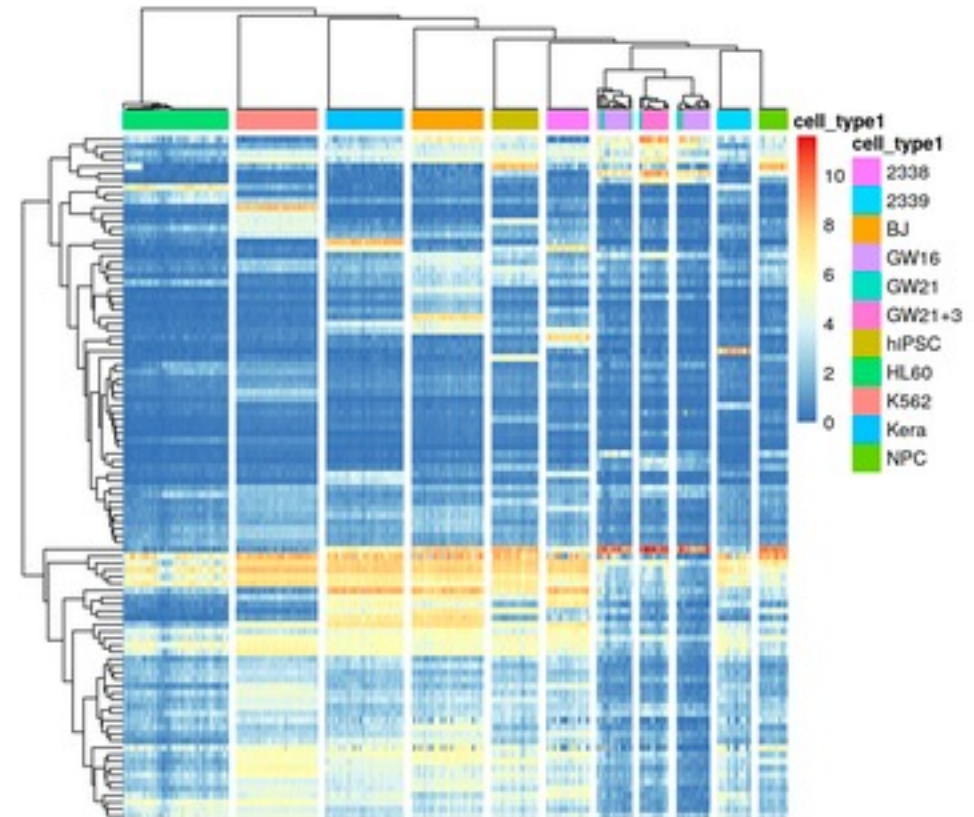
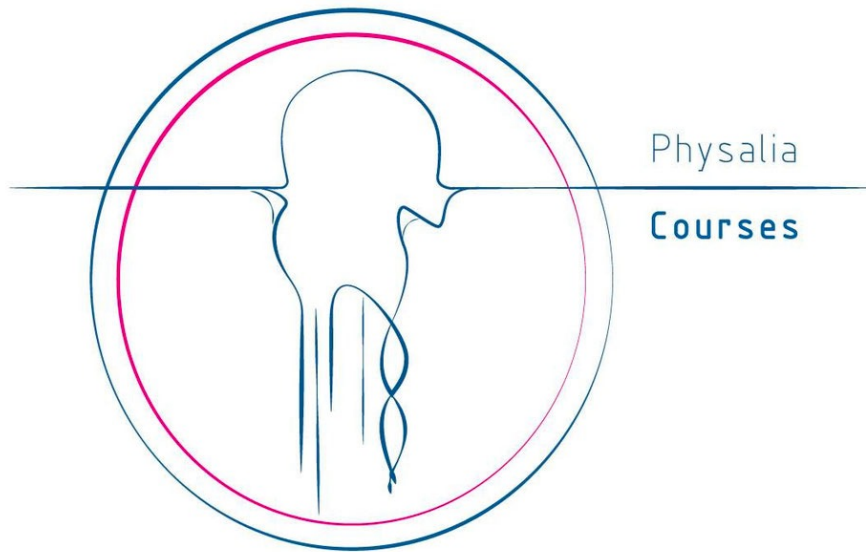


Analysis of Single-Cell RNA-Seq Data: Experimental Design

Orr Ashenberg, Jacques Serizay, Arnav Mehta

June 6, 2022



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_2022/

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...)

Survey results (*n*=27)

Current position: 25% PhD student, 25% postdoc, 50% other (staff scientist, industry, undergraduate, MD, professor...)

Operating system: 33% MacOS, 33% Windows, 22% Linux

Background: strong in molecular/cellular biology and next gen sequencing, less strong in programming and single cell

Advance topic interests: cell trajectories, scATAC-Seq, spatial

Biggest challenge: number of methods available and interpretation, math/stats/ML, programming

Mouse organogenesis studied by single-cell RNA sequencing

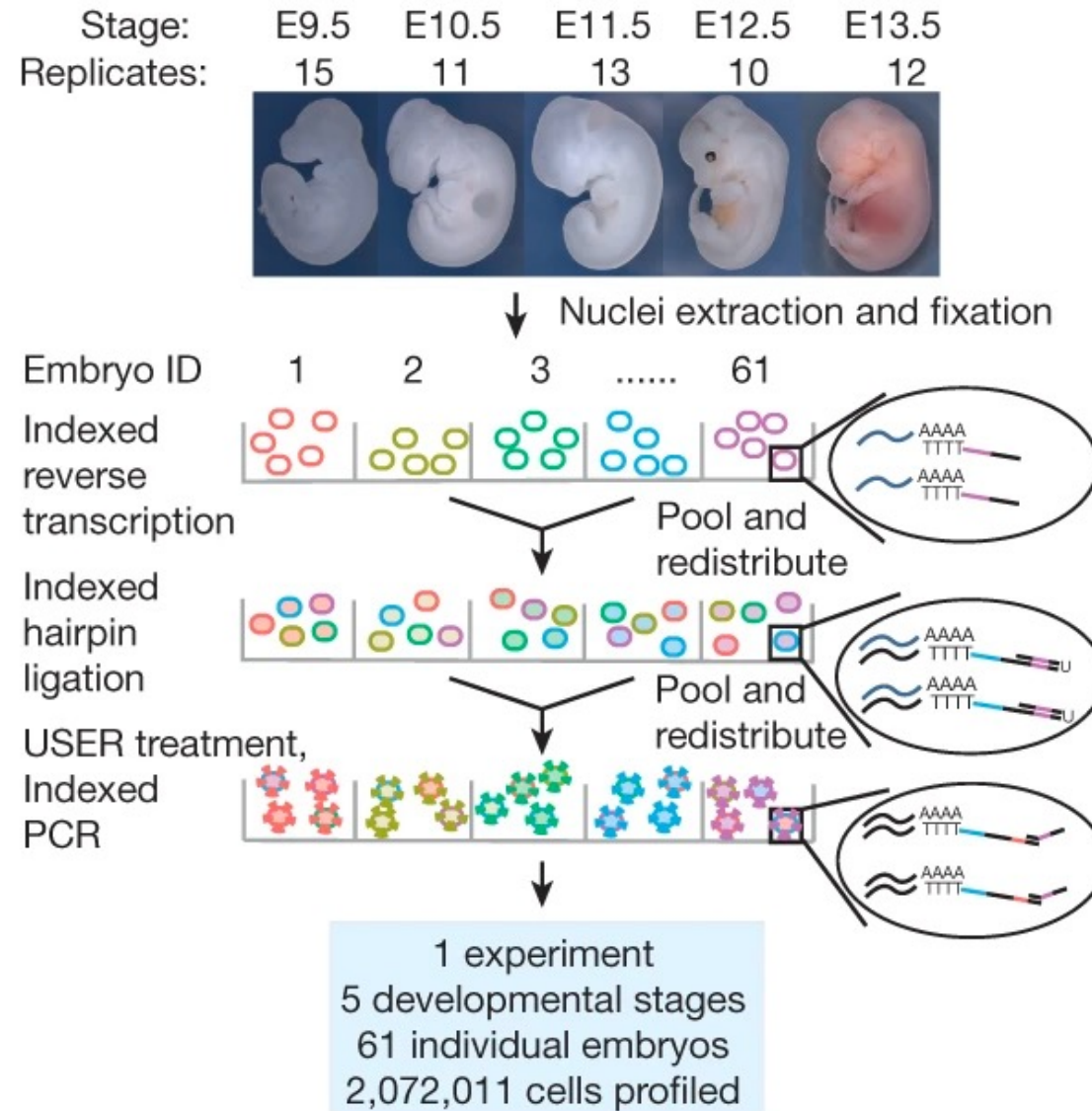
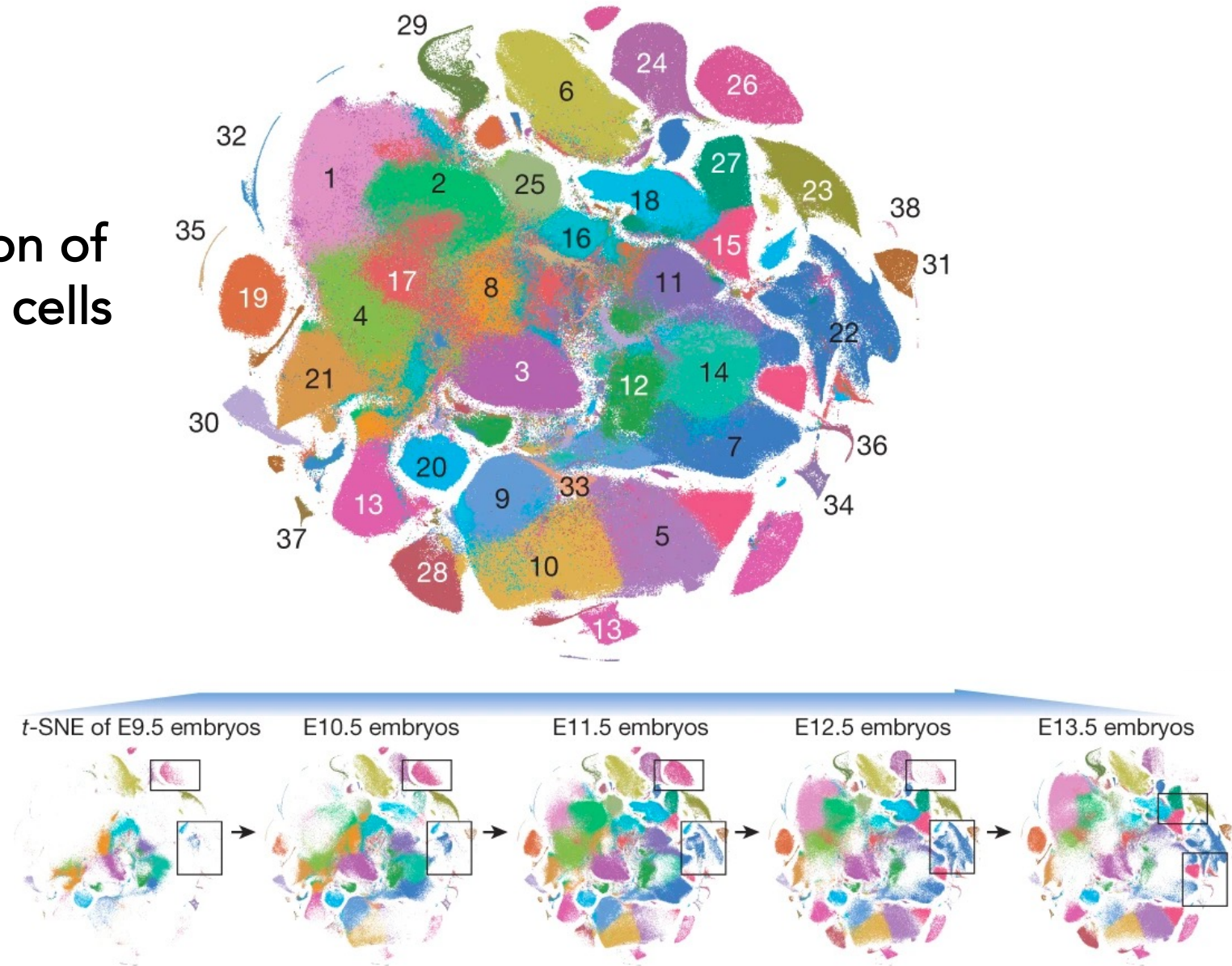


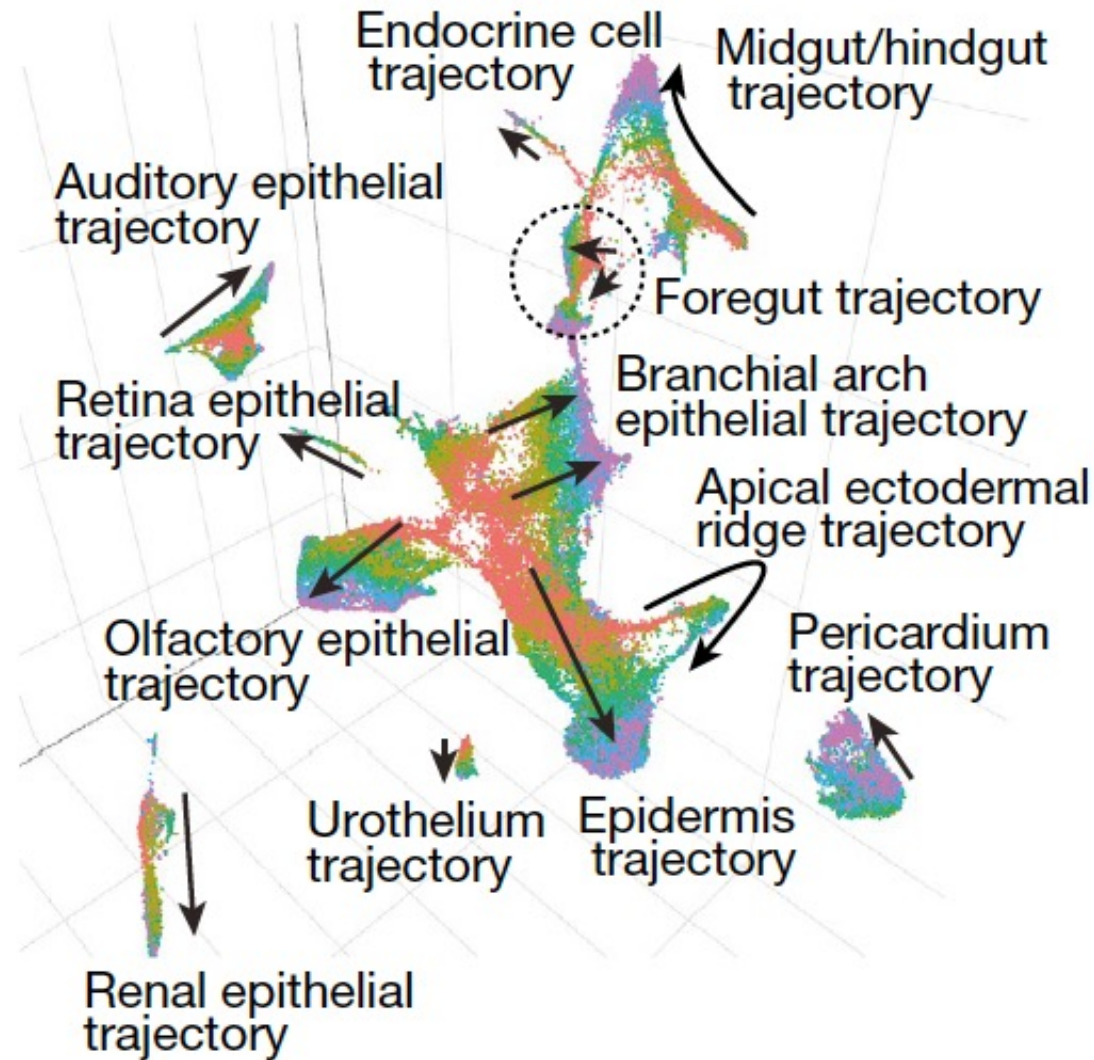
Fig. 1: sci-RNA-seq3 enables profiling of 2,072,011 cells from 61 mouse embryos across 5 developmental stages in a single experiment.

