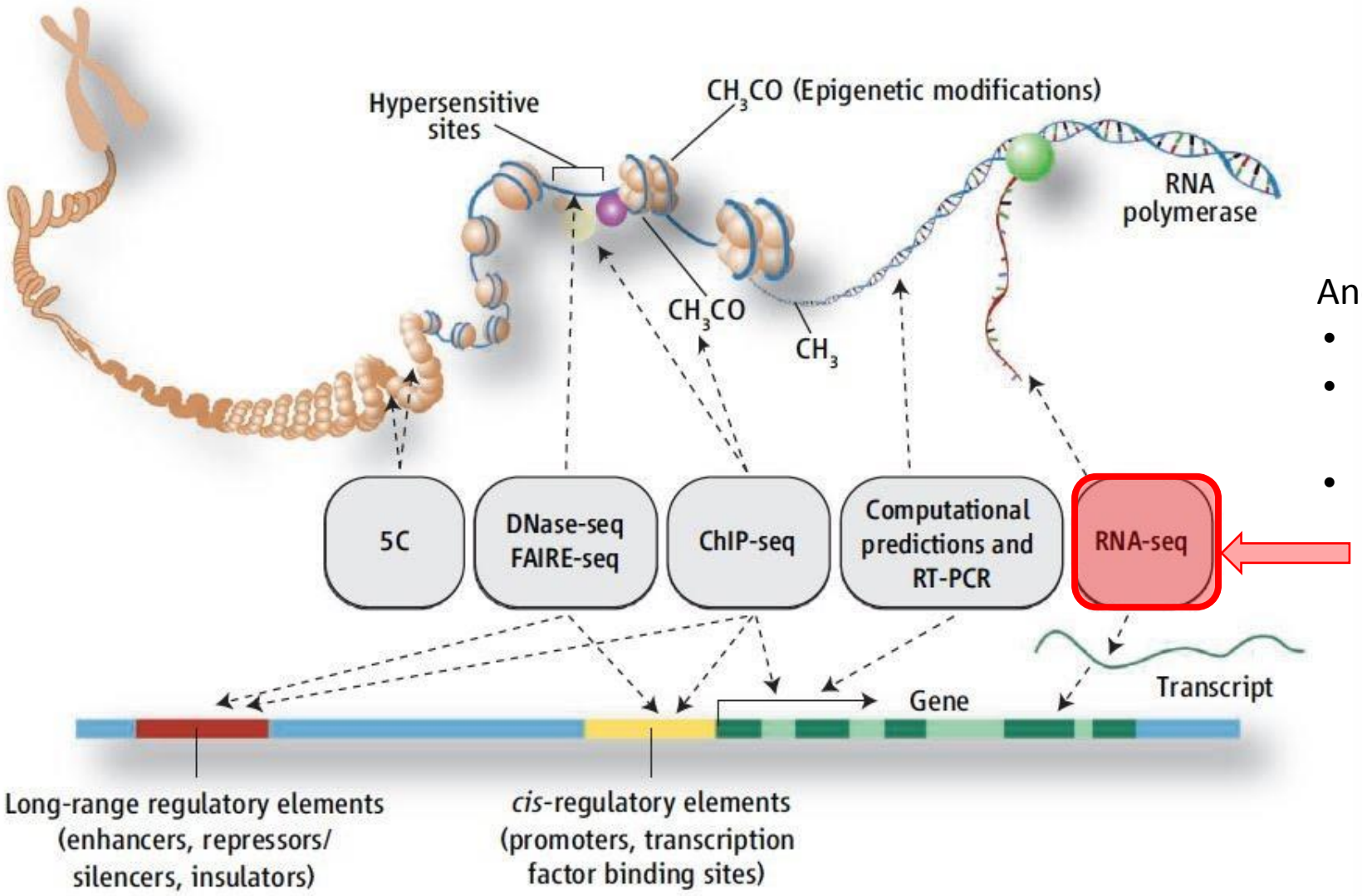
State of the art in the field of single-cell biology

Vincent Gardeux

Deplancke's Laboratory of Systems Biology and Genetics

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

There are multiple genomic layers that can be measured



And

- DNA Methylation
- Nucleosome position (ATAC-seq)
- etc…

One genome: diverse functional outputs

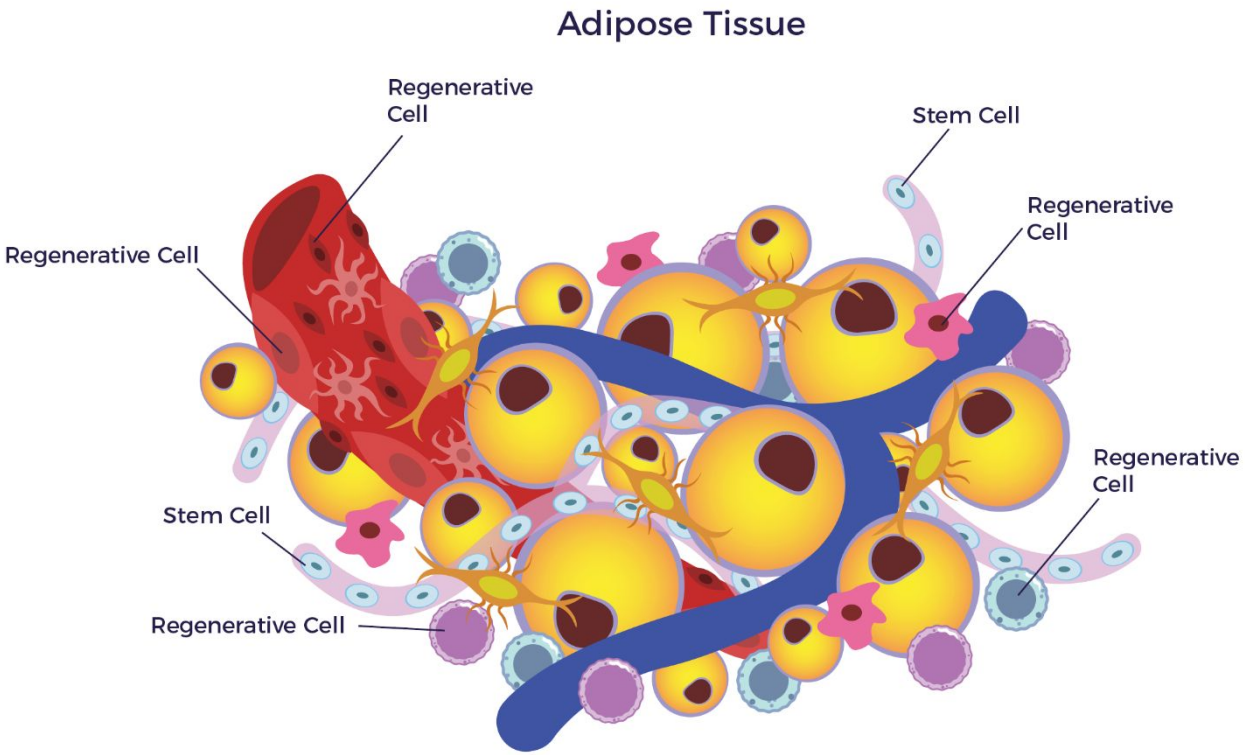
1 genome

One genome gives rise to a multitude of different cell types with highly distinct morphologies & functions

Tissues are generally heterogeneous

Remarkable Cellular Diversity and Specialisation

Human body:
~ 100 trillion cells
~ 200 types of cells
complex
tissue & organ
functions



anatomyandphysiology.com

Bulk RNA-seq: estimate expression of transcripts in a sample

How to measure gene expression?

⇒ Bulk RNA-seq

Technique appeared in 2008

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi[1,*], Zhong Wang[1,*], Karl Waern[1], Chong Shou[2], Debasish Raha[1], Mark Gerstein[2,3], Michael Snyder[1,2,3,†]

Abstract

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

Nagalakshmi et al, Science 2008

Limitations of bulk RNA-seq

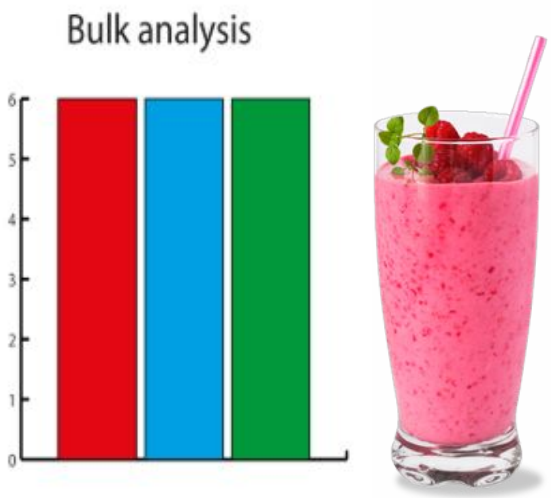
Bulk RNA-seq was a major breakthrough in the late 00's (replaced microarrays)
=> Great advances were made through genomics, but …

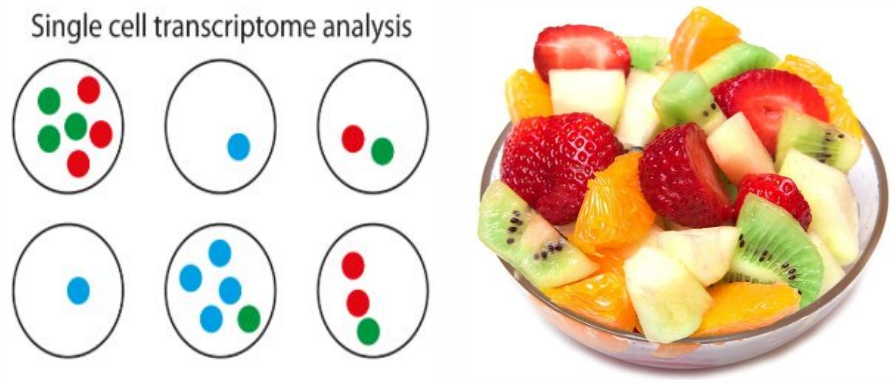
Limitations: minimum starting material requirements
techniques applied on <u>millions of cells</u>

- ***rare cell types & states*** cannot be analyzed (e.g. transitions, circulating tumor cells, etc..)
- ***insufficient*** for studying heterogeneous systems (e.g. complex tissues such as brain)

*! **<u>each sample is an AVERAGE!</u>***
no idea of the underlying values in single cells
of the heterogeneity of the tissue

Single-cell RNA-seq





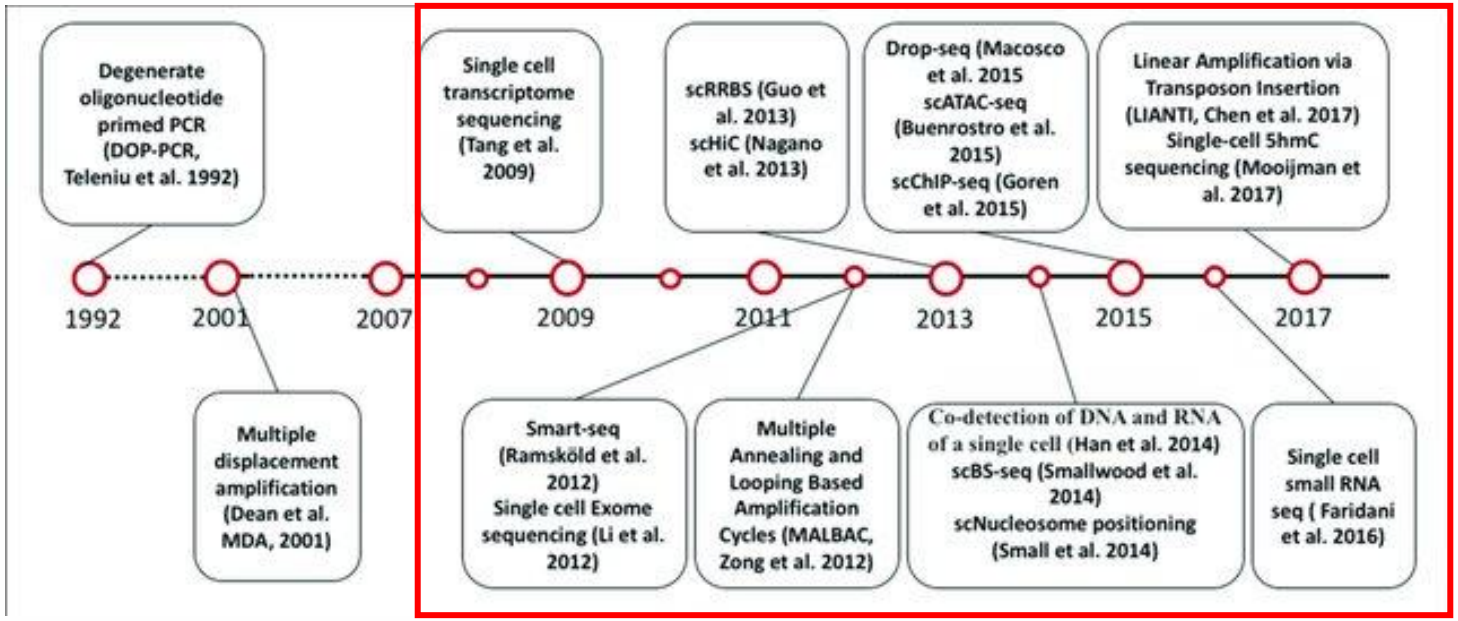
Macaulay IC, Voet T (2014) PLoS Genet

Each black circle is a cell, each colored dot is a transcript, colors encode transcript of the same gene

RARE CELL TYPES (e.g. early development, stem cells, circulating tumor cells)
HETEROGENEITY (e.g. tissue composition, cancer, temporal processes)
GENE REGULATORY NETWORKS (non-confounded correlations)
SINGLE-CELL PHENOMENA* (gene expression stochasticity, mono-allelic expression)

