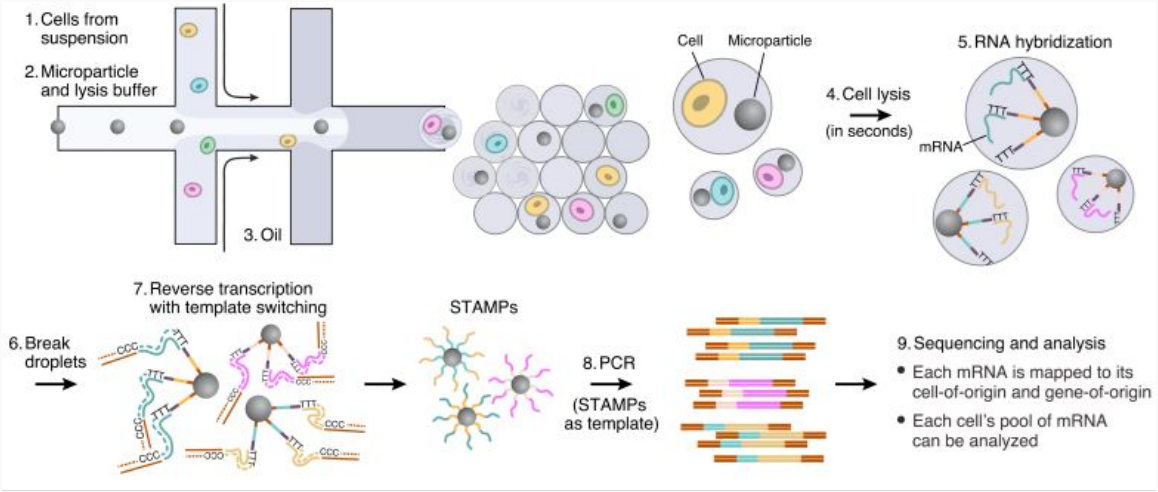


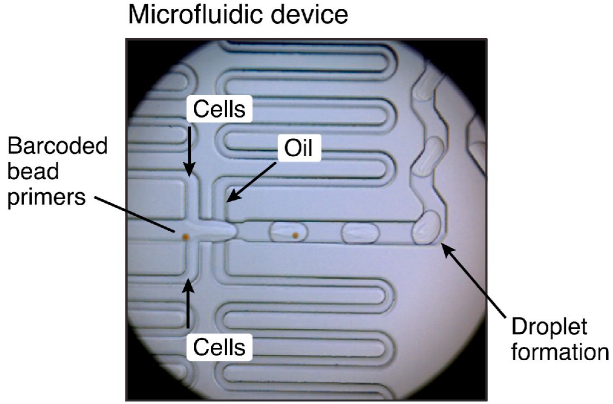
Jaye, David L., Robert A. Bray, Howard M. Gebel, Wayne A. C. Harris, and Edmund K. Waller. 2012. “Translational Applications of Flow Cytometry in Clinical Practice.” *Journal of Immunology* 188 (10): 4715–19.

Drop-seq

- Microfluidics used to pair cells and barcodes/reagents into separate oil droplets
- Concentrations carefully controlled to get 1:1 cell/barcode matches in each oil droplet with high statistical confidence