Mouse organogenesis studied by single-cell RNA sequencing

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



Mouse organogenesis studied by single-cell RNA sequencing

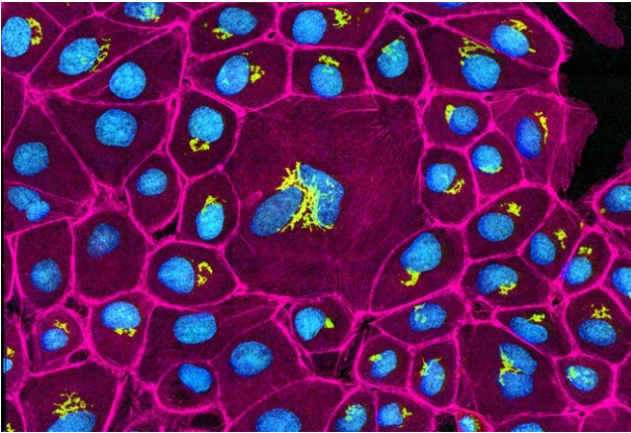


Cao, J., et al. *Nature* 2019

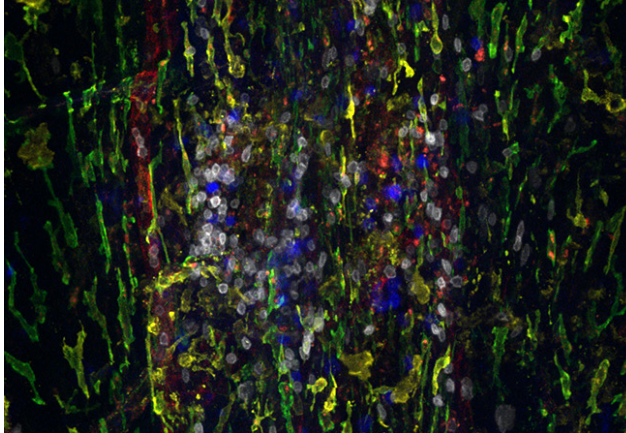
<https://www.youtube.com/watch?v=cWG2CkkDcrM>

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

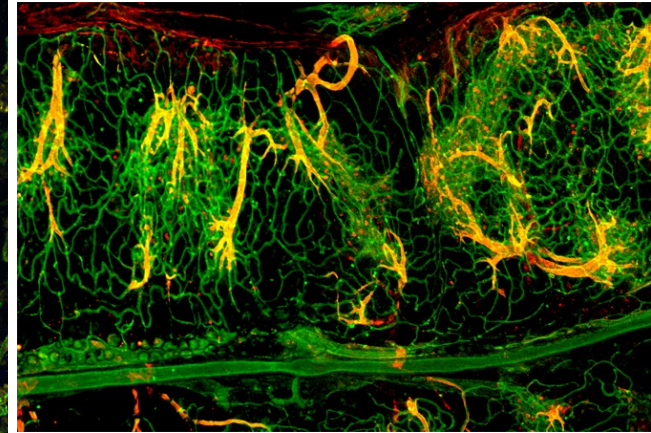
Skin epithelium



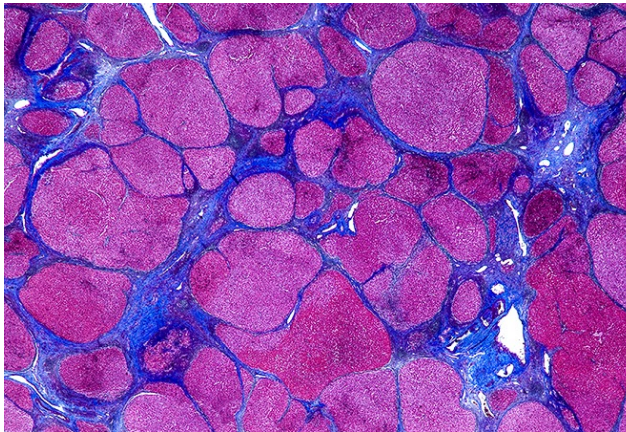
Brain meninges



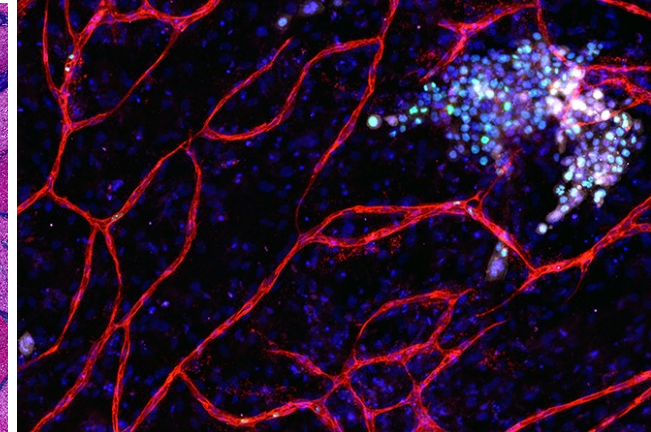
Blood vessels



Small intestine

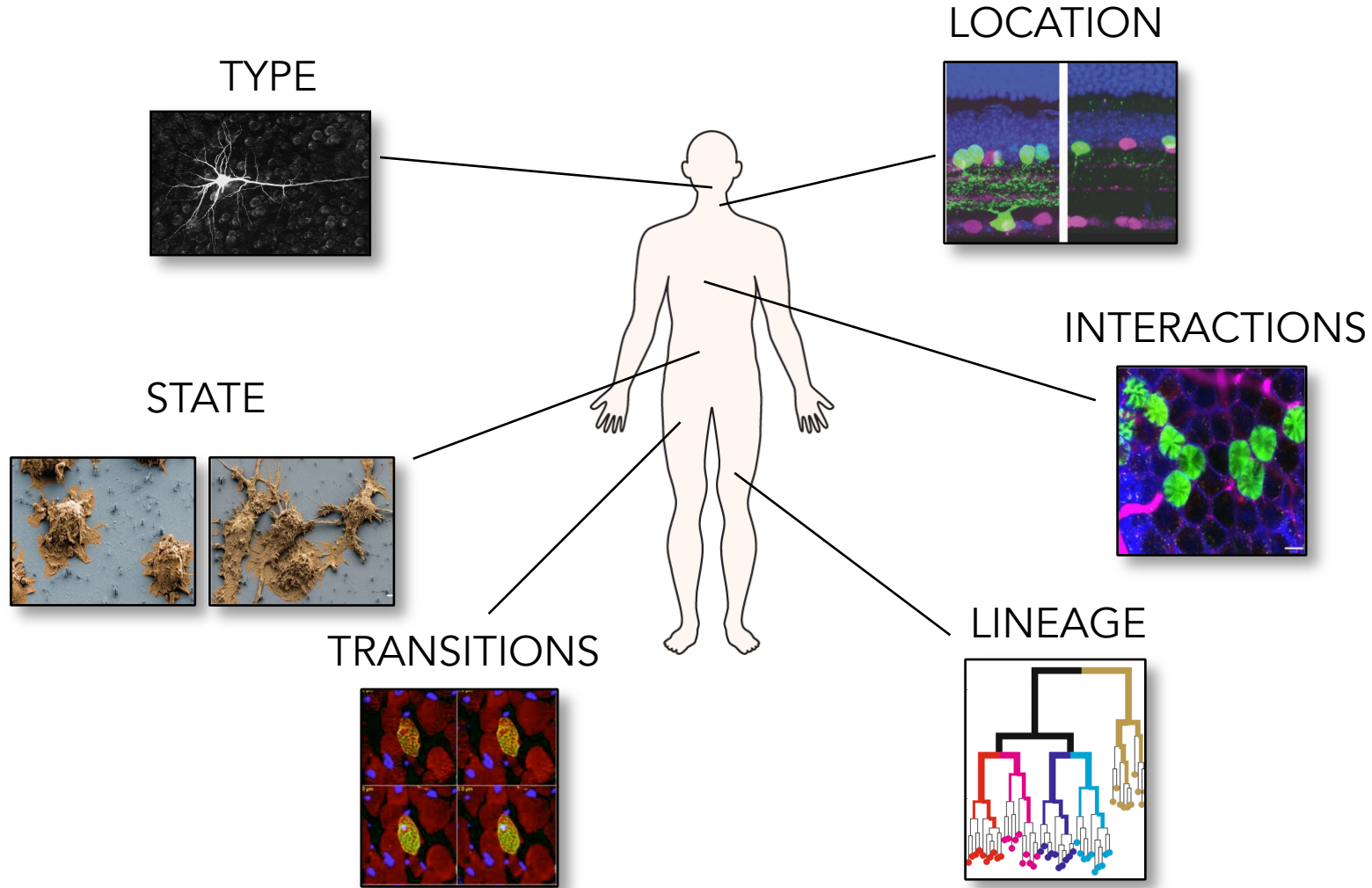


Liver cirrhosis

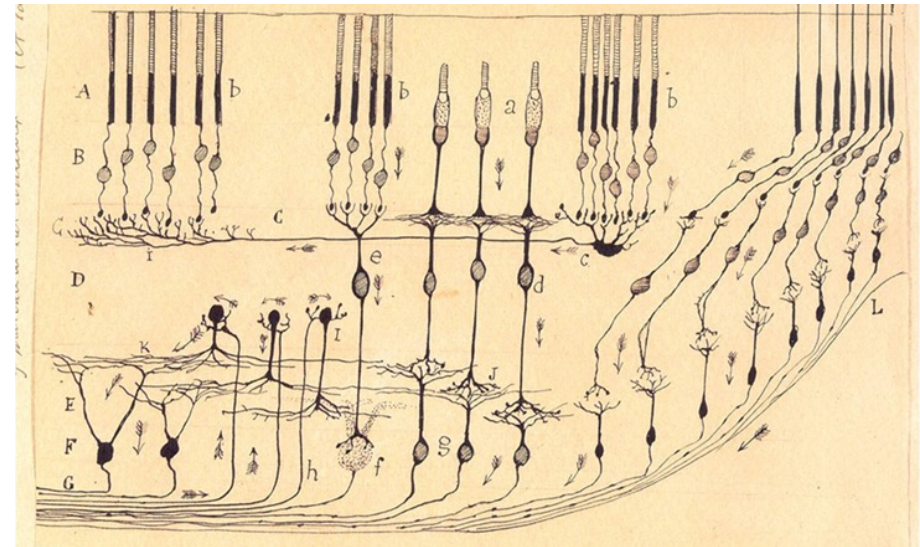


Breast cancer

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

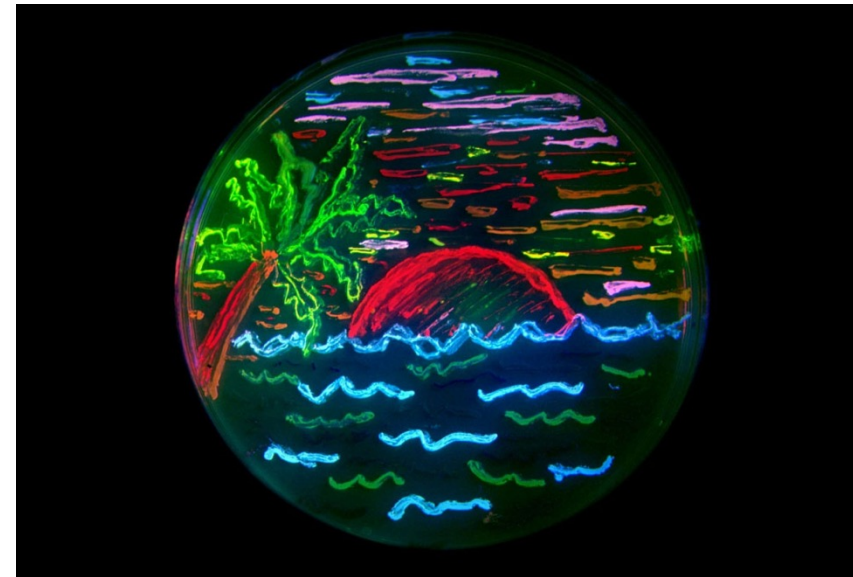
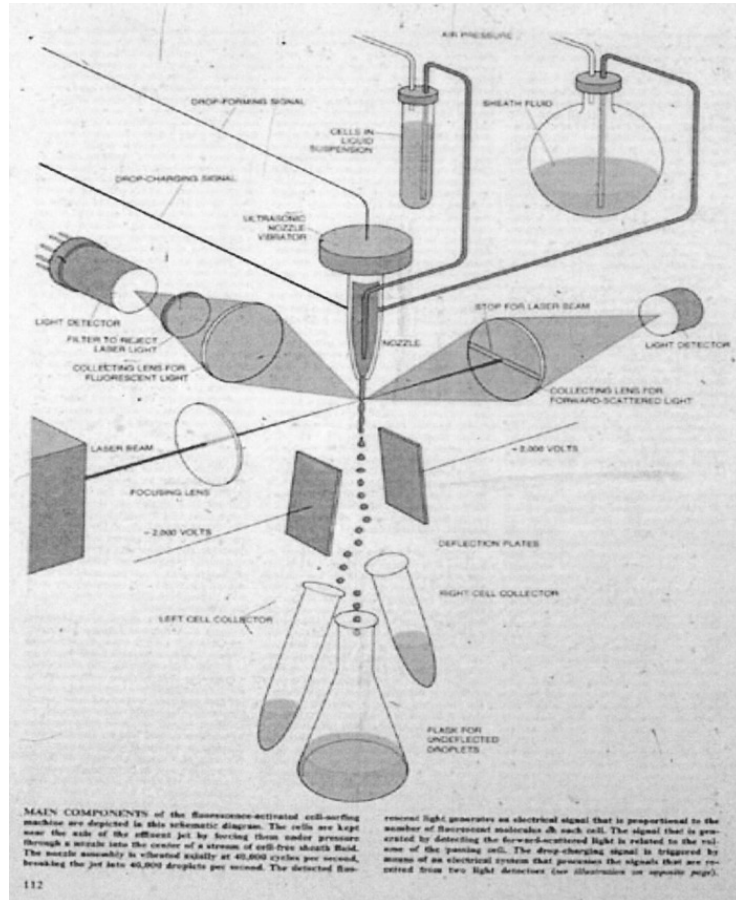