** see also review CoulonLarsonNatRevGenet2013*

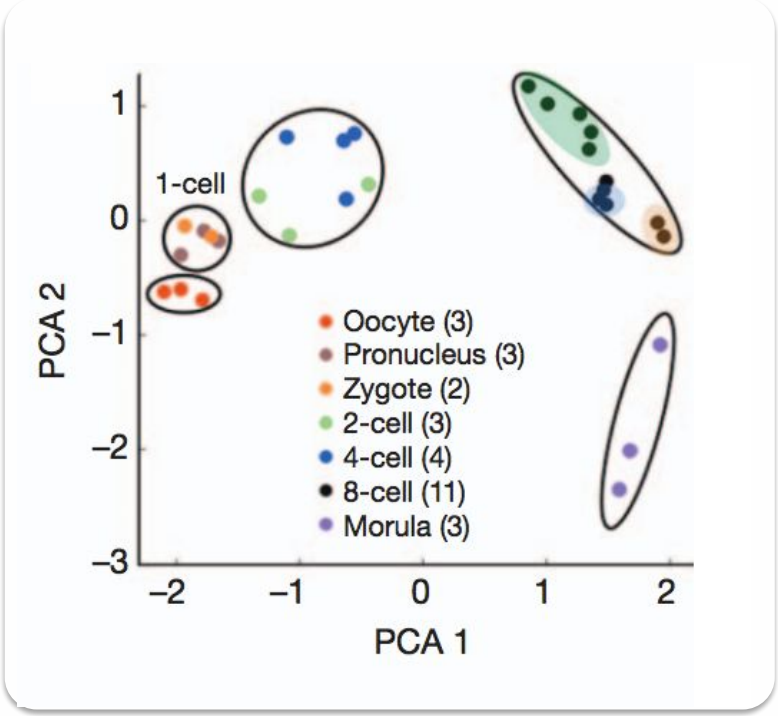
Single-cell timeline

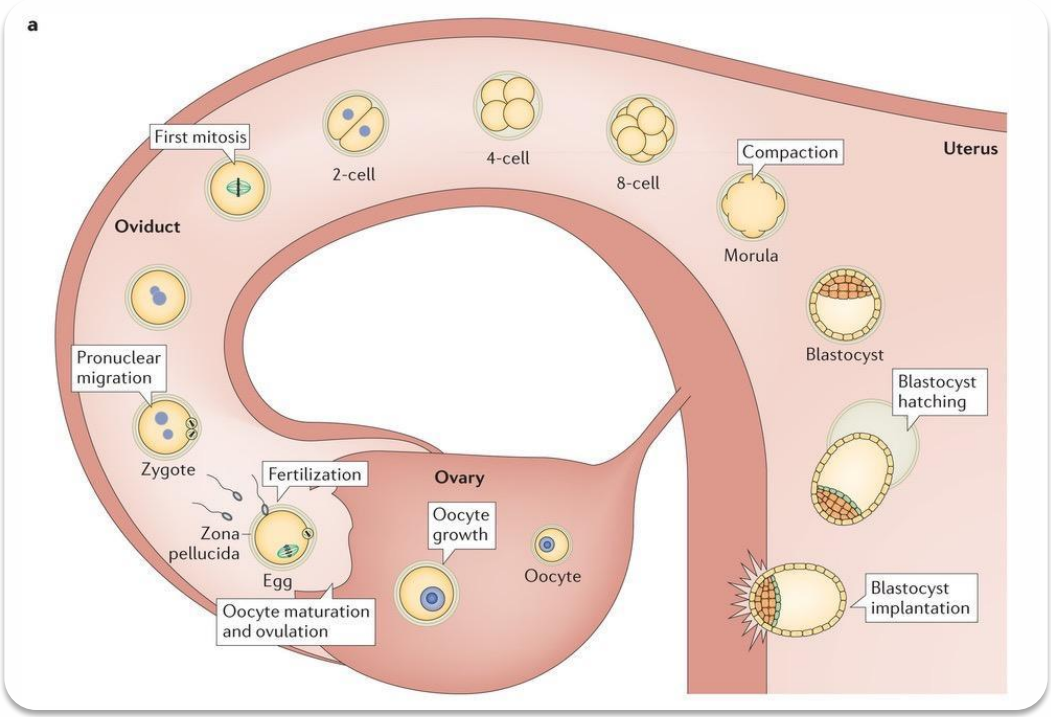


Hu, Y., An, et al (2018). Single cell multi-omics technology: methodology and application. Frontiers in cell and developmental biology, 6, 28

Single-cell transcriptomics (scRNA-seq) applications – Development

Analyzing transcriptome of cells in human and mouse early embryos





⇒ Each developmental stage can be delineated concisely by a small number of functional modules of co-expressed genes

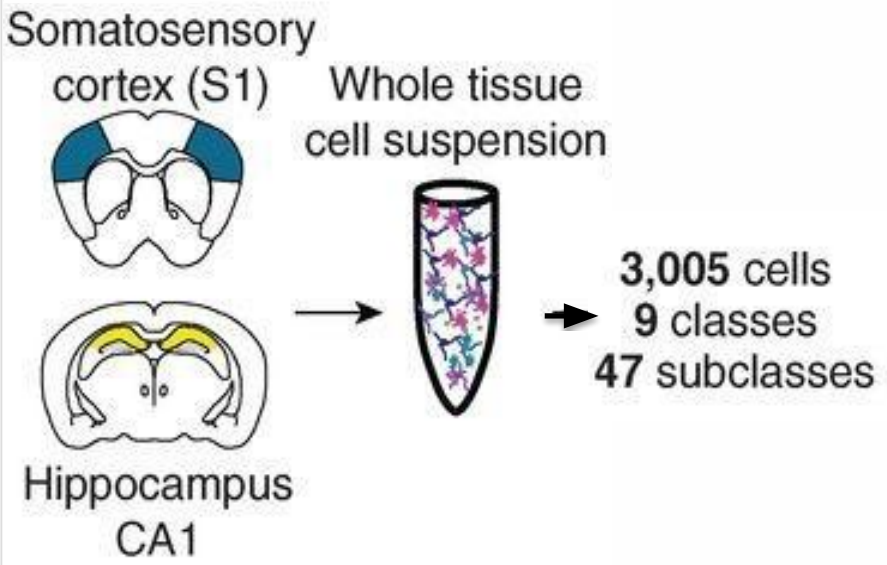
⇒ Temporal developmental pattern different mouse-human

⇒ Conserved key members of human & mouse networks

Xue et al. (2013) Nature

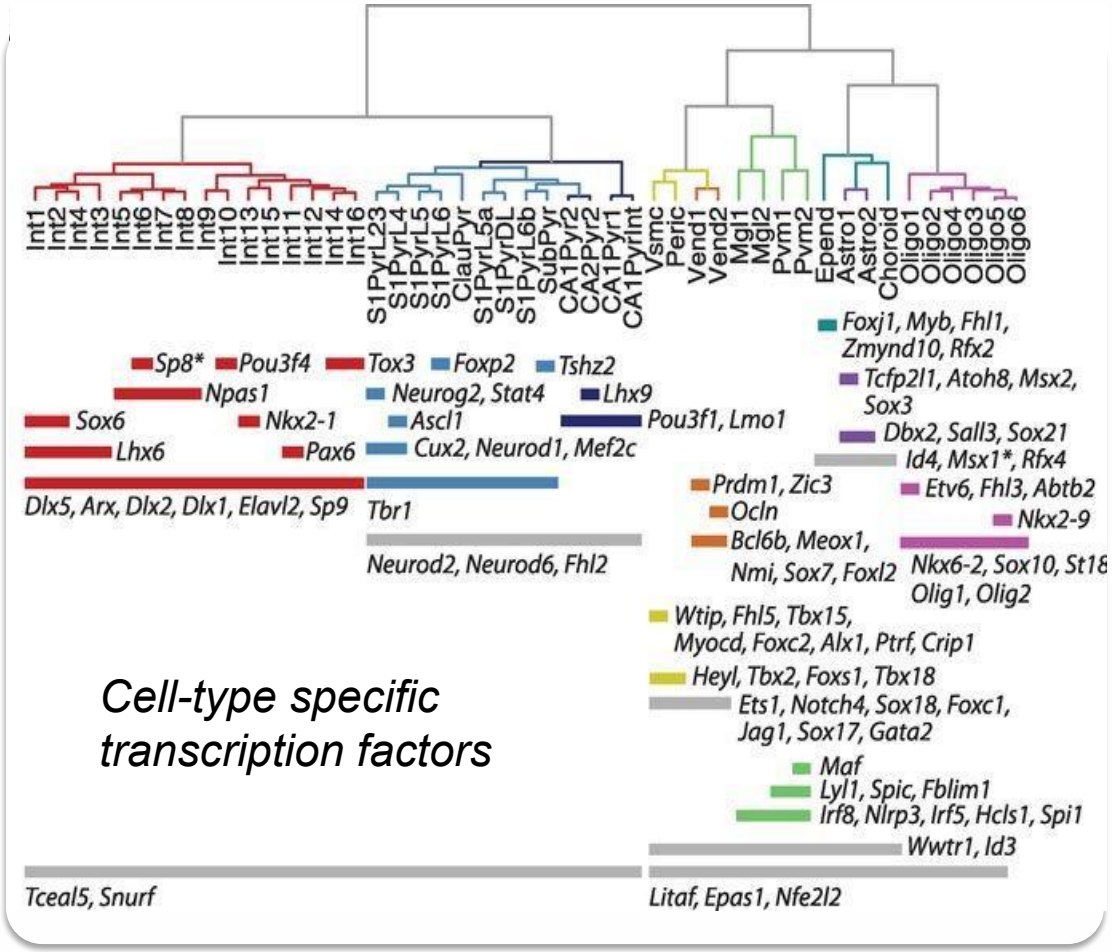
scRNA-seq applications – Tissue heterogeneity

Mapping out cell types in the mouse cortex & hippocampus



High cellular diversity in the brain but morphological similarity
=> difficult to disentangle distinct cells

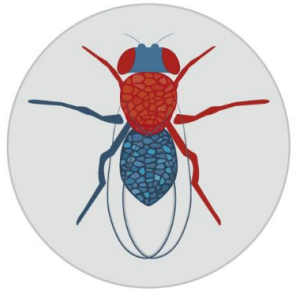
Neuronal cell types hierarchy



Interneurons of similar type exist in dissimilar regions of the brain
Identification of oligodendrocytes subtypes
Microglia associated with blood vessels distinguished from perivascular macrophages

Zeisel et al. (2015) Science

scRNA-seq applications – Creating XXX cell atlases



FLY CELL ATLAS

About

The Fly Cell Atlas will bring together Drosophila researchers interested in single-cell genomics, transcriptomics, and epigenomics, to build comprehensive cell atlases during different developmental stages and disease models.

>> More

HUMAN CELL ATLAS

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

A molecular cell atlas of the human lung from single cell RNA sequencing

Kyle J. Travaglini, Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V. Sit, Stephen Chang, Stephanie D. Conley, Yasuo Mori, Jun Seita, Gerald J. Berry, Joseph B. Shrager, Ross J. Metzger, Christin S. Kuo, Norma Neff, Irving L. Weissman, Stephen R. Quake, Mark A. Krasnow

nature neuroscience

Resource | Published: 06 May 2019

A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment

Hannah Van Hove, Liesbet Martens, Isabelle Scheyltjens, Karen De Vlaminck, Ana Rita Pombo Antunes, Sofie De Prijck, Niels Vandamme, Sebastiaan De Schepper, Gert Van Isterdael, Charlotte L. Scott, Jeroen Aerts, Geert Berx, Guy E. Boeckxstaens, Roosmarijn E. Vandenbroucke, Lars Vereecke, Diederik Moechars, Martin Guilliams, Jo A. Van Ginderachter, Yvan Saeys & Kiavash Movahedi

nature
International journal of science

Article | Published: 10 July 2019

A human liver cell atlas reveals heterogeneity and epithelial progenitors

Nadim Aizarani, Antonio Saviano, Sagar, Laurent Mailly, Sarah Durand, Josip S. Herman, Patrick Pessaux, Thomas F. Baumert & Dominic Grün

Science

RESEARCH ARTICLE

The Malaria Cell Atlas: Single parasite transcriptomes across the complete *Plasmodium* life cycle

Virginia M. Howick[1,*], Andrew J. C. Russell[1,*], Tallulah Andrews[1], Haynes Heaton[1], Adam J. Reid[1], Kedar Natarajan[2], Hellen...

+ See all authors and affiliations

Science 23 Aug 2019:
Vol. 365, Issue 6455, eaaw2619
DOI: 10.1126/science.aaw2619

Cell Atlas of Worm

A Cell Atlas of Worm

The *C. elegans* transcriptome at single cell resolution

In Cao et al. (Science, 2017) we reported single cell RNA-seq of *C. elegans* larvae at ~50x 'shotgun cellular coverage' using a combinatorial indexing approach (sci-RNA-seq).

Cell

Resource

A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer

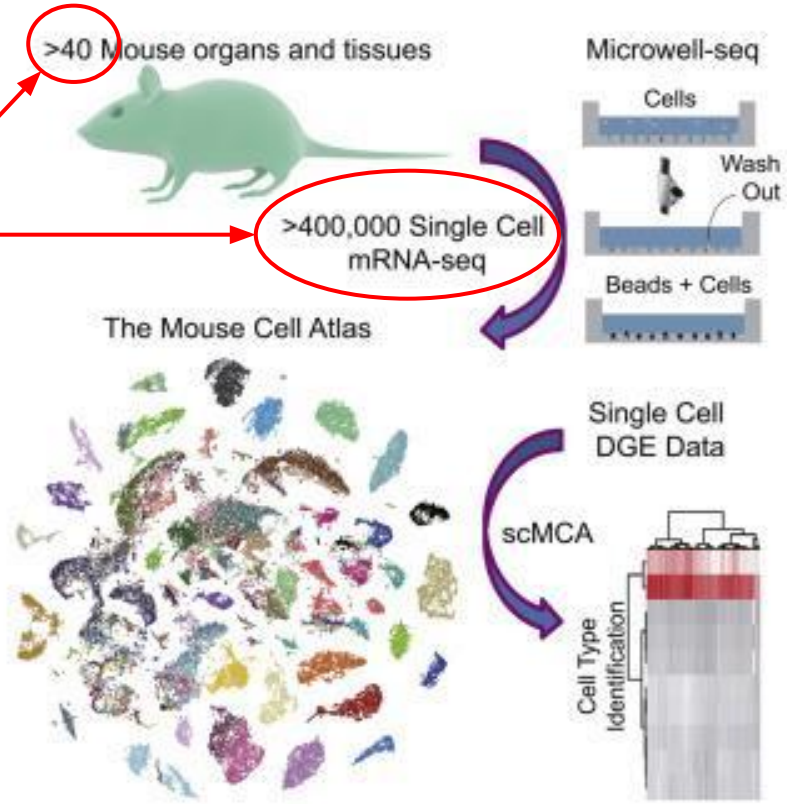
Johanna Wagner,[1,2,14] Maria Anna Rapsomaniki,[3] Stéphane Chevrier,[1,14] Tobias Anzeneder,[4] Claus Langwieder,[5] August Dykgers,[5] Martin Rees,[5] Annette Ramaswamy,[6] Simone Muenst,[7] Savas Deniz Soysal,[8,9] Andrea Jacobs,[1,14] Jonas Windhager,[1,10,14] Karina Silina,[11] Maries van den Broek,[11] Konstantin Johannes Dedes,[12] Maria Rodríguez Martínez,[3,15] Walter Paul Weber,[9,13,15] and Bernd Bodenmiller[1,14,16,*]

Amongst first atlases: *Mus musculus*

Mouse Cell Atlas

(Han, Guo et al., Cell, 2018)

In total

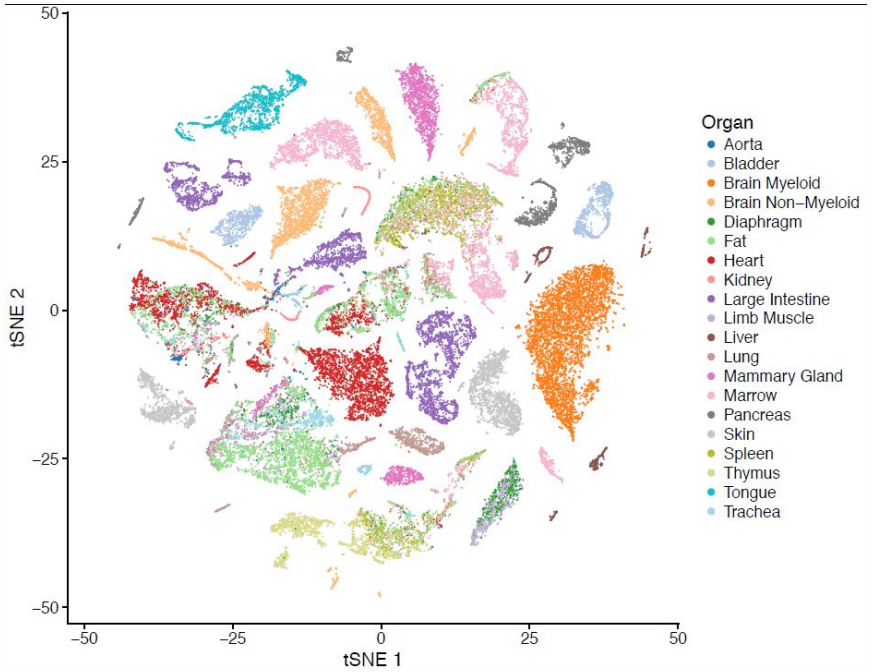


bis.zju.edu.cn/MCA/

CHAN ZUCKERBERG

~50'000 cells from 20 organs and tissues (SMART-Seq2)
~50'000 cells from 10x

(Quake, Darmanis et al., Nature, 2018)