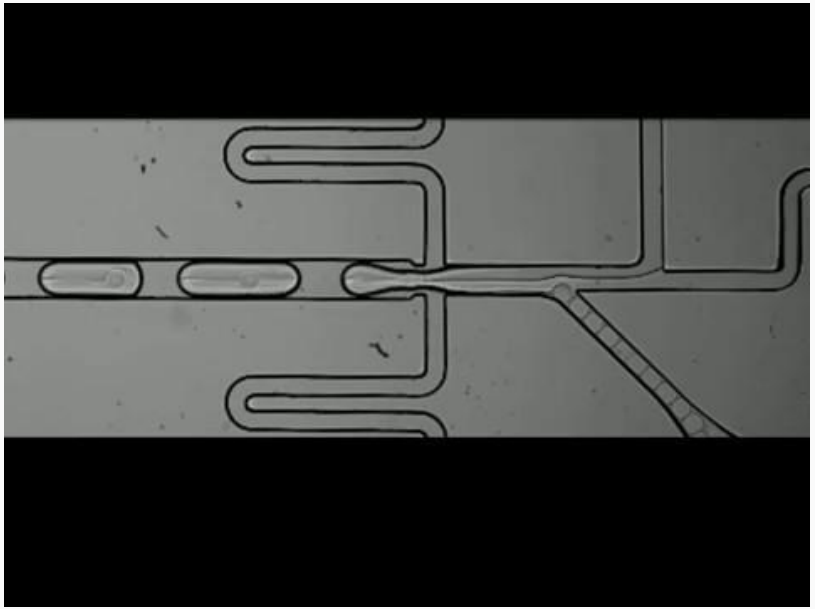


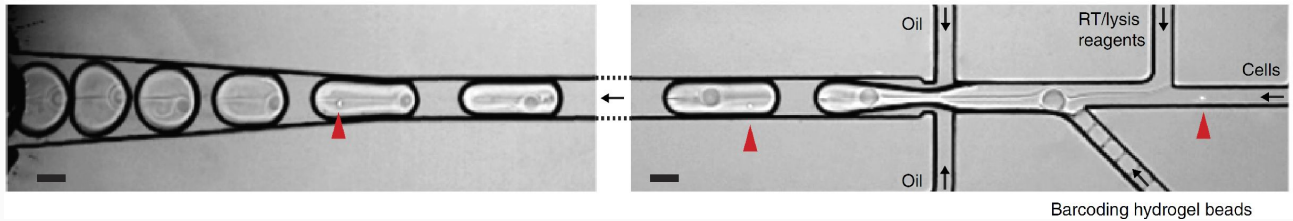




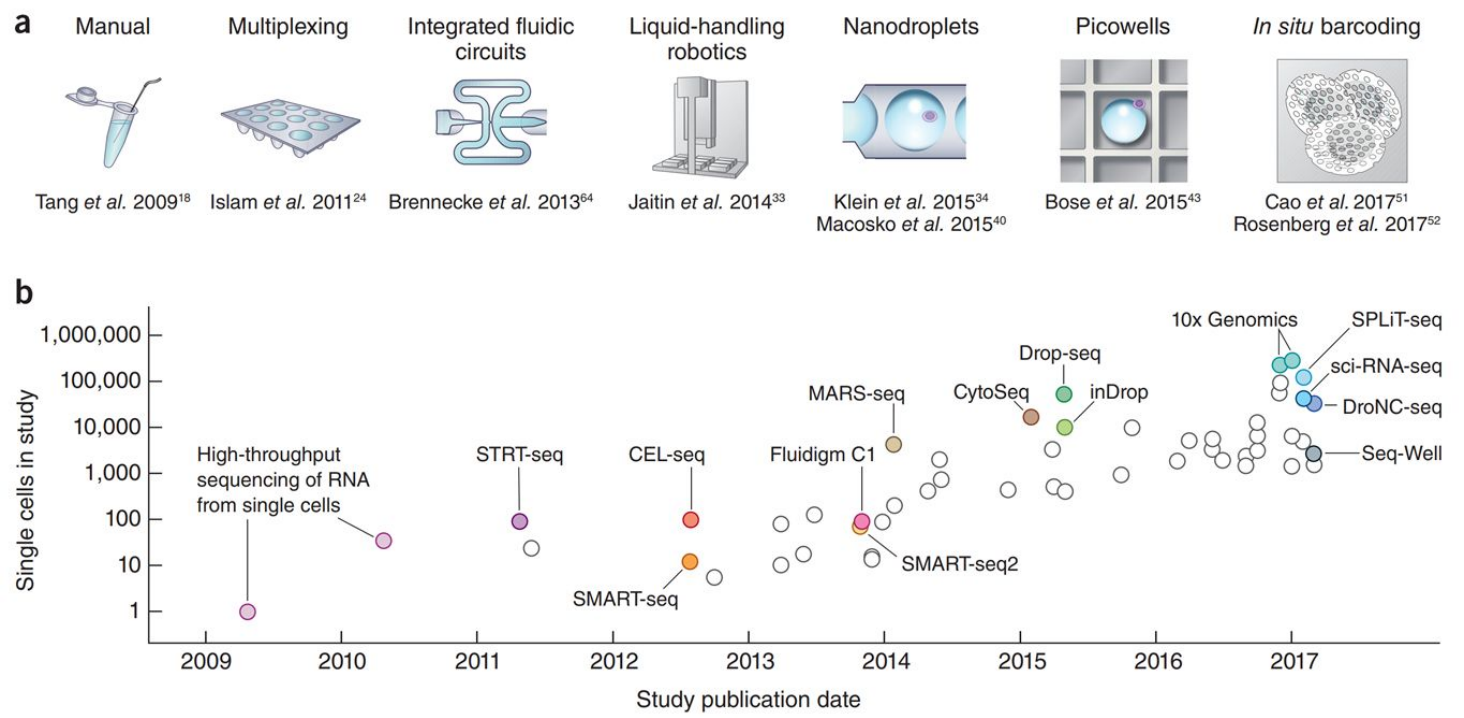
inDrops

- Alternative technology to Drop-Seq
- Essentially the same strategy, but uses hydrogel beads instead of microparticles