Technological innovations allow observation at increasing resolution

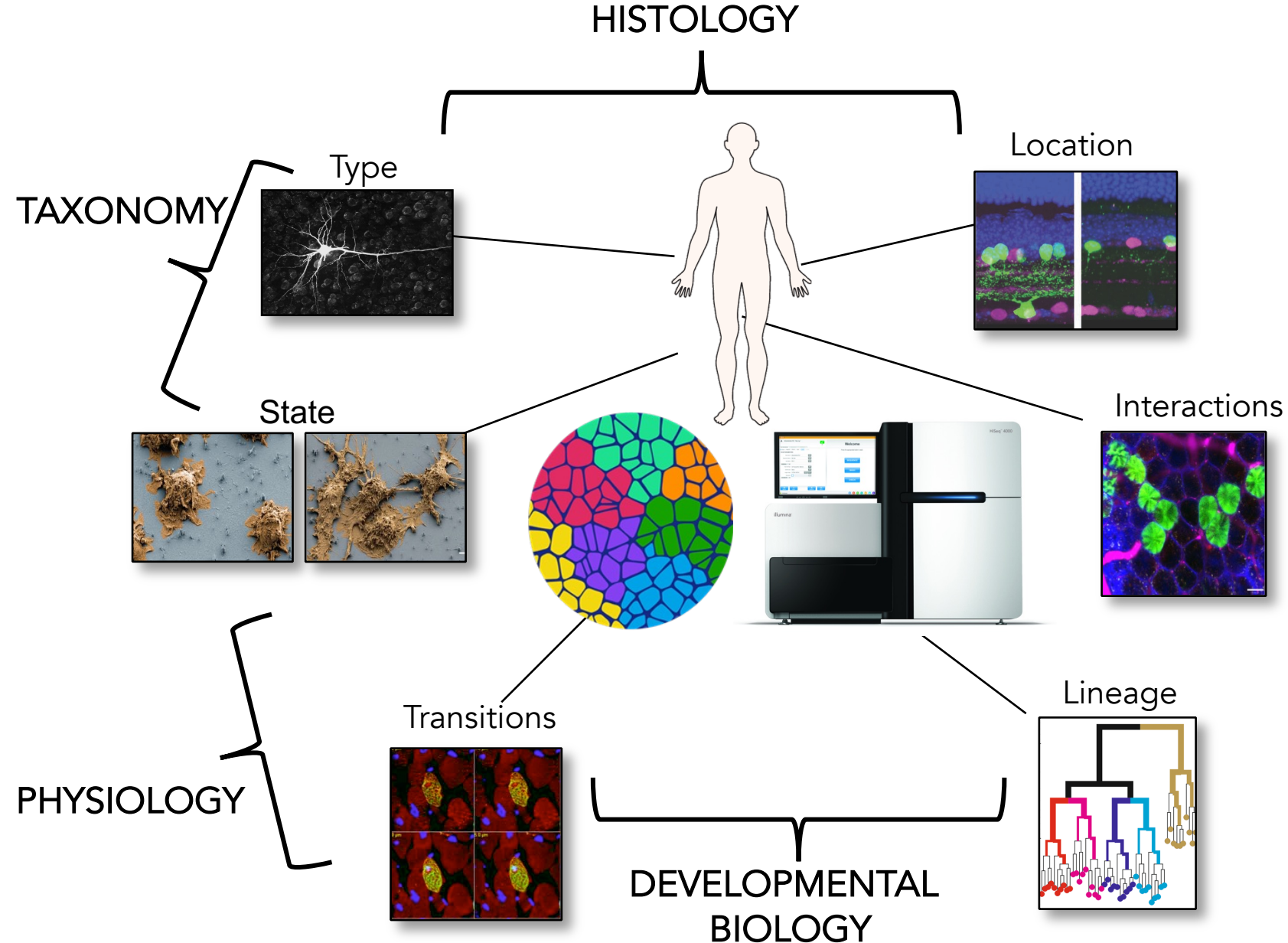


"by the help of Microscopes, there is nothing so small as to escape our inquiry"
Robert Hooke

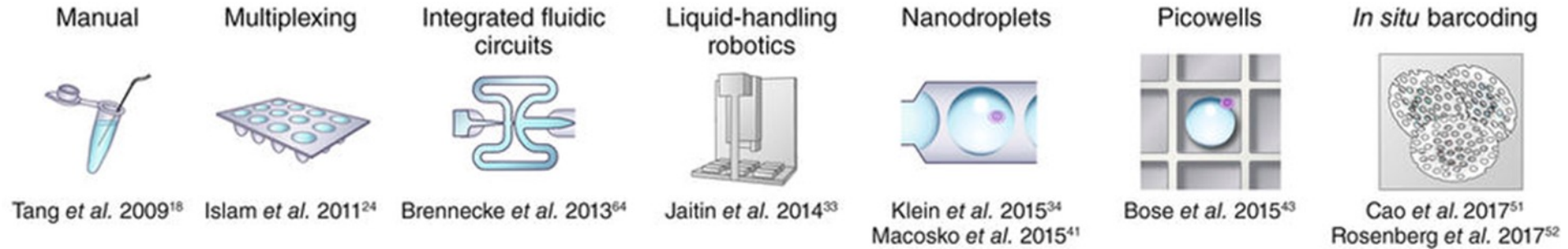
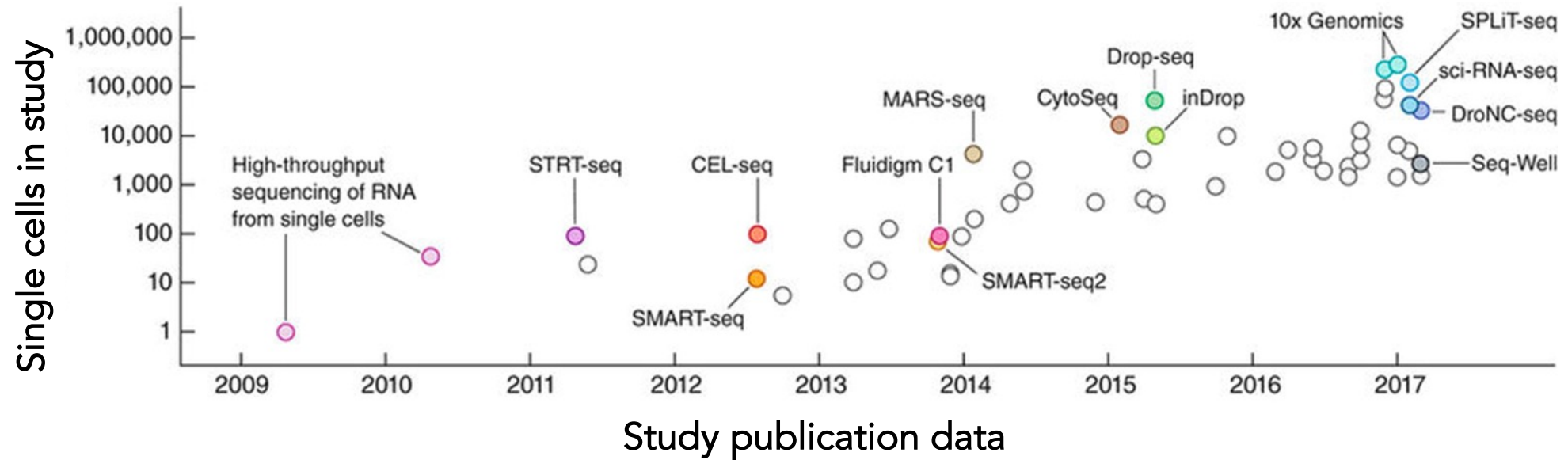
Technological innovations allow observation at increasing resolution



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



Single-cell RNA sequencing has grown exponentially



Human Cell Atlas mission

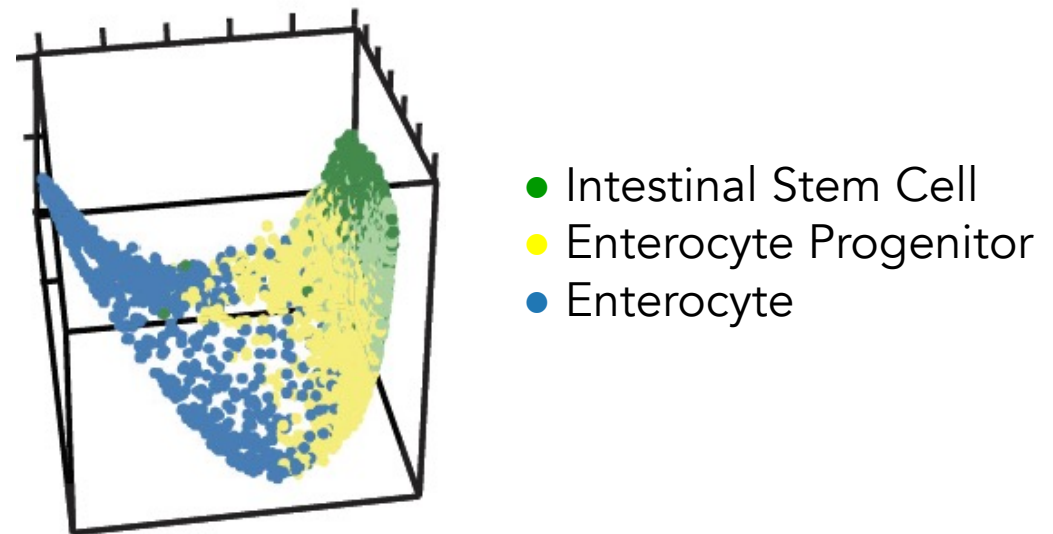
To create a comprehensive reference map of the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease.



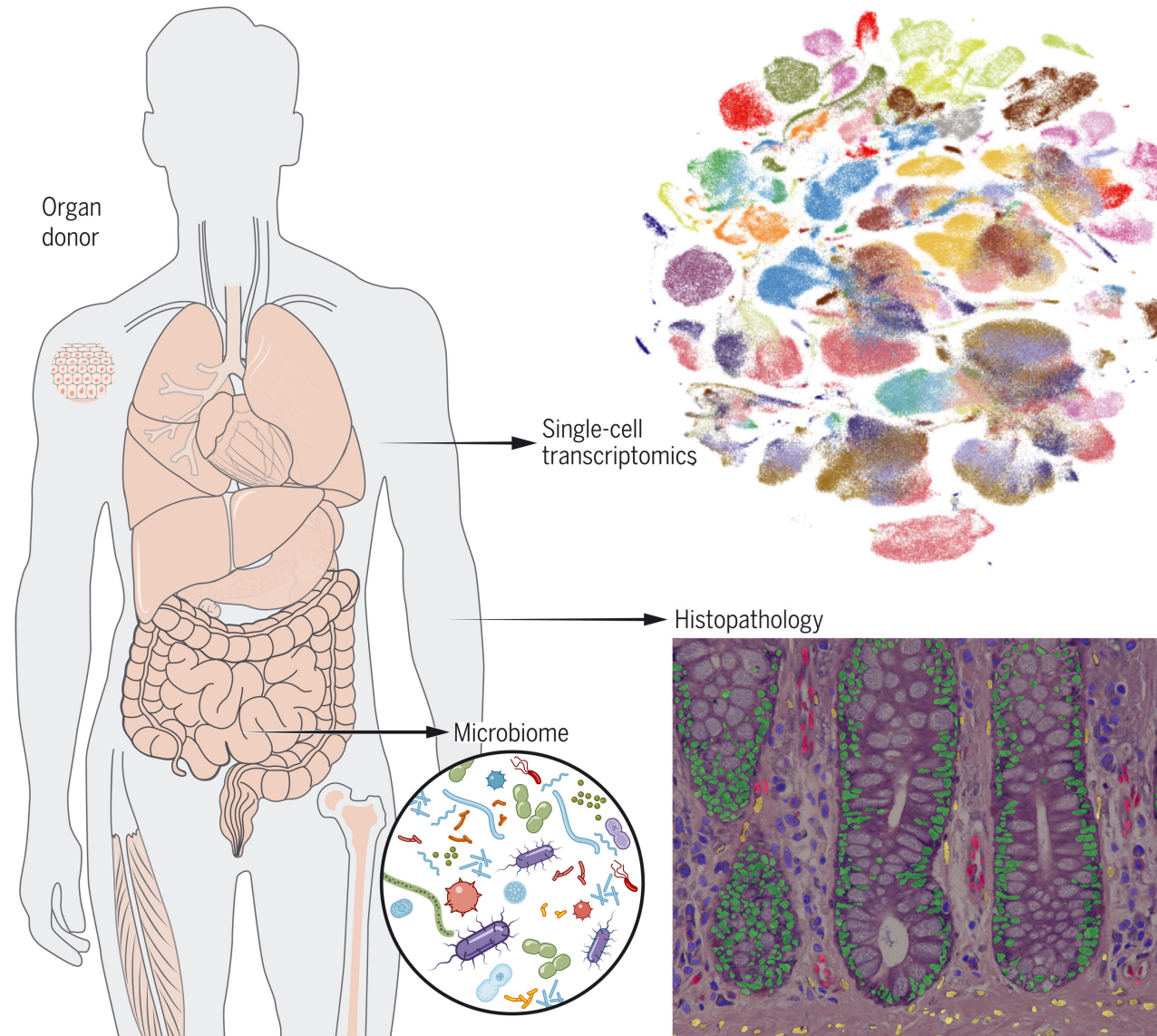
"The vestiges of the rupture reveal themselves, if someone brings forward a map of the world and considers carefully the coasts of the three [continents]."
Dutch map maker Abraham Ortelius (1596)

Human Cell Atlas mission

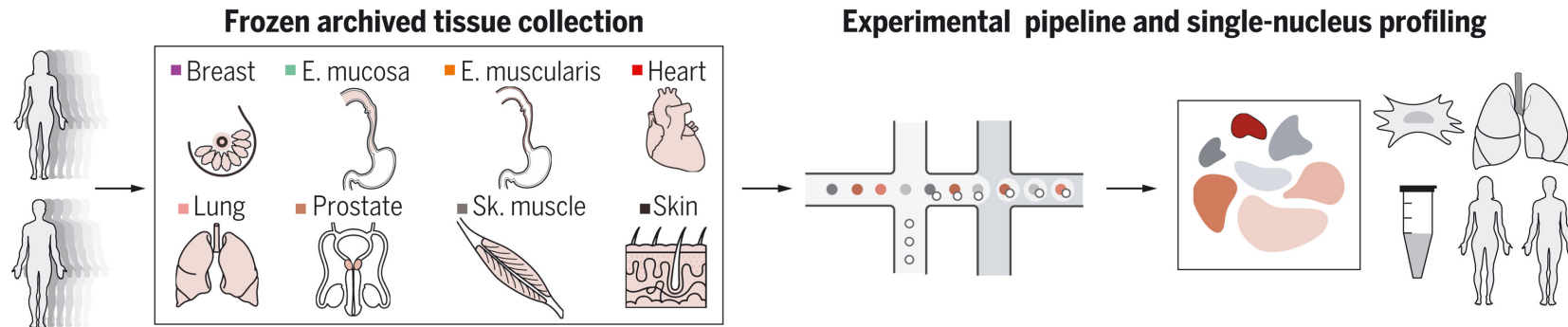
To create a comprehensive reference map of the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease.