⇒ **More challenging** because needs for reproducibility, data sharing, standardization of nomenclature, integration between techs

tabula-muris.ds.czbiohub.org

The Human Cell Atlas

THE HUMAN CELL ATLAS

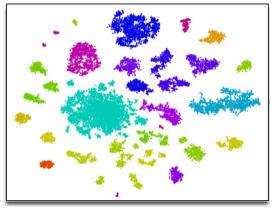
MISSION

To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

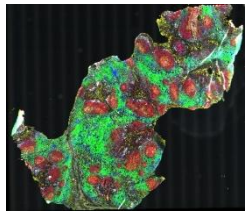
HCA plans to sequence ~1-10 billion cells?
©Dana Pe'er talk at last HCA meeting

Chan Zuckerberg Biohub ($600 M Initiative)

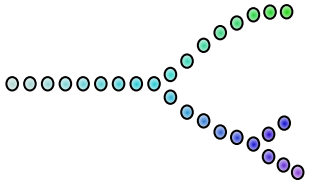
Scope, Scale, Quality and Compatibility



Cell States and Types



Spatial location and architecture



Lineages and transitions

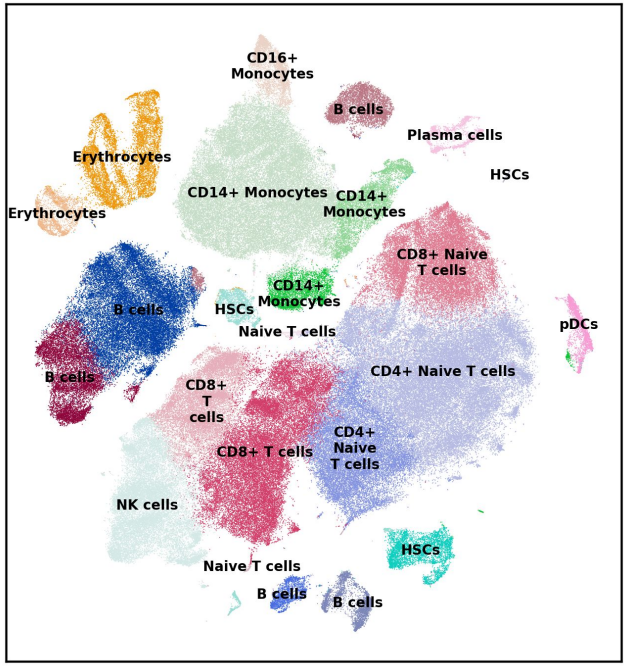
data.humancellatlas.org

https://www.humancellatlas.org/

The Human Cell Atlas PREVIEW: 2 pilot studies

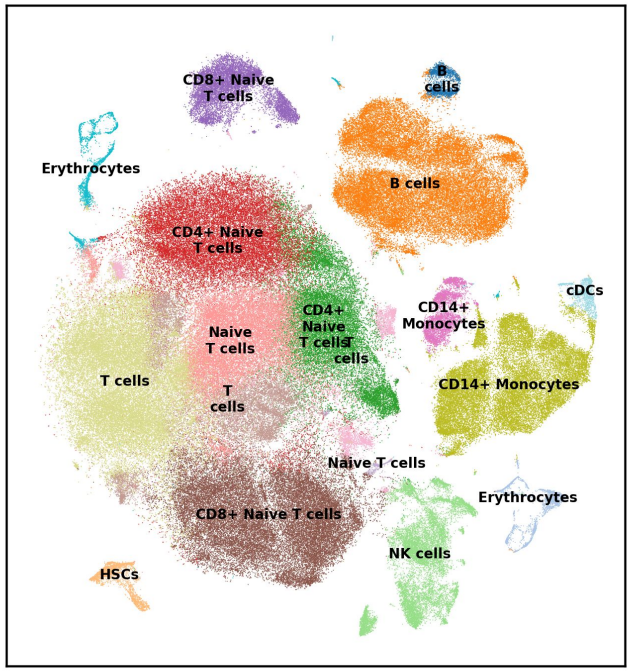
The Immune Cell Atlas 1.0

Bone Marrow



378,000 cells
~500Mb sparse count matrix

Cord Blood



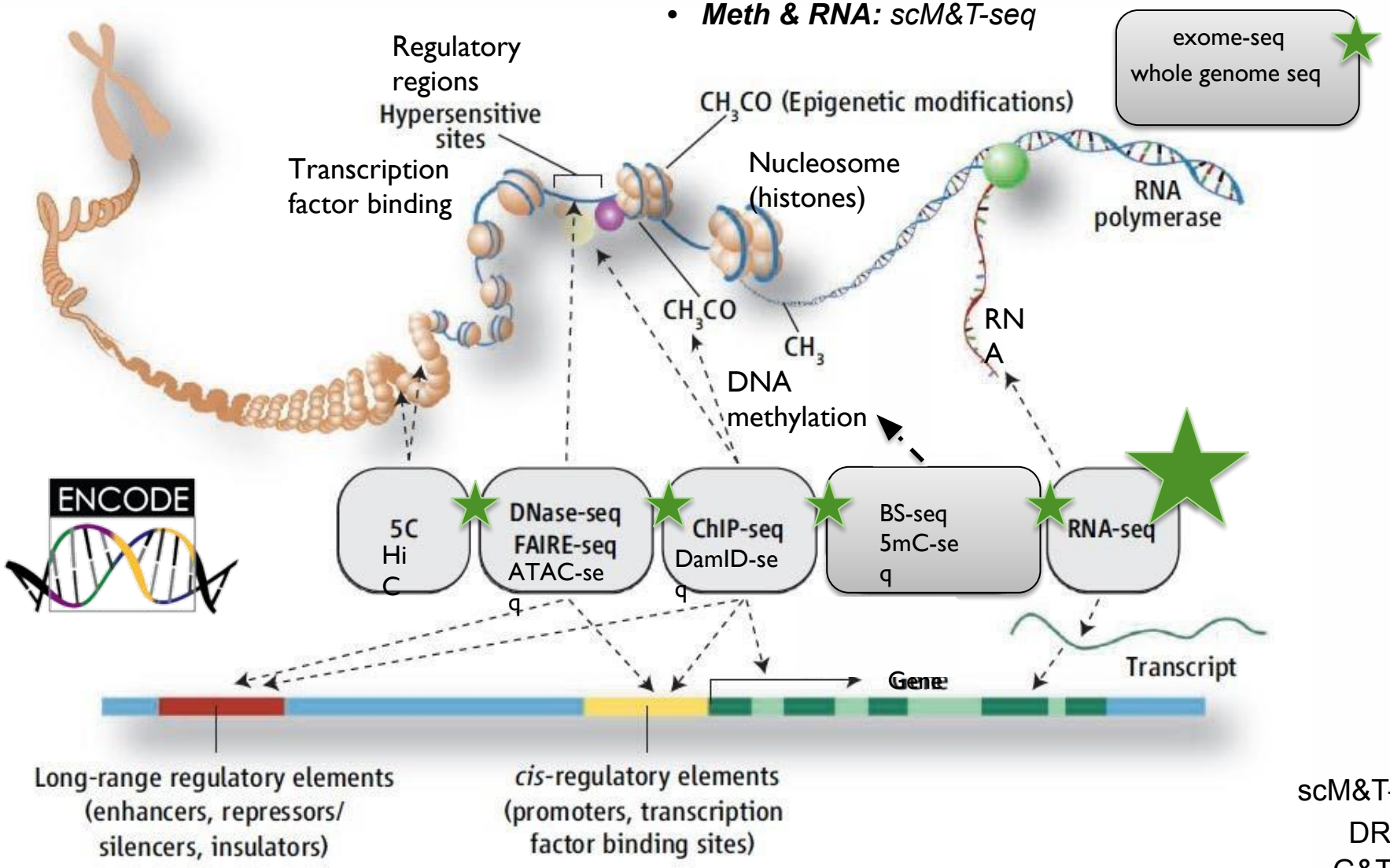
384,000 cells
~430Mb sparse count matrix

Online since April 2018 (1.3Tb w/ .fastq) preview.data.humancellatlas.org

Single-cell genomics can now assess most genomic layers

Some can even be combined in the <u>same cell</u> (multiomics single-cell)

- ***DNA & RNA:*** *DR-seq and G&T-seq*
- ***Meth & RNA:*** *scM&T-seq*



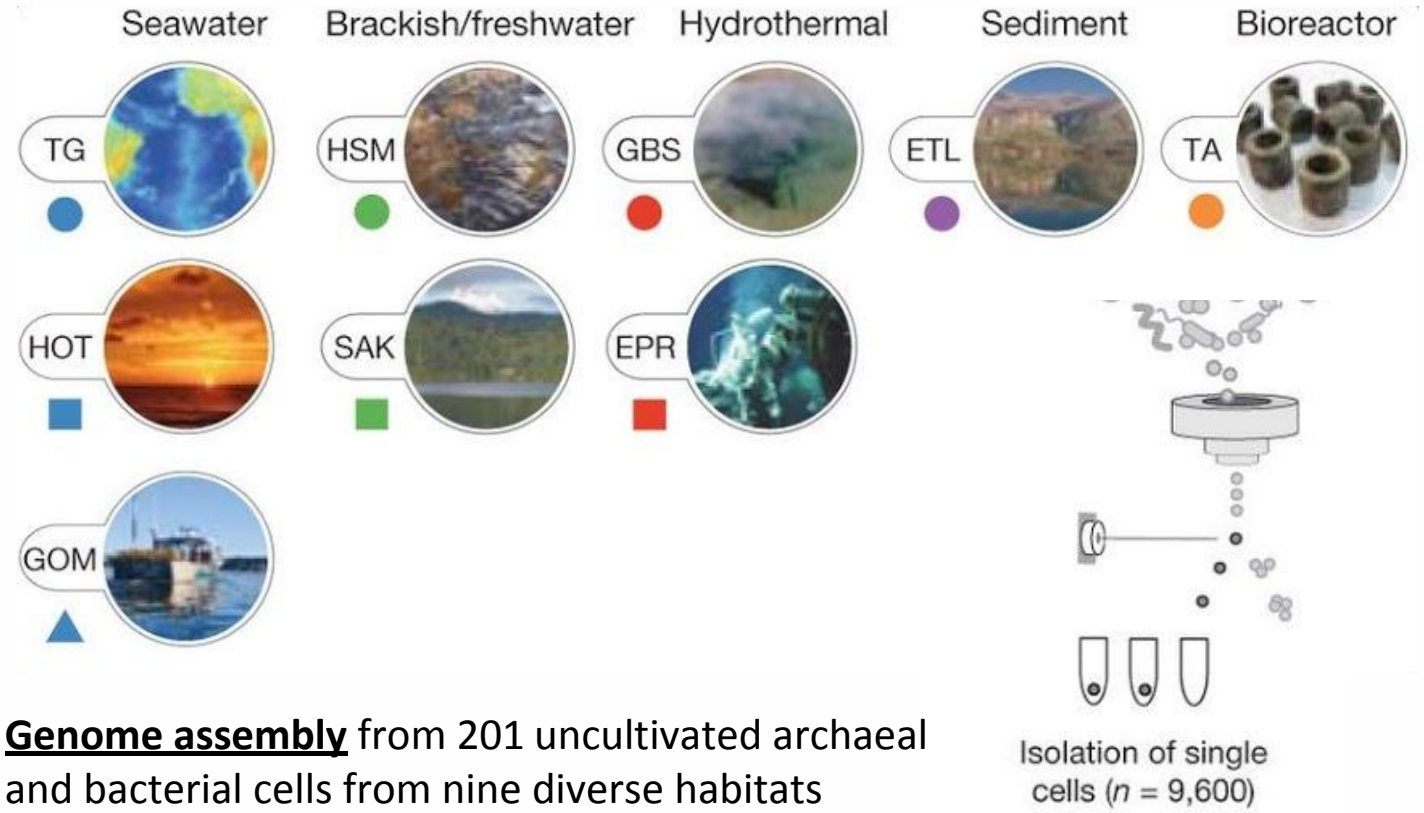
scM&T-seq: <u>Cheow et al., 2015</u>

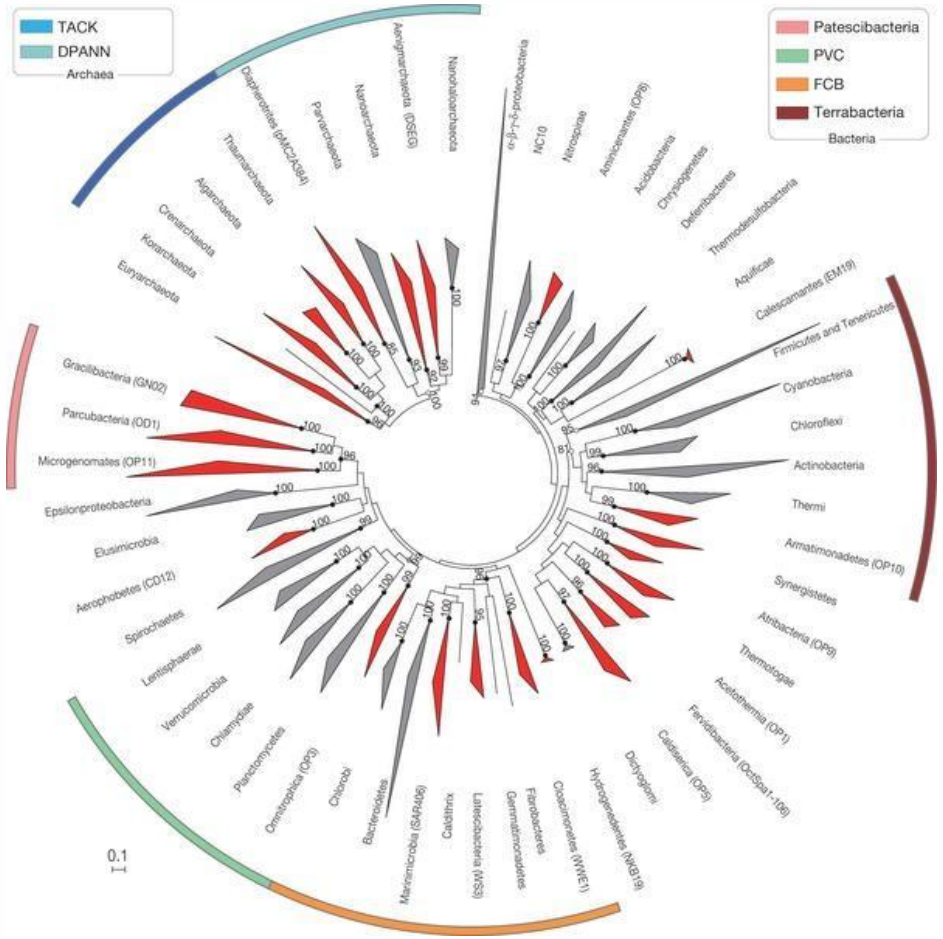
DR-seq: <u>Dey et al., 2015</u>

G&T-seq: <u>Macaulay et al., 2015</u>

Single-cell genetics – Application to microbiology

Mapping out the "microbial dark matter": species that cannot be cultivated





Genome assembly from 201 uncultivated archaeal and bacterial cells from nine diverse habitats

29 major mostly uncharted branches of the tree of life, so-called 'microbial dark matter'
resolved many intra- and inter-phylum-level relationships
novel, unexpected metabolic features (UGA stop codon recoded for Gly, purine synthesis, etc.)

