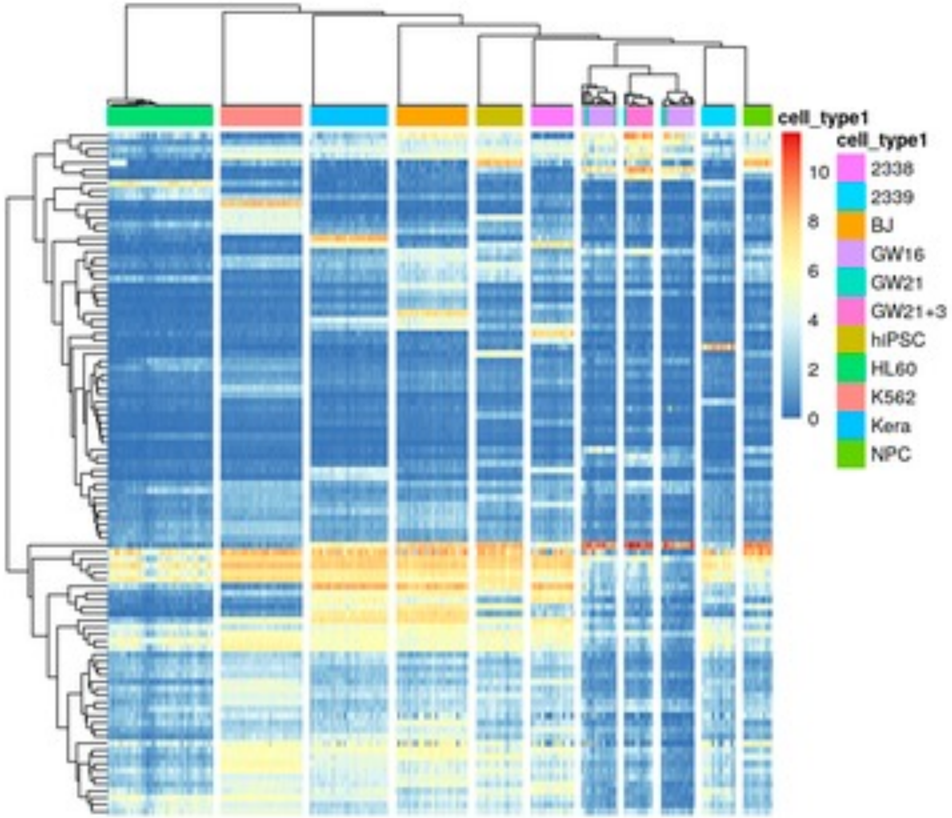
Analysis of Single-Cell RNA-Seq Data: Experimental Design

Orr Ashenberg, Jacques Serizay, Fabrício Almeida-Silva
November 2024

Physalia
Courses



Overall goals

Introduction to the rapidly expanding world of single-cell transcriptomics

Focus less on specific software tools but more on underlying concepts - so down the line, you can make informed choices

Hands on lab exercises analyzing single-cell heterogeneity

Create a fun, learning, collaborative, and interactive environment over the next week

A few organizational notes

https://jserizay.com/scRNAseq_Physalia_2024/

Ask questions or use chat! Raise hand in Zoom (Participants).

Please use video, and mute microphone when not in use.

Write course notes and questions in a shared Google document.

We will send you a link to store your flash talk.

Please be patient with technical issues (network, Zoom, etc…)

Mouse organogenesis studied by single-cell RNA sequencing

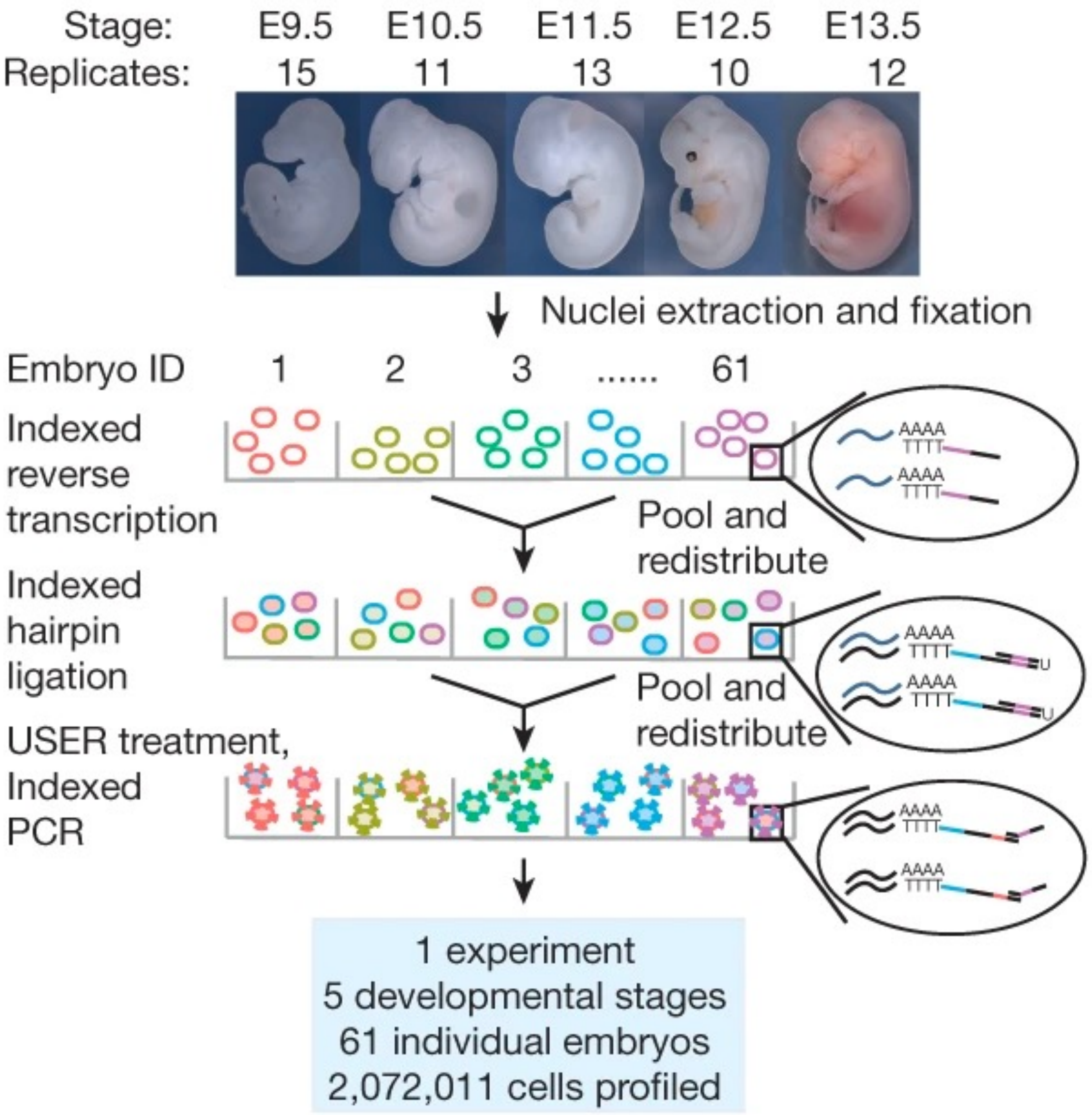
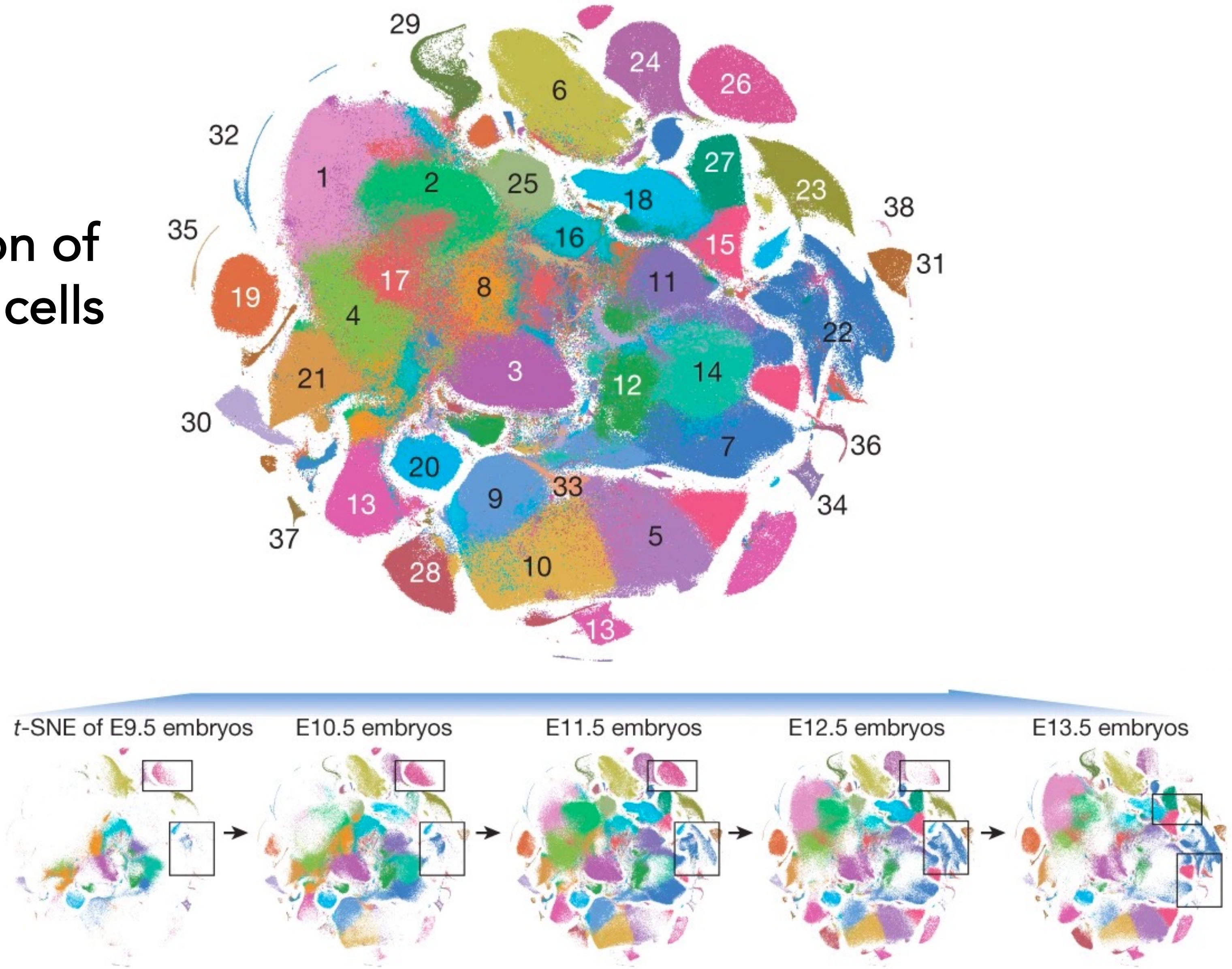


Fig. 1: sci-RNA-seq3 enables profiling of 2,072,011 cells from 61 mouse embryos across 5 developmental stages in a single experiment.

Cao, J., at al. *Nature* 2019

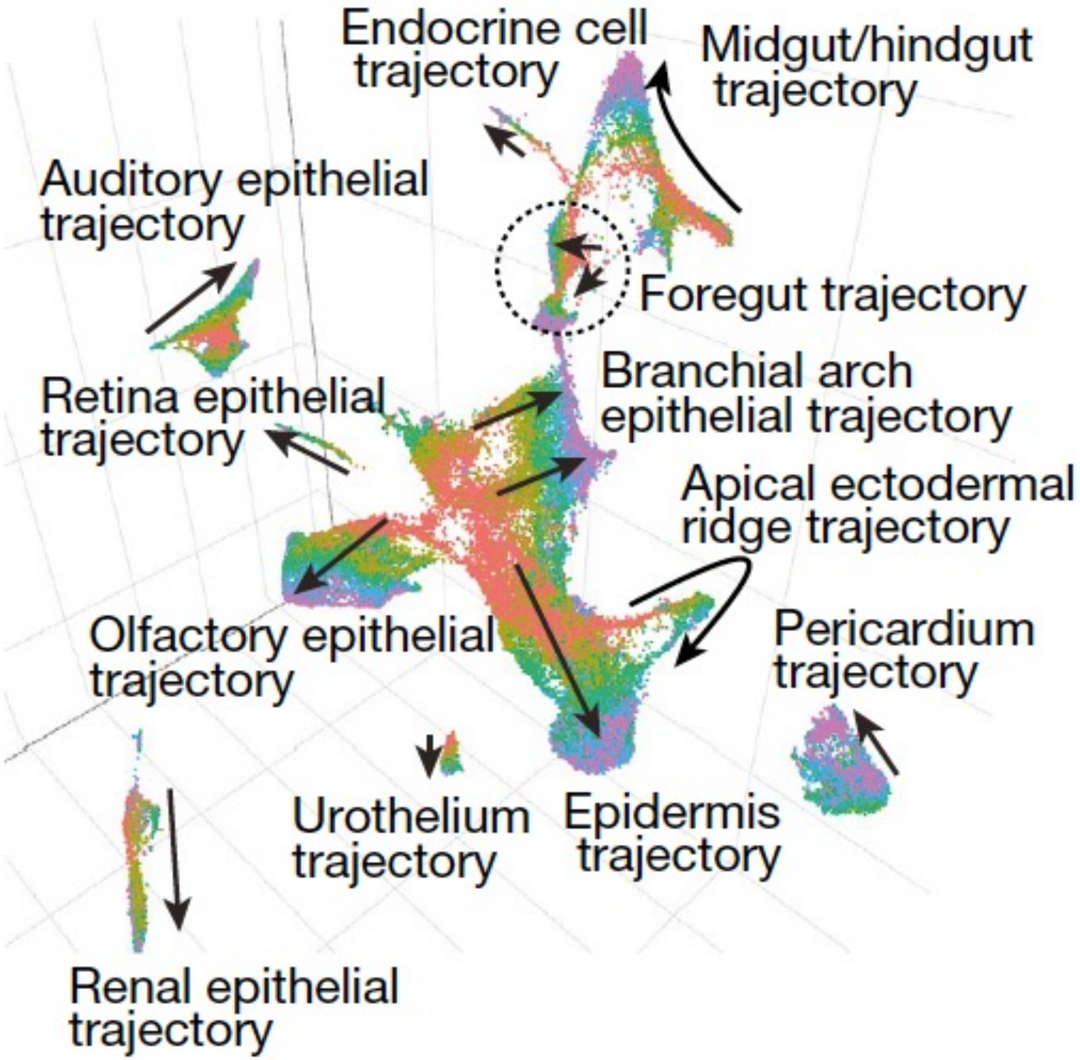
Mouse organogenesis studied by single-cell RNA sequencing

Clustering and visualization of 2,026,641 mouse embryo cells



Cao, J., at al. *Nature* 2019

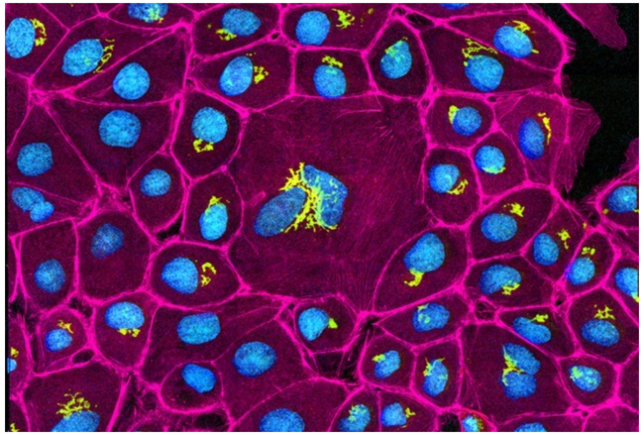
Mouse organogenesis studied by single-cell RNA sequencing



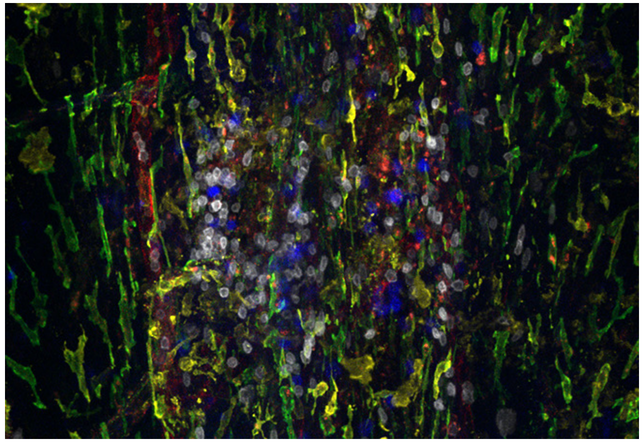
Cao, J., at al. *Nature* 2019
https://www.youtube.com/watch?v=cWG2CkkDcrM

Incredible diversity in cell types, states, and interactions across human tissues