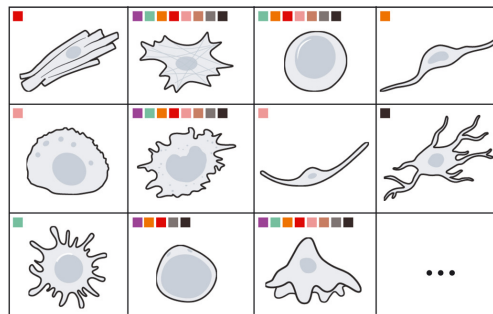
A series of human cell atlases published in May 2022: 500,000 cells from 24 different tissues and organs



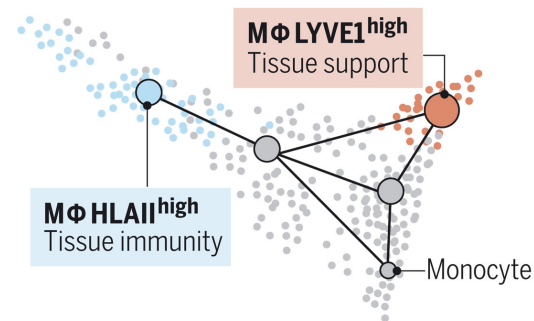
A series of human cell atlases published in May 2022: Association of cell types and states with human disease



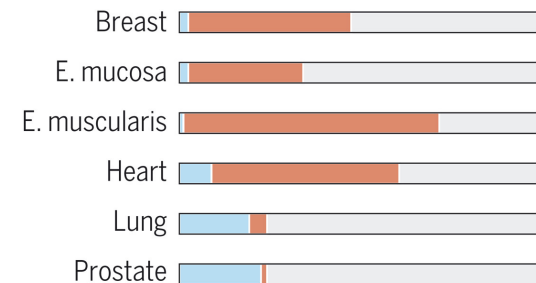
Census of cell types across tissues



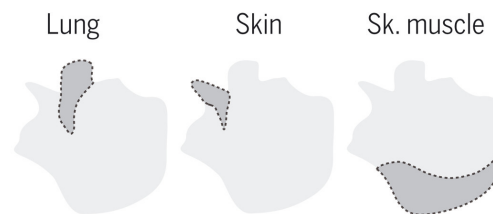
Dichotomous macrophage populations



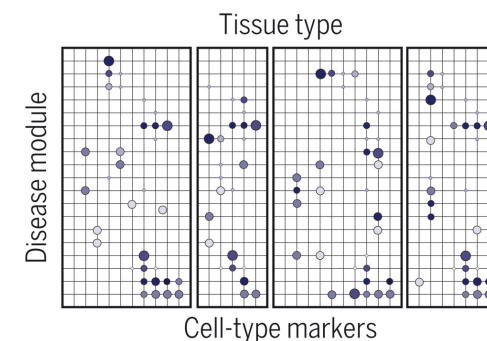
Macrophage composition of tissues



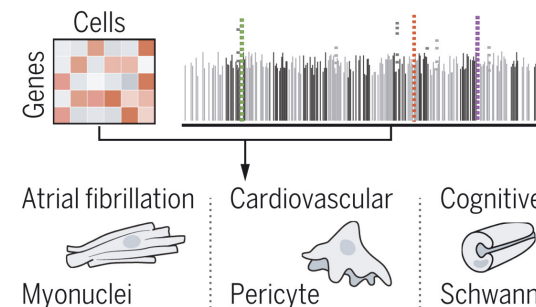
Shared and tissue-specific features of fibroblasts



Monogenic disease gene enrichment

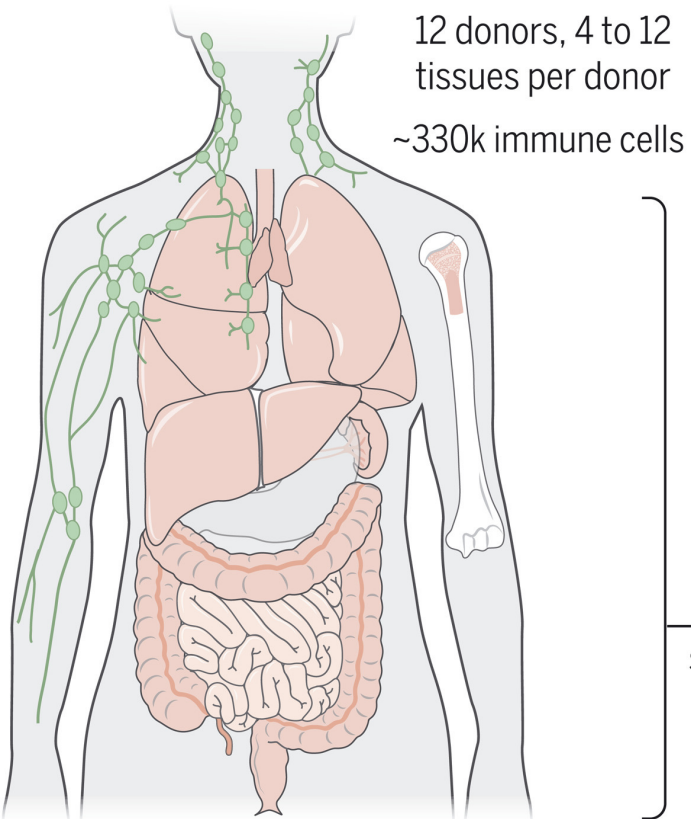


Association of complex phenotypes with cell types

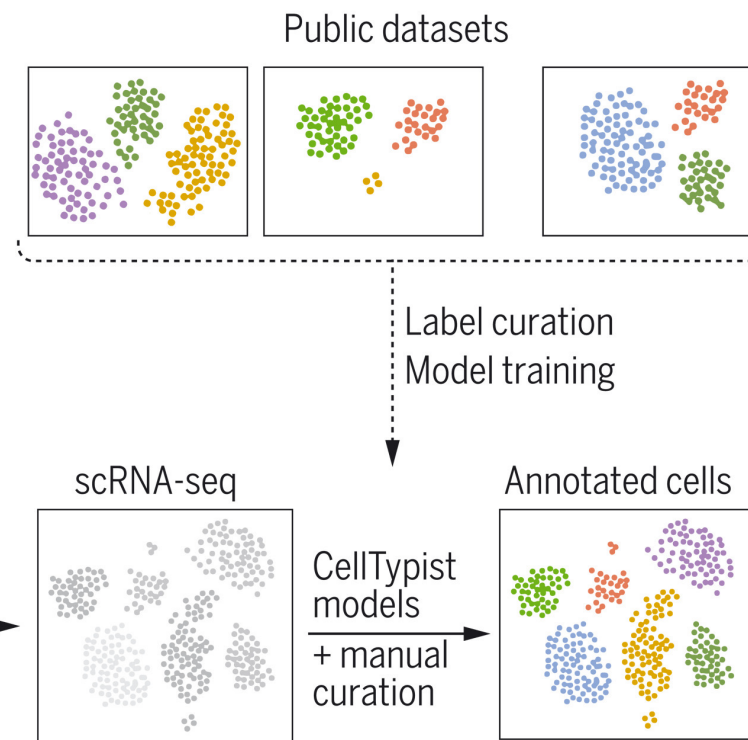


A series of human cell atlases published in May 2022: 300,000 immune cells from 16 different tissues

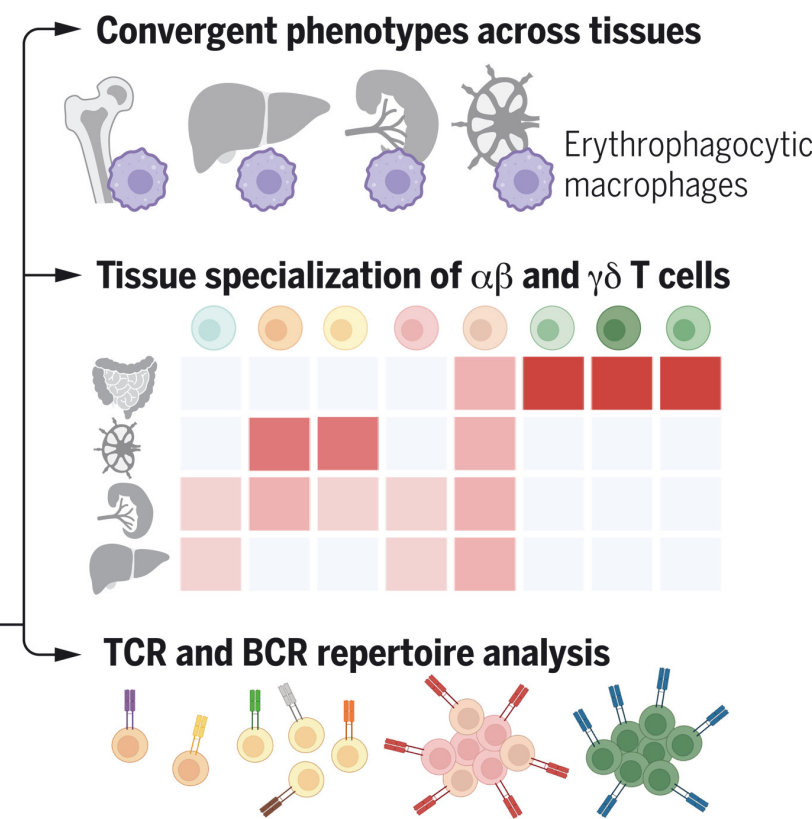
Human immune cells across tissues



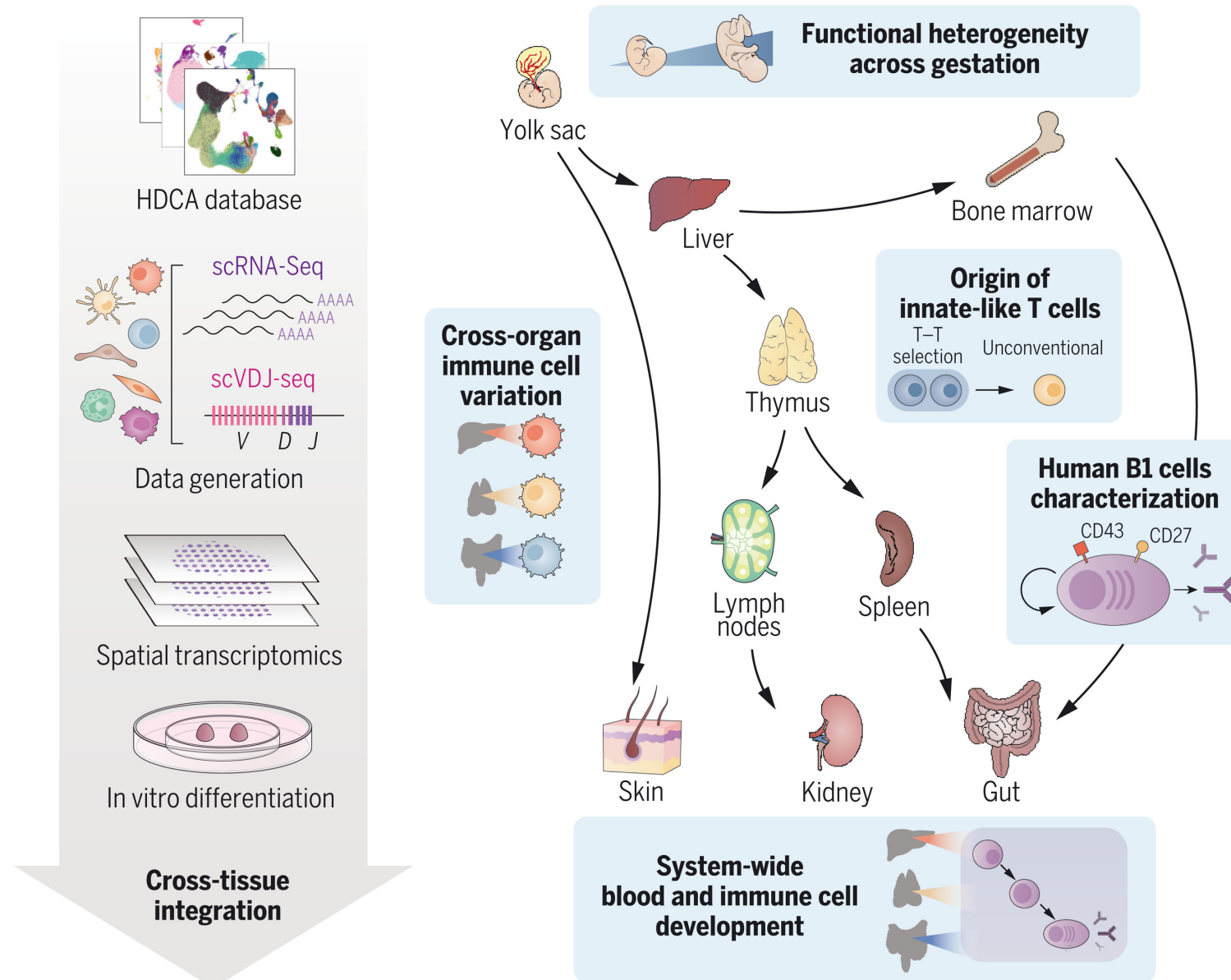
CellTypist: automated cell type annotation



Emerging cross-tissue features



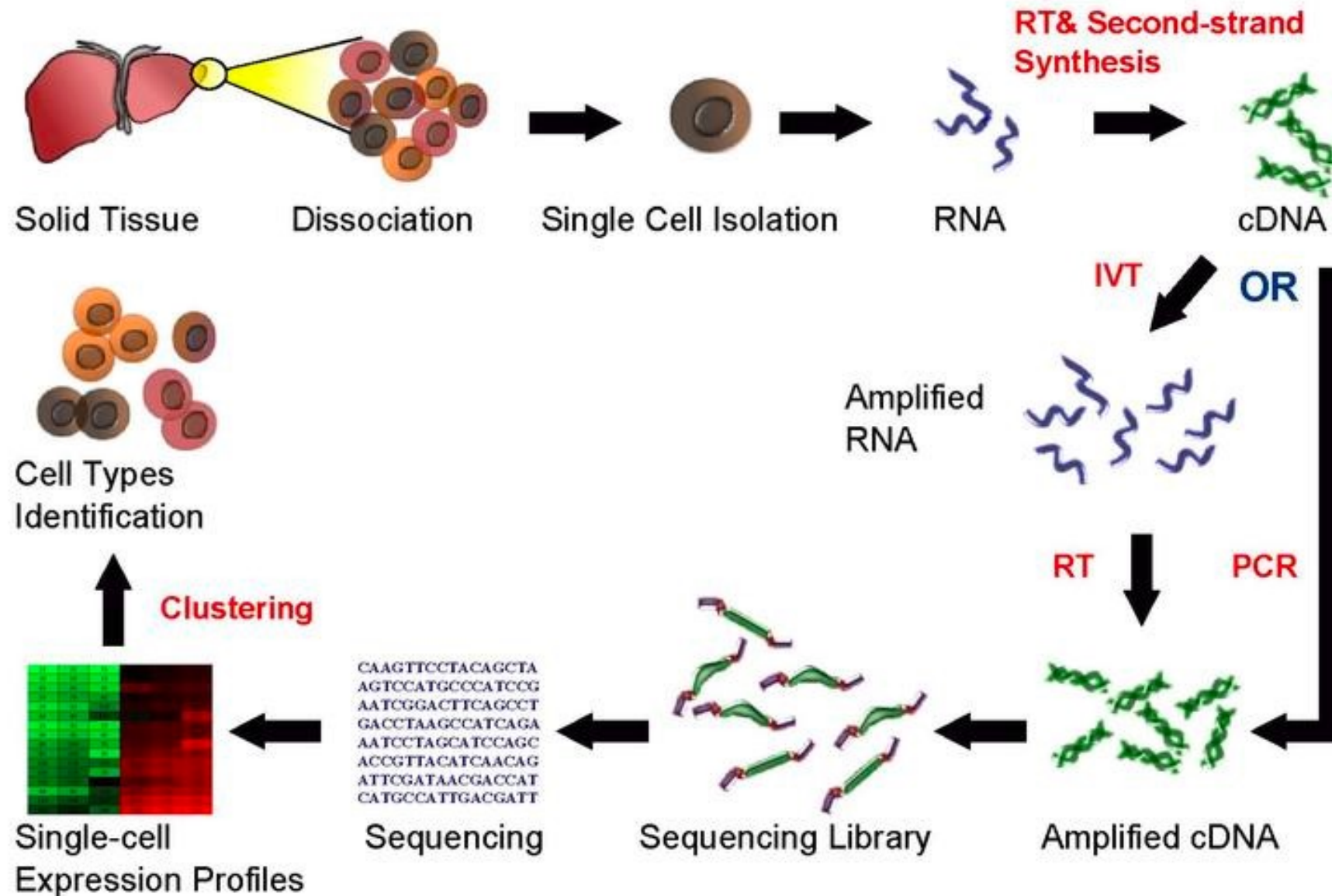
A series of human cell atlases published in May 2022: Developing human immune system



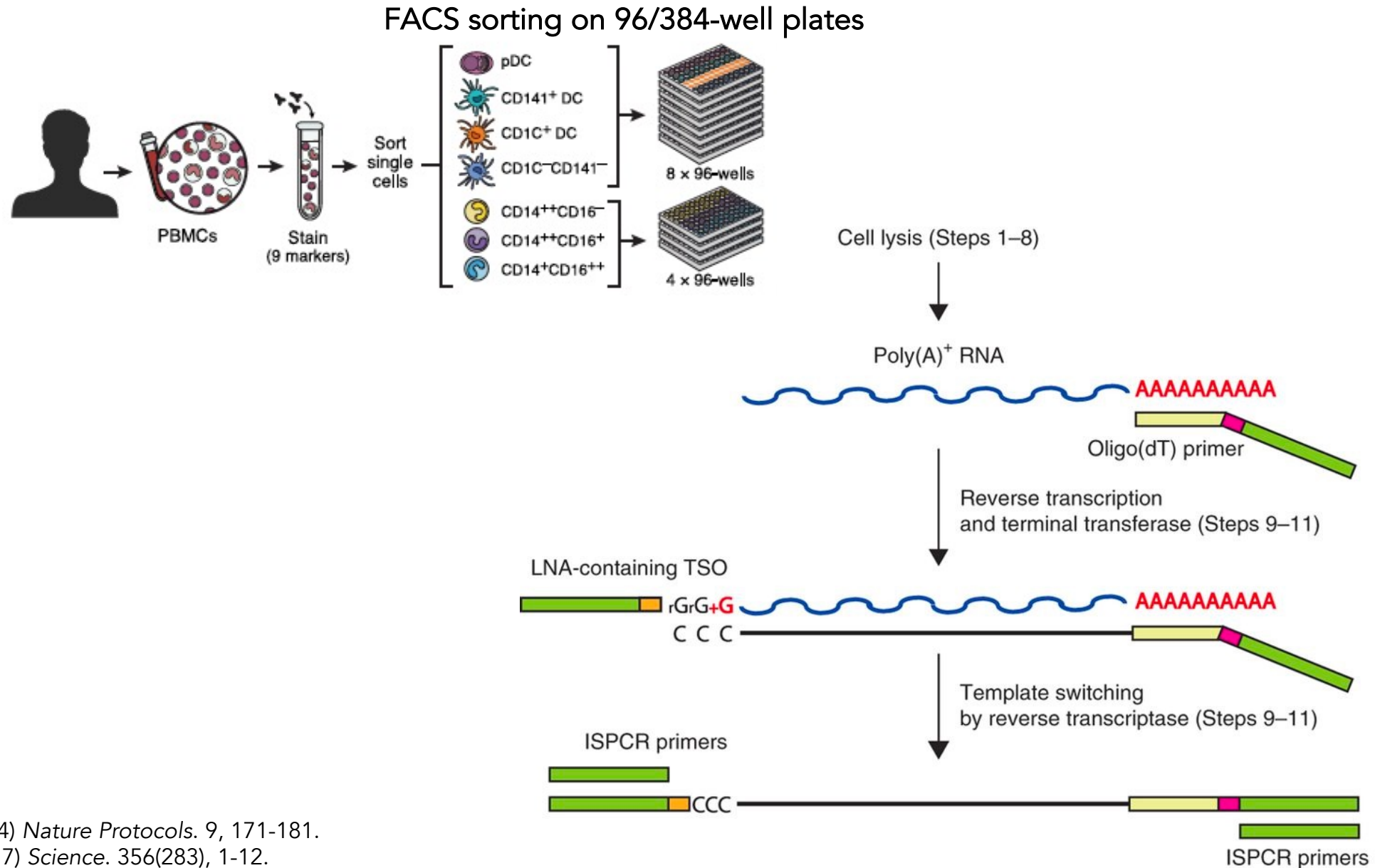
Typical RNA-Seq workflow

 mRNA from a source

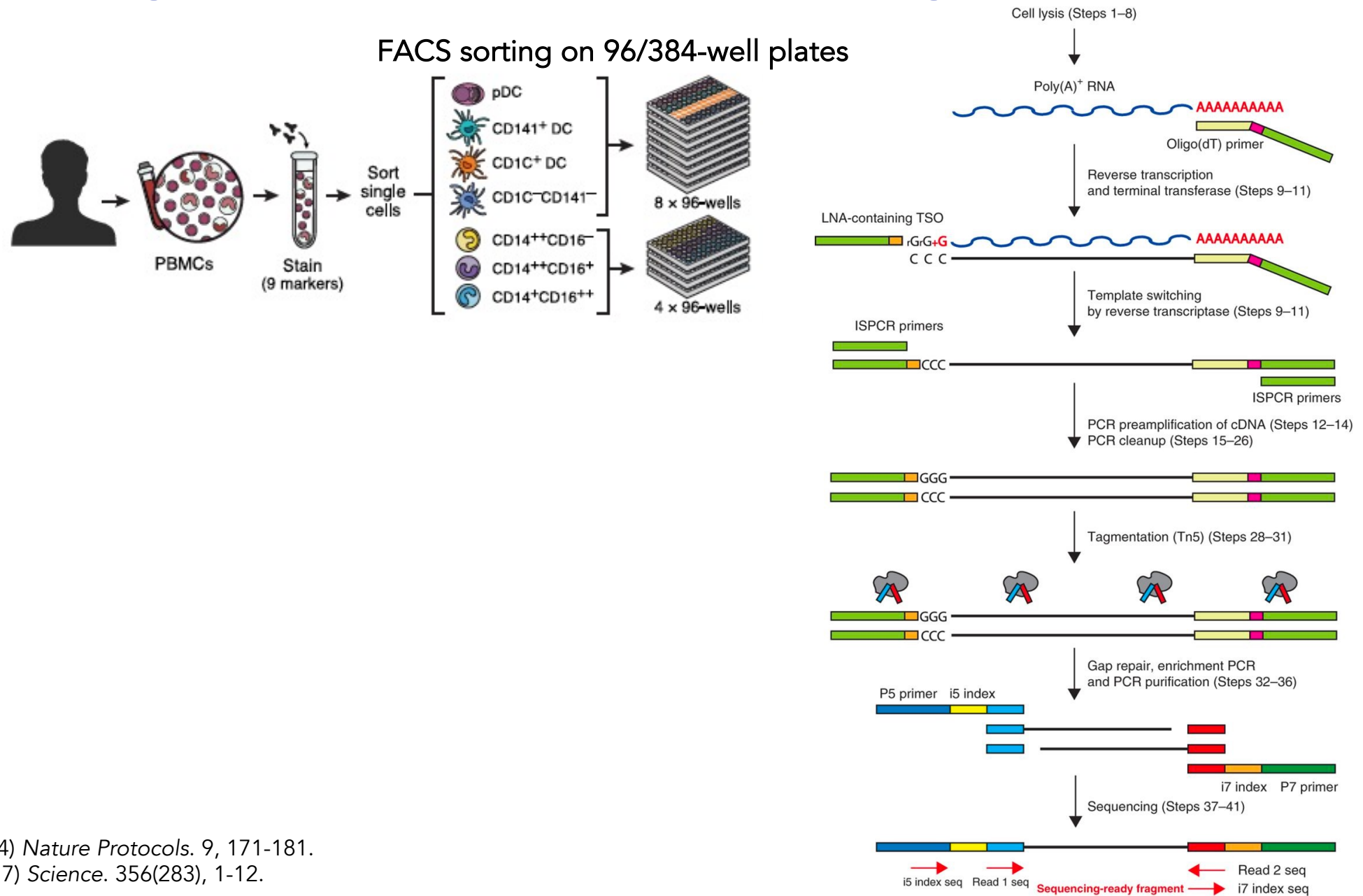
Experimental design: single cell RNA-Seq



Single cell transcriptomics using SMART-Seq2



Single cell transcriptomics using SMART-Seq2



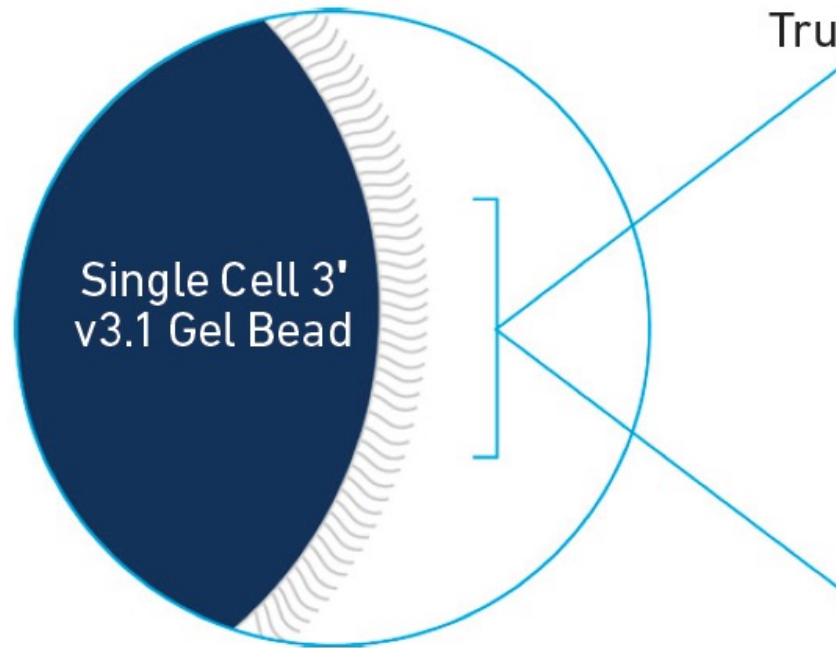
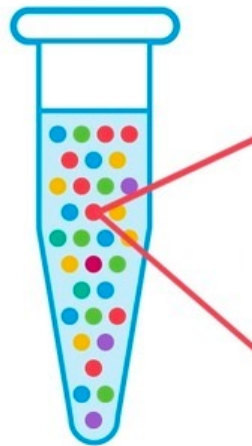
Picelli S. et al. (2014) *Nature Protocols*. 9, 171-181.

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

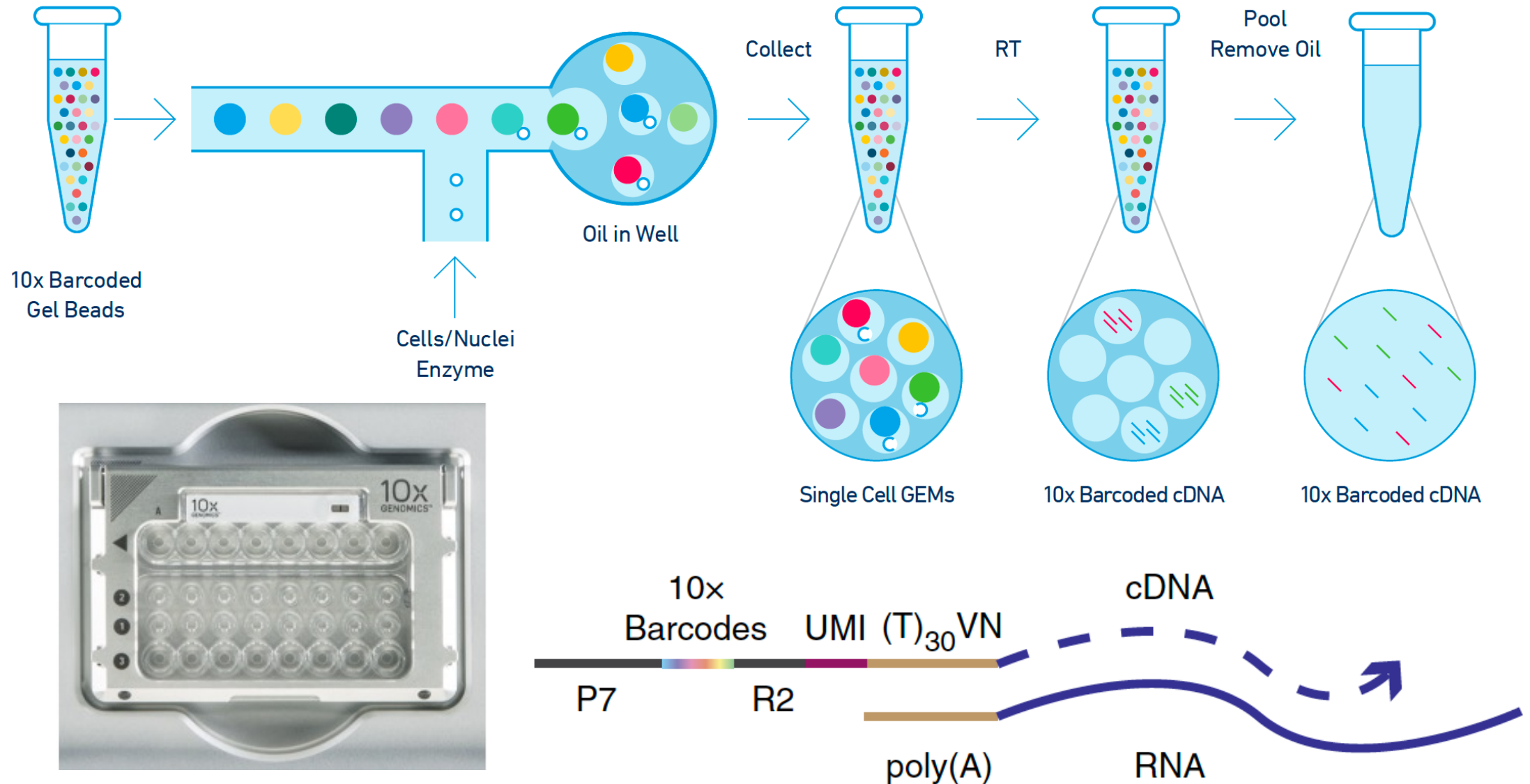
10x Barcoded
Gel Beads



TruSeq Read 1



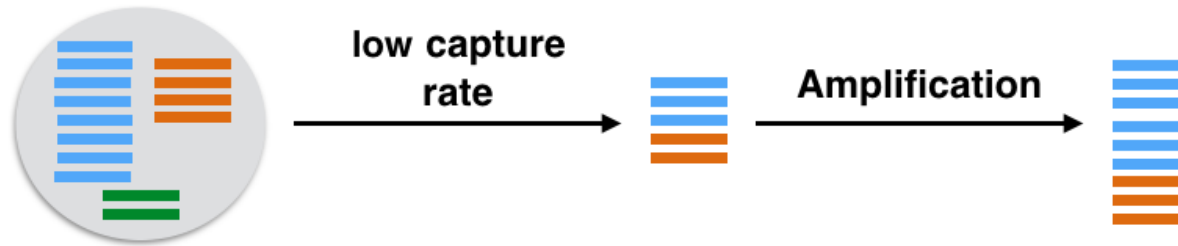
Single cell transcriptomics using 10x Chromium system



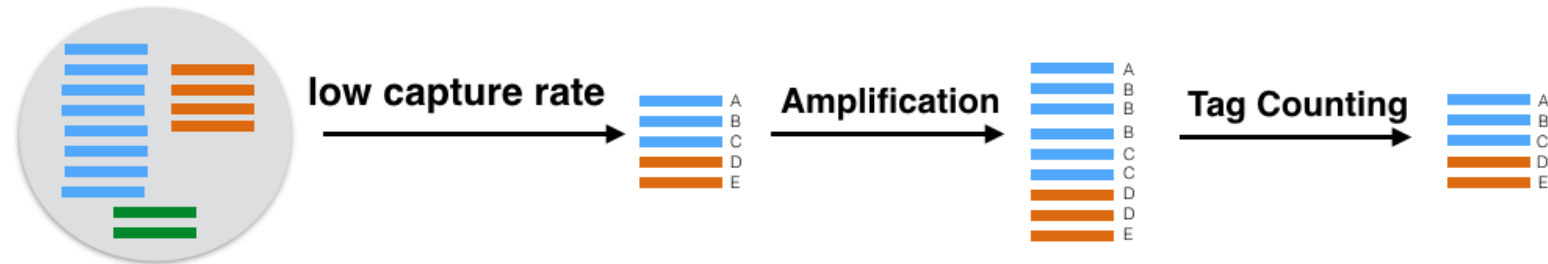
UMI : Unique Molecular Identifiers (Random Molecular Tags)

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

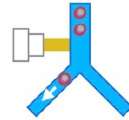


Remember : UMIs do not correct for low-capture rates, which leads to an abundance of false negatives. Capture rates are estimated to 5-20% across various protocols

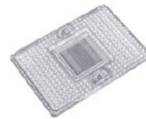
Single-cell RNA-Seq pipeline



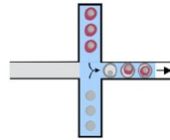
FACS



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

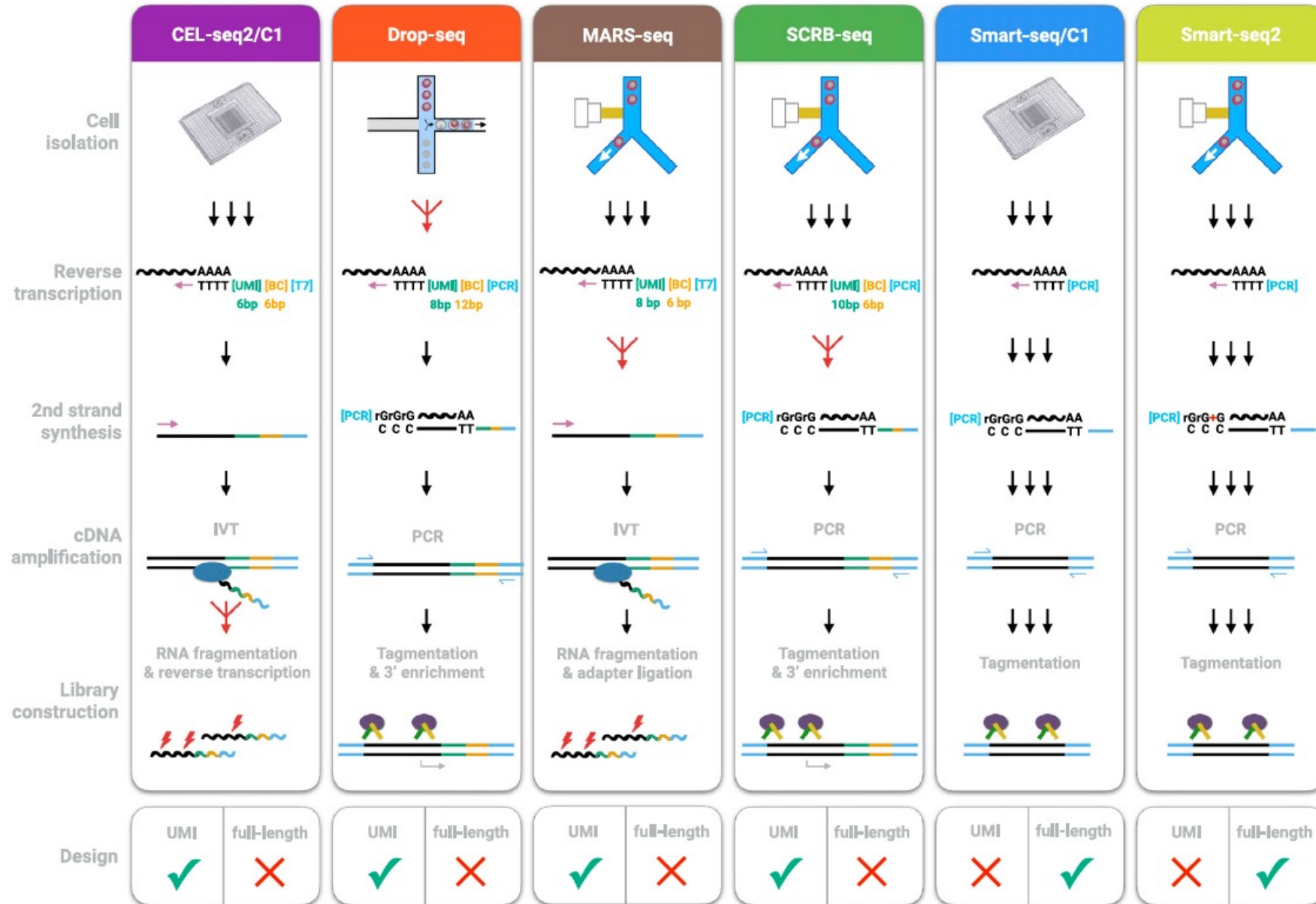
~ 500M reads total

HiSeq 4000

4 billion reads

Slide courtesy of Karthik Shekhar

There are many single-cell RNA sequencing methods



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	10 ²	10 ²	10 ^{2.5}	10 ^{2.5}	10 ^{2.5}	10 ^{3.5}	10 ^{3.5}	10 ^{3.5}	10 ^{3.5}	10 ^{4.5}	10 ^{4.5}

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

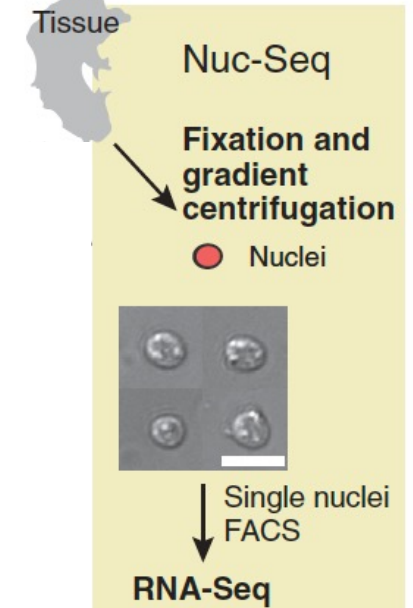
Considerations for single cell RNA-Seq

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

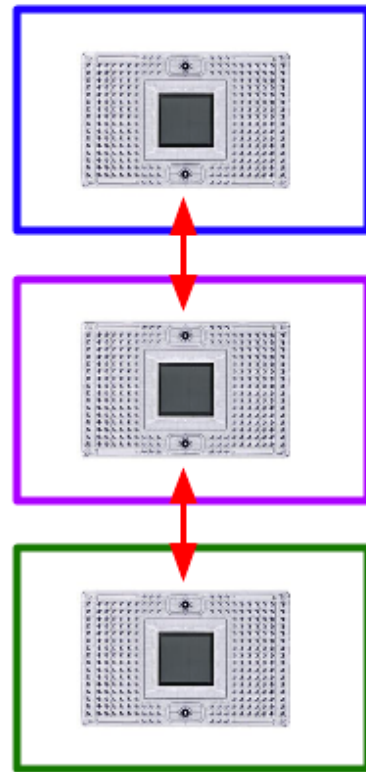
For example :

- 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)



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

Completely Confounded

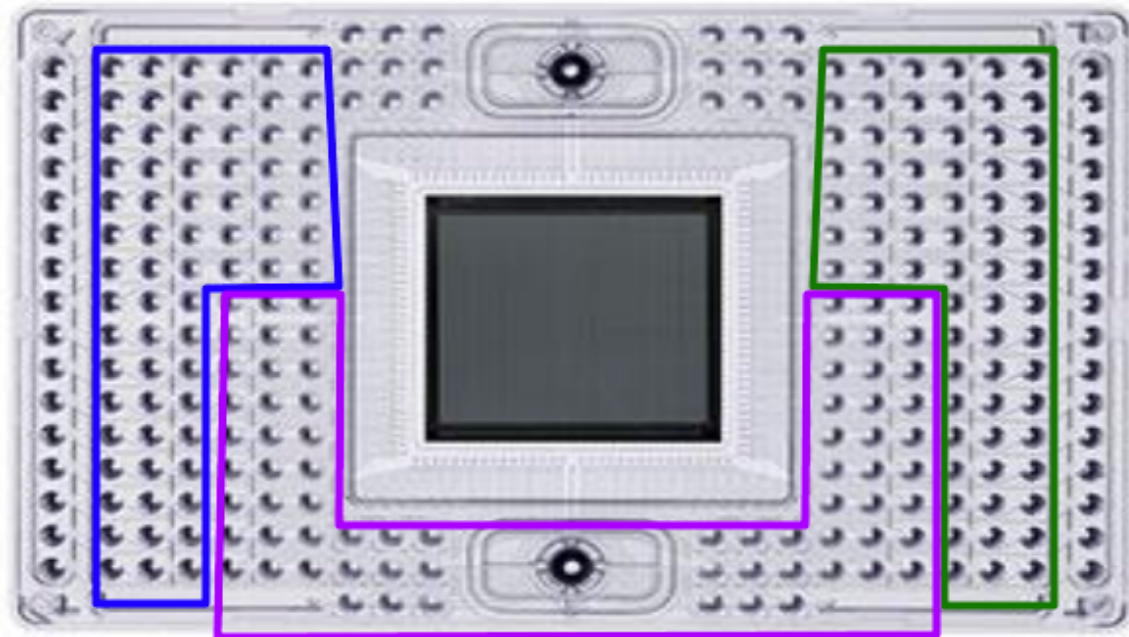


Biological
Condition 1

Biological
Condition 2

Biological
Condition 3

Unconfounded



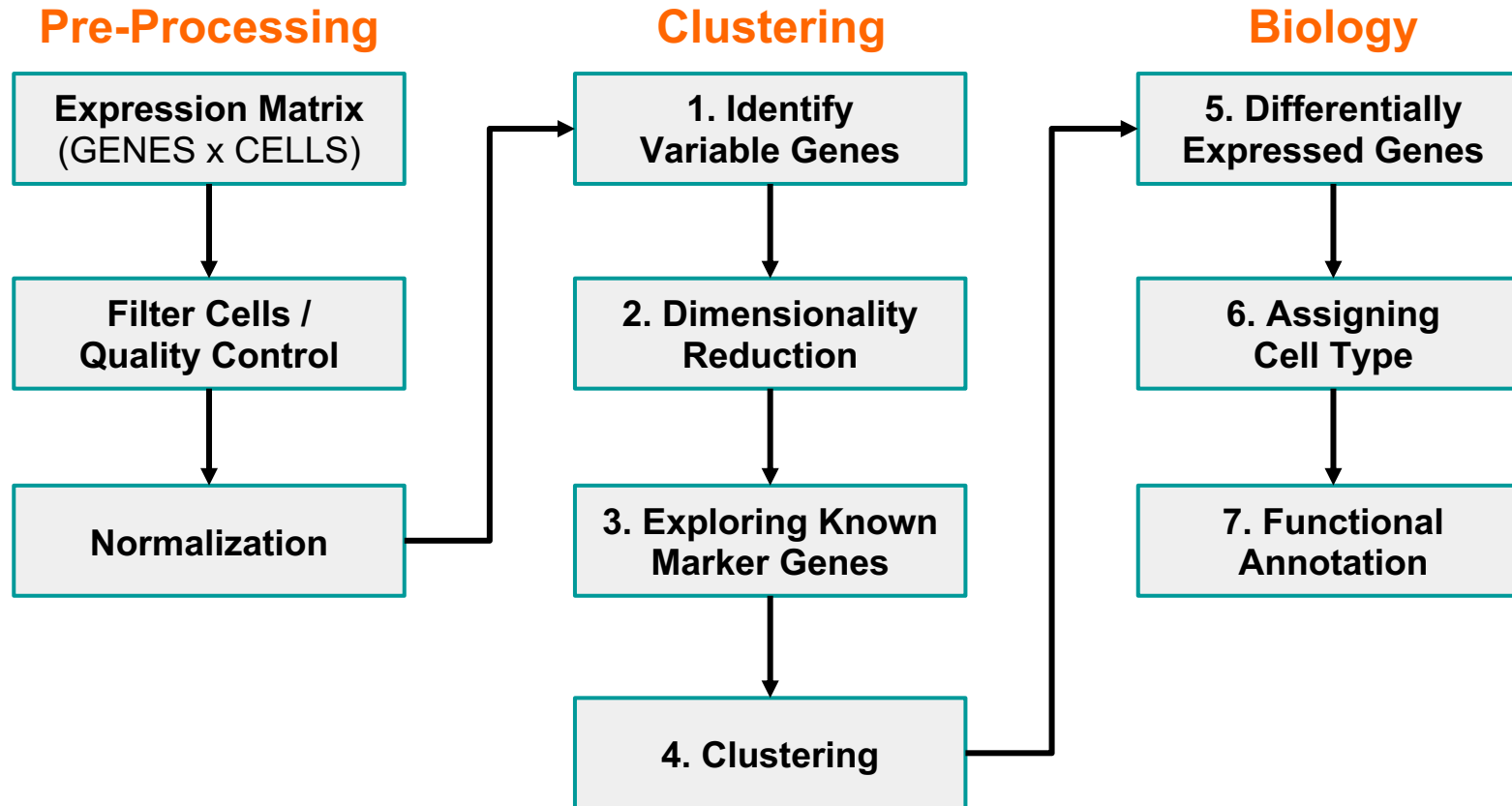
Biological
Condition 1

Biological
Condition 2

Biological
Condition 3

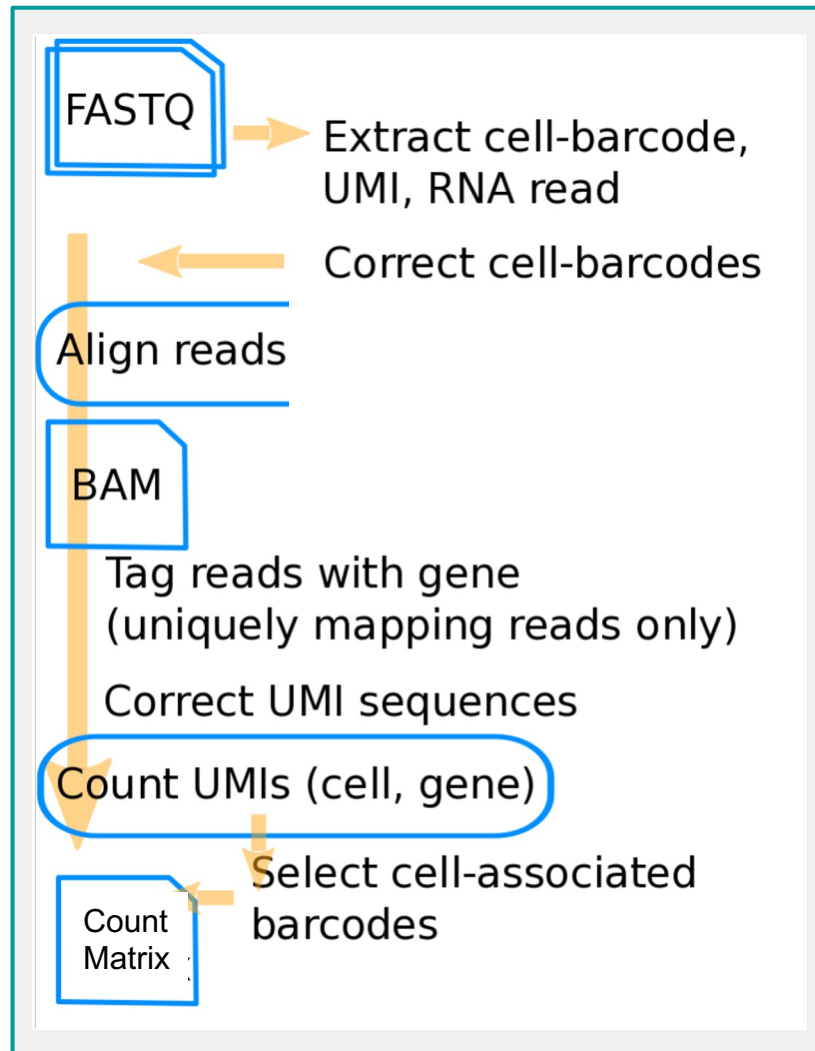
Single-cell RNA-Seq analysis pipeline

Analyzing the expression data



Single-cell RNA-Seq analysis pipeline

Generating the count matrix



Support › Single Cell Gene Expression › Software

SEARCH

Q&A NEW

CONTACT SUPPORT

SOFTWARE › OVERVIEW

CELL RANGER

Introduction

- [What is Cell Ranger?](#)
- [What is Feature Barcoding?](#)
- [Glossary](#)

Quick Start

- [Downloads](#)
- [System Requirements](#)
- [Installing Cell Ranger](#)

Tutorials

- [Getting Started with Cell Ranger](#)
- [Example Data Analysis](#)

Running Pipelines

- [Computing Options](#)
- [mkfastq](#)
- [Specifying Input FASTQs](#)
- [count \(Gene Expression\)](#)
- [count \(Feature Barcoding\)](#)
- [count \(Feature Barcoding Only\)](#)
- [aggr](#)
- [reanalyze](#)
- [Troubleshooting](#)

Understanding Outputs

Algorithms Overview

Advanced

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](#)

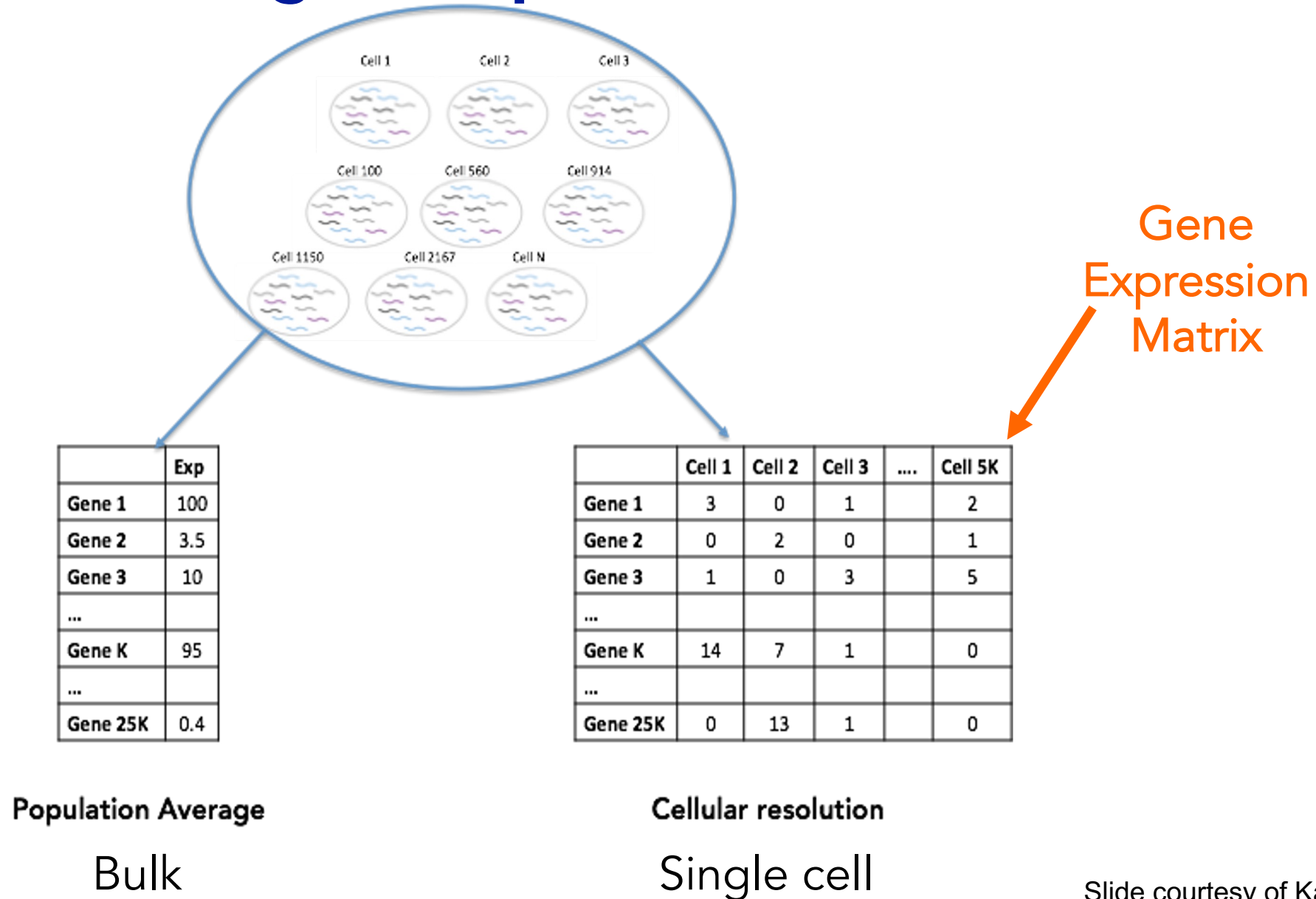
ALGORITHMS OVERVIEW

[Gene Expression](#)
[Antibody](#)
[CRISPR](#)

OTHER RESOURCES

[Questions & Answers](#)
[Datasets](#)

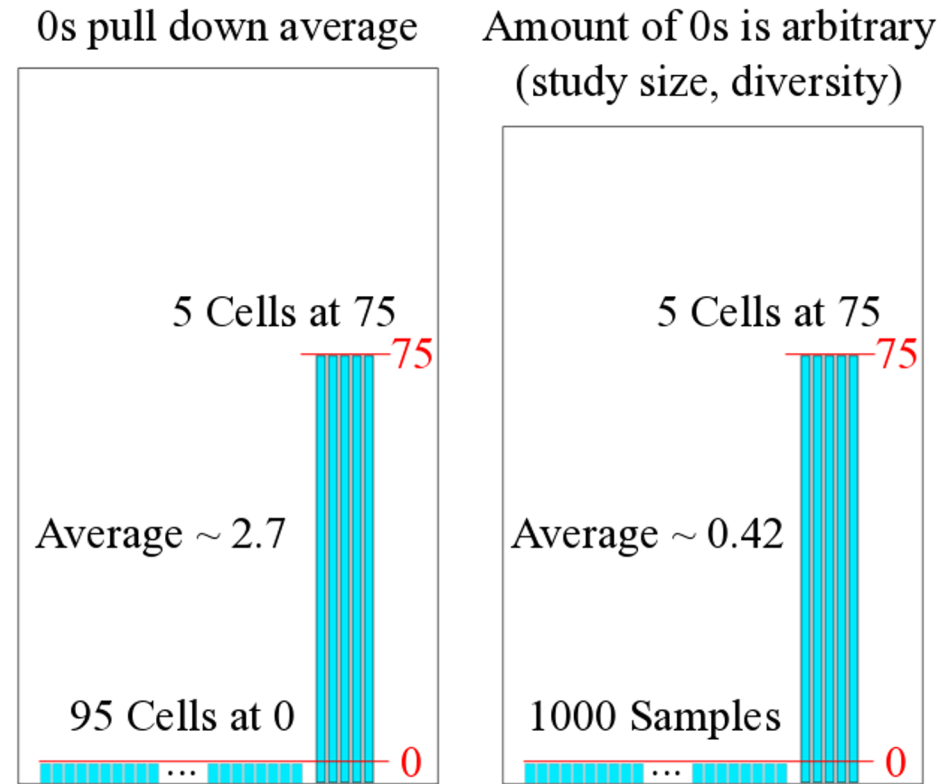
Single-cell gene expression distributions are very different from bulk gene expression distributions



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

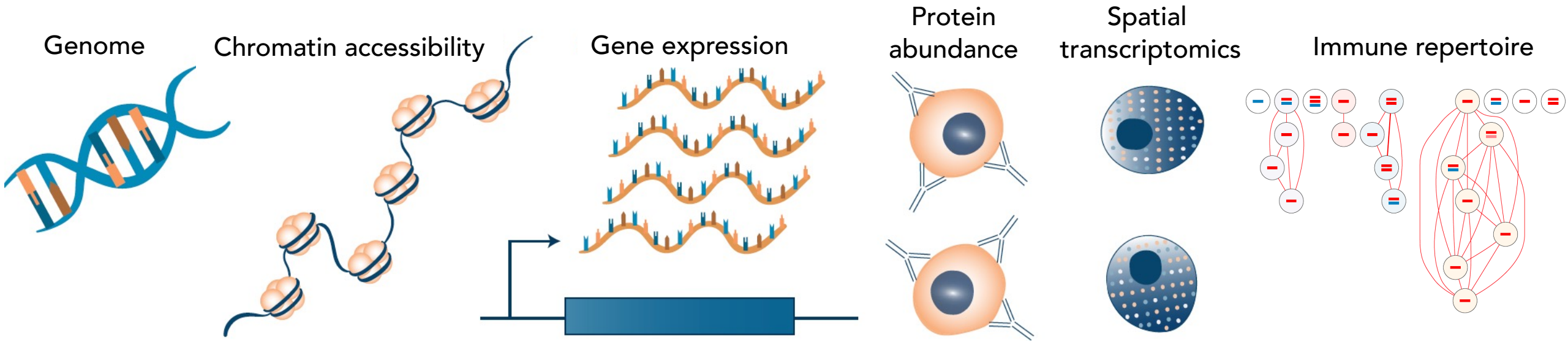
- measures the distribution of expression levels for each gene across a population of single cells
- can study $1e2$ - $1e6$ cells in an experiment

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

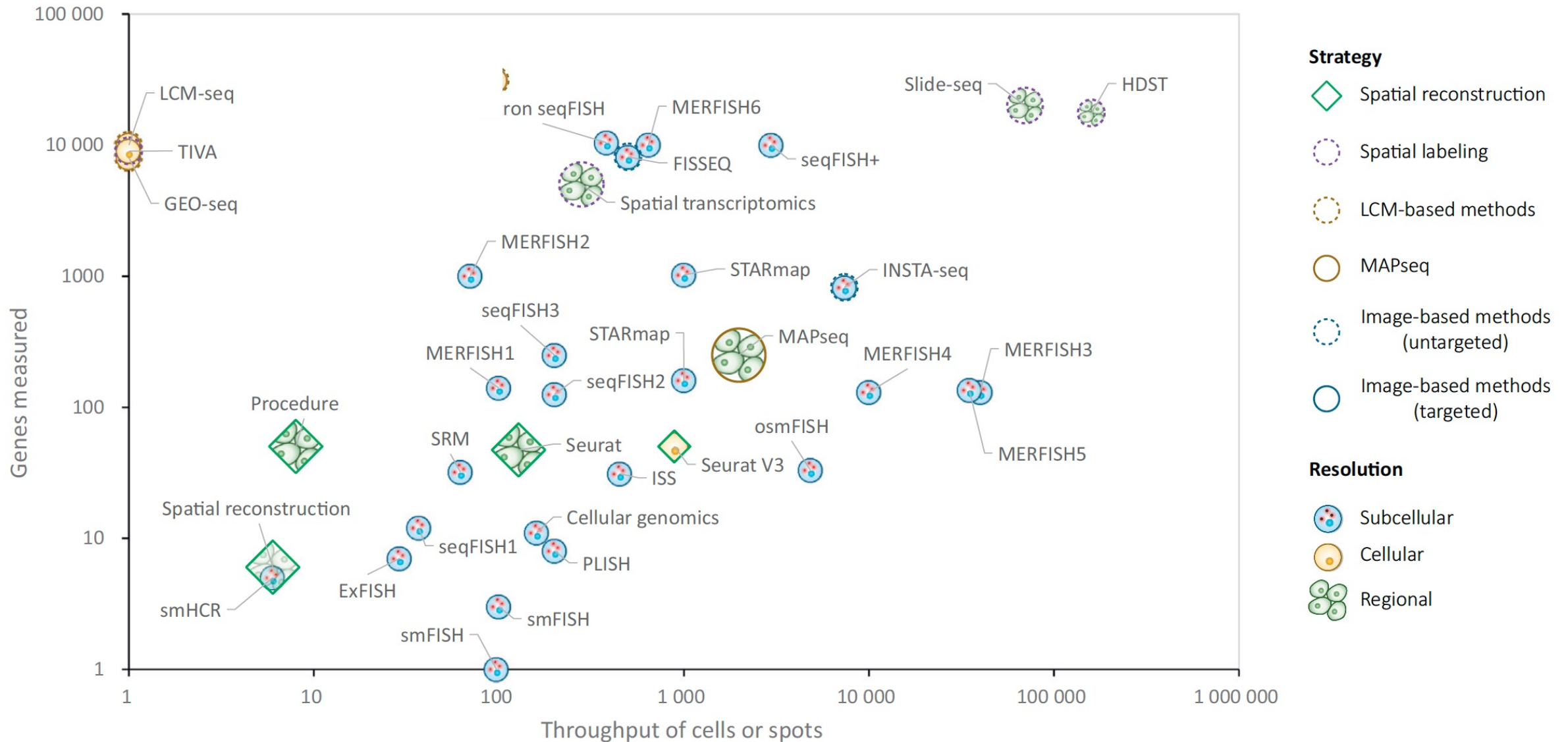
Proteins

Multi-parameter flow
Mass cytometry
Single cell proteomics

'multi-omics'

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

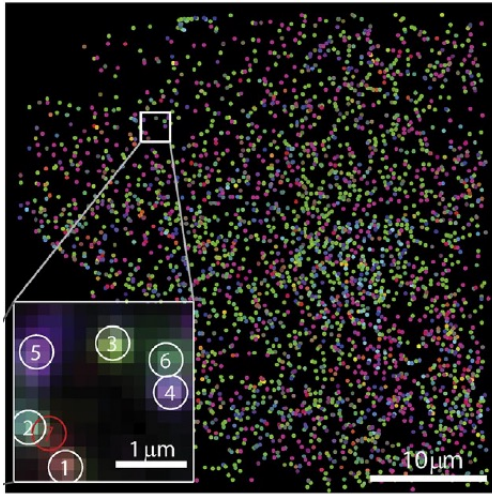
Ongoing developments in single-cell genomics: Growing toolbox for spatial genomics (profiling cells *in situ*)



Ongoing developments in single-cell genomics:

Growing toolbox for spatial genomics (profiling cells *in situ*)

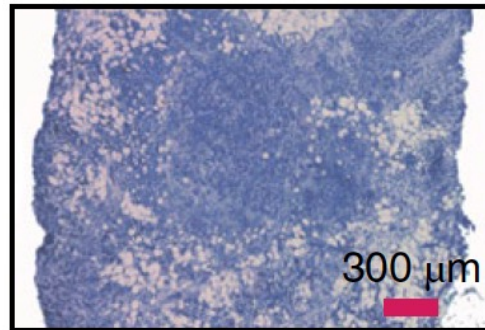
MERFISH (RNA)



breast cancer

High Density Spatial Transcriptomics (RNA)

H&E

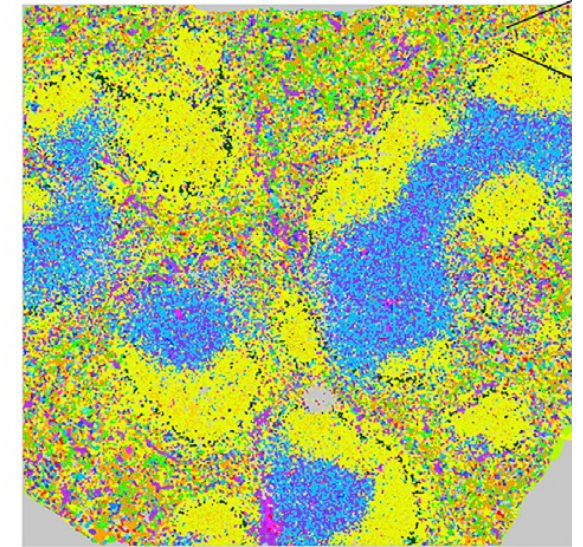


Annotations



- 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

CODEX (protein)



mouse spleen

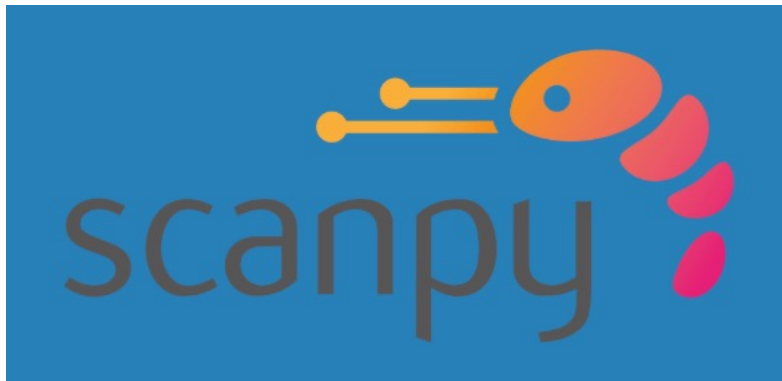
- capsule
- CD31(hi) vascular
- plasma cells
- CD106(-)CD16/32(-)Ly6C(+)CD31(+) stroma
- CD4(+)CD8(-)cDC
- CD4(+)CD8(-)cDC
- CD4(+)MHCII(+) stroma
- CD4(-)CD8(-)cDC
- CD4(-)CD8(-)cDC
- CD106(+)CD16/32(-)Ly6C(+)CD31(+)
- CD106(-)CD16/32(-)Ly6C(+)CD31(-)
- CD106(+)CD16/32(-)Ly6C(-)CD31(+) stroma
- granulocytes
- CD3(+) other markers (-)
- NK cells
- F4/80(+) mps
- erythroblasts
- CD11c(-) B cells
- CD106(+)CD16/32(-)Ly6C(-)CD31(-) stroma
- FDcs
- marginal zone mps
- B cells
- CD8(+) T cells
- CD4(+) T cells
- no id
- B220(+) DN T cells

Moffitt J.R. et al. (2016) *PNAS* 113: 11046-11051.

Goltsev Y. et al. (2018) *Cell*. 174: 968-981.

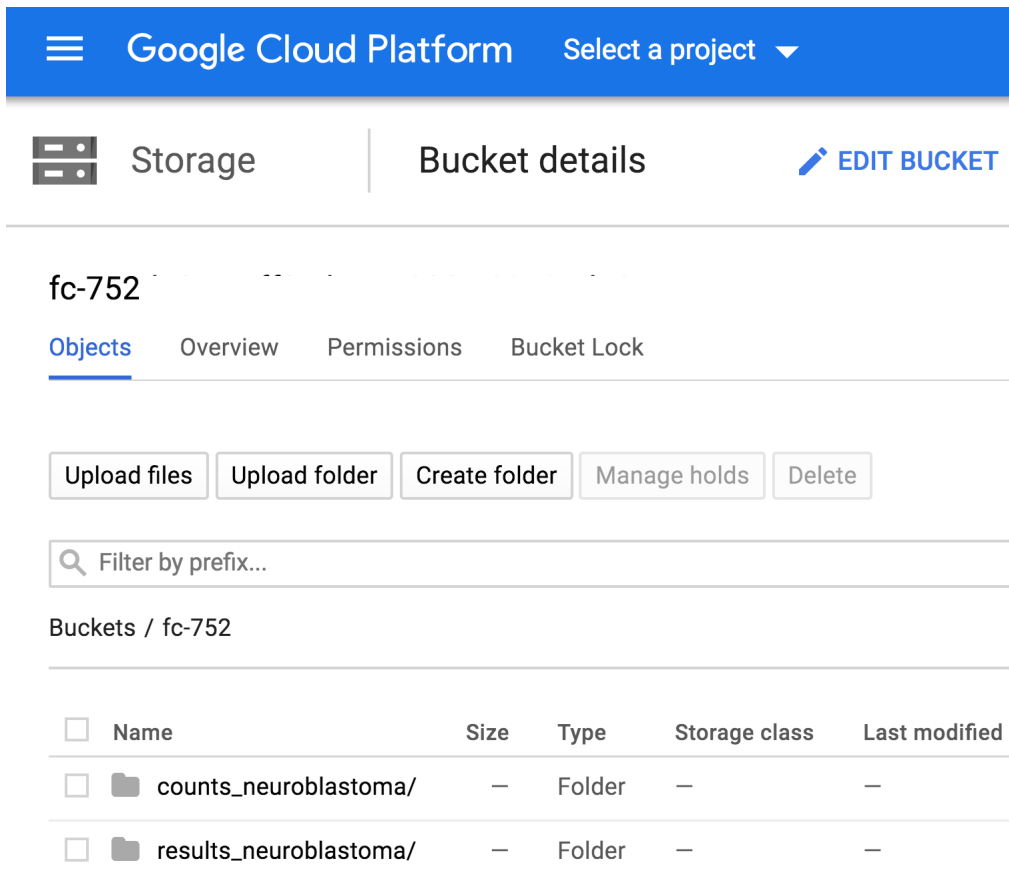
Vickovic S. et al. (2019) *Nature Methods*. 16: 987-990.

Seurat and scanpy: single cell analysis toolkits



Single-cell analysis computations in the Cloud

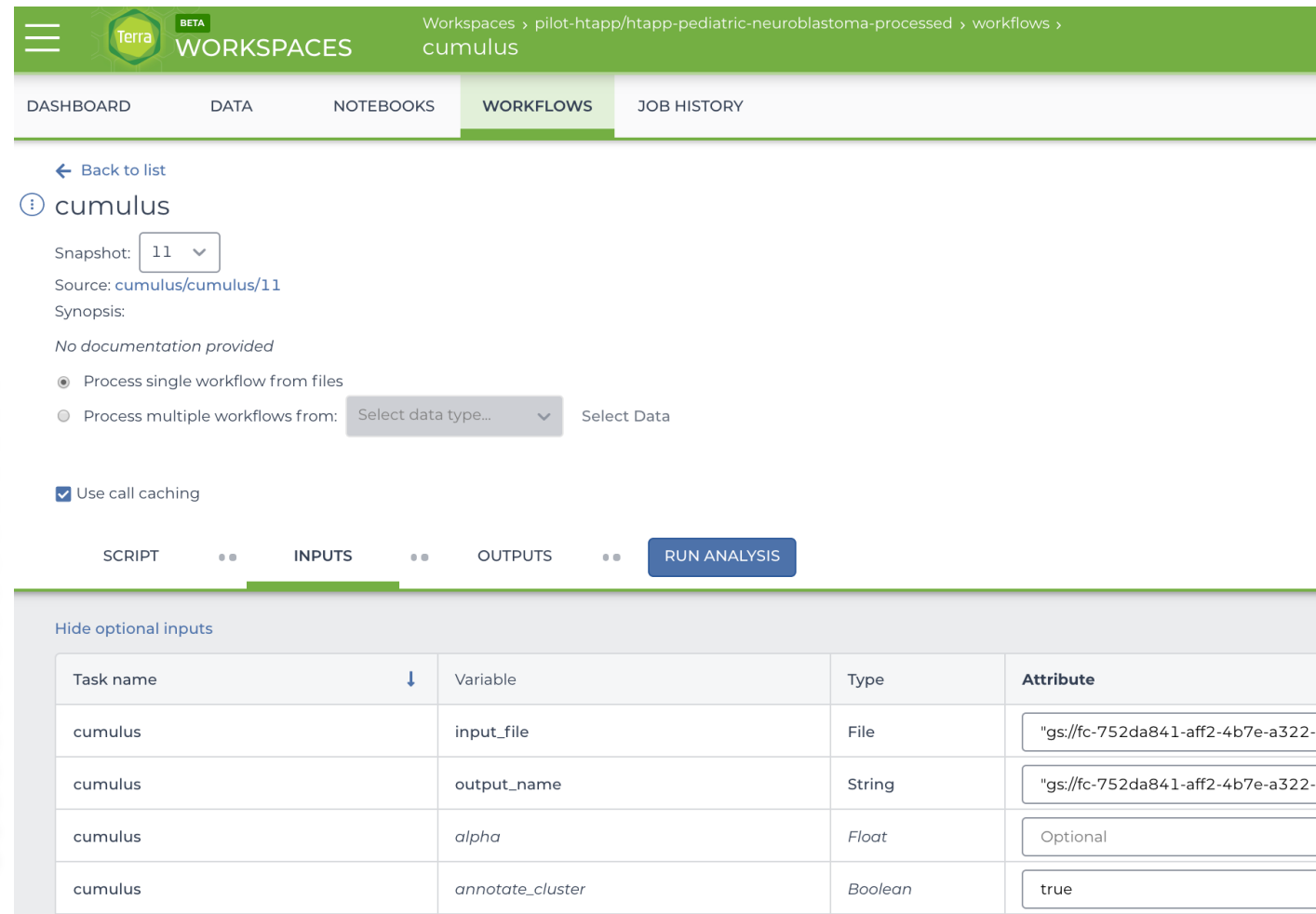
Inputs and outputs are stored in Google Cloud storage buckets



The screenshot shows the Google Cloud Platform Storage interface. At the top, there's a blue header with the Google Cloud Platform logo and a 'Select a project' dropdown. Below this, the 'Storage' section is active, showing 'Bucket details' for 'fc-752'. There are tabs for 'Objects', 'Overview', 'Permissions', and 'Bucket Lock'. Below the tabs, there are buttons for 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete'. A search bar labeled 'Filter by prefix...' is present. Below the search bar, the path 'Buckets / fc-752' is shown. A table lists the contents of the bucket:

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified
<input type="checkbox"/>	counts_neuroblastoma/	—	Folder	—	—
<input type="checkbox"/>	results_neuroblastoma/	—	Folder	—	—

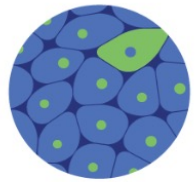
Data processing and analysis are performed in Terra



The screenshot shows the Terra WORKSPACES interface. At the top, there's a green header with the Terra logo and 'WORKSPACES'. Below this, there's a navigation bar with 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. The 'WORKFLOWS' tab is active, showing details for a workflow named 'cumulus'. There's a 'Back to list' link and a 'Snapshot' dropdown set to '11'. The source is 'cumulus/cumulus/11'. The synopsis is 'No documentation provided'. There are two options: 'Process single workflow from files' (selected) and 'Process multiple workflows from:'. Below these, there's a 'Select data type...' dropdown and a 'Select Data' button. There's a checkbox for 'Use call caching' which is checked. At the bottom, there's a 'RUN ANALYSIS' button. Below the button, there's a table with columns 'Task name', 'Variable', 'Type', and 'Attribute':

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

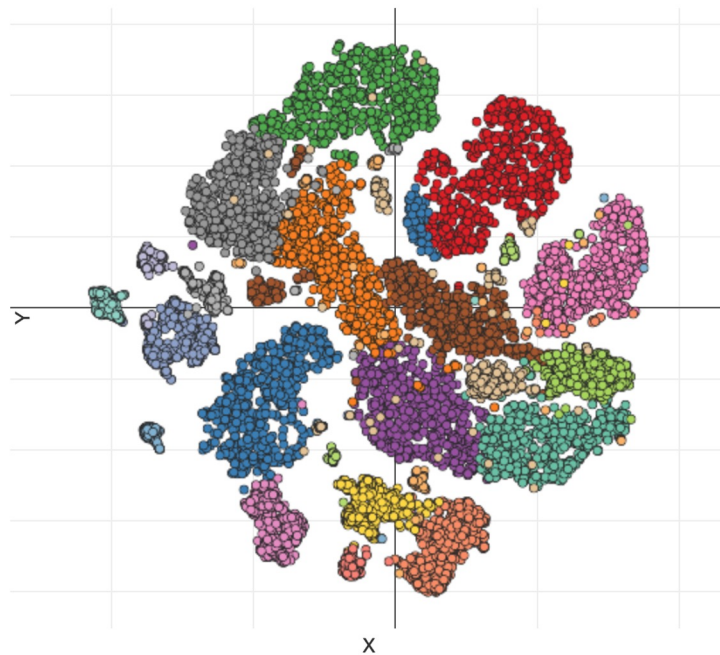
Single Cell Portal facilitates sharing of single-cell studies



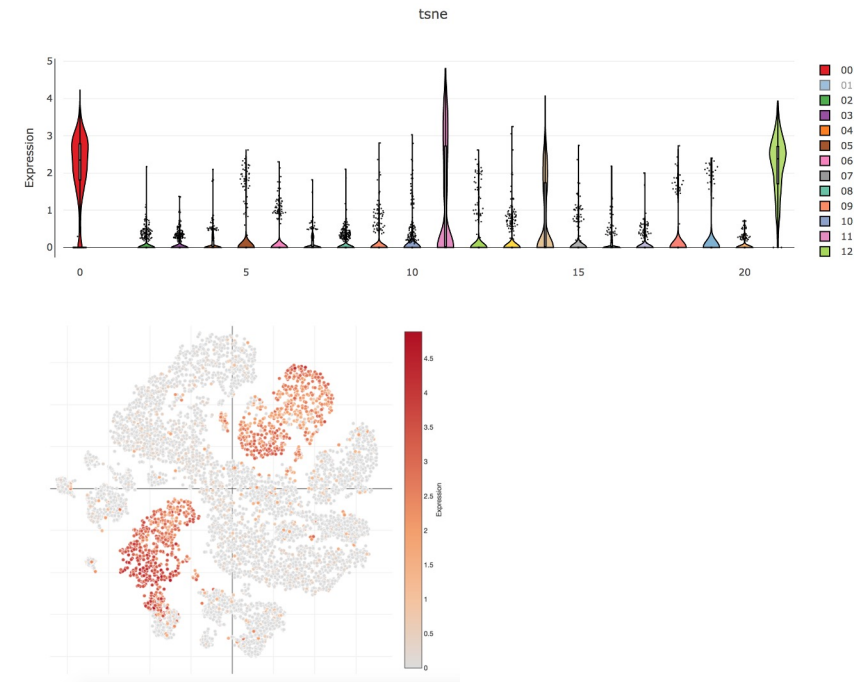
Single Cell
PORTAL

Reducing barriers and accelerating single-cell research

Featuring
482 studies
25,385,357 cells



- clustering: 00 (538 points)
- clustering: 01 (538 points)
- clustering: 02 (538 points)
- clustering: 03 (537 points)
- clustering: 04 (537 points)
- clustering: 05 (537 points)
- clustering: 06 (537 points)
- clustering: 07 (537 points)
- clustering: 08 (537 points)
- clustering: 09 (537 points)
- clustering: 10 (537 points)
- clustering: 11 (537 points)
- clustering: 12 (537 points)
- clustering: 13 (537 points)
- clustering: 14 (537 points)
- clustering: 15 (440 points)
- clustering: 16 (365 points)
- clustering: 17 (291 points)
- clustering: 18 (270 points)
- clustering: 19 (230 points)
- clustering: 20 (204 points)
- clustering: 21 (142 points)



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.