Rinke, C. et al., Nature, (2013)

Single-cell ATAC – and yet another Atlas?

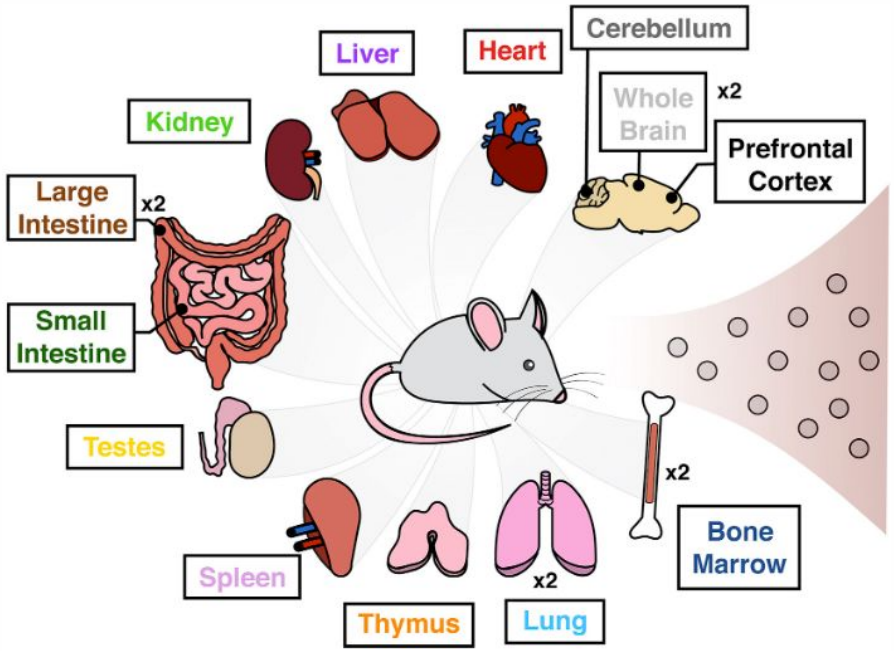
Resource

Cell

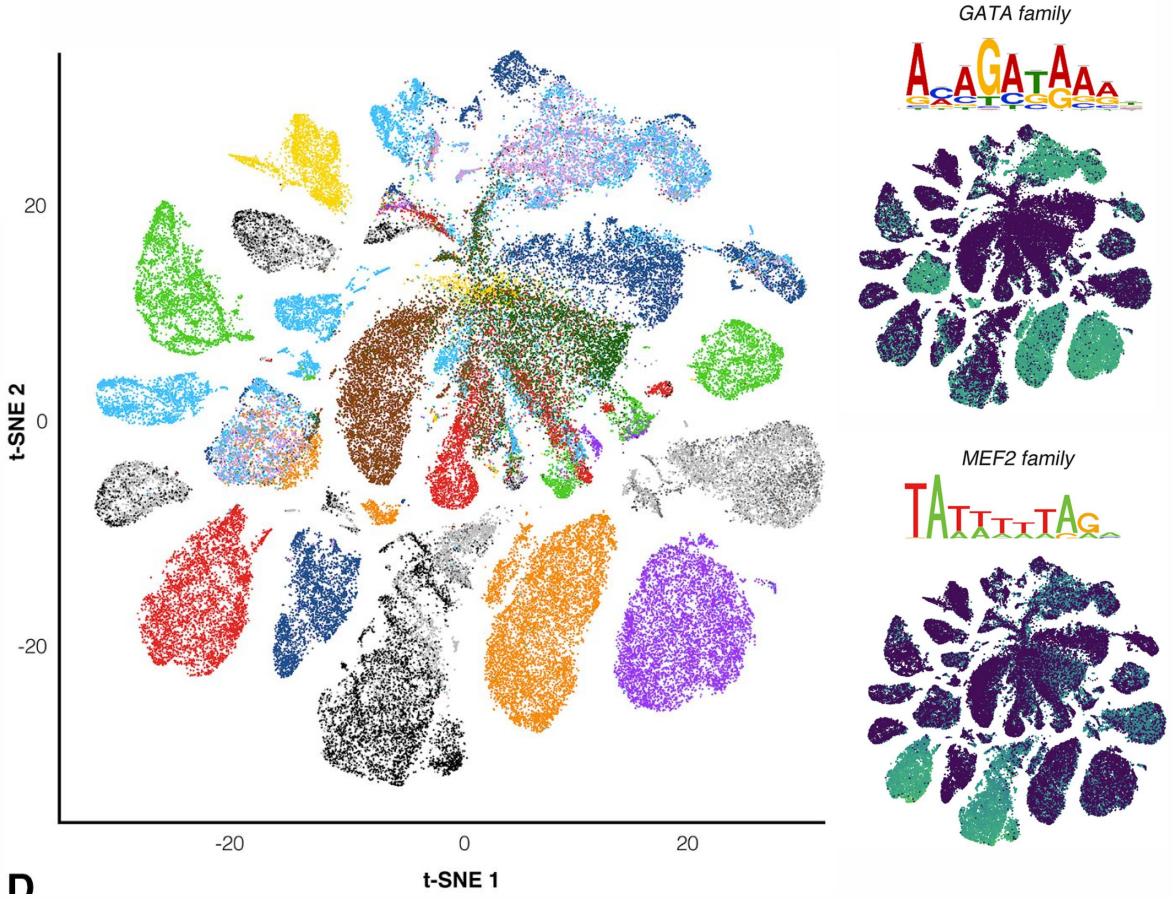
A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility

Authors

Darren A. Cusanovich, Andrew J. Hill, Christine M. Disteche, Cole Trapnell, Jay Shendure, ...,



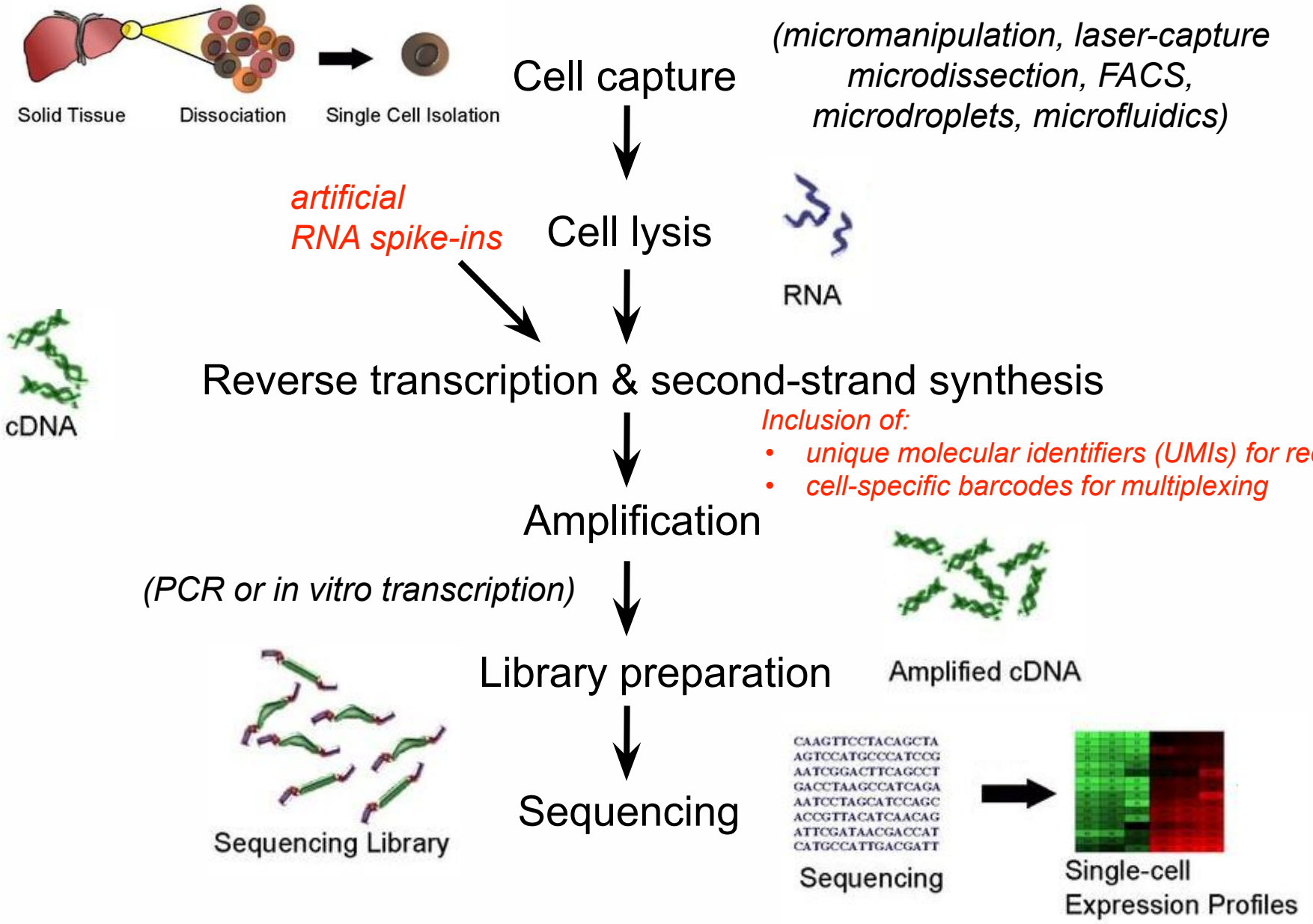
⇒ **13 different tissues**



⇒ **85 distincts chromatin patterns**

Cusanovich, D.A.. et al., Cell, (2018)

Single-cell RNA-seq experimental workflow



adapted from Wikipedia

Single-cell biology challenges

* **Cell Capture:** throughput, automation, cell stress

* **Small quantities:** obtain enough material for an accurate readout without introducing biases

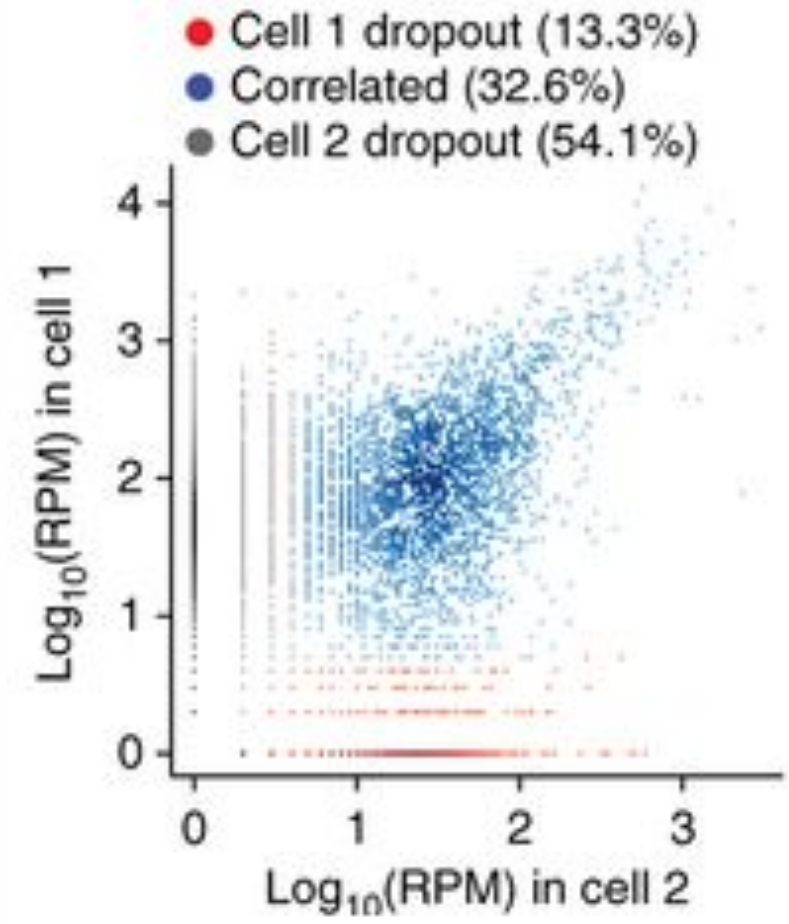
⇒ Amplification requires many more cycles than bulk methods

* **Data analysis & interpretation:** sparseness, noise, high dimensionality, batch effect, doublets, …

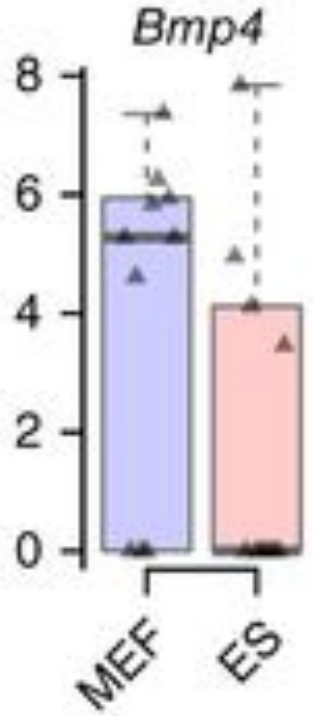
⇒ Gene 'dropouts' in which a gene is observed at a moderate expression level in one cell but is not detected in another cell

Kharchenko et al., Nature (2014)

Dropout effect



BMP4 is part of the top DE genes between 10 mouse embryonic fibroblast (MEF) vs 10 mouse embryonic stem cells (ES)



2 cells of same type: mouse embryonic fibroblast (MEF)

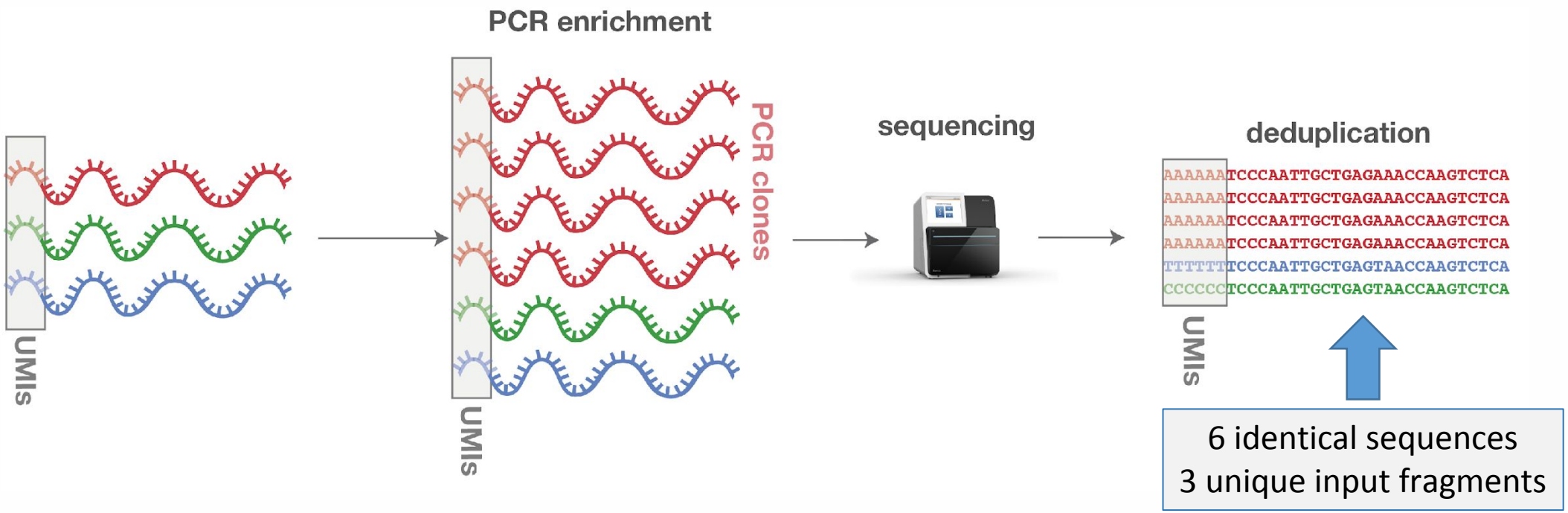
Kharchenko et al., Nature (2014)