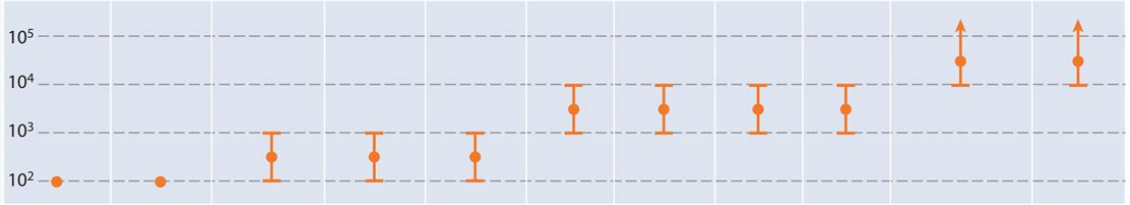
A decade of single cell RNA-seq



Svensson et al. 2018. DOI:10.1038/nprot.2017.149

Slide by Lior Pachter and Matt Thomson

A decade of single cell RNA-seq

	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											

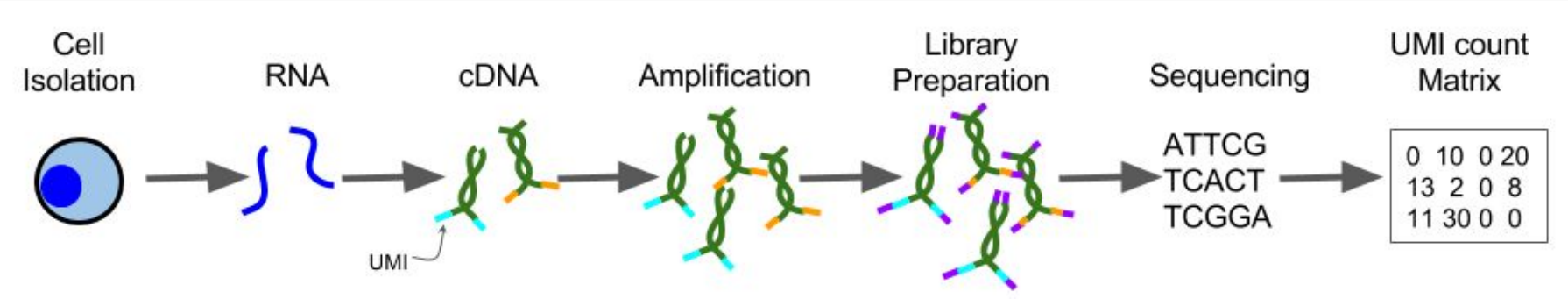
Chen et al. 2018. DOI:10.1146/annurev-biodatasci-080917-013452

Nucleic Acid Extraction + Processing

- femto- to picograms of input material
- Each *cell* is:
  - Assigned a unique DNA barcode
  - Optionally treated with UMIs
  - Amplified by one of:
    - Reverse transcriptase (RNA)
    - Multiple displacement amplification (DNA)
    - In vitro transcription (RNA)

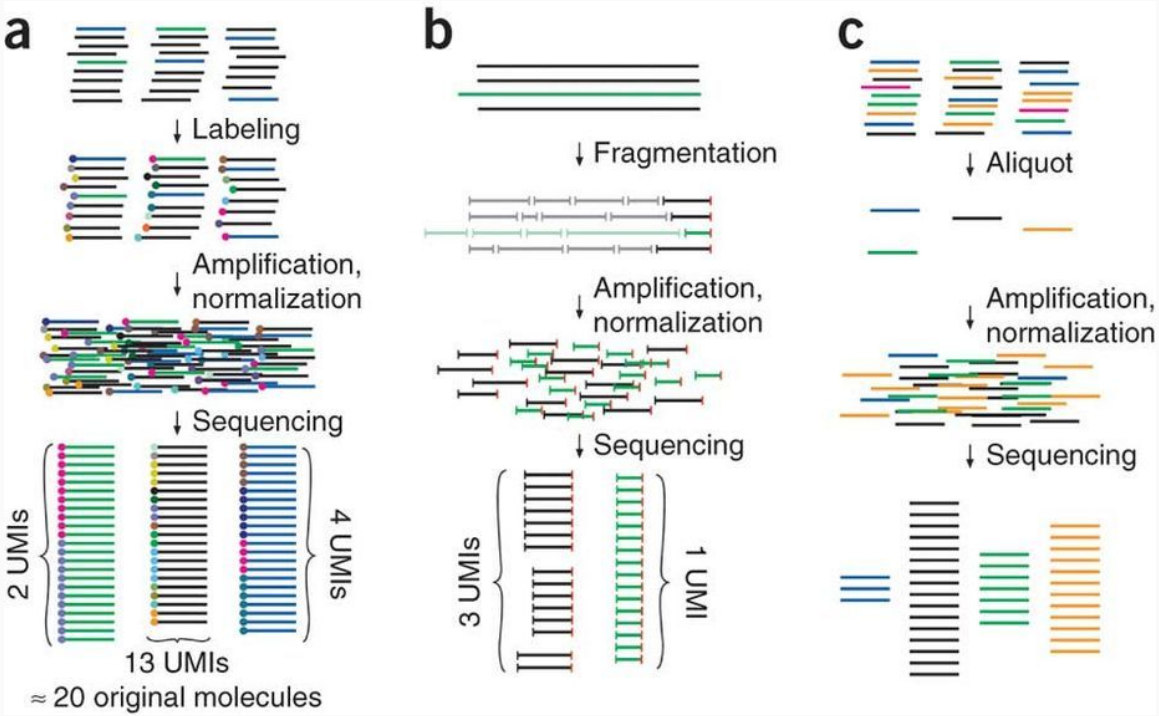
Unique Molecular Identifiers (UMIs)

- Low input material may cause amplification bias
- UMIs are sequences that correspond to *one fragment*
- Sequenced reads with the same UMI are from the same fragment
- Unique sequences collapsed/deduplicated for counting



Unique Molecular Identifiers (UMIs)

Strategies for counting individual molecules



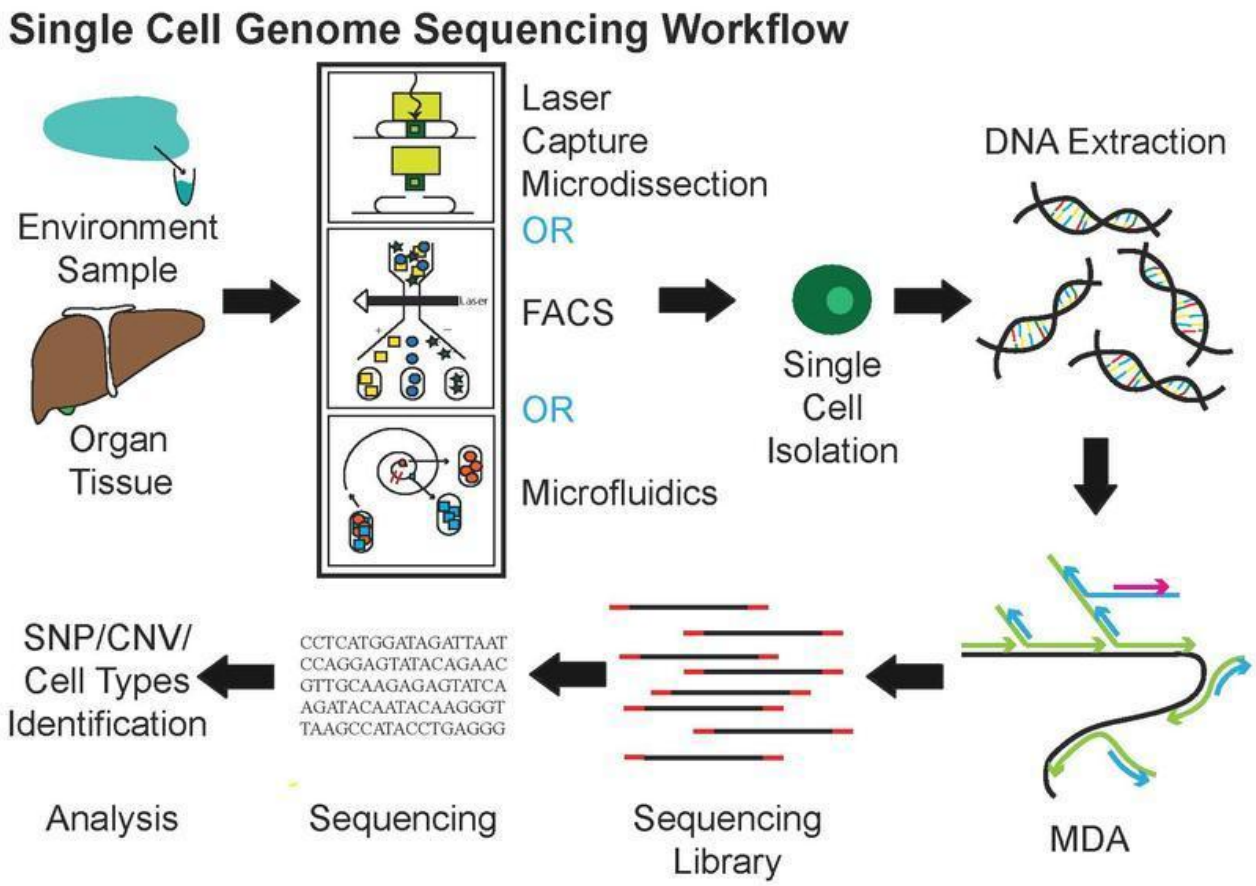
Kivioja et al. 2011. “Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers.” *Nature Methods* 9 (1): 72–74.

Sequencing Library Prep and Sequencing

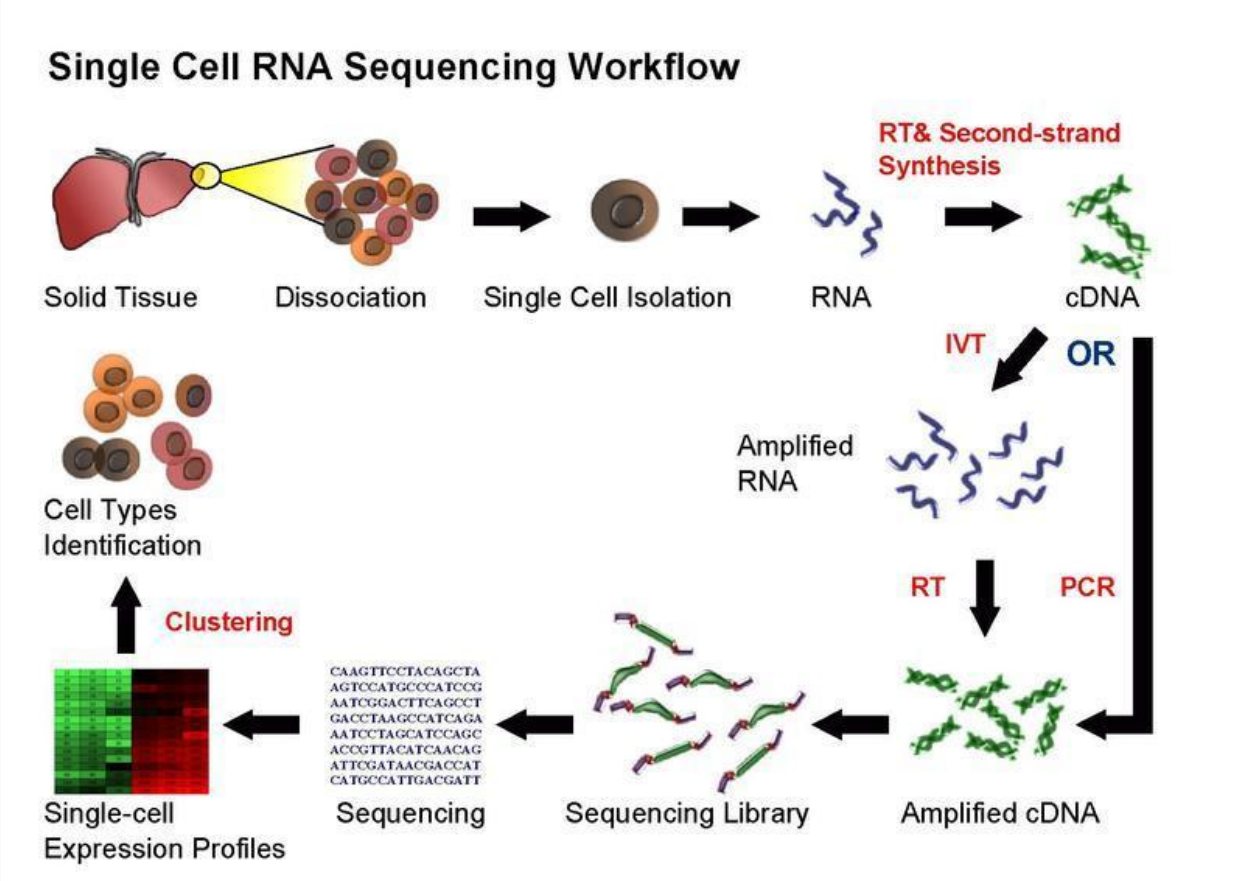
- Previous protocols typically include sequencing primers
- Sequencing depth = (# of cells) x (required depth):
  - RNA - *50k paired end reads / cell* for cell type classification
  - RNA - *.25M-1M paired reads / cell* for transcriptome coverage
  - DNA - 30-100x per cell
- e.g. 1000 cell scRNA-Seq = 250M-1B reads per sample!
- Sequences in one PE fastq file are entirely barcodes
- Read length > 50bp for annotated genome
- Single cell sequencing is still *very expensive*

Rizzetto, et al. 2017. “Impact of Sequencing Depth and Read Length on Single Cell RNA Sequencing Data of T Cells.” *Scientific Reports* 7 (1): 12781.

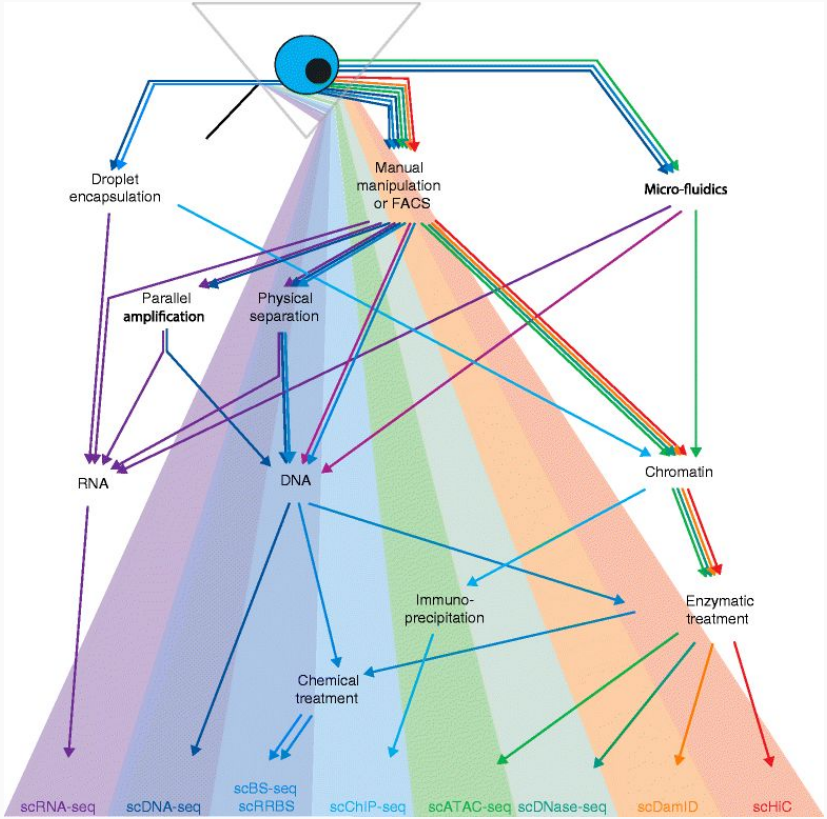
scDNA-Seq Workflow



scRNA-Seq Workflow



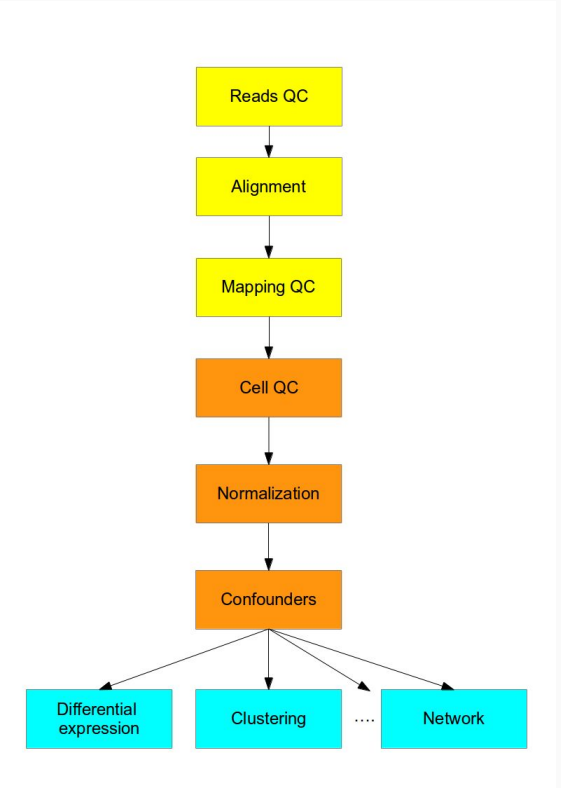
Different Types of Single Cell Sequencing



Clark, Stephen J., Heather J. Lee, Sébastien A. Smallwood, Gavin Kelsey, and Wolf Reik. 2016. “Single-Cell Epigenomics: Powerful New Methods for Understanding Gene Regulation and Cell Identity.” *Genome Biology* 17 (April): 72.

Analysis Overview

1. Sequence QC
   a. Demultiplex
   b. UMI Collapsing
2. Alignment
3. Quantification
4. Normalization
5. DE, Clustering, etc



https://hemberg-lab.github.io/scRNA.seq.course/introduction-to-single-cell-rna-seq.html

scruff R/Bioconductor package

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help | Developers | About

Search:

Home » Bioconductor 3.8 » Software Packages » scruff

scruff

platforms all | rank unknown | posts 0 | in Bioc < 6 months
build ok | updated < 1 month

DOI: 10.18129/B9.bioc.scruff

Single Cell RNA-Seq UMI Filtering Facilitator (scruff)

Bioconductor version: Release (3.8)

A pipeline which processes single cell RNA-seq (scRNA-seq) reads from CEL-seq and CEL-seq2 protocols. Demultiplex scRNA-seq FASTQ files, align reads to reference genome using Rsubread, and generate UMI filtered count matrix. Also provide visualizations of read alignments and pre- and post-alignment QC metrics.

Author: Zhe Wang [aut, cre], Junming Hu [aut], Joshua Campbell [aut]

Maintainer: Zhe Wang <zhe at bu.edu>

Citation (from within R, enter `citation("scruff")`):

Wang Z, Hu J, Campbell J (2019). *scruff: Single Cell RNA-Seq UMI Filtering Facilitator (scruff)*. R package version 1.0.3.

Installation

To install this package, start R (version "3.5") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("scruff", version = "3.8")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("scruff")
```

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

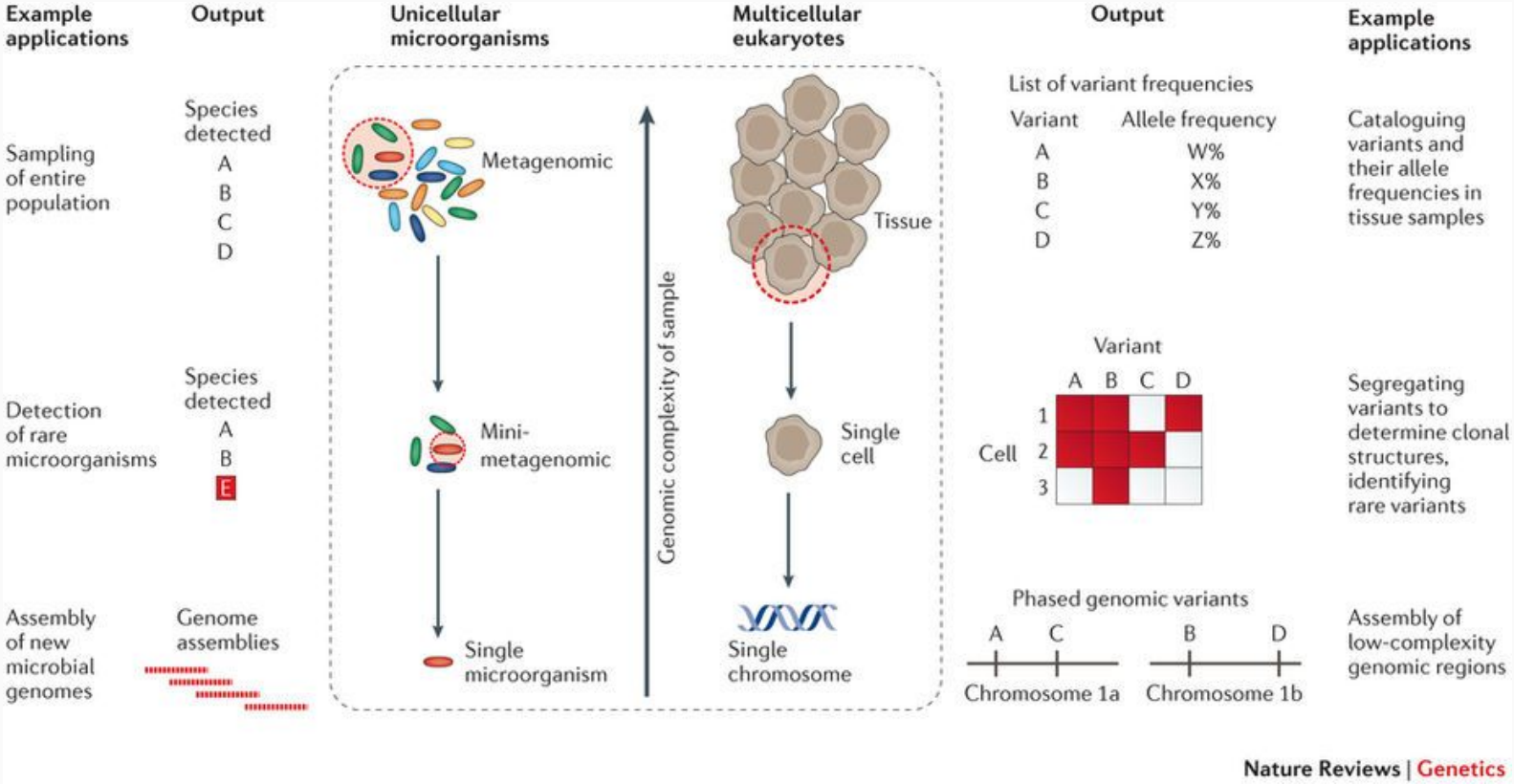
Sequence QC

- One sample is 100s or 1000s of cells
  - i.e. ~1,000 fastq files *per sample*
  - May or may not be already demultiplexed by core
- UMI-Tools - open source UMI software
- Normal fastq processing and QC:
  - Adapter and quality trimming
  - fastqc, multiqc

Alignment

- Standard tools and QC
- Alignment: STAR, bwa, bowtie, etc
- QC: RSeQC, multiqc, etc

scDNA-Seq Analysis



Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews. Genetics* 17 (3): 175–88.

scDNA-Seq Analysis

- Genome assembly
  - Bacteria genomes
  - Mosaic/chimeric genomes (e.g. tumors)
- Cell lineage-specific
  - SNPs
  - Structural variants
  - Ploidy

Quantification

- STAR+htseq-count, kallisto, salmon, etc
- Each sample has a different # of cells
- Each cell has the same number of measurements (e.g. genes)
  - = (# of samples) x (# of cells) x (# of genes)
  - Sparse: most will be zero!

The Counts Matrix

Matrix is cells x genes

Needs to be filtered:

- gene3 - all zeros
- gene5 - mostly zeros
- cell3 - failed/rare cell
- cell5 - failed/overamplified cell

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Count Matrix Normalization

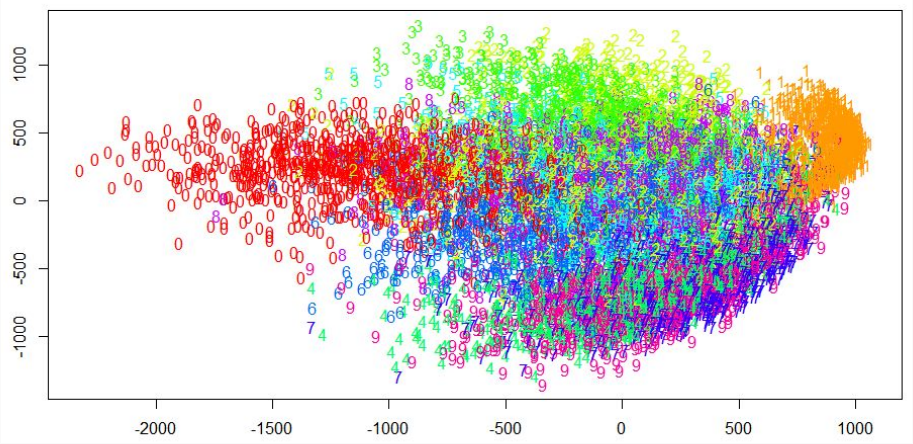
- Each cell is like an independent dataset
- Normalization to compare between cells
- No consensus yet, many methods (CPM, FPKM, upper quartile, downsampling etc...)
- Just *within* sample normalization, between sample is even harder
- Confounding factors are...confounding

Analysis: Dimensionality Reduction

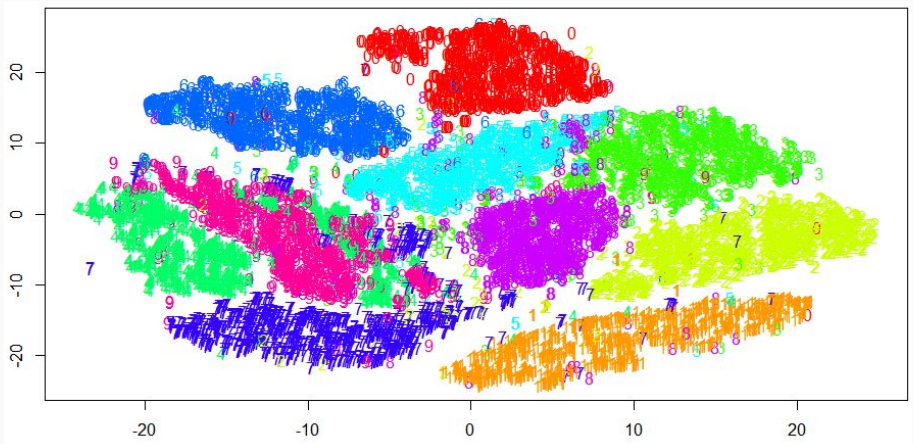
Digit Recognizer (Kaggle)

10,000 hand written digits,
784 pixels per digit.
Pixel-value is an integer
between 0 and 255,
inclusive.

PCA



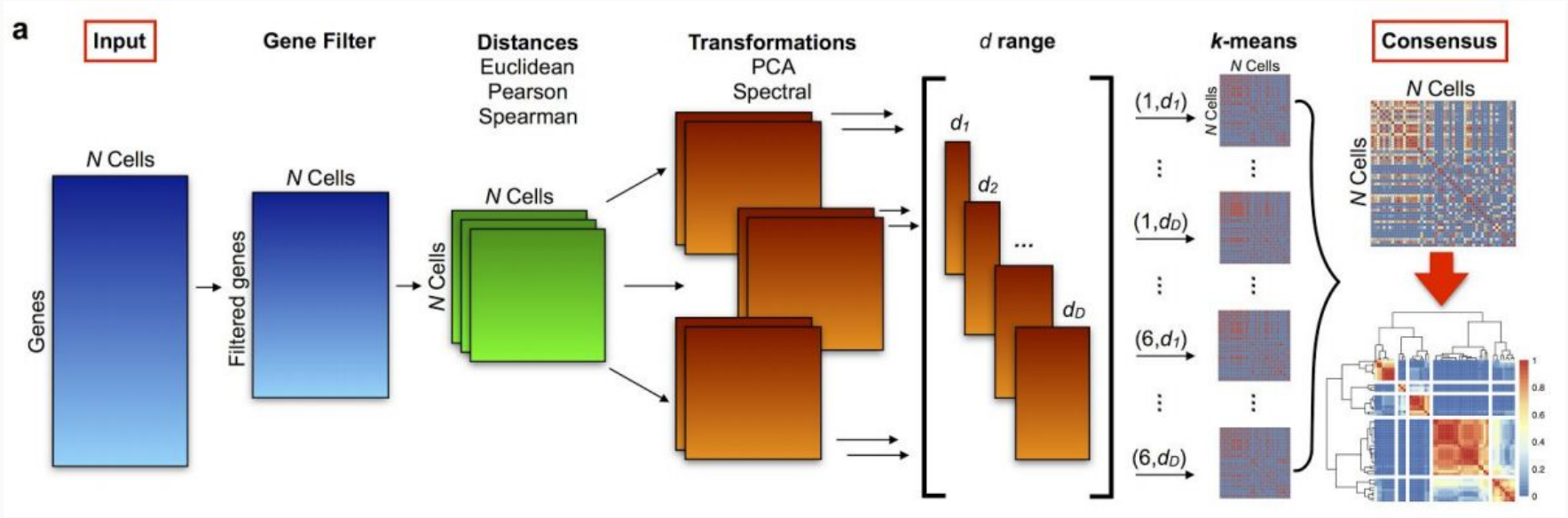
tSNE



Analysis: *de novo* Clustering

- Wish to identify clusters of cells using similarity of transcript abundance
- *Unsupervized clustering* methods
- *A priori* no knowledge of number of clusters
- Current methods:
  - PCA/Spectral analysis
  - Hierarchical or k-means clustering
  - Graph based

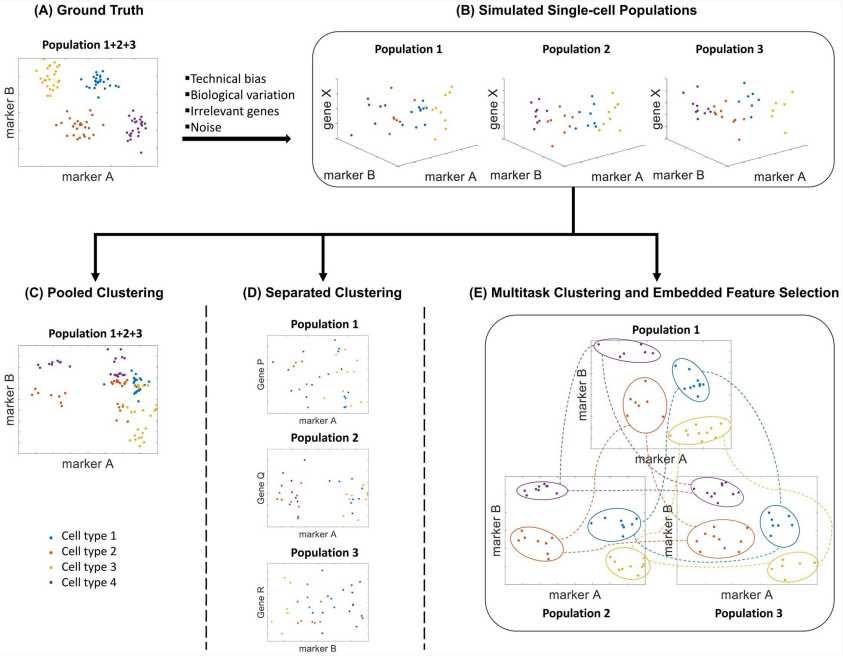
Analysis: *de novo* Clustering (SC3)



Kiselev, Vladimir Yu, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, et al. 2017. “SC3: Consensus Clustering of Single-Cell RNA-Seq Data.” *Nature Methods* 14 (5): 483–86.

Analysis: Cross-cell-population clustering

- How to compare cell clusters across samples?
- Each sample may have differences in:
  - Number of clusters
  - Cell type composition
  - Cell type behavior
- Extremely open problem
- This method (scVDMC) was published on Monday 4/9/2018



Zhang et al. 2018. “A Multitask Clustering Approach for Single-Cell RNA-Seq Analysis in Recessive Dystrophic Epidermolysis Bullosa.” *PLoS Computational Biology* 14 (4): e1006053.

Analysis: Differential Expression

- Goal: identify gene expression differences between cell types
- Current methods are similar to those used in bulk RNA-Seq
- *Dropouts:* genes with zero abundance due to technical limitations (sequencing depth)
- Extremely open problem

Other Analyses

- Feature Selection
  - Eliminating “noisy” genes
- Pseudotime Analysis
  - Inferring cellular development/change over time
- Imputation
  - Infer expression values for “dropout” genes

Current Software

- Seurat - http://satijalab.org/seurat/install.html
- SingleCellExperiment - bioconductor
- scater - bioconductor
- Many others now
- Millions of others soon...