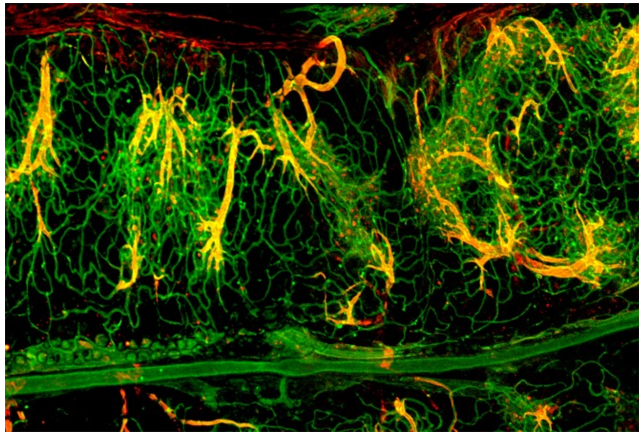
Skin epithelium



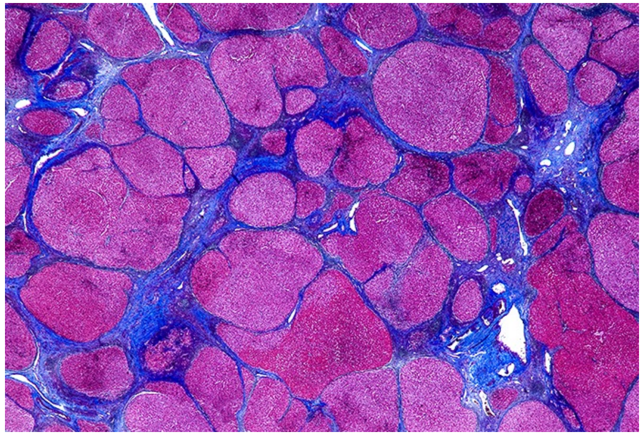
Brain meninges



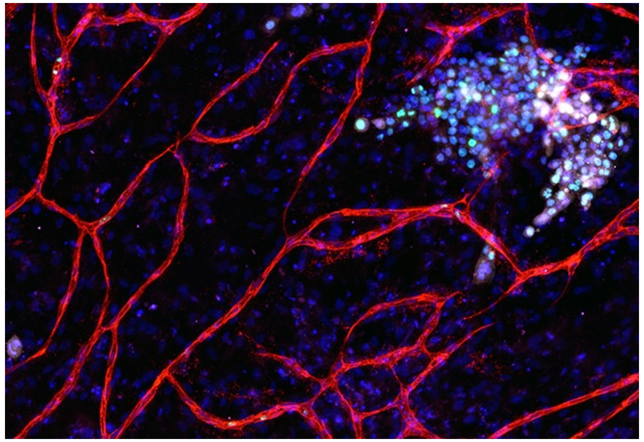
Blood vessels



Small intestine



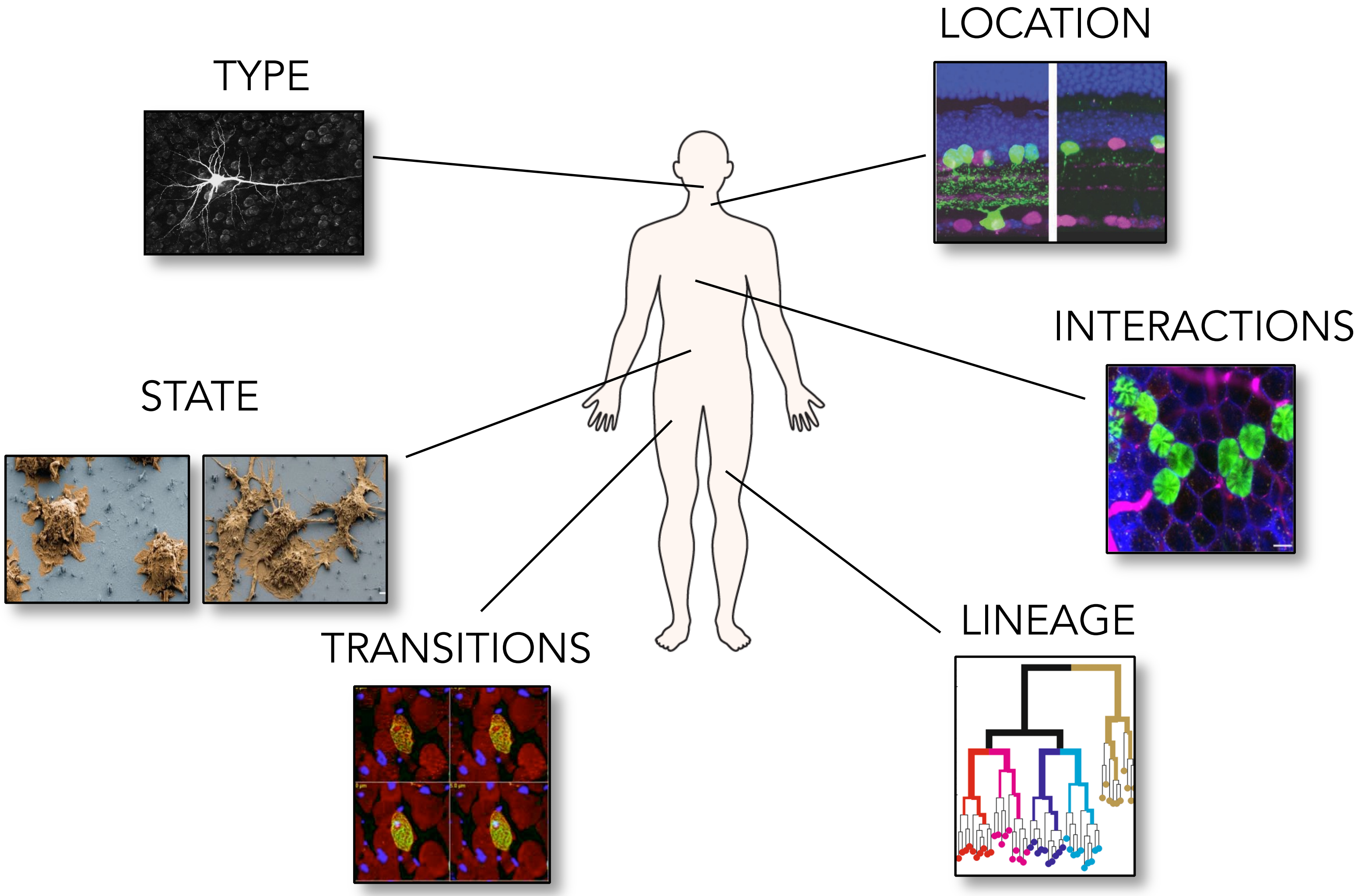
Liver cirrhosis



Breast cancer

http://www.cell.com/pictureshow/skin
https://library.med.utah.edu/WebPath/webpath.html

A cell's identity and fate are shaped by many features



TYPE

LOCATION

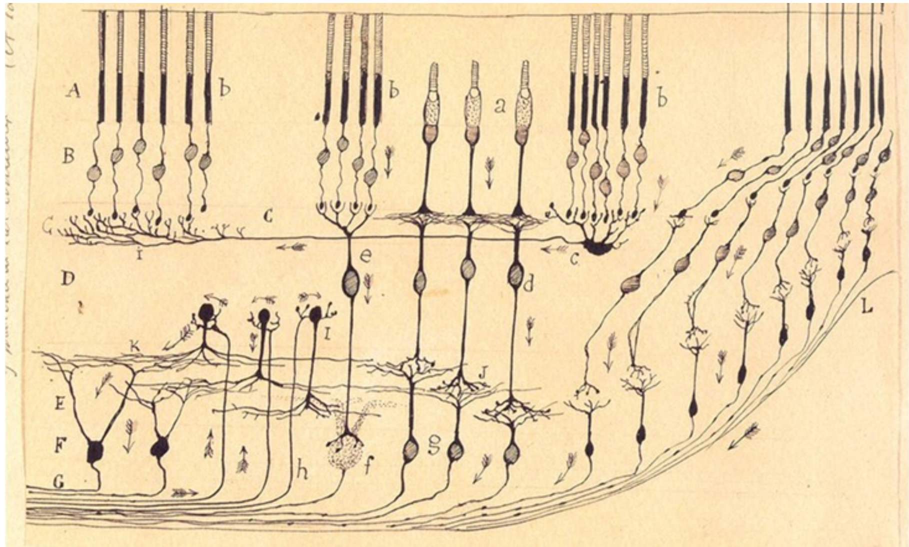
INTERACTIONS

STATE

TRANSITIONS

LINEAGE

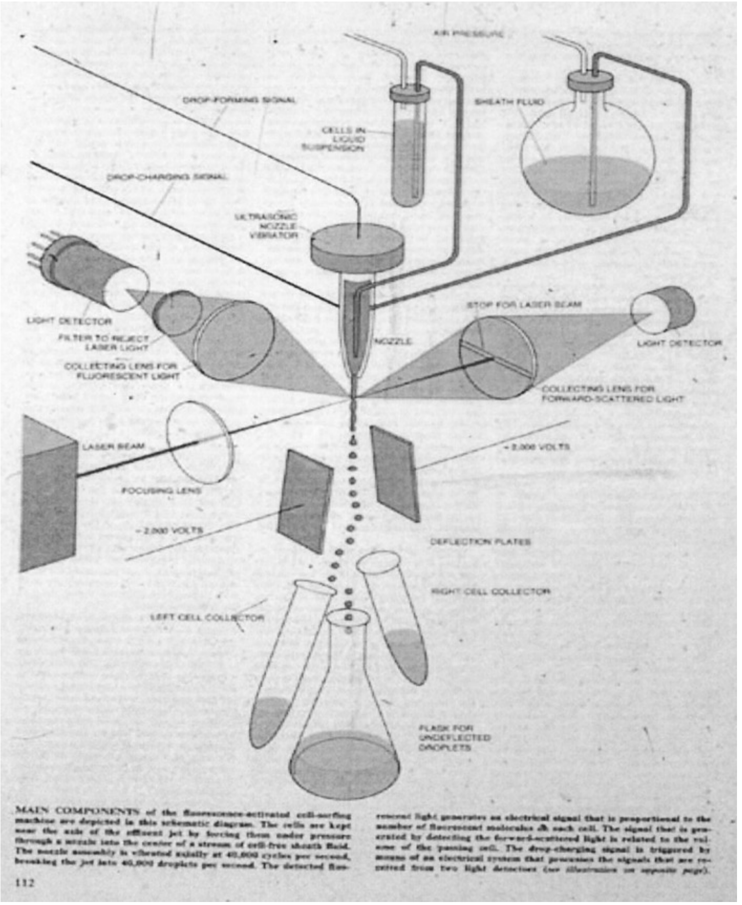
Technological innovations allow observation at increasing resolution



"by the help of Microscopes, there is nothing so small as to escape our inquiry"
Robert Hooke

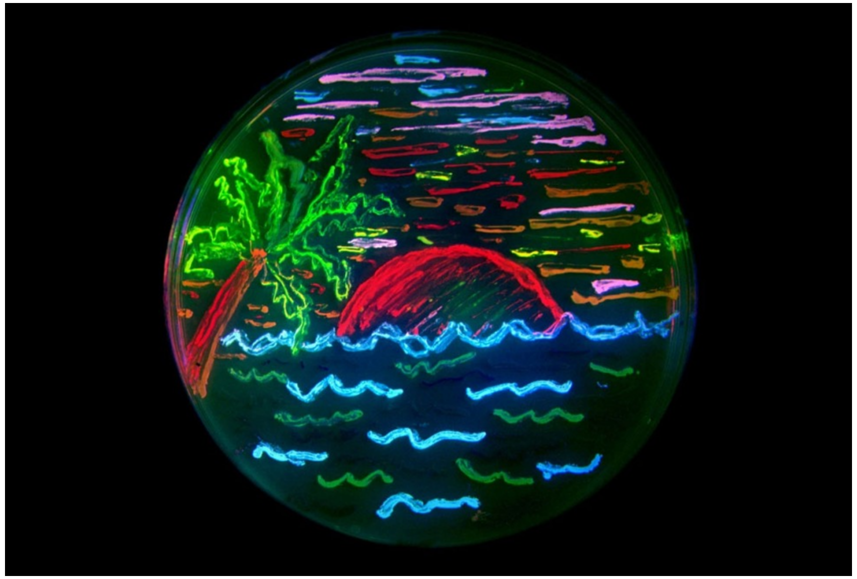
www.microscopehistory.com
nikoninstruments.com
http://braintour.harvard.edu

Technological innovations allow observation at increasing resolution



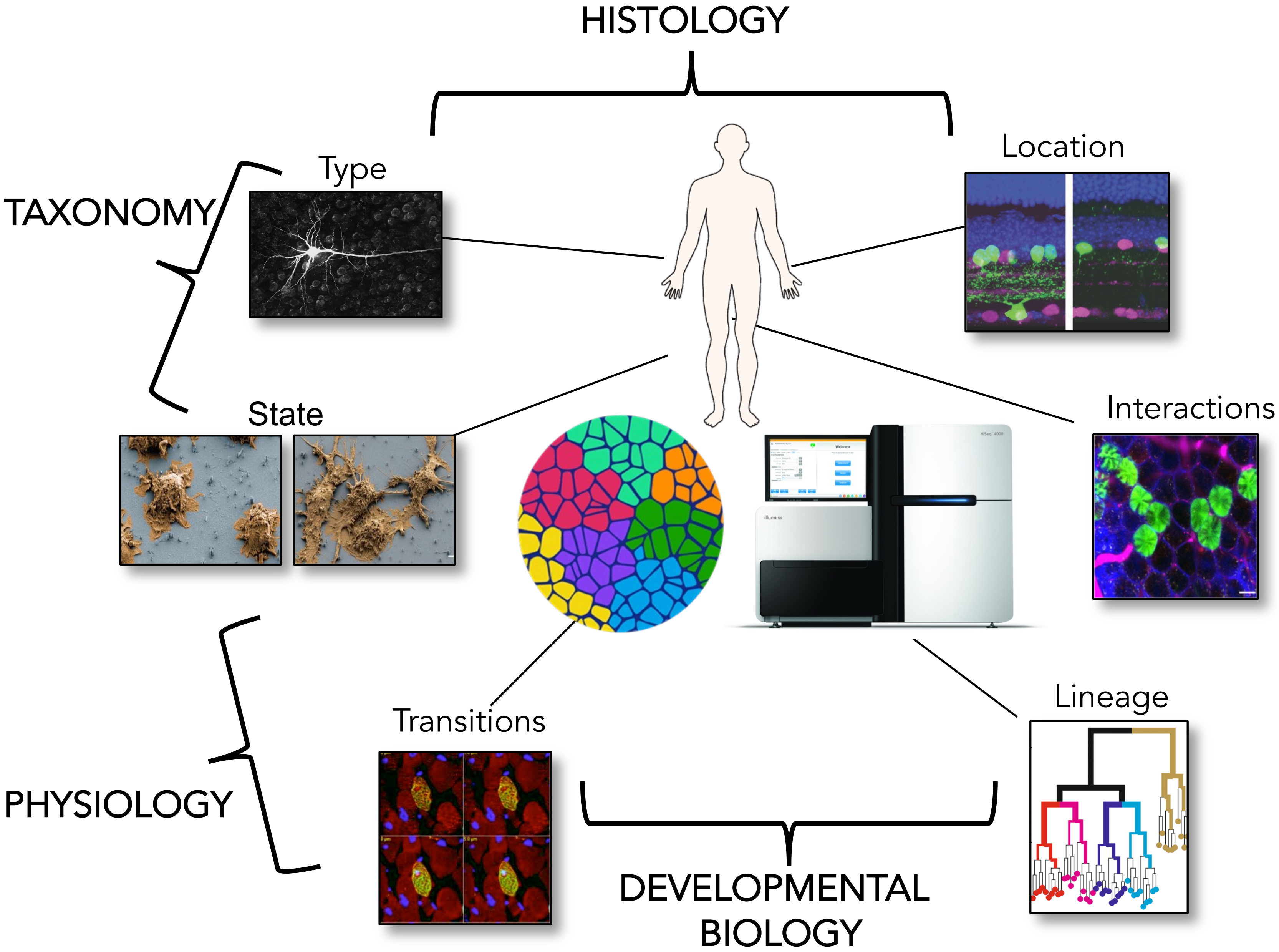
MAIN COMPONENTS of the fluorescence-activated cell-sorting machine are depicted in this schematic diagram. The cells are kept near the axis of the effluent jet by forcing them under pressure through a nozzle into the center of a stream of cell-free sheath fluid. The nozzle assembly is vibrated axially at 40,000 cycles per second, breaking the jet into 40,000 droplets per second. The detected fluorescent light generates an electrical signal that is proportional to the number of fluorescent molecules on each cell. The signal that is generated by detecting the forward-scattered light is related to the volume of the passing cell. The drop-charging signal is triggered by means of an electrical system that processes the signals that are received from two light detectors *(see illustration on opposite page)*.