Unique Molecular Identifier (UMI) for dealing with amplification bias

UMIs are random barcodes that are attached to the fragments prior amplification

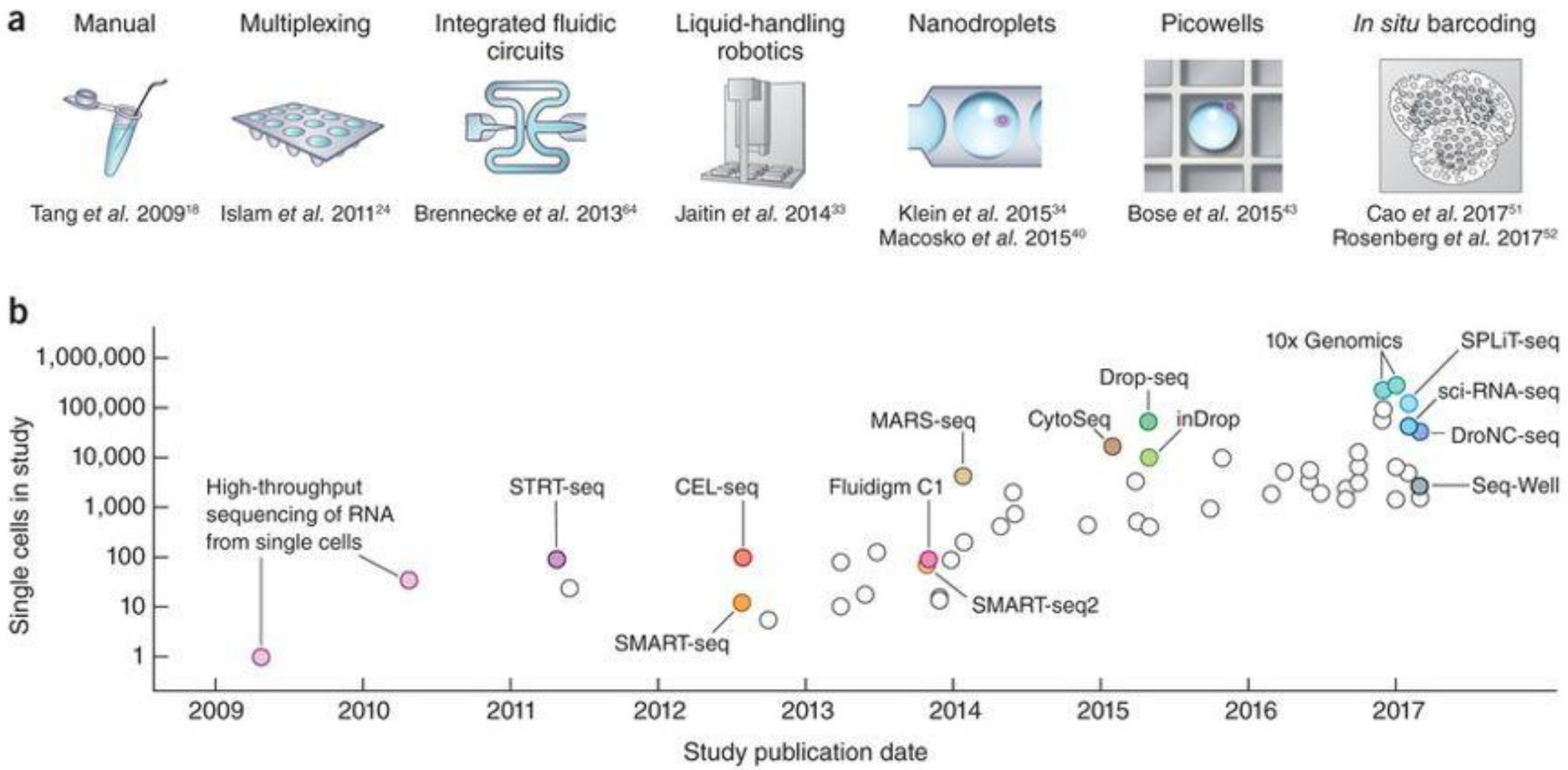
Individual transcripts can thus be detected in the final output by removing the duplicated barcodes/UMIs

⇒ Reduce the technical noise
⇒ Simpler statistical models (vs read counts)
⇒ Better approximation of duplicates as compared to standard Picard/Samtools tools (e.g. for variant calling)



Single-cell genomics, available protocols

scRNA-seq method development paving the way for other genomics techniques

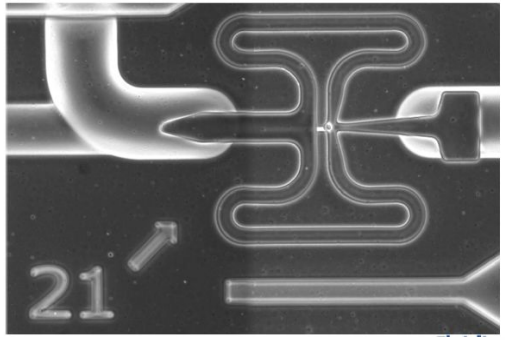


Svensson …Teichman, Nature protocols, 2018

Microfluidics

- **making biology more quantitative:** development of **MICROFLUIDICS*** techniques with applications in structural biology, drug discovery, molecular affinity, diagnostics

*reviewed in *SackmanBeebeNature2014*



FLUIDIGM

C1 Single-Cell Auto-prep system

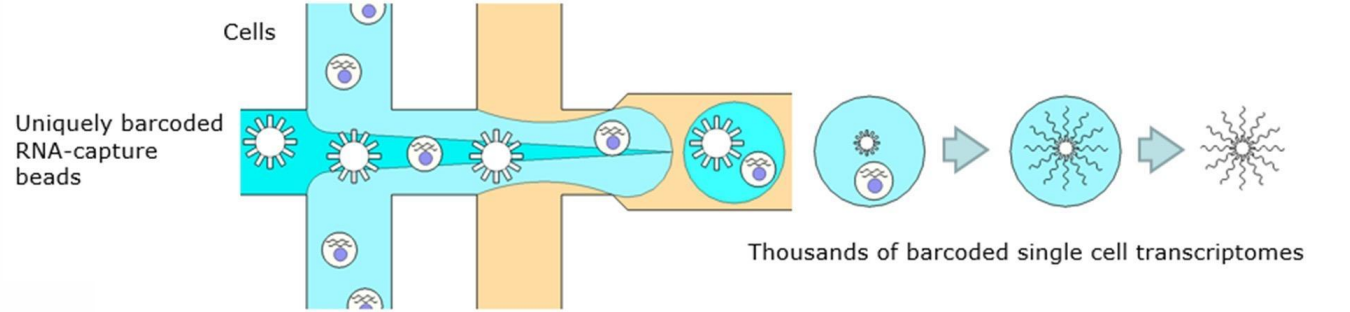
www.fluidigm.com

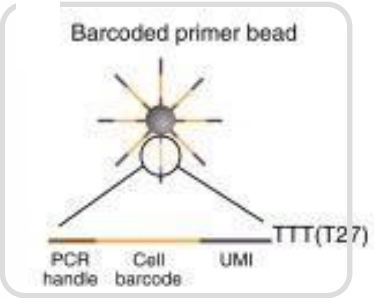
First commercially available, fully automated scRNA-seq workflow
Cells are captured using integrated fluidic circuits (up to 800 cells/experiment)

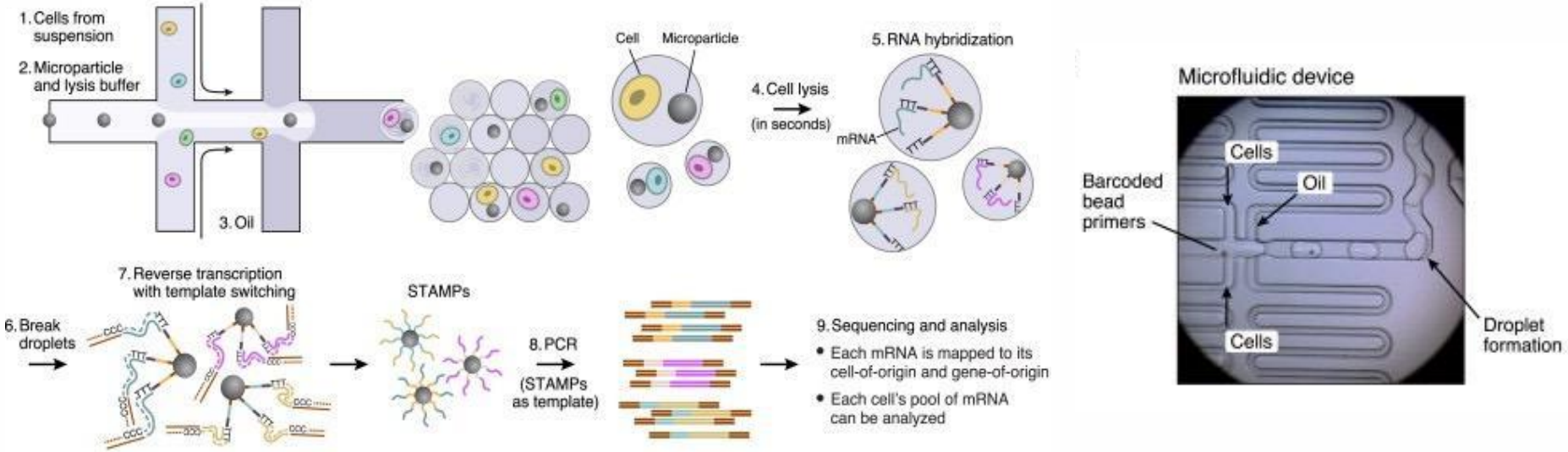
https://www.youtube.com/watch?v=TF4NJRE4Xg4

Droplet-based microfluidics

Highly Parallel Genome-wide Expression Profiling of Individual cells







DROP-seq

https://www.youtube.com/watch?v=YCup_5njeS4

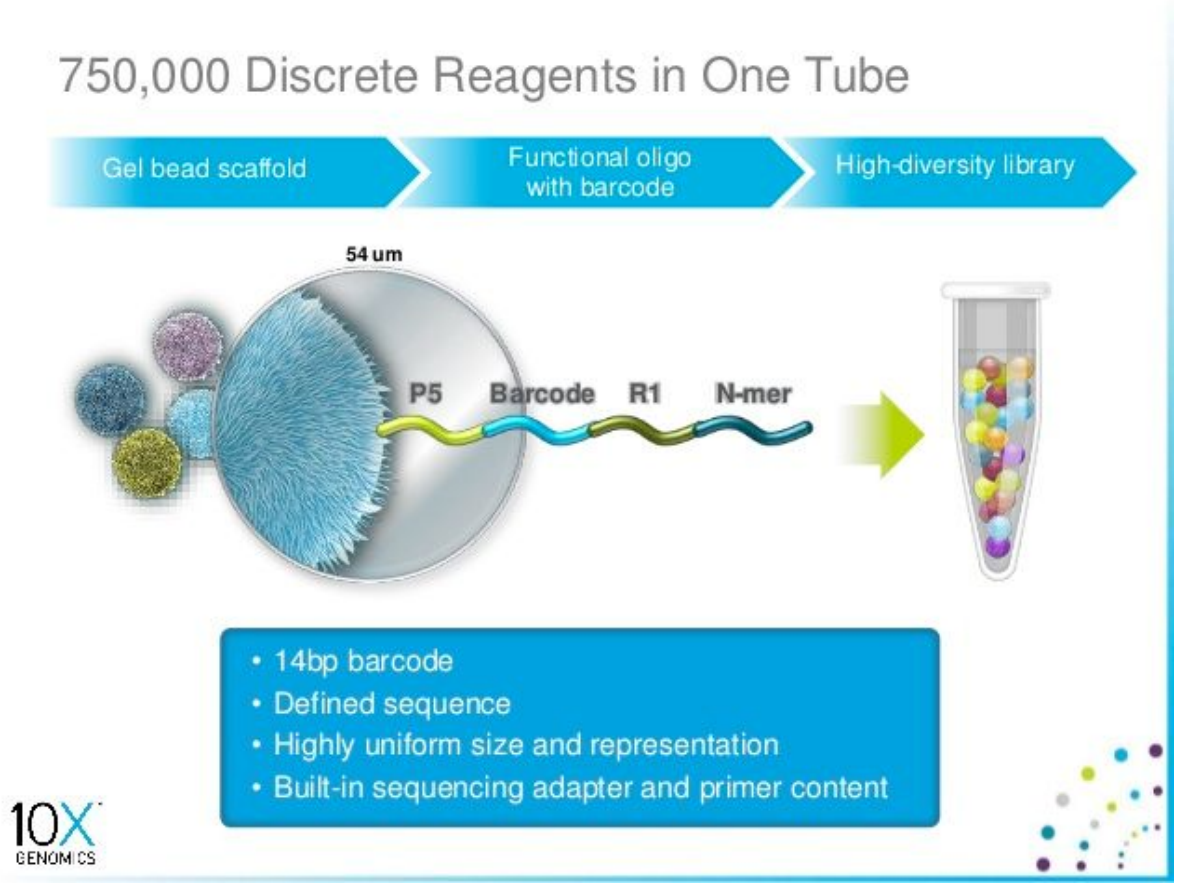
— cell capture rate (currently ~5-10%)
— no cell selection

Macosko … McCaroll Cell 2015

Now commercialised by Dolomite Microfluidics

alternative similar tech Klein .. Kischner Cell 2015

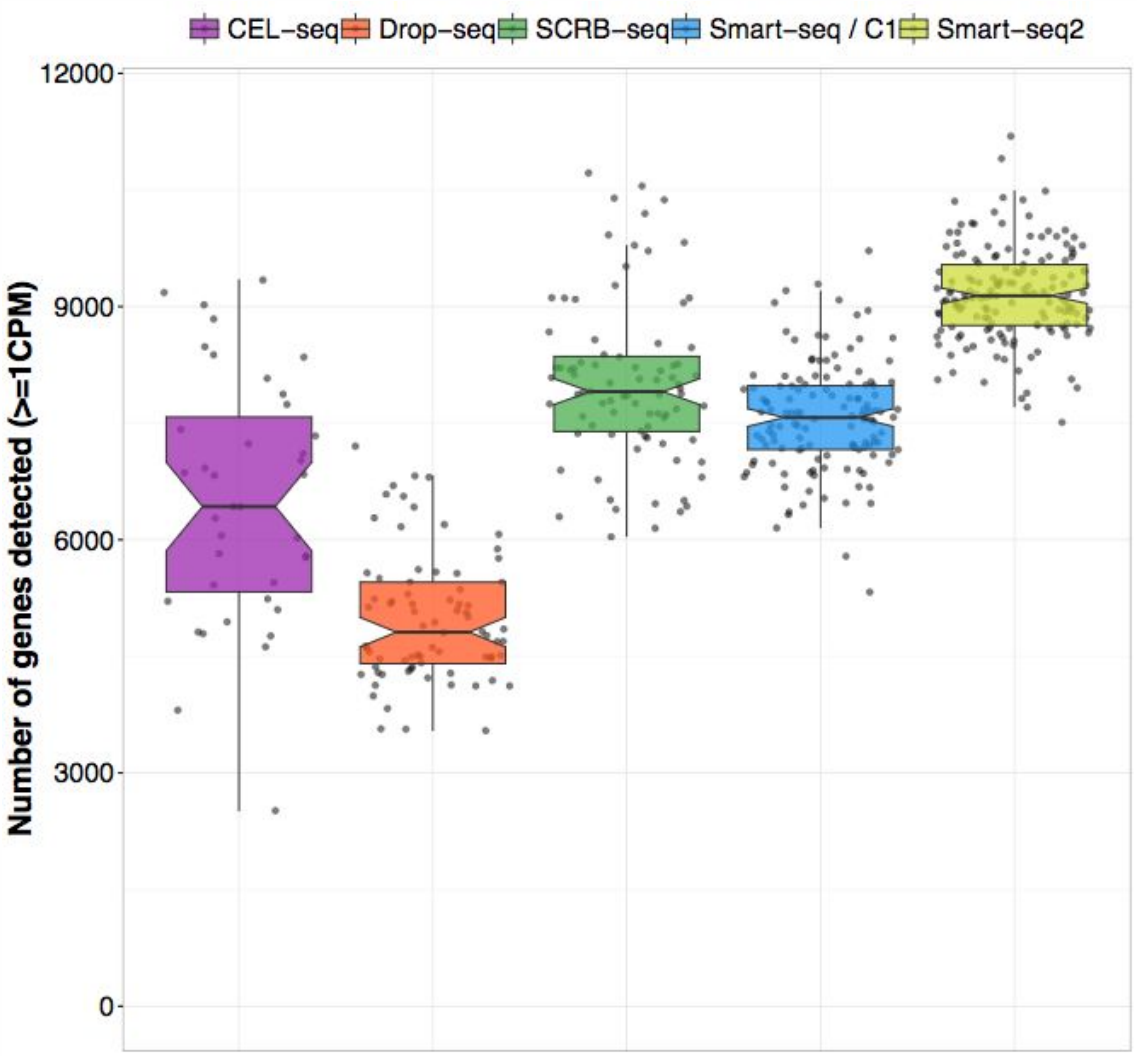
10X Genomics

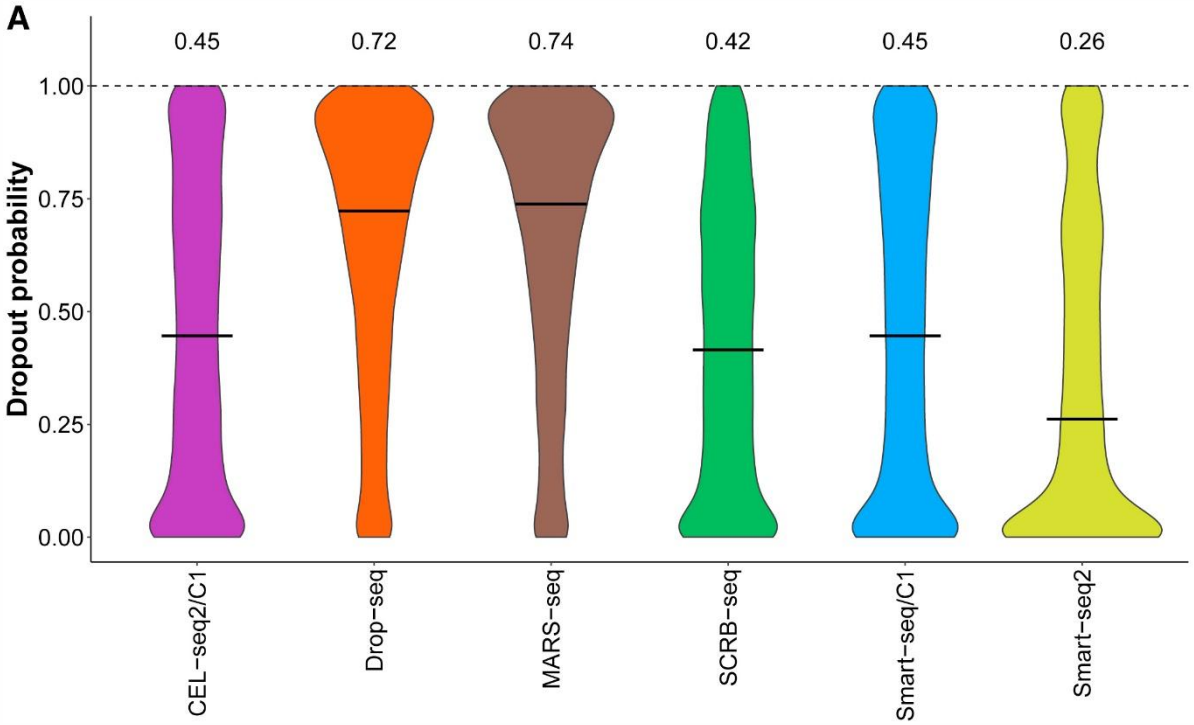


10X Genomics (Commercial Solution)

- Hydrogel bead dissolves in droplet
  □ uniform distribution of oligo's
- RT in drop
- Overall, higher data quality compared to DROP-seq

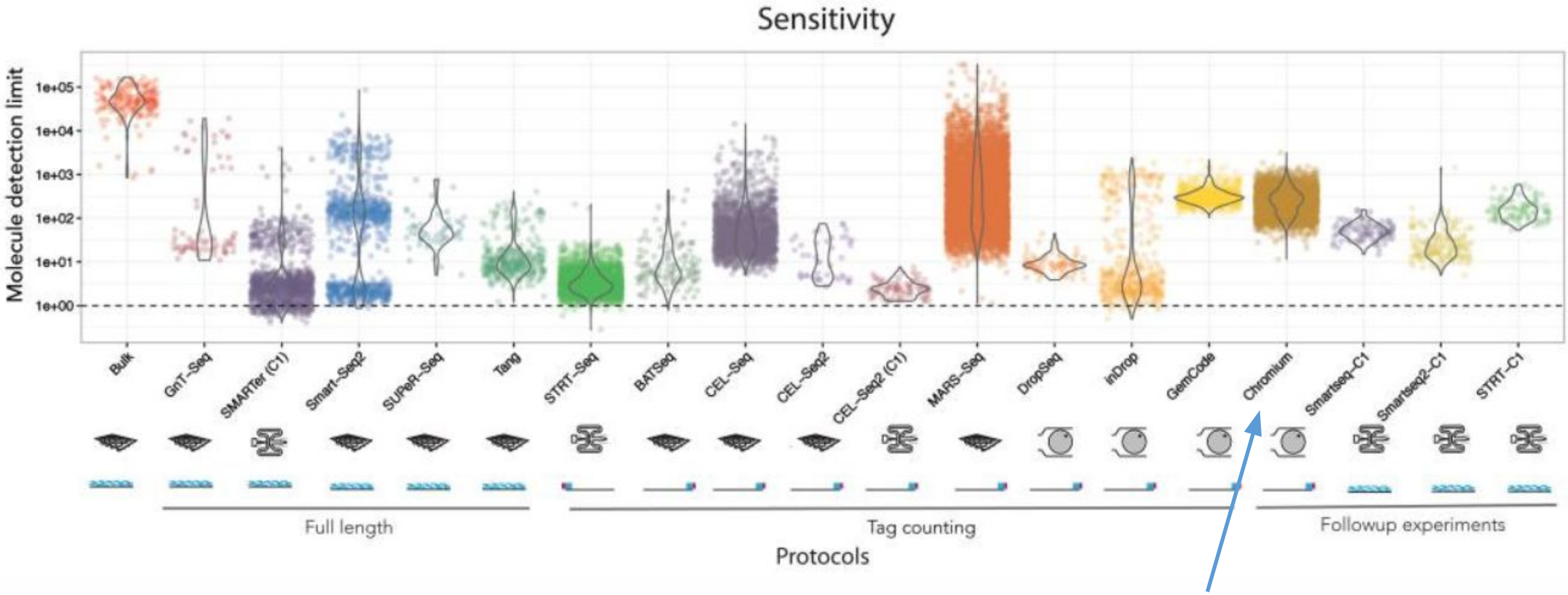
Which protocol to use?





Ziegenhain et al., Molecular Cell (2017)

Which protocol to use?



10x Chromium

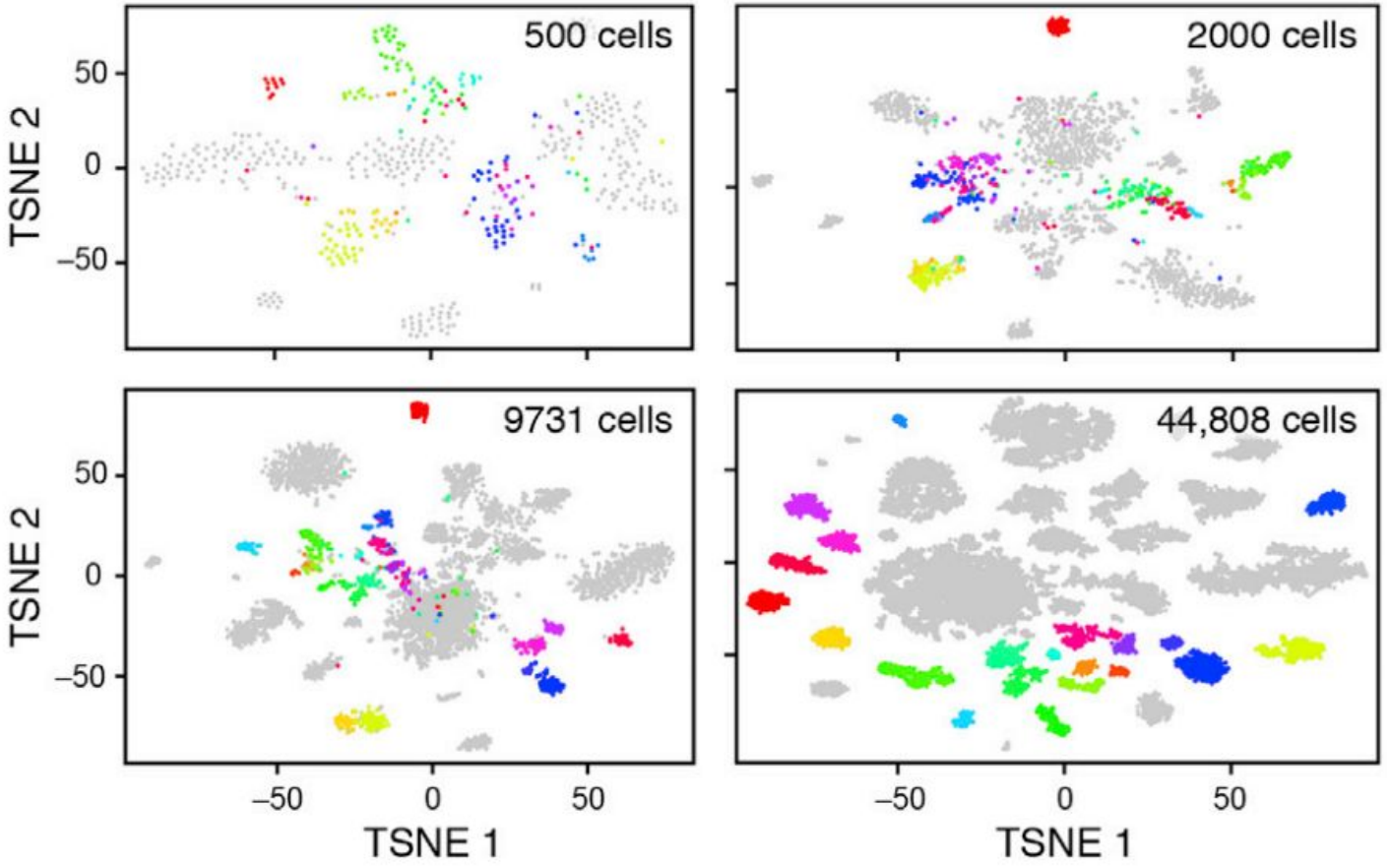
Svensson et al., Nature Methods (2017)

Which protocol to use?

⇒ Currently, the decision is a trade-off between accuracy and number of cells

⇒ Smart-seq2 (Smart-seq3) is the most sensitive / 10x is more accessible & can process more cells

e.g. DROP-seq maps out retinal cell types

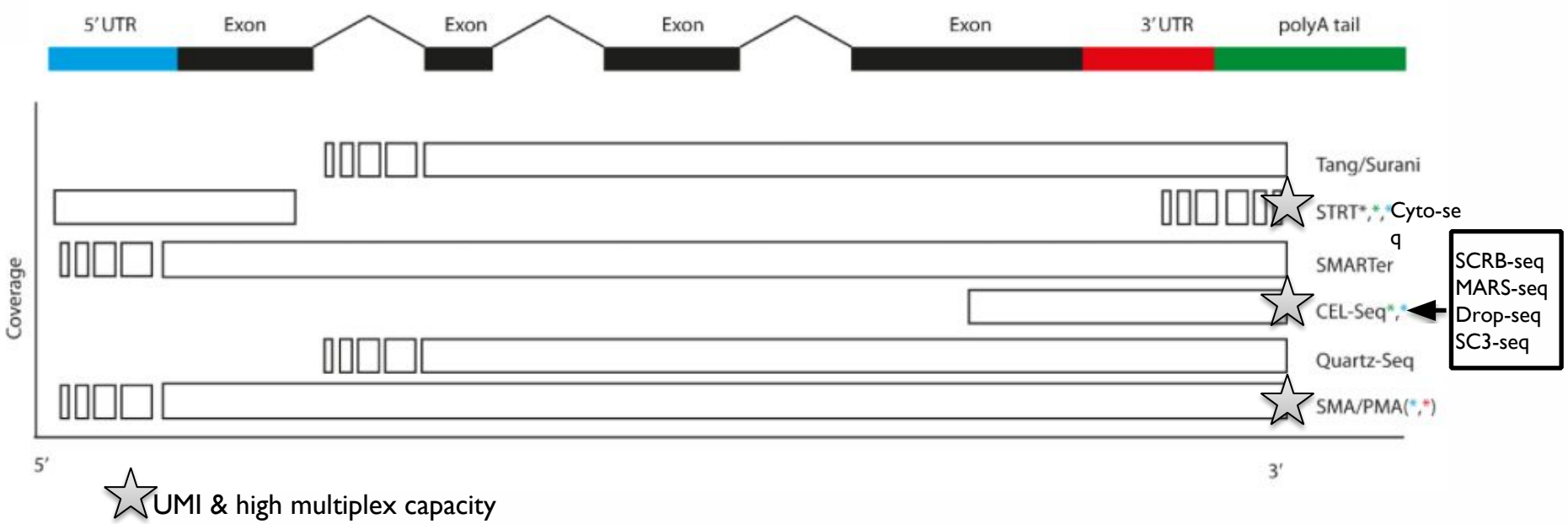


⇒ The more cells you study, the better you are able to find and characterize sub populations (rare cell types, here amacrine cells)

Macosko … McCarroll Cell 2015

Single-cell RNA-seq coverage over gene body

Different methods cover distinct parts of the transcript



Full-length techniques enable study of alternative splicing and allele-specific expression
3' or 5' techniques enable multiplexing of a large number of cells => easy handling, low cost
UMI incorporation allows reduction of PCR-introduced biases in amplification