112



Herzenberg LA, at al. *Sci Am* 1976;**234**:108-117.
Nature doi:10.1038/nature.2016.20532

Single-cell genomics makes possible high-resolution characterization of cells



HISTOLOGY

TAXONOMY

Type

Location

State

Interactions

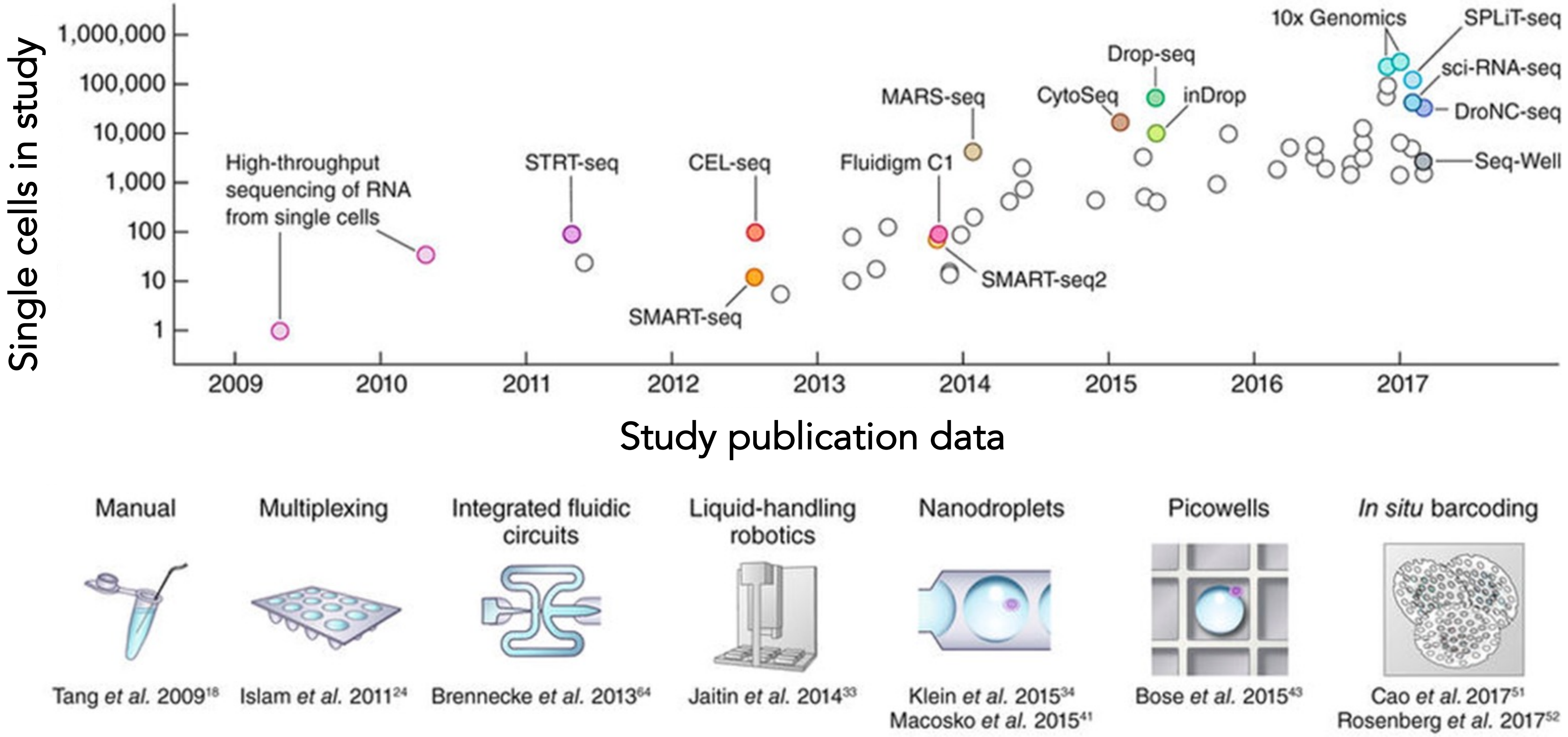
Transitions

Lineage

PHYSIOLOGY

DEVELOPMENTAL BIOLOGY

Single-cell RNA sequencing has grown exponentially



Svensson V et al. (2018) *Nature Protocols* 13:599–604.

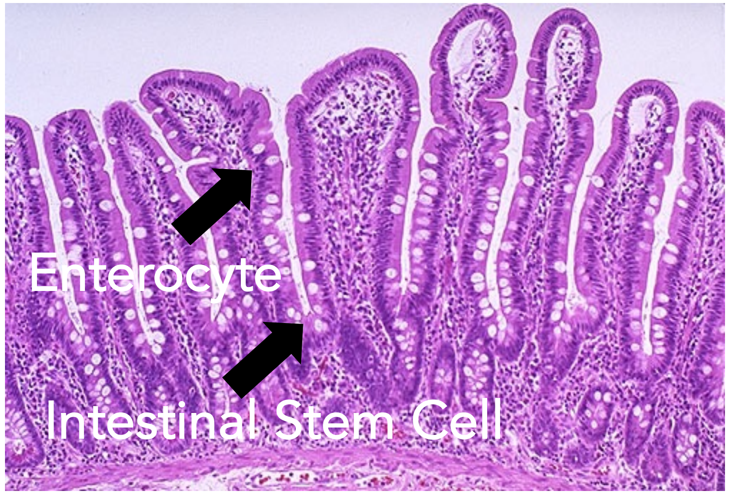
Human Cell Atlas mission

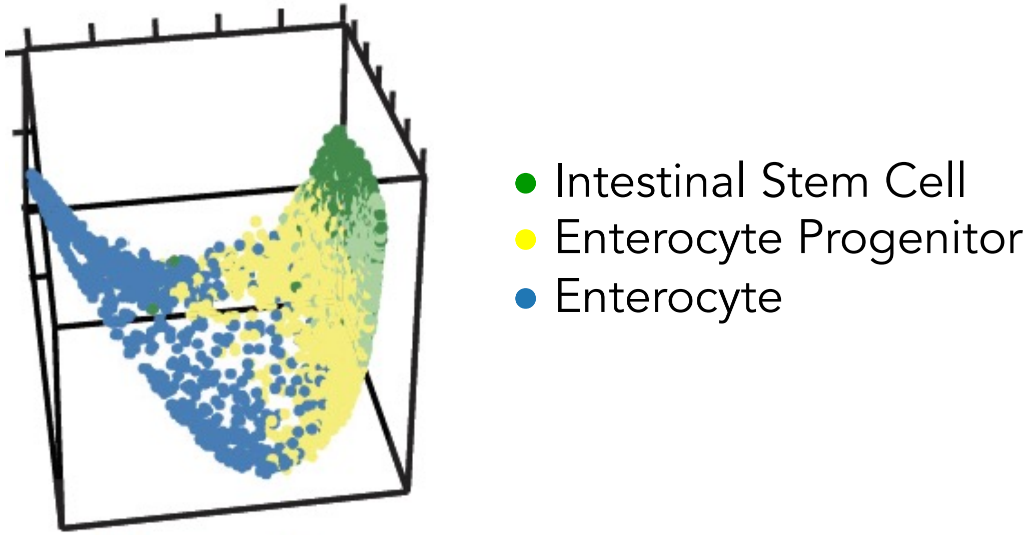
To create a comprehensive reference map of the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease.

"The vestiges of the rupture reveal themselves, if someone brings forward a map of the world and considers carefully the coasts of the three [continents]."
Dutch map maker Abraham Ortelius (1596)

Human Cell Atlas mission

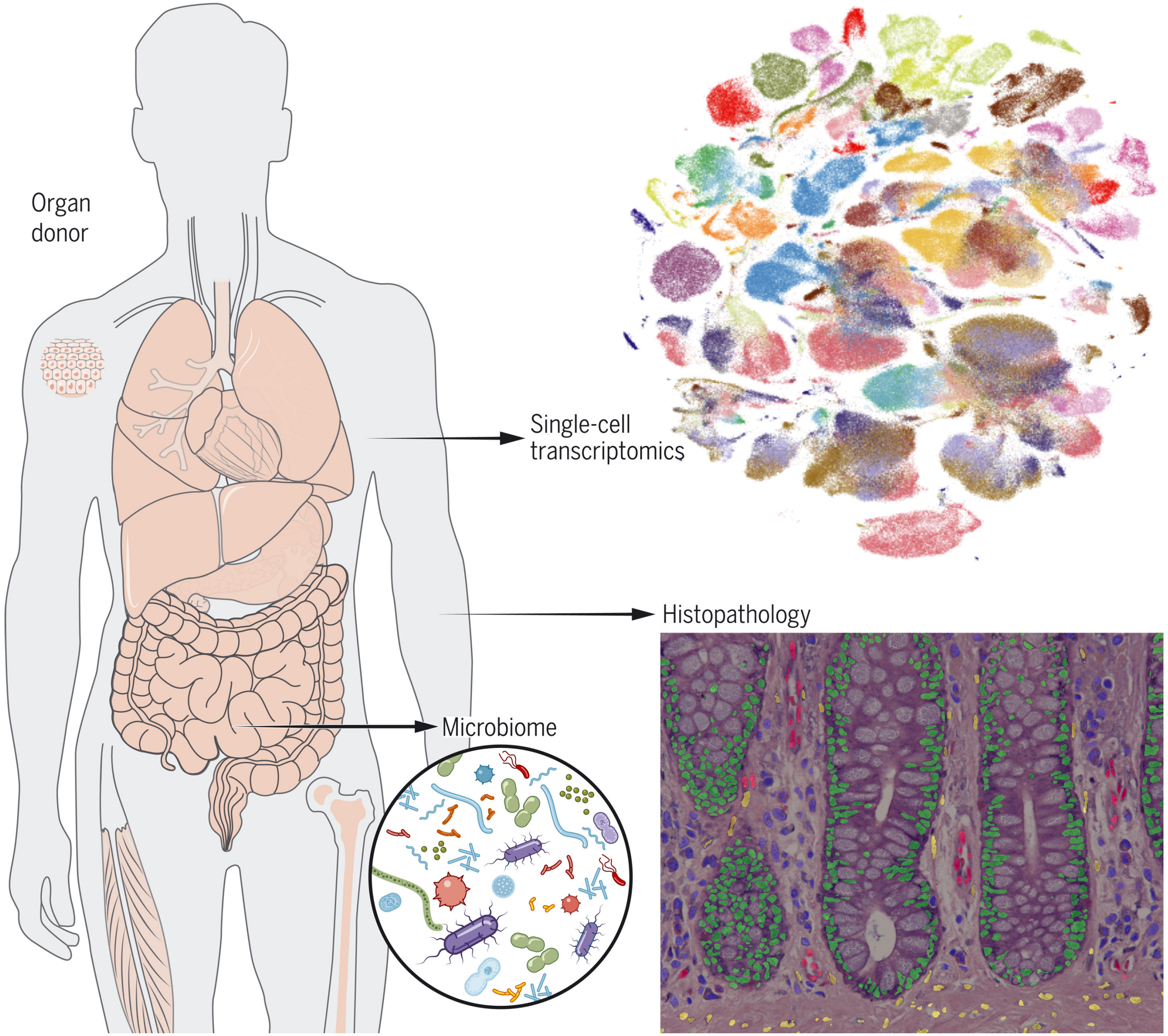
To create a comprehensive reference map of the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease.





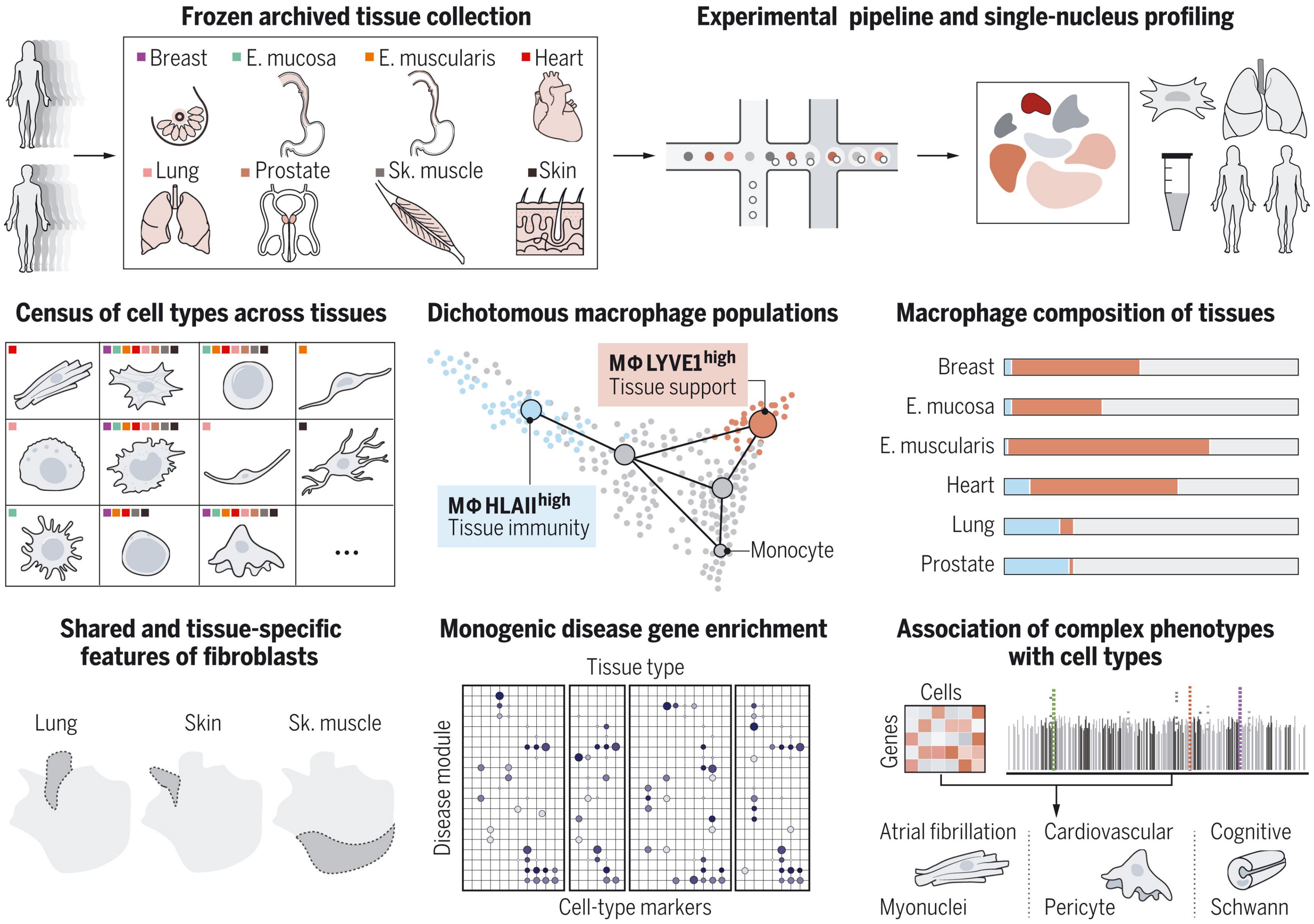
Haber, A.L., Biton, M., Rogel, N. et al. (2017) *Nature* 551: 333-339.

A series of human cell atlases published in May 2022: 500,000 cells from 24 different tissues and organs



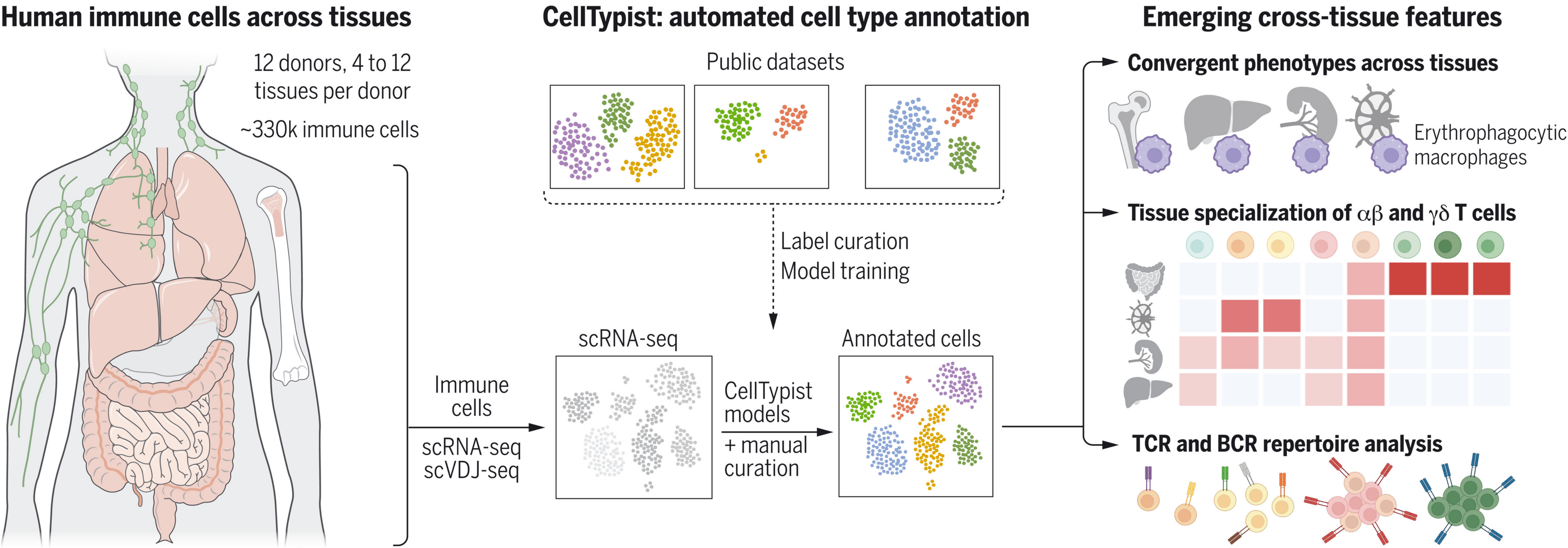
The Tabula Sapiens Consortium. *Science* 2022.

A series of human cell atlases published in May 2022: Association of cell types and states with human disease



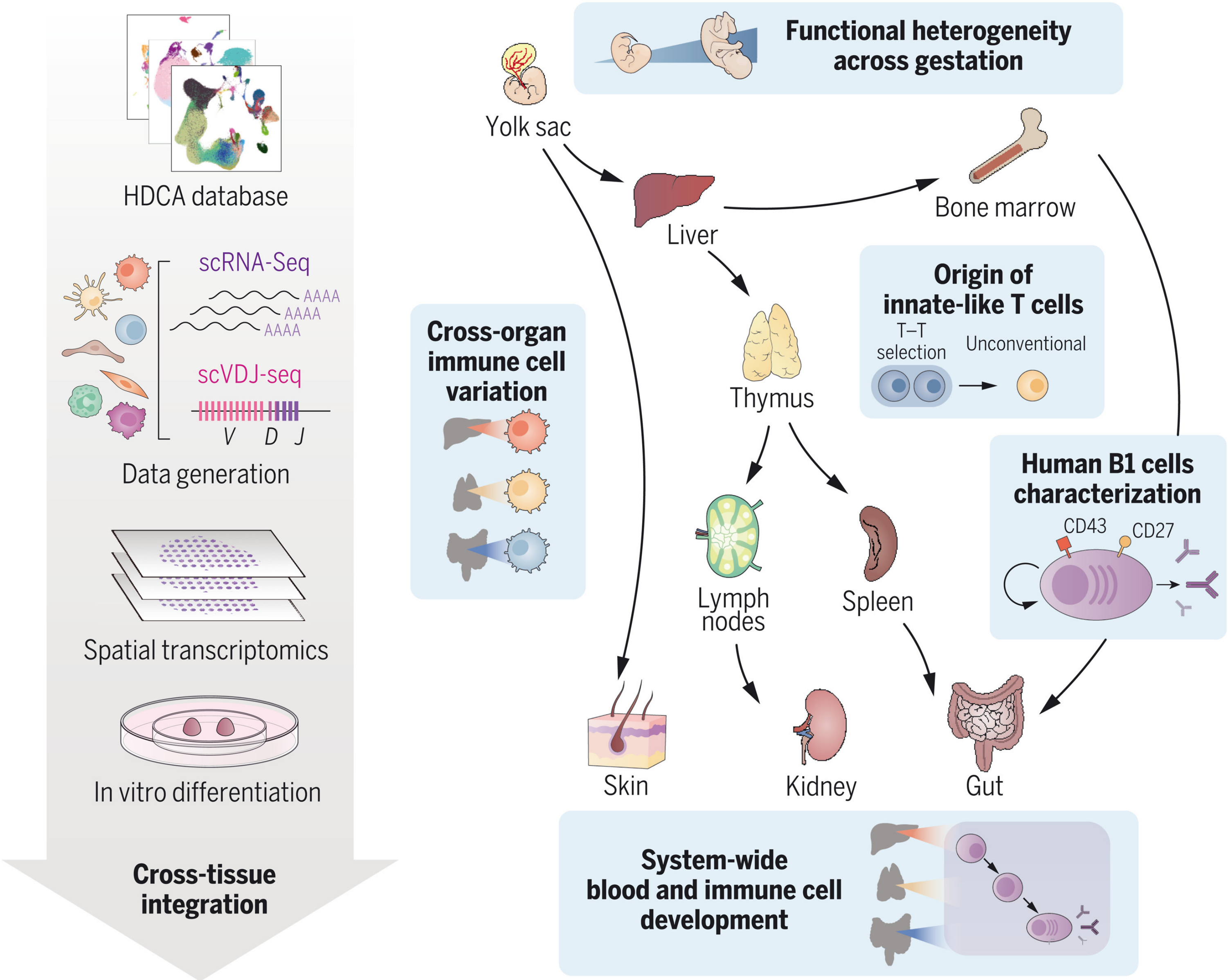
Eraslan, G., at al. *Science* 2022.

A series of human cell atlases published in May 2022: 300,000 immune cells from 16 different tissues



Domínguez Conde, C., at al. *Science* 2022.

A series of human cell atlases published in May 2022: Developing human immune system



Suo, C., at al. *Science* 2022.

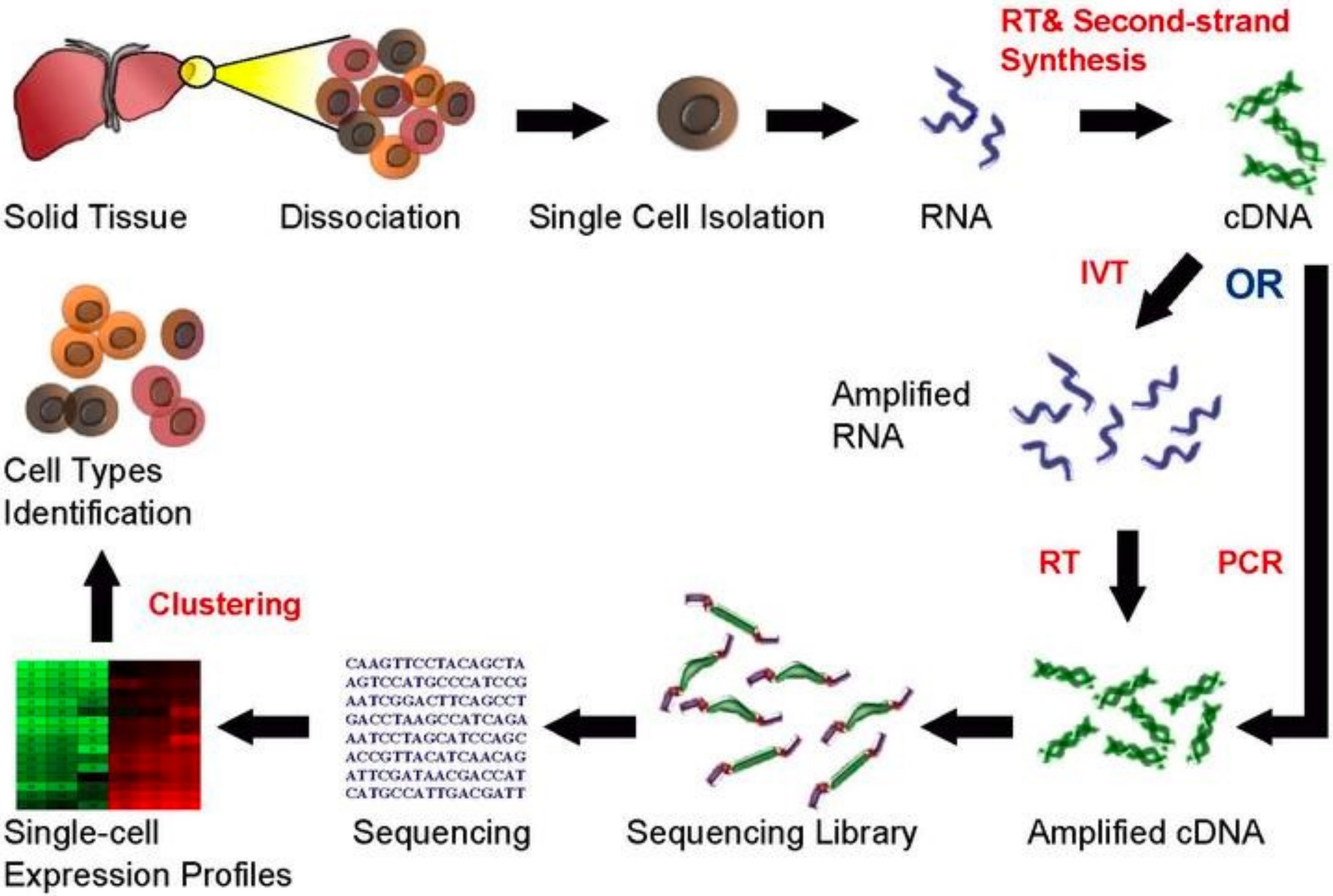
Typical RNA-Seq workflow

mRNA from a source

Slide courtesy of Karthik Shekhar

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

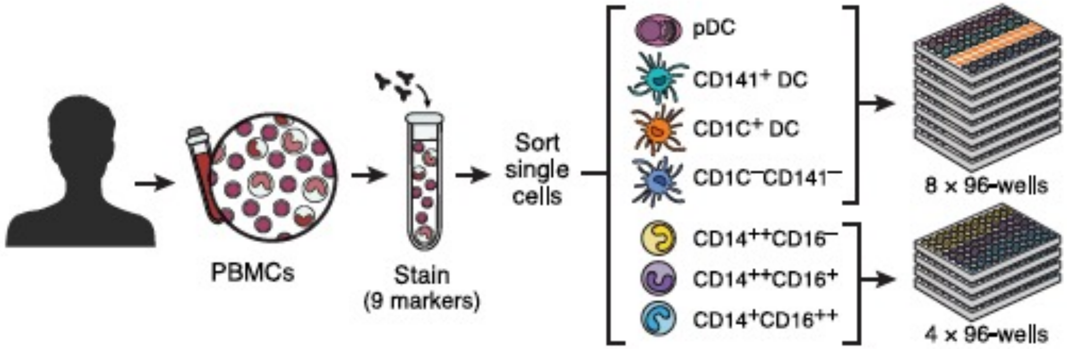
Experimental design: single cell RNA-Seq



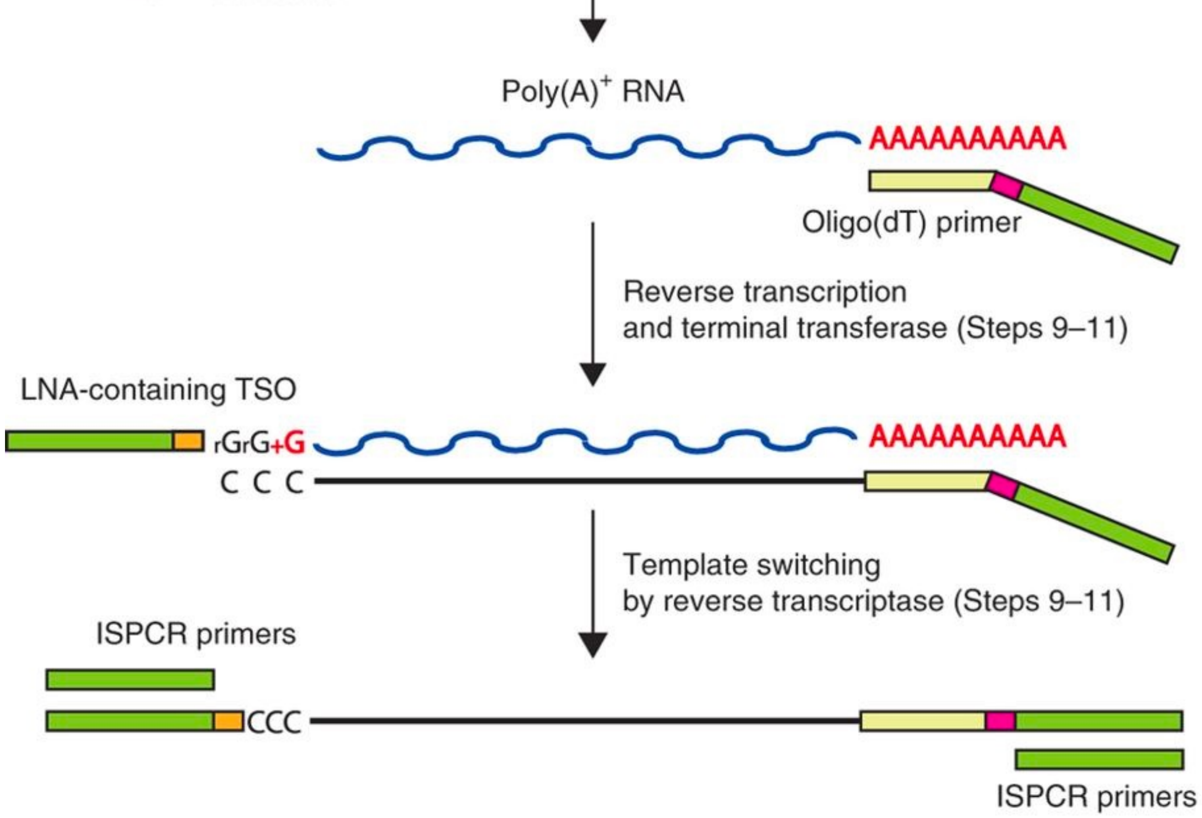
https://en.wikipedia.org/wiki/Single_cell_sequencing

Single cell transcriptomics using SMART-Seq2

FACS sorting on 96/384-well plates



Cell lysis (Steps 1–8)



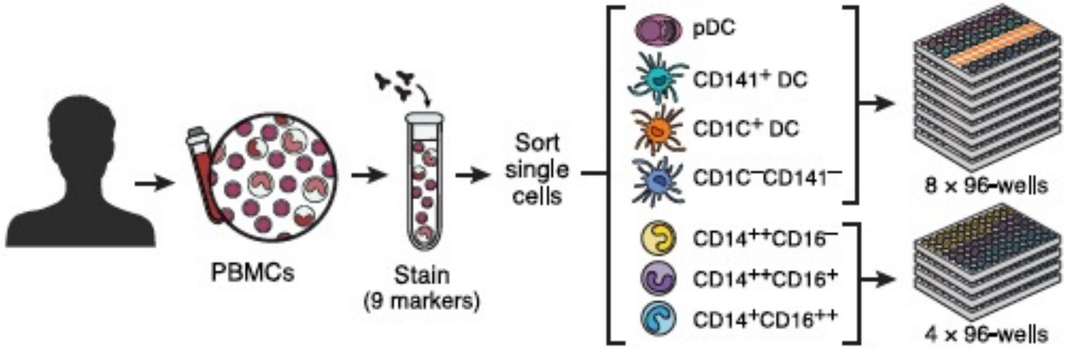
Reverse transcription
and terminal transferase (Steps 9–11)

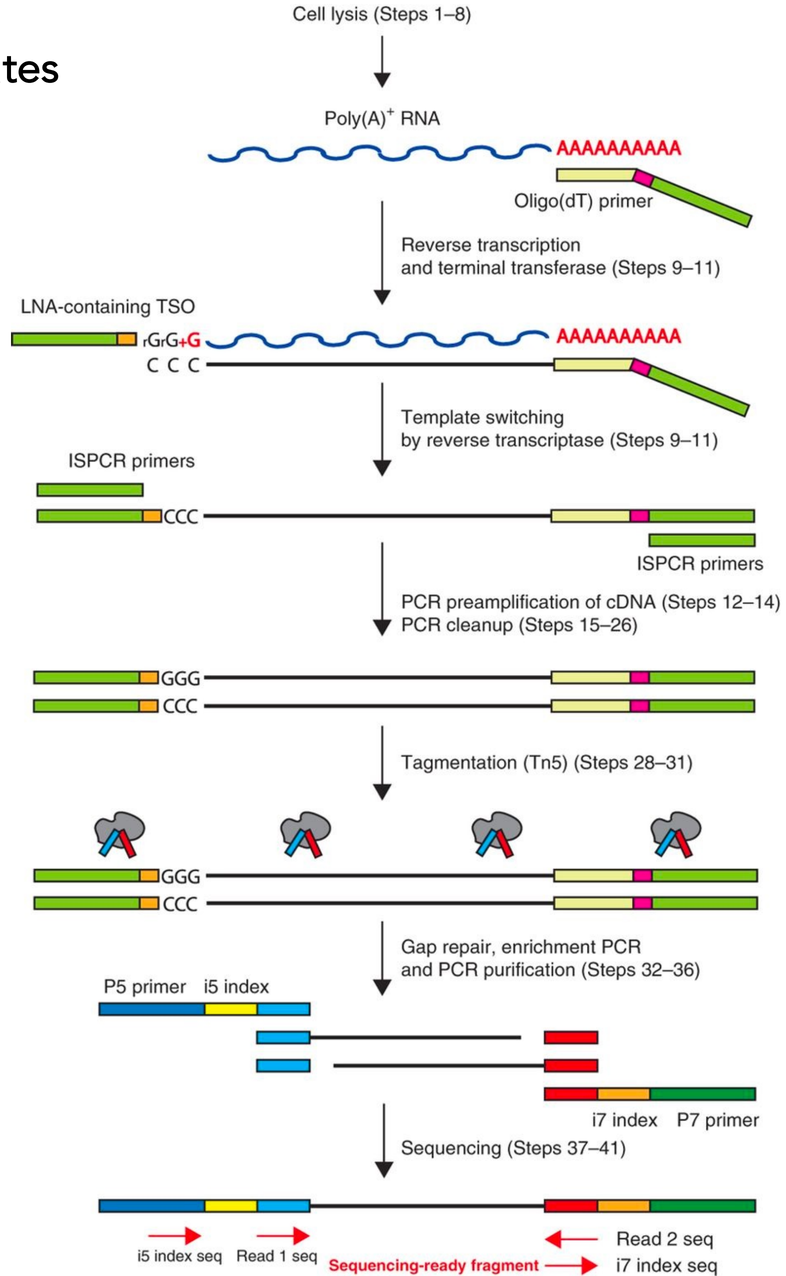
Template switching
by reverse transcriptase (Steps 9–11)

Picelli S. et *al*. (2014) *Nature Protocols*. 9, 171-181.
Villani A. et *al*. (2017) *Science*. 356(283), 1-12.

Single cell transcriptomics using SMART-Seq2

FACS sorting on 96/384-well plates

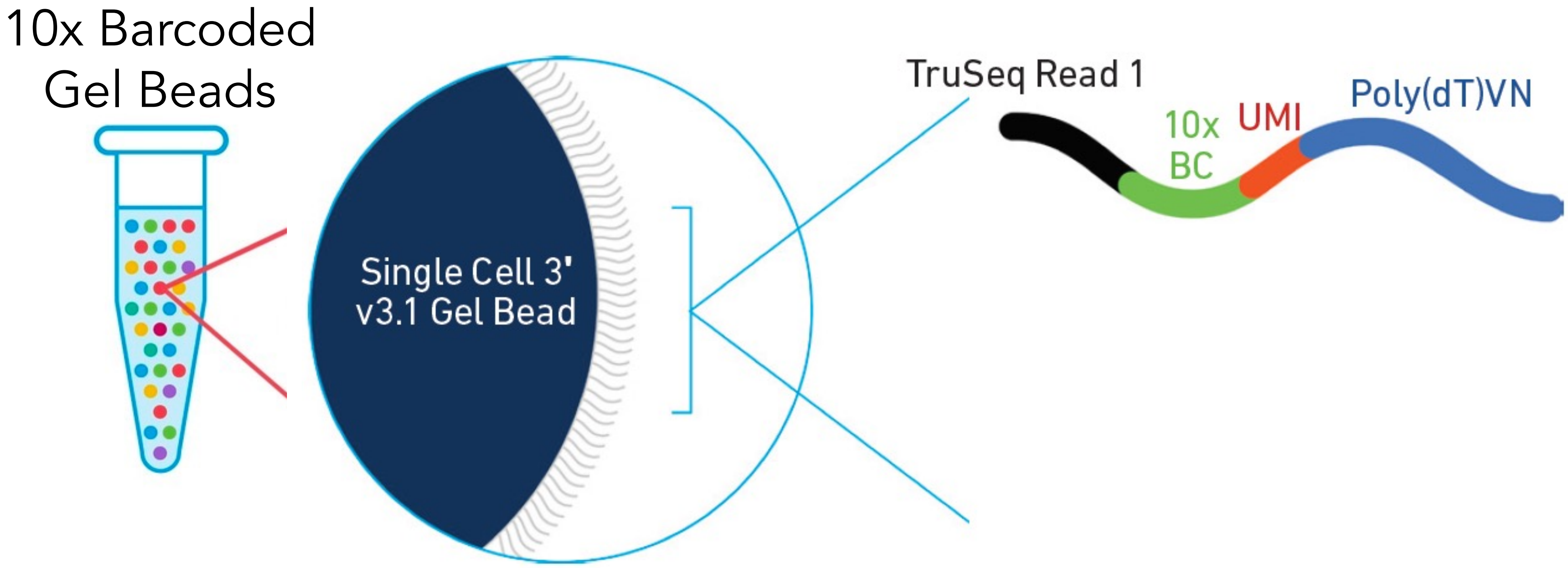




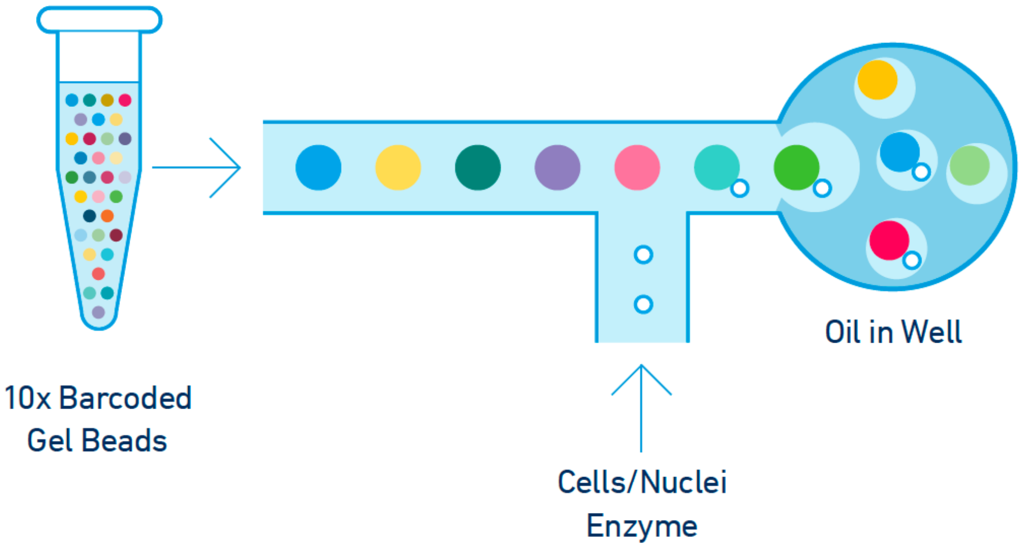
Picelli S. et *al*. (2014) *Nature Protocols*. 9, 171-181.
Villani A. et *al*. (2017) *Science*. 356(283), 1-12.

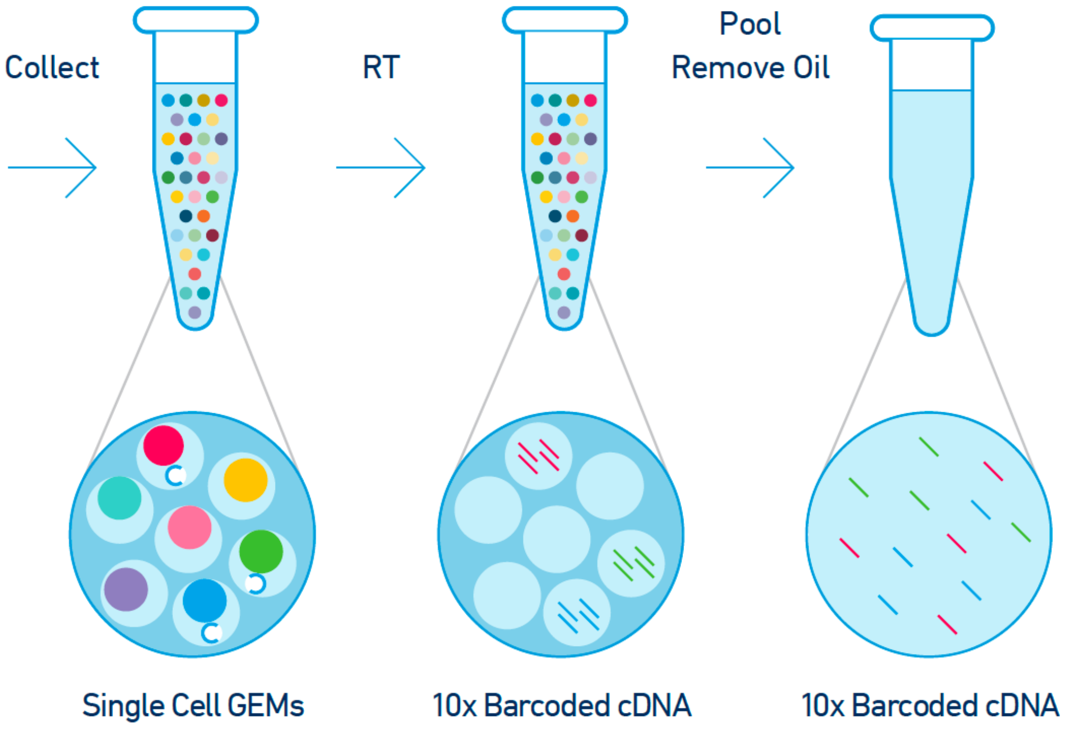
Single cell transcriptomics using droplets and microfluidics

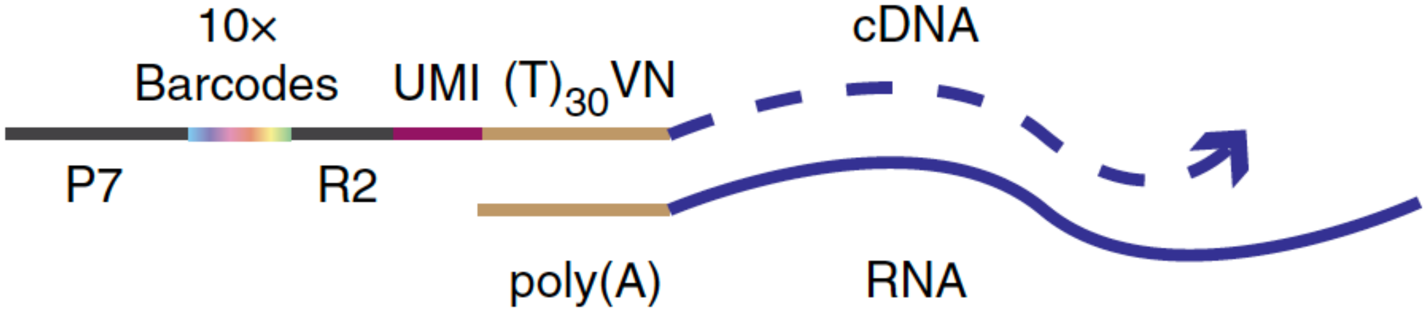
10x Next GEM samples a pool of ~3,500,000 10x Barcodes to separately index each cell's transcriptome



Single cell transcriptomics using 10x Chromium system





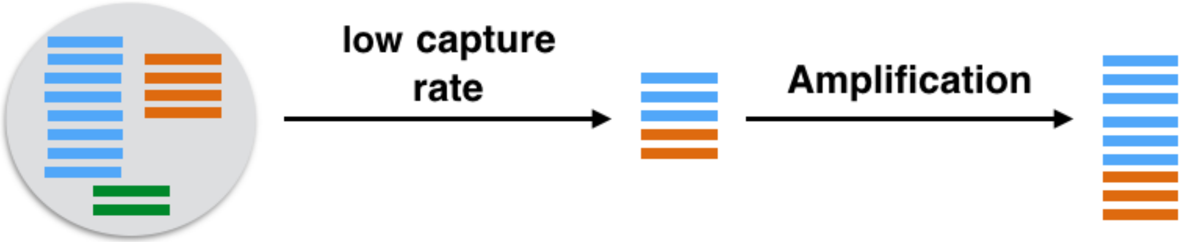


Visit https://liorpachter.wordpress.com/2019/02/07/sub-poisson-loading-for-single-cell-rna-seq/

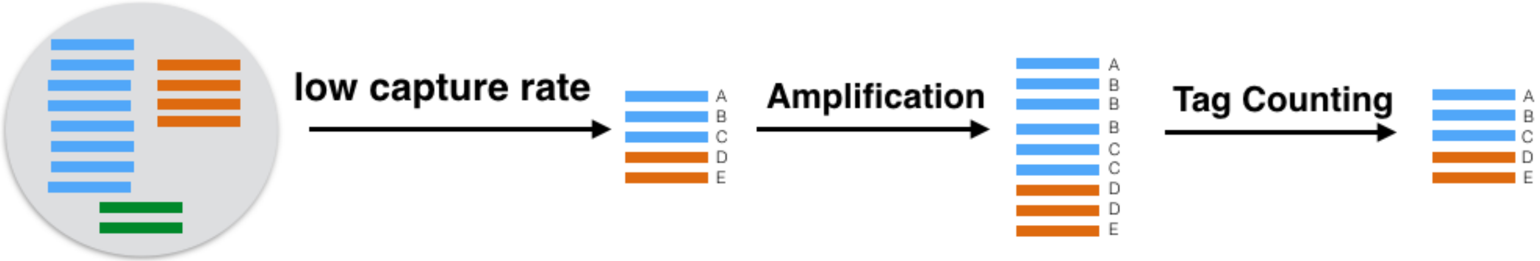
UMI : Unique Molecular Identifiers (Random Molecular Tags)

Early labeling of mRNA molecules with random nucleotide tags enables amplification biases to be corrected

- Low input amount -> transcript dropout + PCR amplification bias



- Unique Molecular Identifiers (UMIs) can correct for PCR bias



Remember : UMIs do not correct for low-capture rates, which leads to an abundance of false negatives. Capture rates are estimated to 5-20% across various protocols

Slide courtesy of Karthik Shekhar

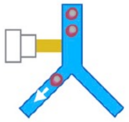
Single-cell RNA-Seq pipeline


Cell capture and Barcoding

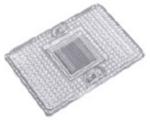
Library Prep

Sequencing

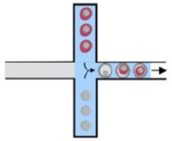
FACS



Micro-capture

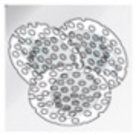


Droplet



Nanowell

Split/pool barcoding



PolyA vs random priming

3′/5′ end tagging vs full-length

PCR vs *in vitro* transcription

…

Miseq

~ 20M reads total

Nextseq

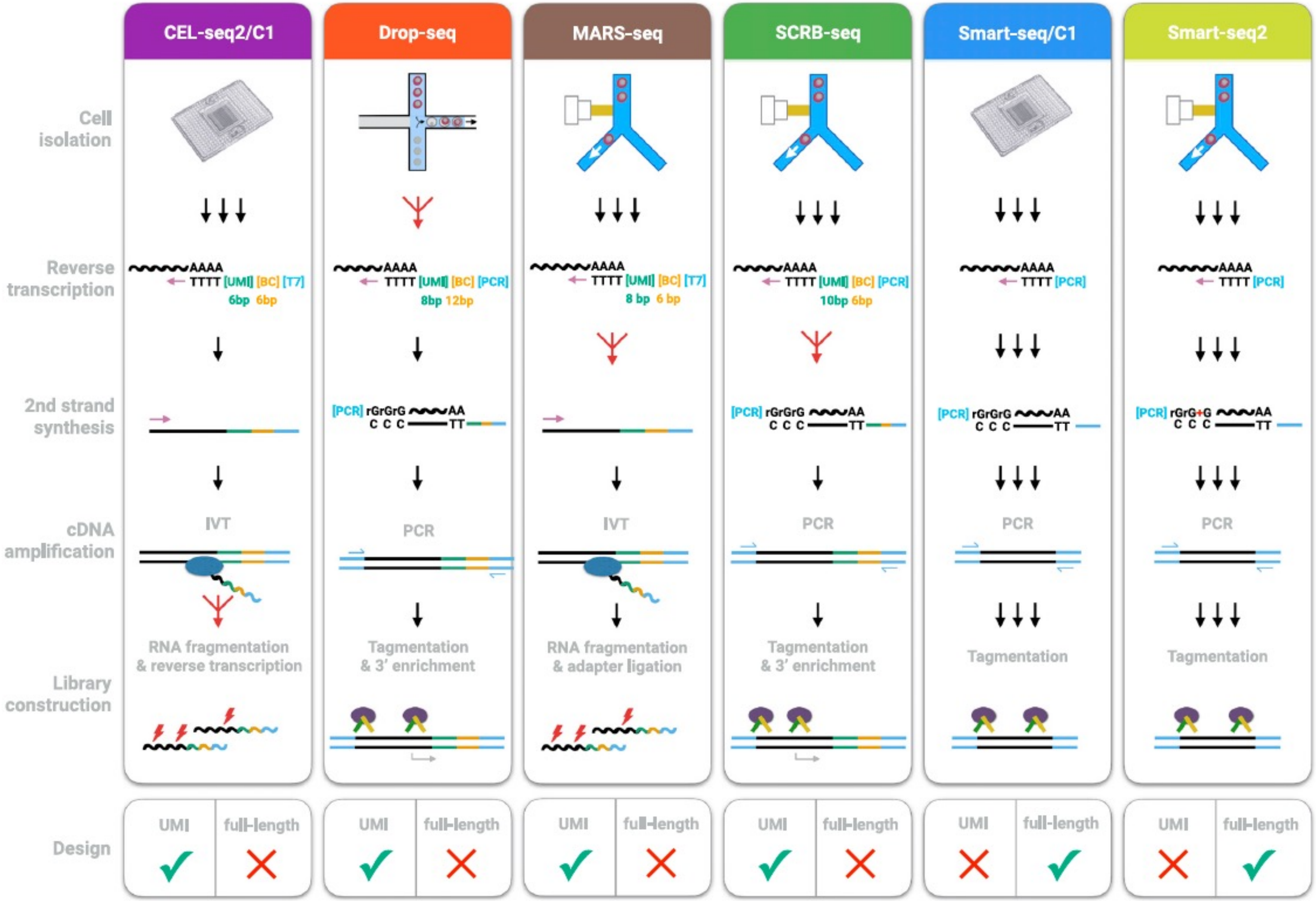
~ 500M reads total

HiSeq 4000

4 billion reads

Slide courtesy of Karthik Shekhar

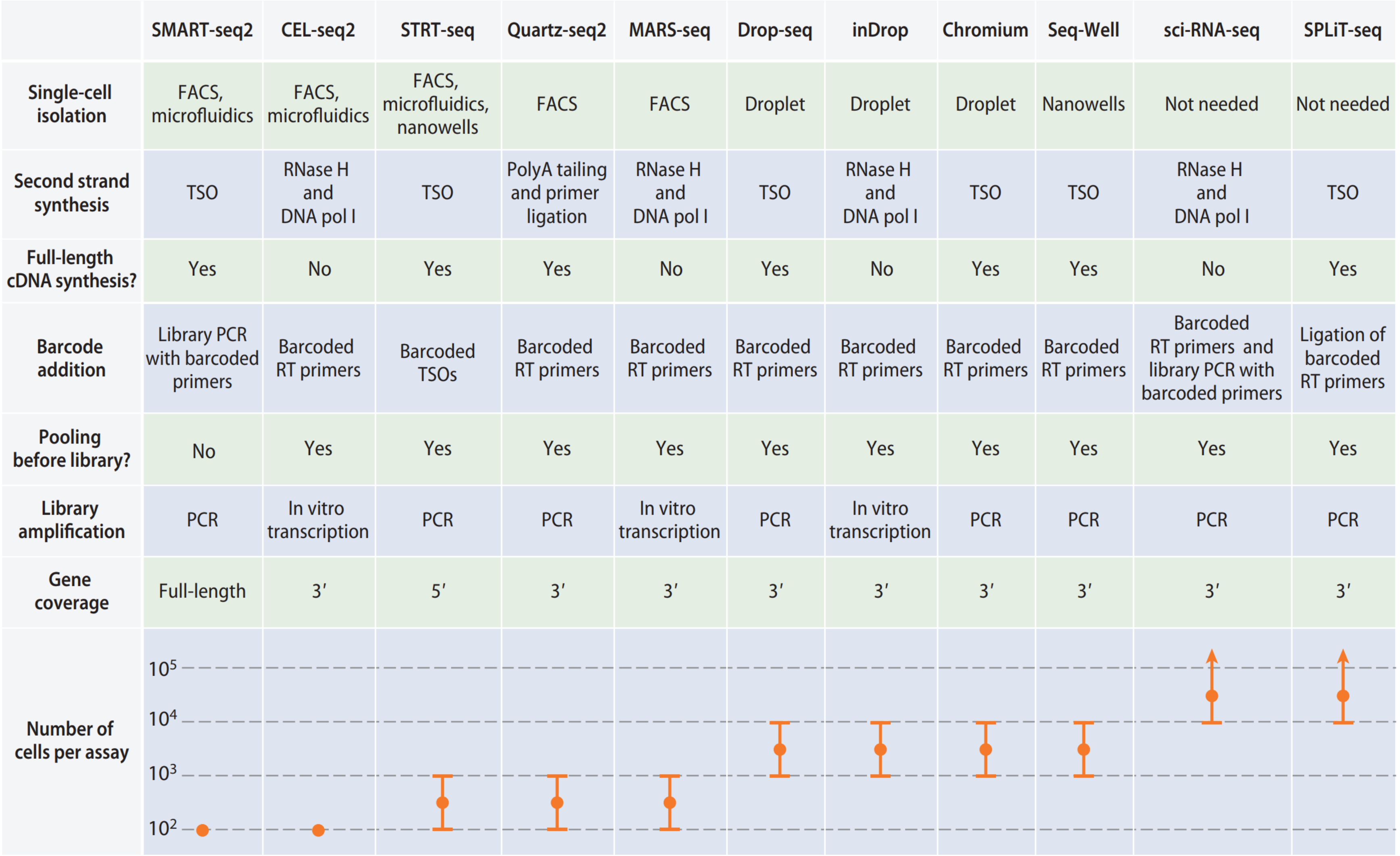
There are many single-cell RNA sequencing methods



	CEL-seq2/C1	Drop-seq	MARS-seq	SCRB-seq	Smart-seq/C1	Smart-seq2
Cell isolation						
Reverse transcription	AAAA TTTT [UMI] [BC] [T7] 6bp 6bp	AAAA TTTT [UMI] [BC] [PCR] 8bp 12bp	AAAA TTTT [UMI] [BC] [T7] 8 bp 6 bp	AAAA TTTT [UMI] [BC] [PCR] 10bp 6bp	AAAA TTTT [PCR]	AAAA TTTT [PCR]
2nd strand synthesis		[PCR] rGrGrG AA C C C TT		[PCR] rGrGrG AA C C C TT	[PCR] rGrGrG AA C C C TT	[PCR] rGrG+G AA C C C TT
cDNA amplification	IVT	PCR	IVT	PCR	PCR	PCR
Library construction	RNA fragmentation & reverse transcription	Tagmentation & 3' enrichment	RNA fragmentation & adapter ligation	Tagmentation & 3' enrichment	Tagmentation	Tagmentation
Design: UMI	✓	✓	✓	✓	✗	✗
Design: full-length	✗	✗	✗	✗	✓	✓

Ziegenhain C et *al.* (2017) *Mol Cell.* 65(4):631-643.

There are many single-cell RNA sequencing methods



	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											

Each protocol has **advantages** and **limitations**. What one ends up using is often dictated by multiple features - the **biological context, cost, objective** etc.

Chen X. et *al.* (2018) *Annual Review of Biomedical Data Science* Vol. 1:29-51

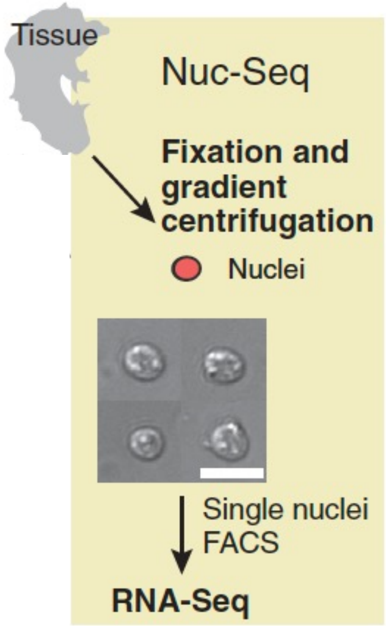
Considerations for single cell RNA-Seq

Choose protocol based on :

- Throughput (# of cells / reaction)
- Sample of origin
- Cost / Labor / Time limitations
- Gene body coverage - 5′, 3′ biased, or full-length?
- UMI vs no-UMI
- Sequencing depth / cell

For example :

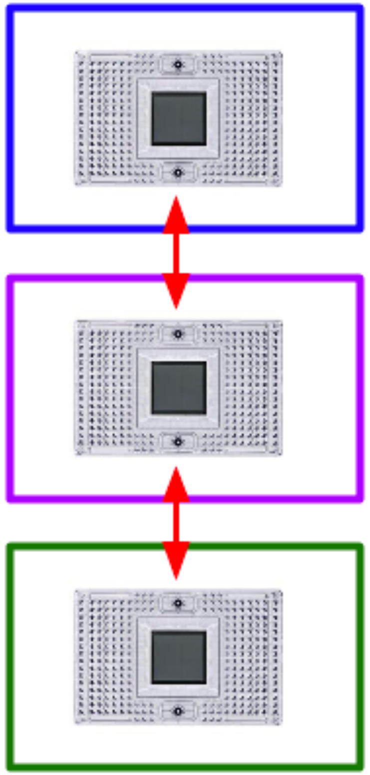
- If I want to classify all cell types in a diverse tissue (e.g. brain), I need high throughput
- If I want to re-annotate the transcriptome and discover new isoforms, I need full-length coverage
- If I only have access to archival human samples, I will need to use a method that permits fixed cells (or nuclei)



Slide modified from Karthik Shekhar
Habib N. et *al*. (2016) *Science*. 353, 925-928.

The most powerful way to control batch effects is with careful experimental design

Completely Confounded

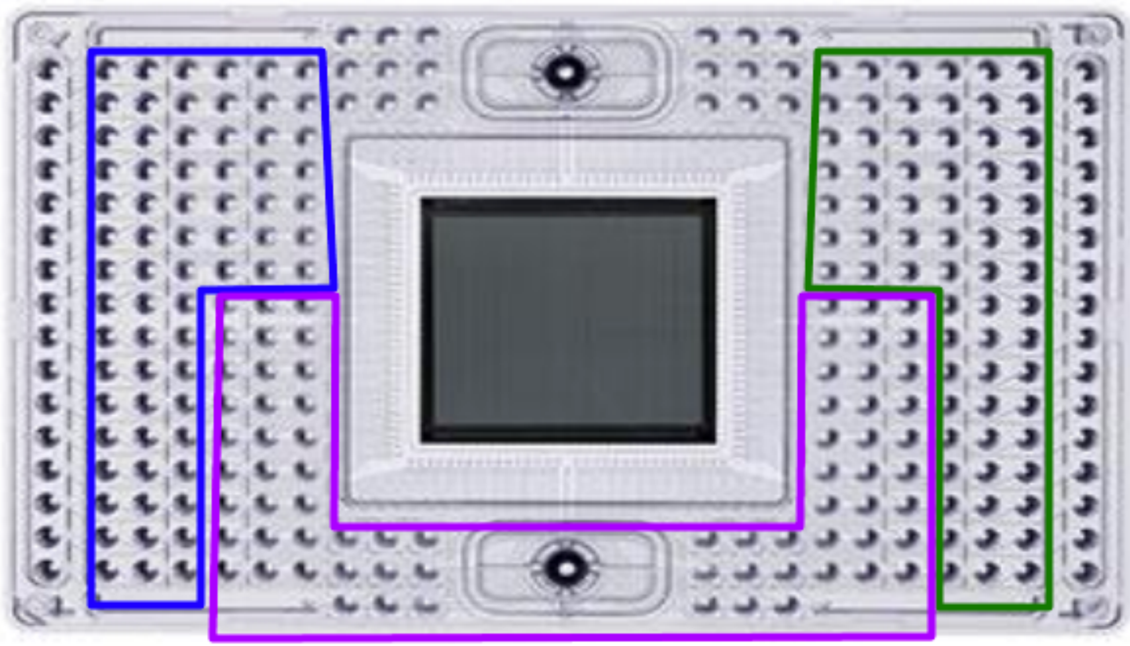


Biological Condition 1

Biological Condition 2

Biological Condition 3

Unconfounded



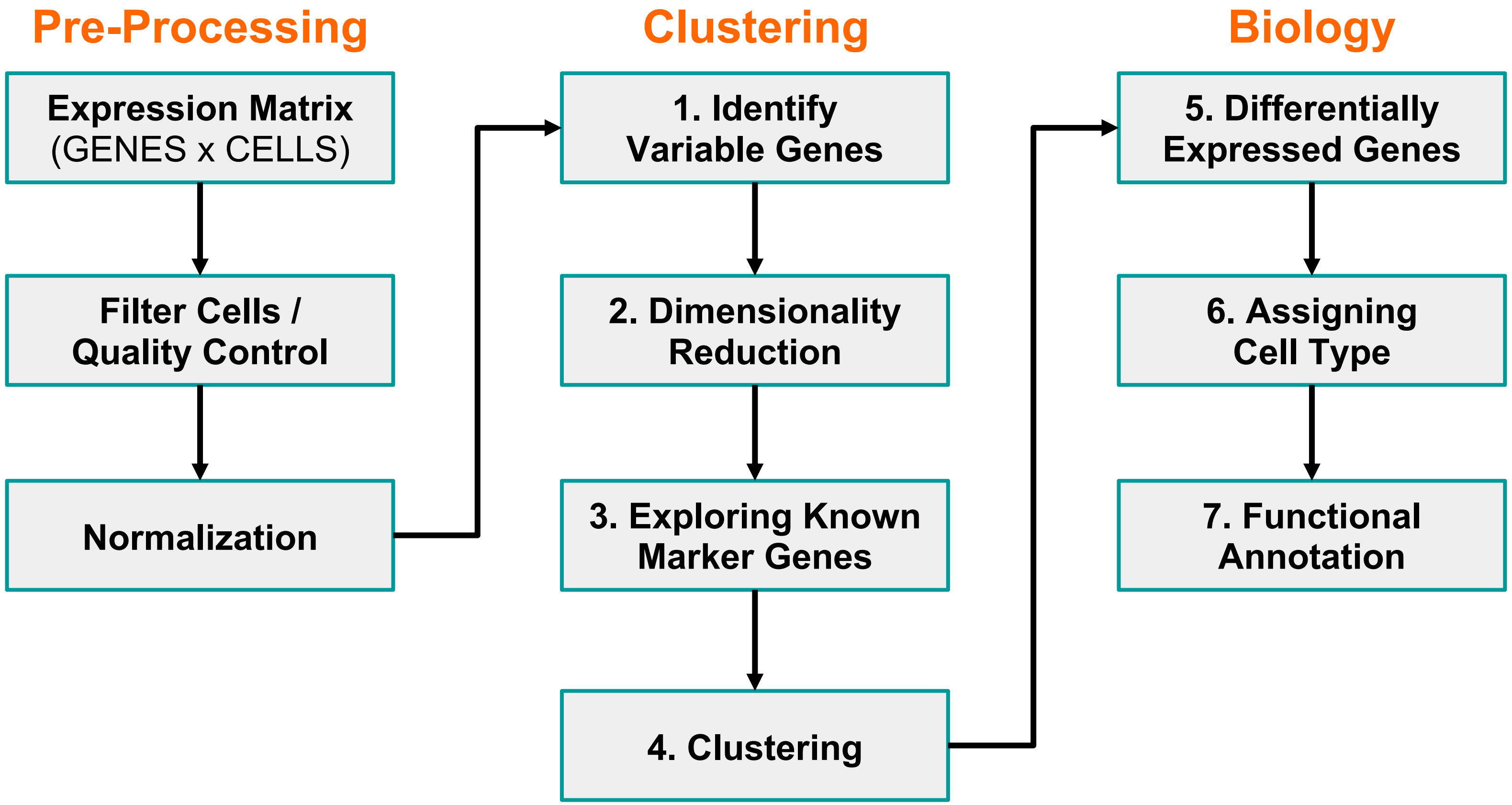
Biological Condition 1

Biological Condition 2

Biological Condition 3

https://scrnaseq-course.cog.sanger.ac.uk/website/ideal-scrnaseq-pipeline-as-of-oct-2017.html

Single-cell RNA-Seq analysis pipeline
Analyzing the expression data



Pre-Processing

Expression Matrix
(GENES x CELLS)

Filter Cells / Quality Control

Normalization

Clustering

1. Identify Variable Genes

2. Dimensionality Reduction

3. Exploring Known Marker Genes

4. Clustering

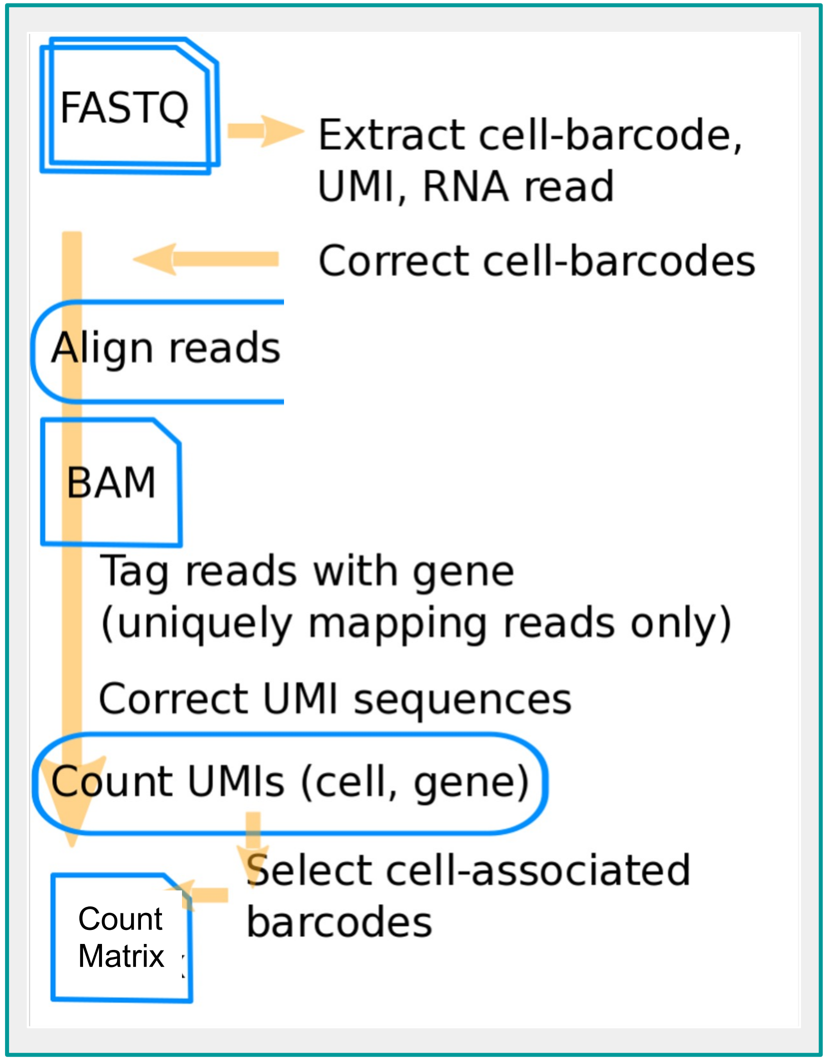
Biology

5. Differentially Expressed Genes

6. Assigning Cell Type

7. Functional Annotation

Single-cell RNA-Seq analysis pipeline
Generating the count matrix



FASTQ

Extract cell-barcode, UMI, RNA read

Correct cell-barcodes

Align reads

BAM

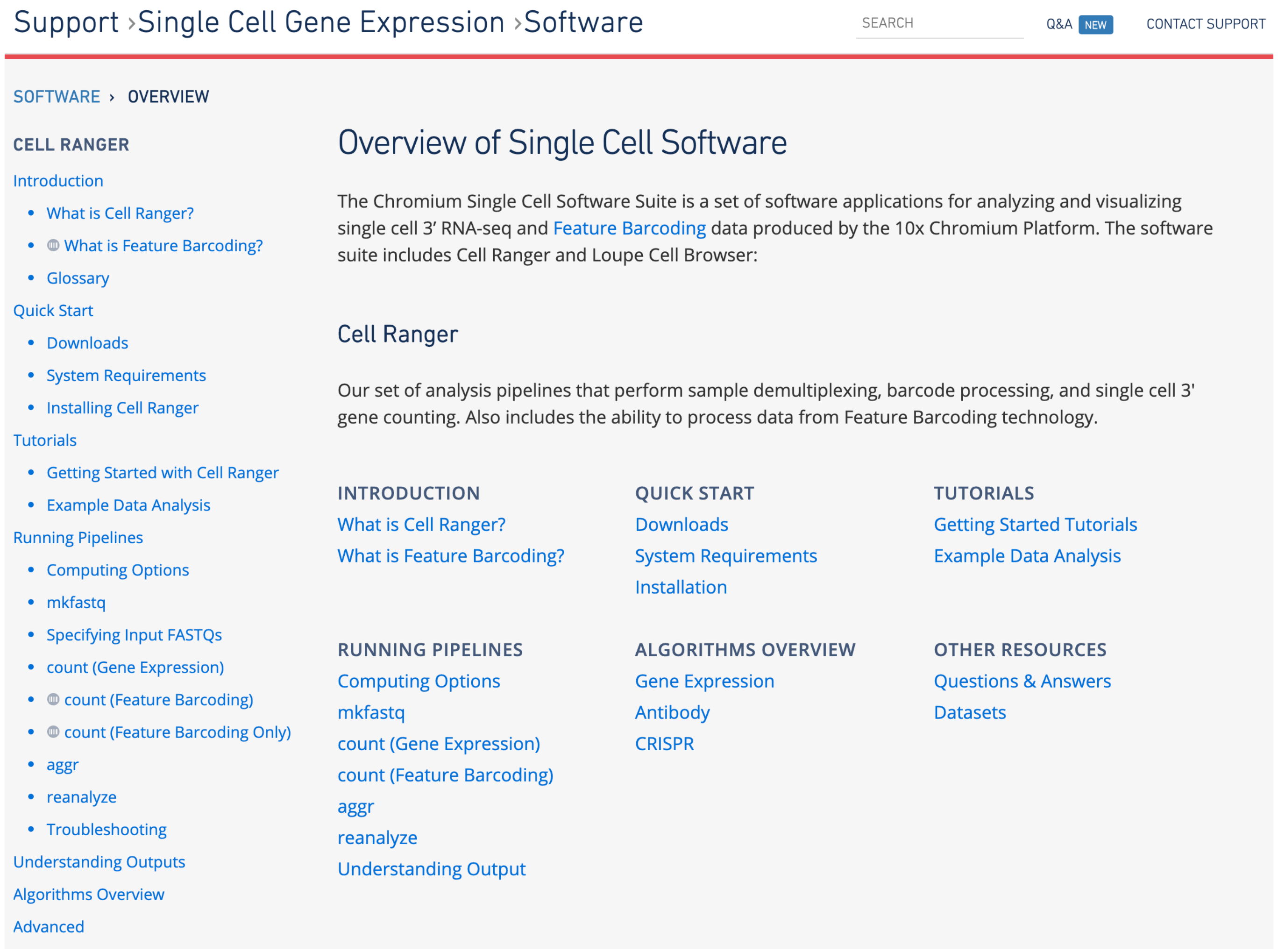
Tag reads with gene (uniquely mapping reads only)

Correct UMI sequences

Count UMIs (cell, gene)

Select cell-associated barcodes

Count Matrix



Overview of Single Cell Software

The Chromium Single Cell Software Suite is a set of software applications for analyzing and visualizing single cell 3′ RNA-seq and Feature Barcoding data produced by the 10x Chromium Platform. The software suite includes Cell Ranger and Loupe Cell Browser:

Cell Ranger

Our set of analysis pipelines that perform sample demultiplexing, barcode processing, and single cell 3' gene counting. Also includes the ability to process data from Feature Barcoding technology.

INTRODUCTION
- What is Cell Ranger?
- What is Feature Barcoding?

QUICK START
- Downloads
- System Requirements
- Installation

TUTORIALS
- Getting Started Tutorials
- Example Data Analysis

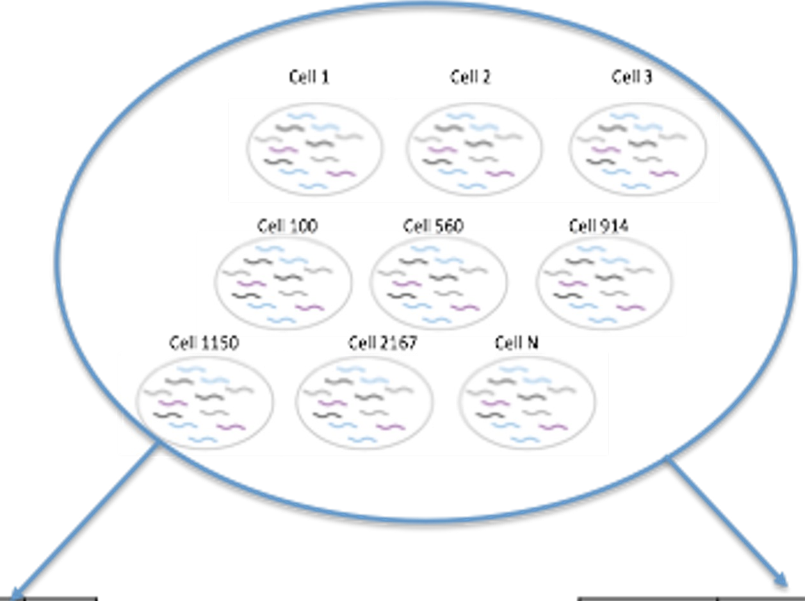
RUNNING PIPELINES
- Computing Options
- mkfastq
- count (Gene Expression)
- count (Feature Barcoding)
- aggr
- reanalyze
- Understanding Output

ALGORITHMS OVERVIEW
- Gene Expression
- Antibody
- CRISPR

OTHER RESOURCES
- Questions & Answers
- Datasets

https://support.10xgenomics.com/single-cell-gene-expression

Single-cell gene expression distributions are very different from bulk gene expression distributions



	Exp
Gene 1	100
Gene 2	3.5
Gene 3	10
...	
Gene K	95
...	
Gene 25K	0.4

Population Average

Bulk

Gene Expression Matrix

	Cell 1	Cell 2	Cell 3	....	Cell 5K
Gene 1	3	0	1		2
Gene 2	0	2	0		1
Gene 3	1	0	3		5
...					
Gene K	14	7	1		0
...					
Gene 25K	0	13	1		0

Cellular resolution

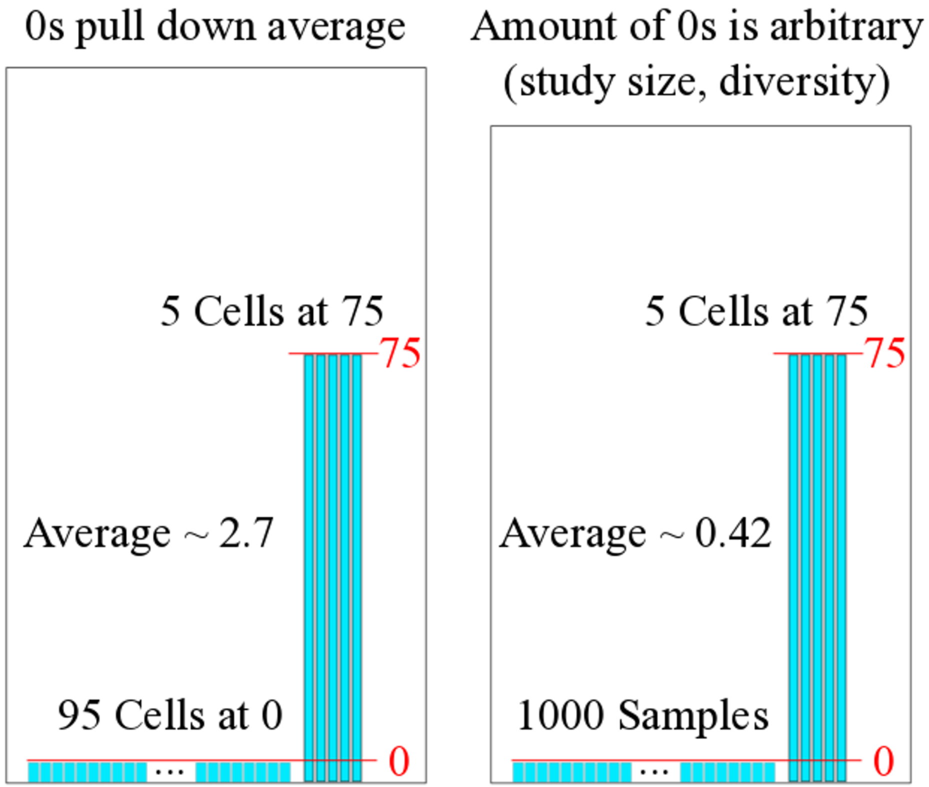
Single cell

Slide courtesy of Karthik Shekhar

Some single-cell RNA-Seq data challenges to remember

- **Drop out:** data has an excessive amount of zeros due to limiting mRNA

Zero expression doesn't mean the gene isn't on



Some unique features and challenges of single cell RNA-Seq

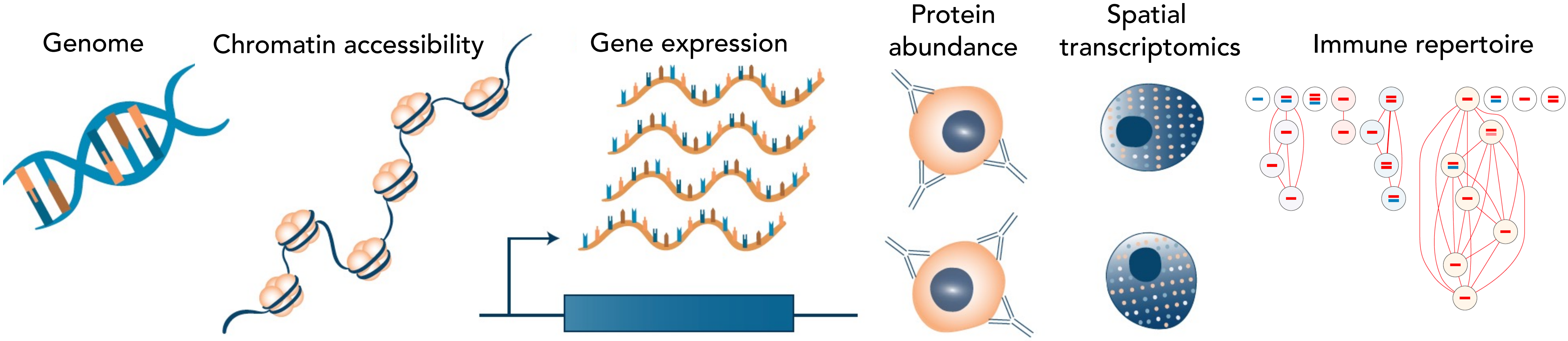
Features

- measures the distribution of expression levels for each gene across a population of single cells
- can study 1e2-1e6 cells in an experiment

Challenges

- amplification during library preparation
- gene dropout
- experimental design and computational analysis

Ongoing developments in single-cell genomics: Many other molecules from single cells may be profiled



DNA and epigenome

- Single cell genomes (WES, WGS)
- Single cell epigenomics (HiC, ChIP, ATAC, mC)

RNA

- Full length (mRNA, total RNA)
- 5’ and 3’ end counting

Proteins

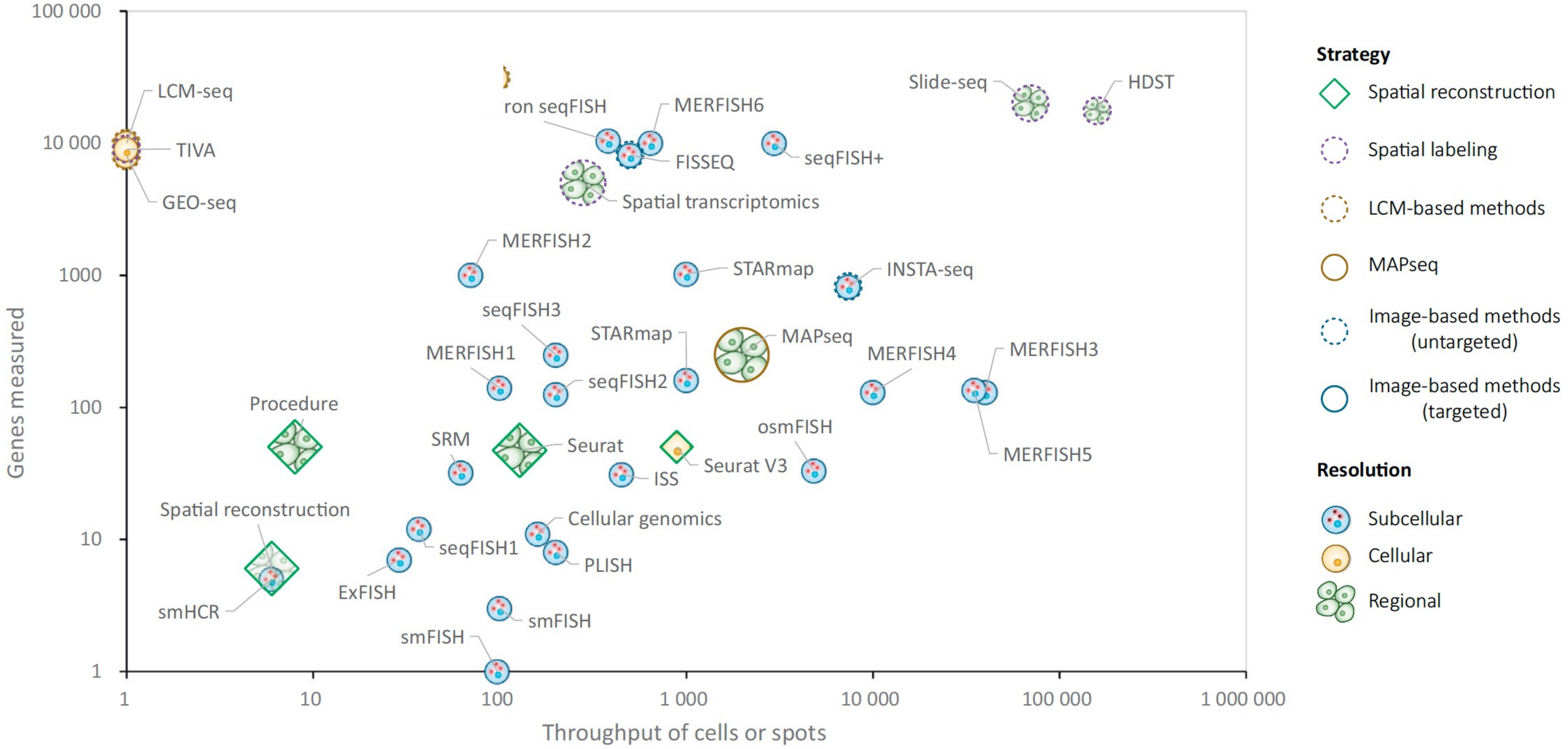
- Multi-parameter flow
- Mass cytometry
- Single cell proteomics

‘multi-omics’

- DNA+RNA (G+T)
- RNA+protein (T+P)
- Epigenome + RNA

Efremova M. et al. (2020) *Nature Methods.* 17:11-20.

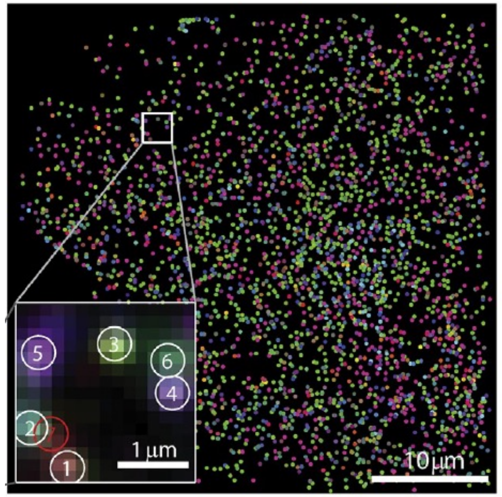
Ongoing developments in single-cell genomics: Growing toolbox for spatial genomics (profiling cells *in situ*)



Liao, J. et al *Trends in Biotechnology* (2021).

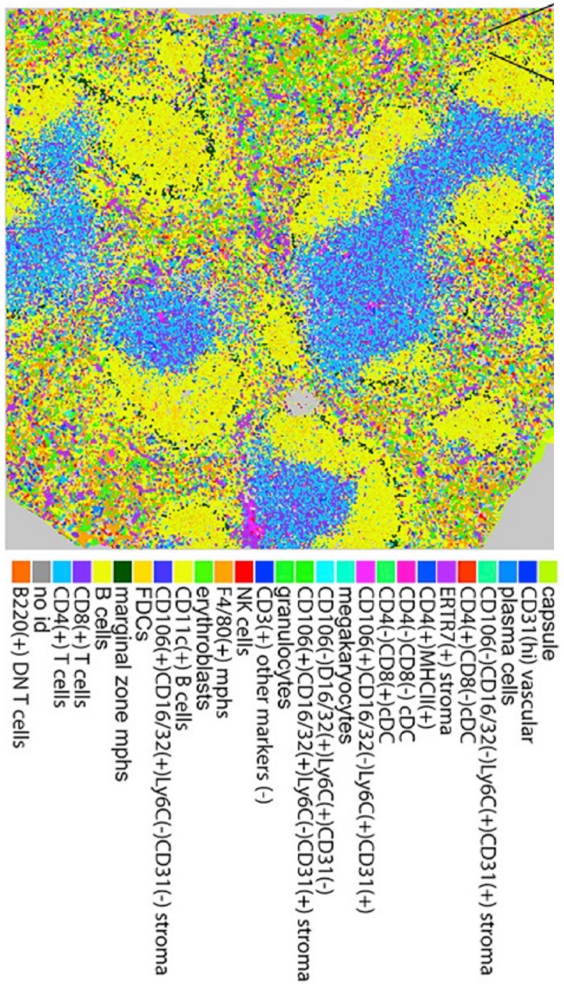
Ongoing developments in single-cell genomics: Growing toolbox for spatial genomics (profiling cells *in situ*)

MERFISH (RNA)



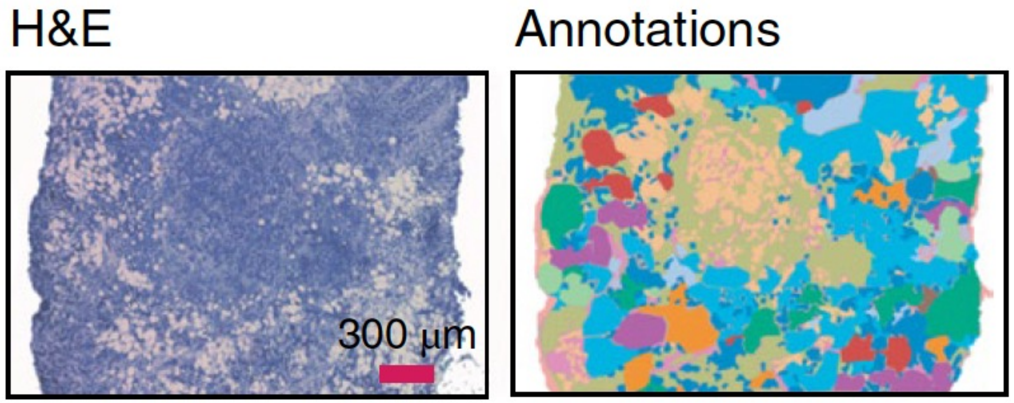
CODEX (protein)

mouse spleen



High Density Spatial Transcriptomics (RNA)

breast cancer


H&E

Annotations

- ● Fatty tissue, immune/lymphoid
- ● Fatty tissue, invasive cancer
- ● Fibrous tissue, invasive cancer
- ● Fibrous tissue, immune/lymphoid
- ● Invasive cancer, immune/lymphoid
- ● Immune/lymphoid
- ● Fatty tissue, fibrous tissue, invasive cancer
- ● Fibrous tissue
- ● Fibrous tissue, invasive cancer, immune/lymphoid
- ● Fatty tissue
- ● Fatty tissue, fibrous tissue, invasive cancer, immune/lymphoid
- ● Fatty tissue, invasive cancer, immune/lymphoid
- ● Invasive cancer

Moffitt J.R. et al. (2016) *PNAS* 113: 11046-11051.
Goltsev Y. et al. (2018) *Cell.* 174: 968-981.
Vickovic S. et al. (2019) *Nature Methods.* 16: 987–990.

Seurat and scanpy: single cell analysis toolkits

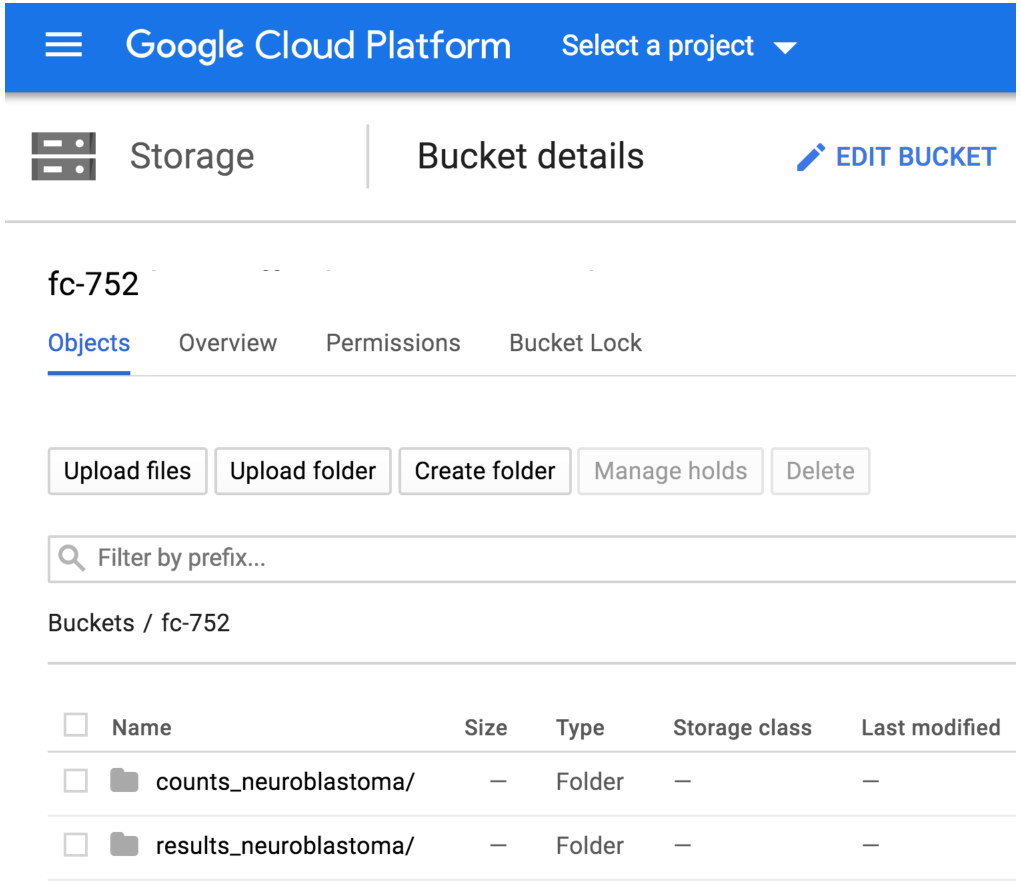

SEURAT R toolkit for single cell genomics

scanpy

https://satijalab.org/seurat/
https://scanpy.readthedocs.io/en/stable/

Single-cell analysis computations in the Cloud

Inputs and outputs are stored in Google Cloud storage buckets

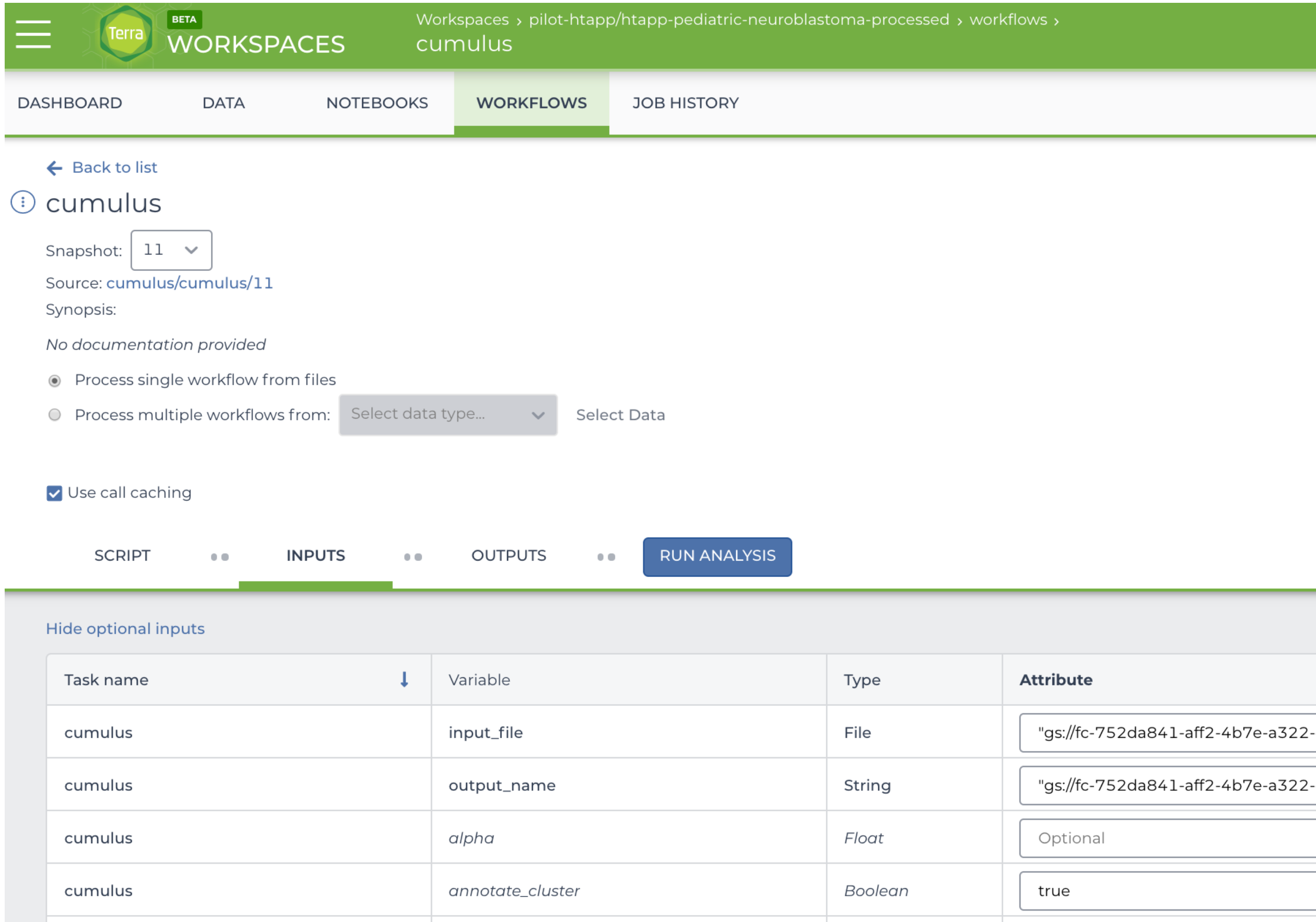


fc-752

Buckets / fc-752

☐	Name	Size	Type	Storage class	Last modified
☐	counts_neuroblastoma/	—	Folder	—	—
☐	results_neuroblastoma/	—	Folder	—	—

Data processing and analysis are performed in Terra



cumulus

Snapshot: 11

Source: cumulus/cumulus/11

Synopsis:

No documentation provided

● Process single workflow from files

○ Process multiple workflows from: Select data type... Select Data

☑ Use call caching

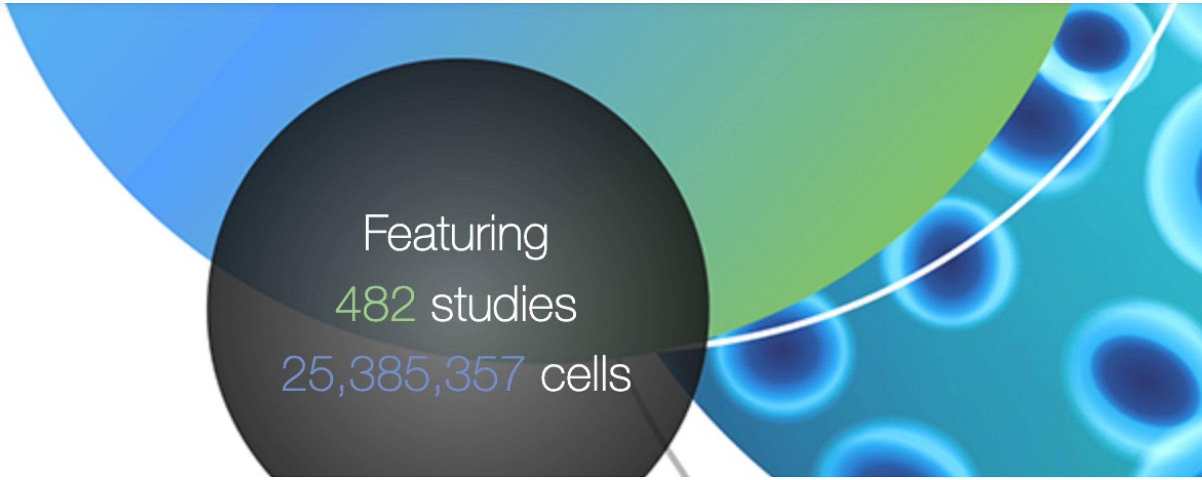
Task name	Variable	Type	Attribute
cumulus	input_file	File	"gs://fc-752da841-aff2-4b7e-a322-
cumulus	output_name	String	"gs://fc-752da841-aff2-4b7e-a322-
cumulus	*alpha*	*Float*	Optional
cumulus	*annotate_cluster*	*Boolean*	true

https://cumulus.readthedocs.io/en/latest/index.html

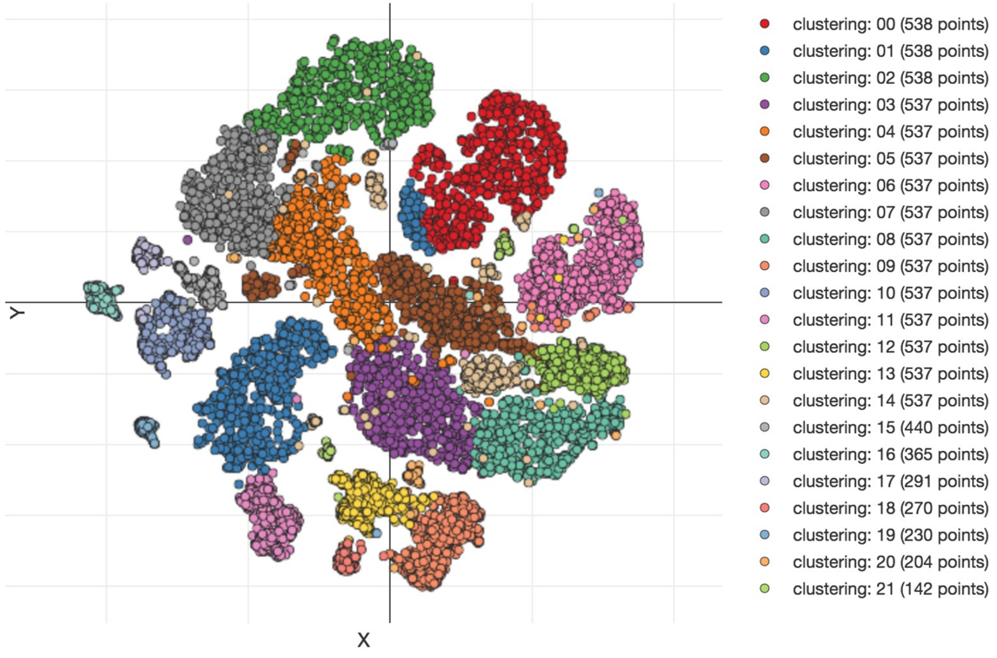
Single Cell Portal facilitates sharing of single-cell studies

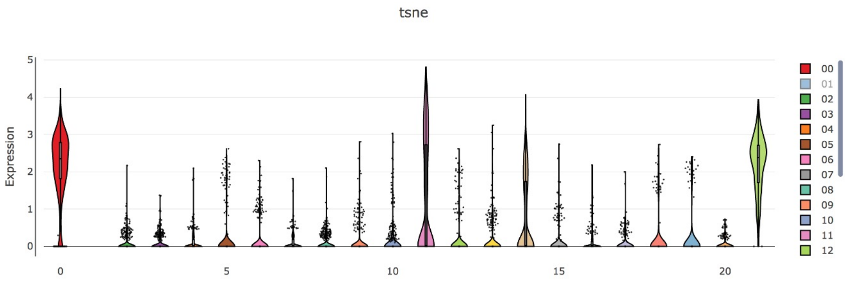
Single Cell
PORTAL

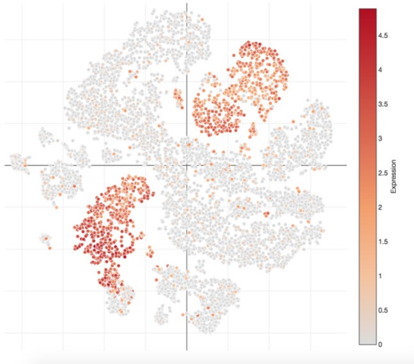
Reducing barriers and accelerating single-cell research



Featuring
482 studies
25,385,357 cells







A few single-cell resources

Comprehensive list of single-cell resources
https://github.com/seandavi/awesome-single-cell
https://www.scrna-tools.org/

Computational packages for single-cell analysis
https://satijalab.org/seurat/
https://scanpy.readthedocs.io/
http://bioconductor.org/packages/devel/workflows/html/simpleSingleCell.html

eLife Commentary on the Human Cell Atlas (HCA) https://elifesciences.org/articles/27041
Nature Commentary on the HCA https://www.nature.com/news/the-human-cell-atlas-from-vision-to-reality-1.22854

Online courses
https://www.singlecellcourse.org/
https://bioconductor.org/books/release/OSCA/
https://lmweber.org/OSTA-book/
https://uppsala.instructure.com/courses/52011

Single cell data repositories
http://jinglebells.bgu.ac.il/.
https://www.nxn.se/single-cell-studies/gui
https://data.humancellatlas.org/
https://tabula-sapiens-portal.ds.czbiohub.org/
https://tabula-muris.ds.czbiohub.org/

Flash talk instructions

Each one of us will give a 1 minute, 1 slide presentation on a question that they are interested in, which could benefit from single-cell RNA-Seq. It can be an area you are currently investigating, something you are broadly interested in and wish to explore in the near future, or an idea for future projects. The goal of this is for all of us to get to know each other a little more, and for you to see the diversity of research possibilities in this area.

Please prepare a 1 slide pdf for Tuesday, and we will send you a link to place the pdf in a Google drive. Include your name in the filename, and please ask for any clarifications.