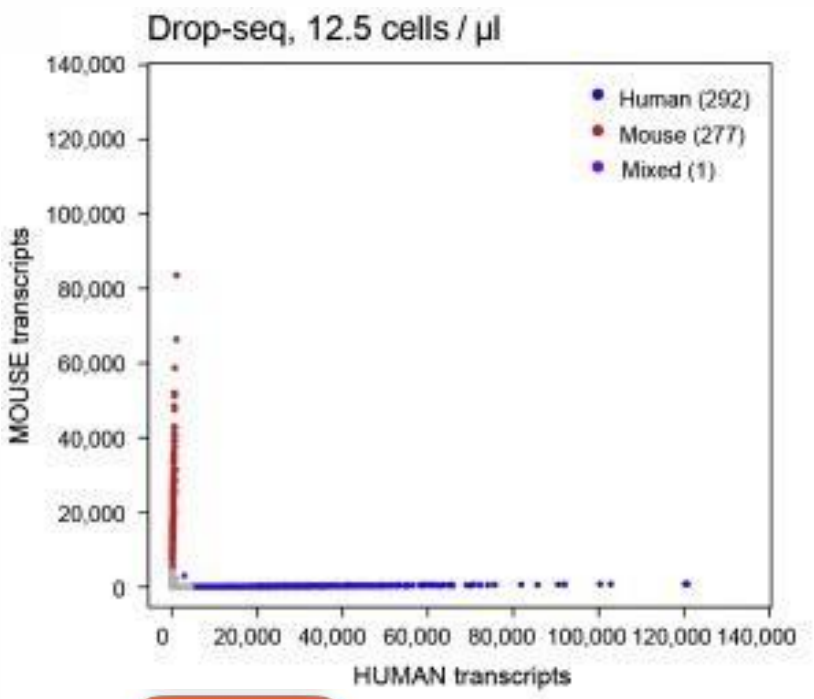
In contrast to population RNA-seq, scRNA-seq much more samples + amplification biases
=> early multiplexing & molecular counting very important

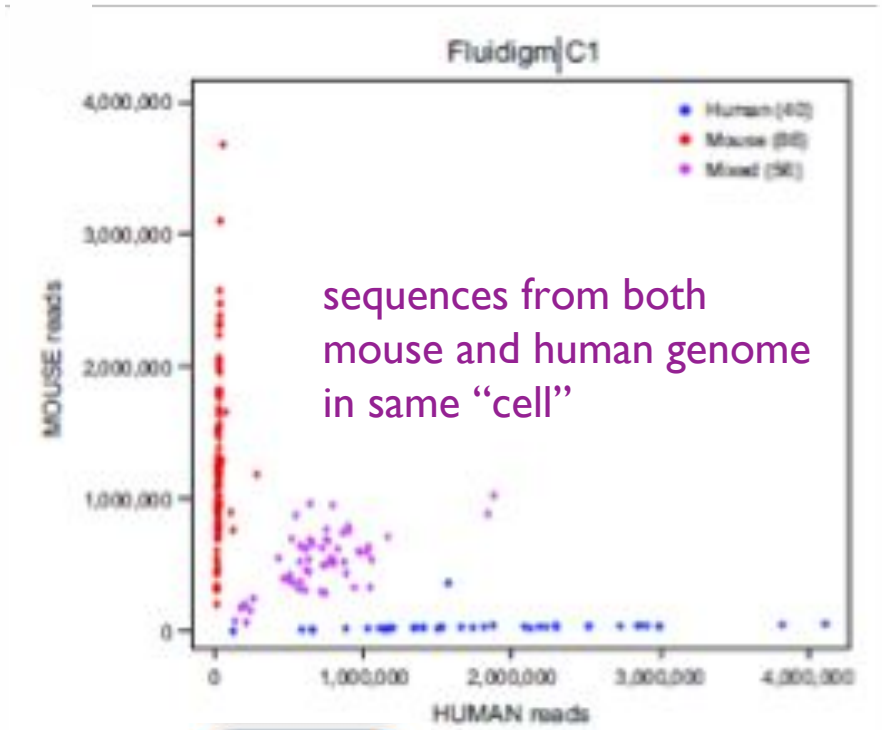
adapted from Macaulay .. Voet PLoS Genetics 2014

Single-cell RNA-seq – really single-cell?

Cell doublets can be assessed by species-mixing experiments

High rate (~ 40% of total) of duplicate cells in Fluidigm C1 experiments (medium chip)





Drop-seq

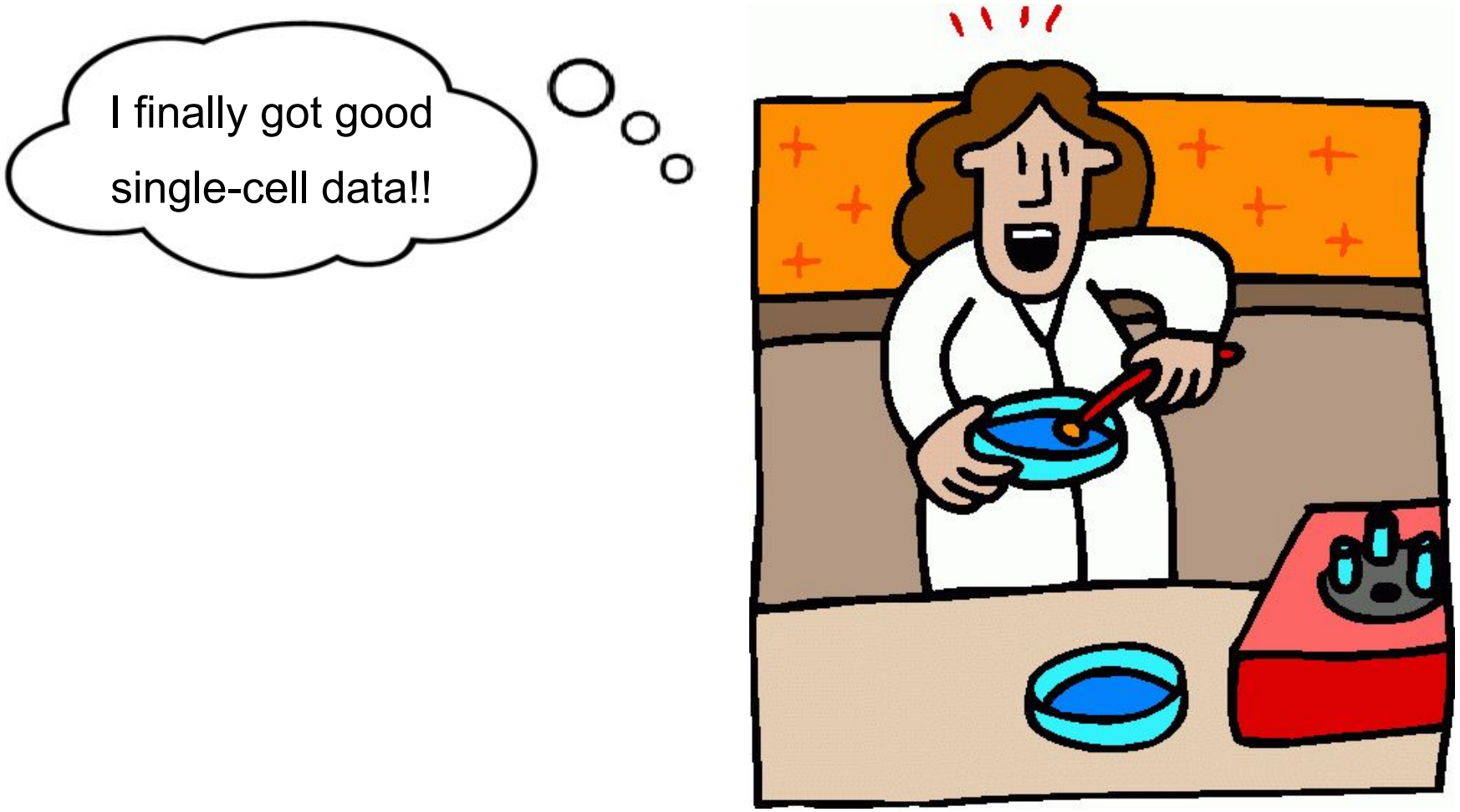
3' counting
8 bp UMI
Droplets
2 Flows

Smart-seq/C1

full-length
no UMI
Fluidigm C1
2 Chips

Macosko … McCarroll Cell 2015

Community needs: « I want to do single-cell ! »



Community needs: « But how do I analyze single-cell data? »

Same as bulk data pipelines?
Which tools to use?

⇒ Typical bottleneck

Estimated Number of Cells

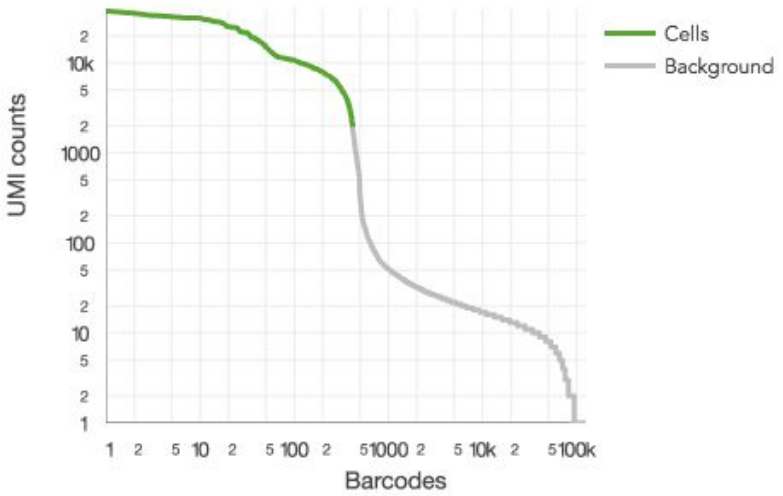
416

Mean Reads per Cell	Median Genes per Cell
2,757,470	3,395

Sequencing

Number of Reads	1,147,107,630
Valid Barcodes	98.4%
Reads Mapped Confidently to Transcriptome	57.3%
Reads Mapped Confidently to Exonic Regions	60.6%
Reads Mapped Confidently to Intronic Regions	17.2%
Reads Mapped Confidently to Intergenic Regions	5.6%
Sequencing Saturation	99.2%
Q30 Bases in Barcode	98.3%
Q30 Bases in RNA Read	76.1%
Q30 Bases in Sample Index	96.6%
Q30 Bases in UMI	98.2%

Cells



Estimated Number of Cells	416
Fraction Reads in Cells	87.1%
Mean Reads per Cell	2,757,470
Median Genes per Cell	3,395
Total Genes Detected	17,479
Median UMI Counts per Cell	7,735

Sample

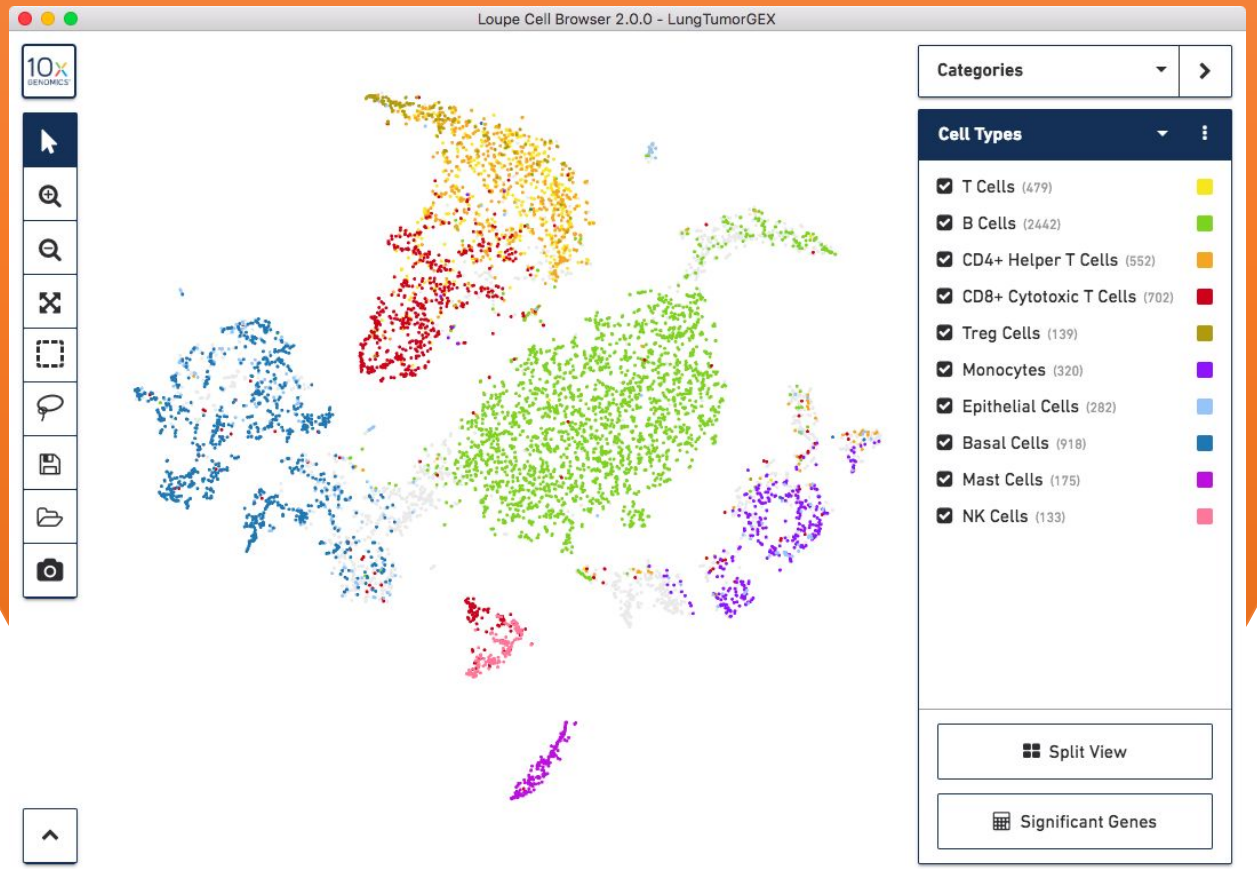
Name	NPY2
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	1.2.1

specific:
demultiplexing
unting
of cell

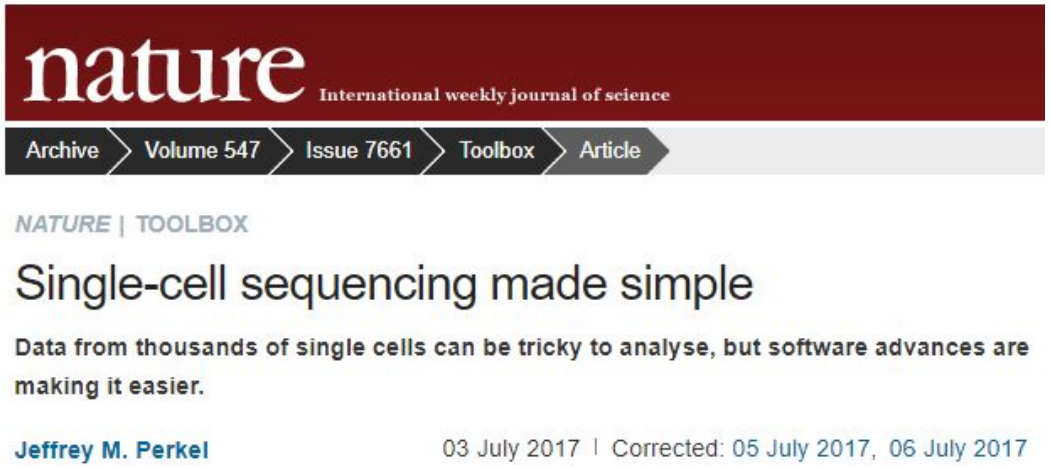
l specific:
out handling
cle removal

scRNA-seq analysis pipeline

"Black box" fixed pipeline?
e.g. 10x cell Ranger pipeline
+ Loupe Visualization



scRNA-seq pipeline may not be applicable to all datasets


nature International weekly journal of science

Archive > Volume 547 > Issue 7661 > Toolbox > Article

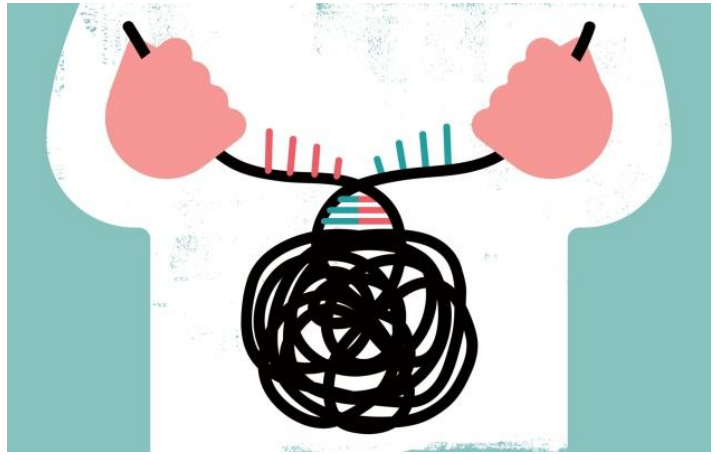
NATURE | TOOLBOX

Single-cell sequencing made simple

Data from thousands of single cells can be tricky to analyse, but software advances are making it easier.

Jeffrey M. Perkel

03 July 2017 | Corrected: 05 July 2017, 06 July 2017



"The tools aren't perfect for every situation"

⇒ "A pipeline that excels at identifying cell types, for instance, might stumble with pseudo-time analysis"

"Appropriate methods are 'very data-set dependent'", says Sandrine Dudoit, (biostatistician at the University of California, Berkeley).

⇒ "The methods and tuning parameters may need to be adjusted to account for variables such as sequencing length"

scRNA-seq pipeline may not be applicable to all datasets

Solutions

- Standardized but parametrizable pipelines (e.g. Seurat)

⇒ Probably best for bioinformaticians

- User-friendly automated analysis portals (e.g. ASAP, Scope, FastGenomics, …)

=> Good first glance at results for bioinformaticians. Sufficient for non-bioinformaticians.

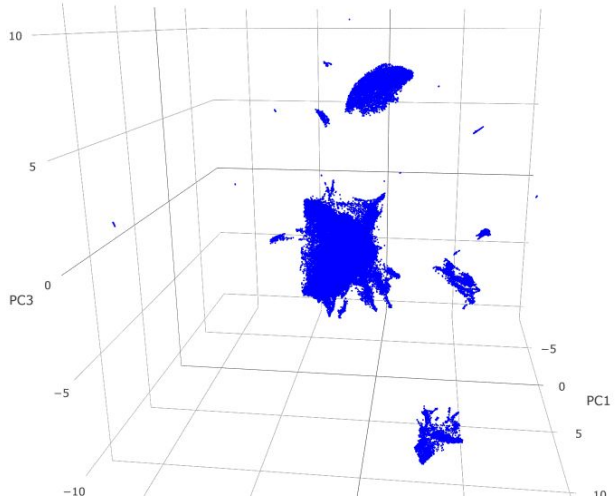
Web portal for the interactive analysis of single-cell RNA-seq data

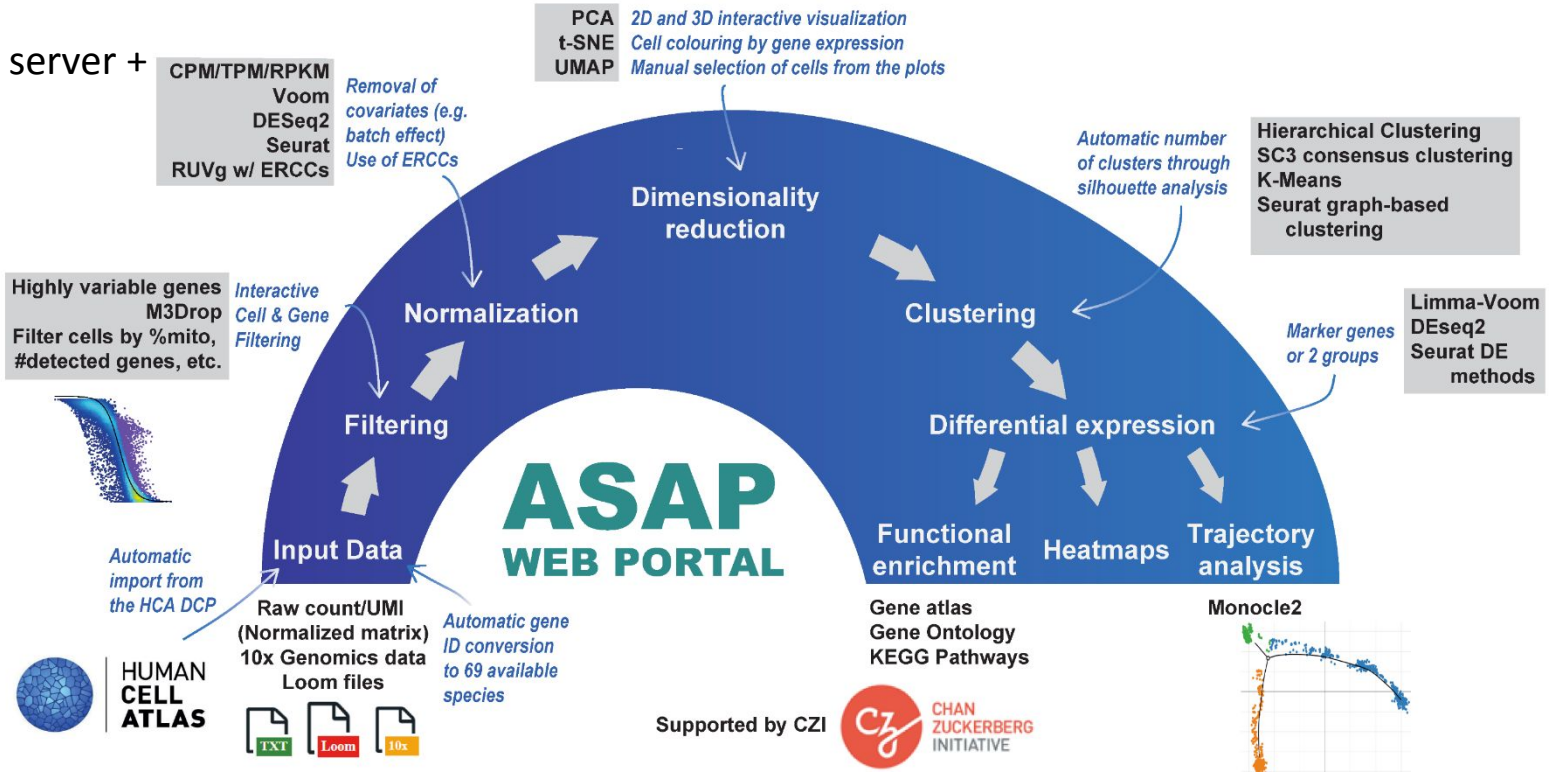
asap.epfl.ch asap-beta.epfl.ch

NEW

Web-tool for the interactive analysis of single-cell RNA-seq data

- ⇒ Provide **HCA** with a common interface for **reproducible** and **interactive** analysis of data
- ⇒ **Centralized computational resources**: Ruby-on-rails server + R/Python/Java code
- ⇒ **Job queuing management:** *delayed-jobs* gem
- ⇒ Currently 1430 projects & 400 registered users





Gardeux *et al., Bioinformatics,* (2017

Uniformization: how to handle/store a single-cell experiment…

Many different file formats exist to store all the information of single-cell RNA-seq project:

- CSV / TSV text file (dense Matrix)
- MTX text file (sparse Matrix)
- SingleCellExperiment R object (sparse Matrix)
- Seurat R object (sparse Matrix)
- Loom file (dense Matrix)
- …

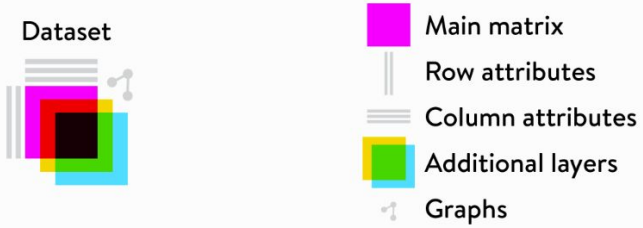
Loom Files?

Loom 2.0 loompy.org GitHub Contents

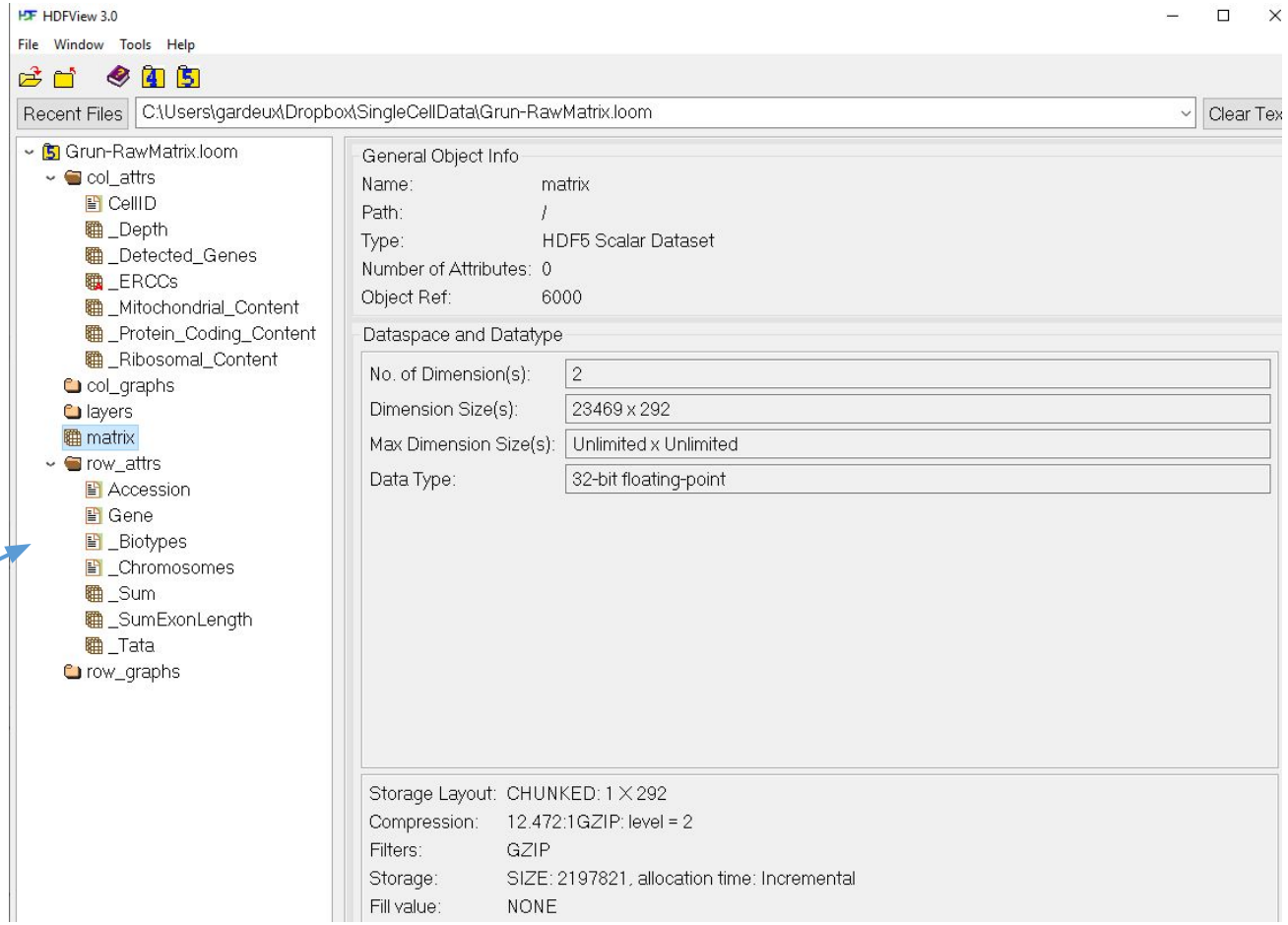
Loompy documentation

Loom is an efficient file format for very large omics datasets, consisting of a main matrix, optional additional layers, a variable number of row and column annotations, and sparse graph objects. We use loom files to store single-cell gene expression data: the main matrix contains the actual expression values (one column per cell, one row per gene); row and column annotations contain metadata for genes and cells, such as `Name`, `Chromosome`, `Position` (for genes), and `Strain`, `Sex`, `Age` (for cells). Graph objects are used to store nearest-neighbor graphs used for graph-based clustering.

The Loom logo illustrates how all the parts fit together:



Loom files (`.loom`) are created in the HDF5 file format, which supports an internal collection of numerical multidimensional datasets. HDF5 is supported by many computer languages, including Python, R, MATLAB, Mathematica, C, C++, Java, and Ruby.



<u>Example</u>:

Current methodological challenges for single-cell analysis…

- **Uniform storage / representation of a single-cell dataset**

- **Manifold alignment:** Define novel methods for integration of multiomics/multiplatform datasets (e.g. batch effect)

- **Scaling:** HCA plans to generate datasets of > 10 billions cells. How to t-SNE that??

⇒ Cloud computing, scalable methods (scanpy, Seurat?), out-of-RAM computations

- **Compression**

⇒ Standardized data format for .fastq/BAM files (MPEG-G)

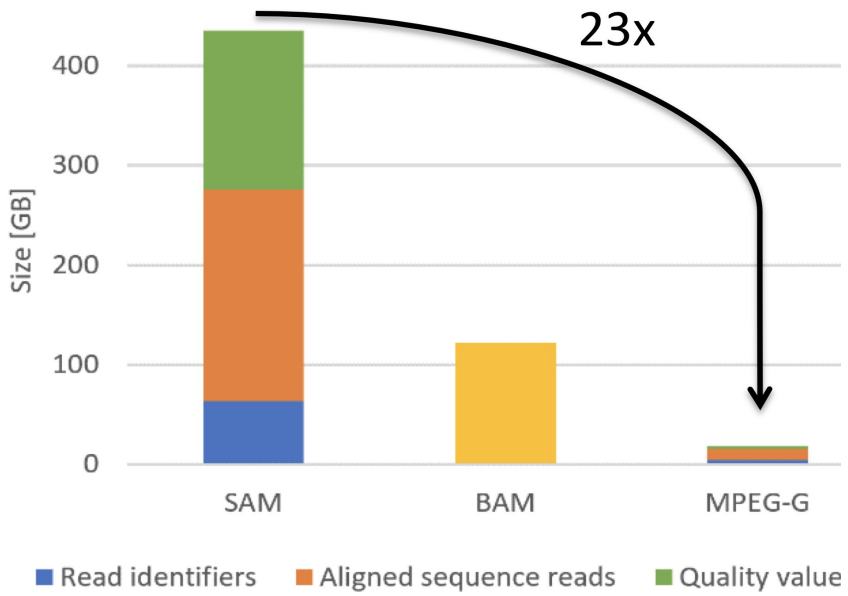
⇒ Lossless compression or not

- **Imputation**

⇒ Solving the dropout issue by replacing the 0s?

- **Trajectory analysis**

⇒ Find scalable methods



SciLifeLab

SIB

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

CHAN ZUCKERBERG INITIATIVE

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE