# Introduction into Single-cell RNA-seq

Konstantin “Kostya” Zaitsev, ITMO University

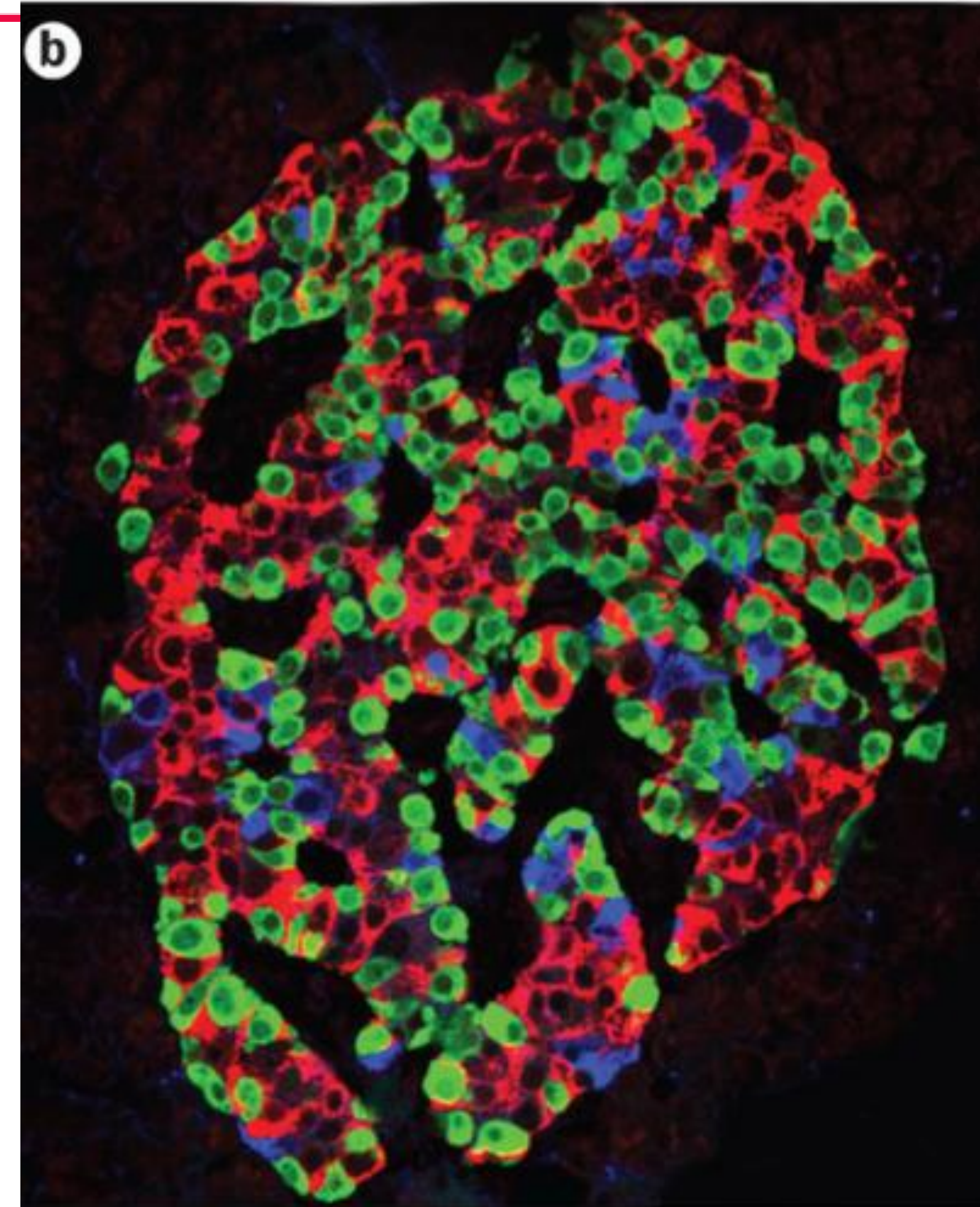
Systems biology workshop, Nice, Sep 22th

# Cell is the fundamental unit

- Microscopy
- FACS (fluorescence activated cell sorting) / CyTOF (Cytometry by Time Of Flight)
- scRNA-seq (single-cell RNA-sequencing)
- Single cell genomics and epigenetics

# Single cell RNA-seq

- RNA-seq is a snapshot of
  what is happening in the sample
- Sample consists of many different cells
  and cell types
- Single cell RNA-seq –
  thousand of individual snapshots of many cells
  to capture the whole picture



# Why single-cell RNA-seq

Heterogenous populations:

- New cell subpopulations discovery
- Comparison of similar cell subpopulations
- Marker selection for cell subpopulations

Homogenous populations:

- Understanding heterogeneity
- Cellular states and cellular processes

Tracking of cell differentiation

PROTOCOL

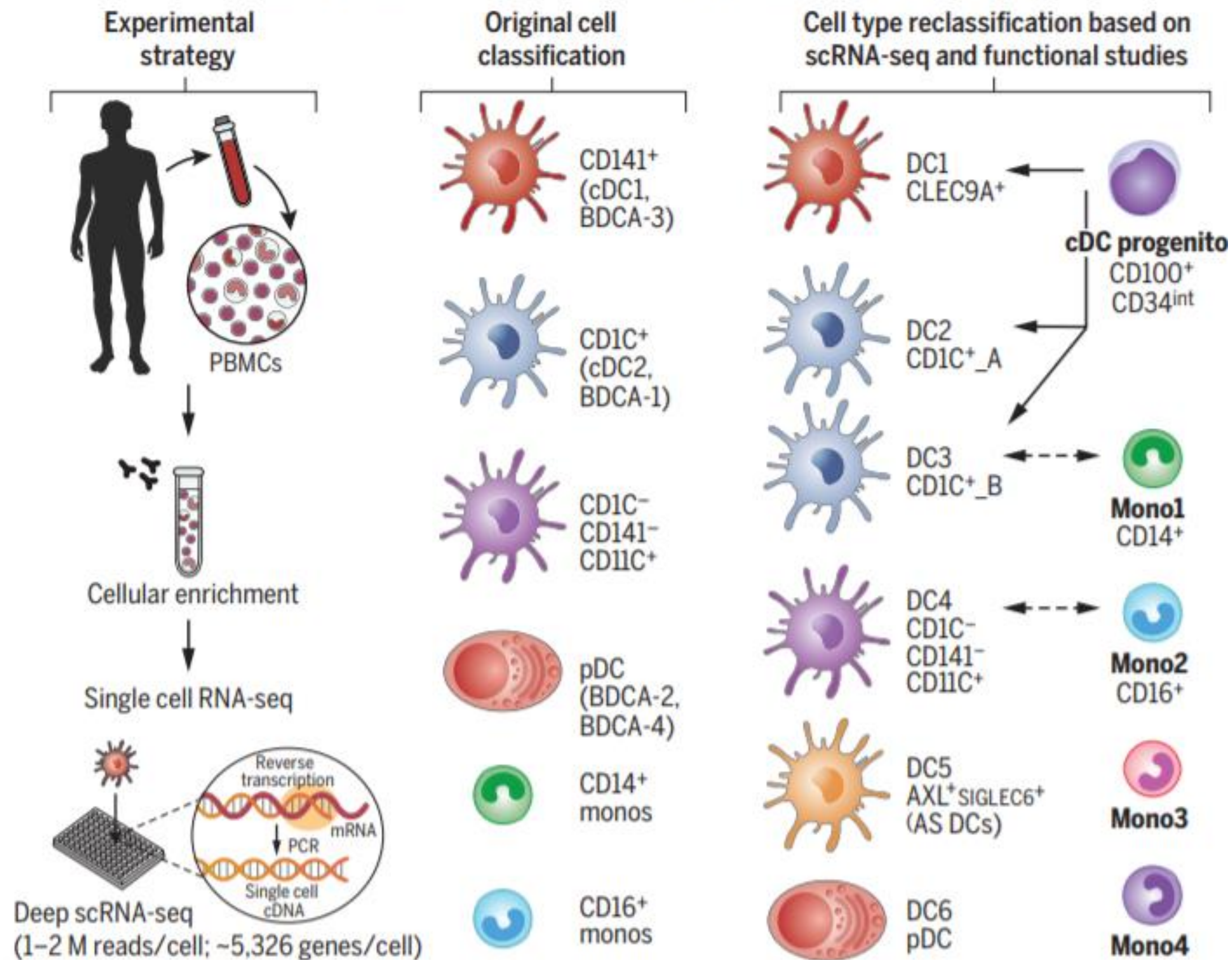
# Full-length RNA-seq from single cells using Smart-seq2

Simone Picelli[1], Omid R Faridani[1], Åsa K Björklund[1,2], Gösta Winberg[1,2], Sven Sagasser[1,2] & Rickard Sandberg[1,2]

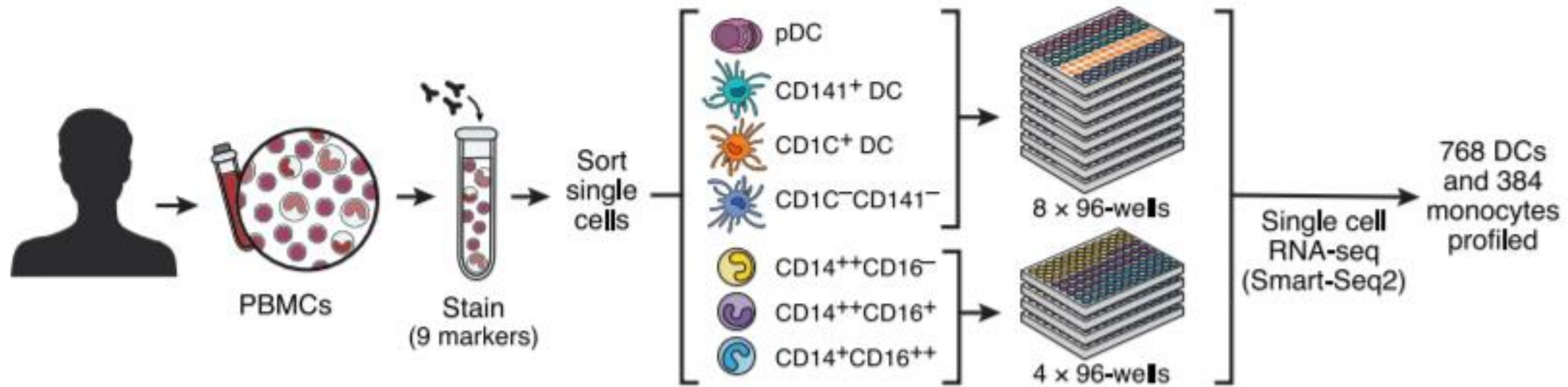
[1]Ludwig Institute for Cancer Research, Stockholm, Sweden. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

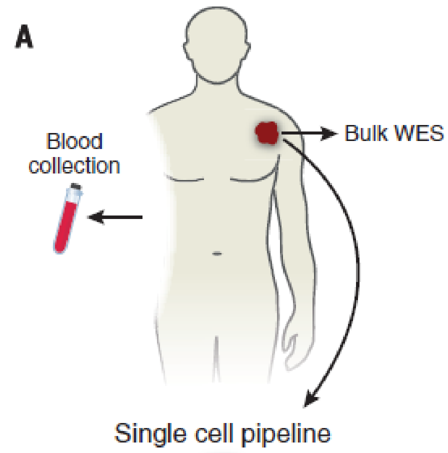
Published online 2 January 2014; doi:10.1038/nprot.2014.006 (2014)

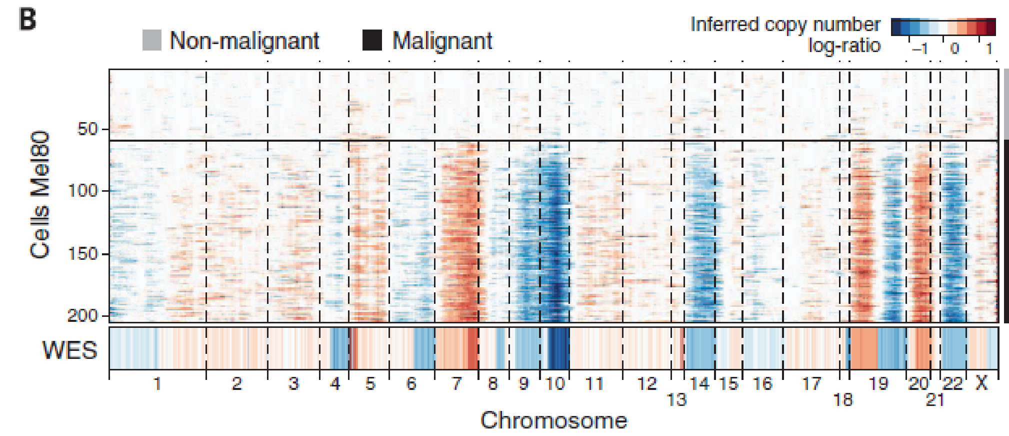
**Emerging methods for the accurate quantification of gene expression in individual cells hold promise for revealing the extent, function and origins of cell-to-cell variability. Different high-throughput methods for single-cell RNA-seq have been introduced that vary in coverage, sensitivity and multiplexing ability. We recently introduced Smart-seq for transcriptome analysis from single cells, and we subsequently optimized the method for improved sensitivity, accuracy and full-length coverage across transcripts. Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents. The entire protocol takes ~2 d from cell picking to having a final library ready for sequencing; sequencing will require an additional 1–3 d depending on the strategy and sequencer. The current limitations are the lack of strand specificity and the inability to detect nonpolyadenylated (polyA$^-$) RNA.**

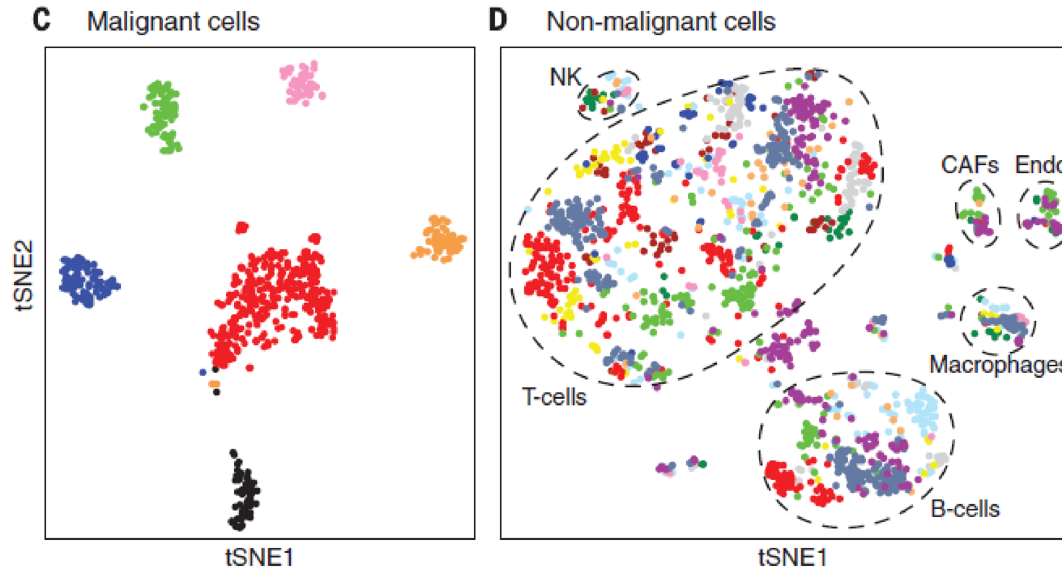



Villani, Satija et al
Science, 2017


A
PBMCs
Stain
(9 markers)
Sort
single
cells
pDC
CD141⁺ DC
CD1C⁺ DC
CD1C⁻CD141⁻
CD14⁺⁺CD16⁻
CD14⁺⁺CD16⁺
CD14⁺CD16⁺⁺
8 × 96-wells
4 × 96-wells
Single cell
RNA-seq
(Smart-Seq2)
768 DCs
and 384
monocytes
profiled




Tirosh, Izar et al, Science 2016

~ 4645 cells

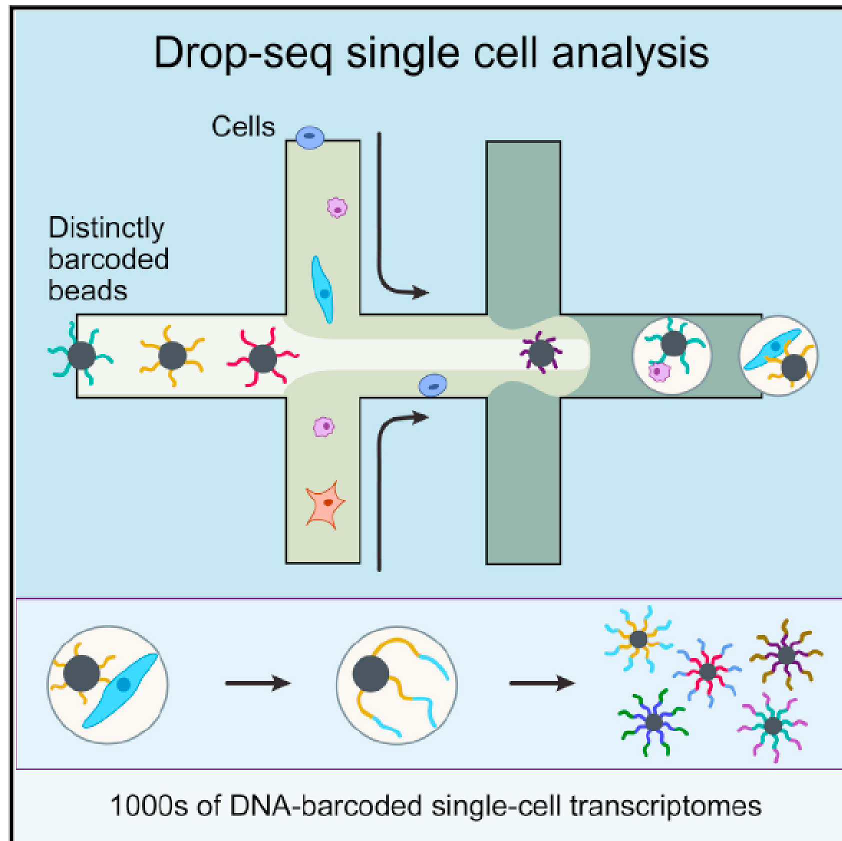
Cell

Resource

# Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Cell, 2015

## Graphical Abstract




## Authors

Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

## Correspondence

emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

## In Brief

Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

nature COMMUNICATIONS

ARTICLE

Received 20 Sep 2016 | Accepted 23 Nov 2016 | Published 16 Jan 2017

DOI: 10.1038/ncomms14049 OPEN

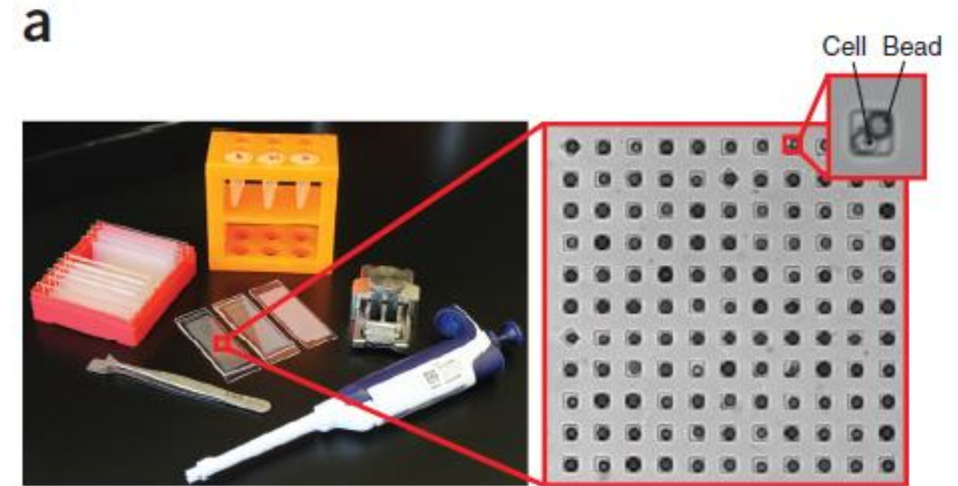
# Massively parallel digital transcriptional profiling of single cells

~ 68 000 cells

Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1], Stefanie Y. Nishimura[1], Michael Schnall-Levin[1], Paul W. Wyatt[1], Christopher M. Hindson[1], Rajiv Bharadwaj[1], Alexander Wong[1], Kevin D. Ness[1], Lan W. Beppu[4], H. Joachim Deeg[4], Christopher McFarland[5], Keith R. Loeb[4,6], William J. Valente[2,7,8], Nolan G. Ericson[2], Emily A. Stevens[4], Jerald P. Radich[4], Tarjei S. Mikkelsen[1], Benjamin J. Hindson[1] & Jason H. Bielas[2,6,8,9]

# Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

Todd M Gierahn[1,8], Marc H Wadsworth II[2–4,8], Travis K Hughes[2–4,8], Bryan D Bryson[4,5], Andrew Butler[6,7], Rahul Satija[6,7], Sarah Fortune[4,5], J Christopher Love[1,3,4,9] & Alex K Shalek[2,3,4,9]



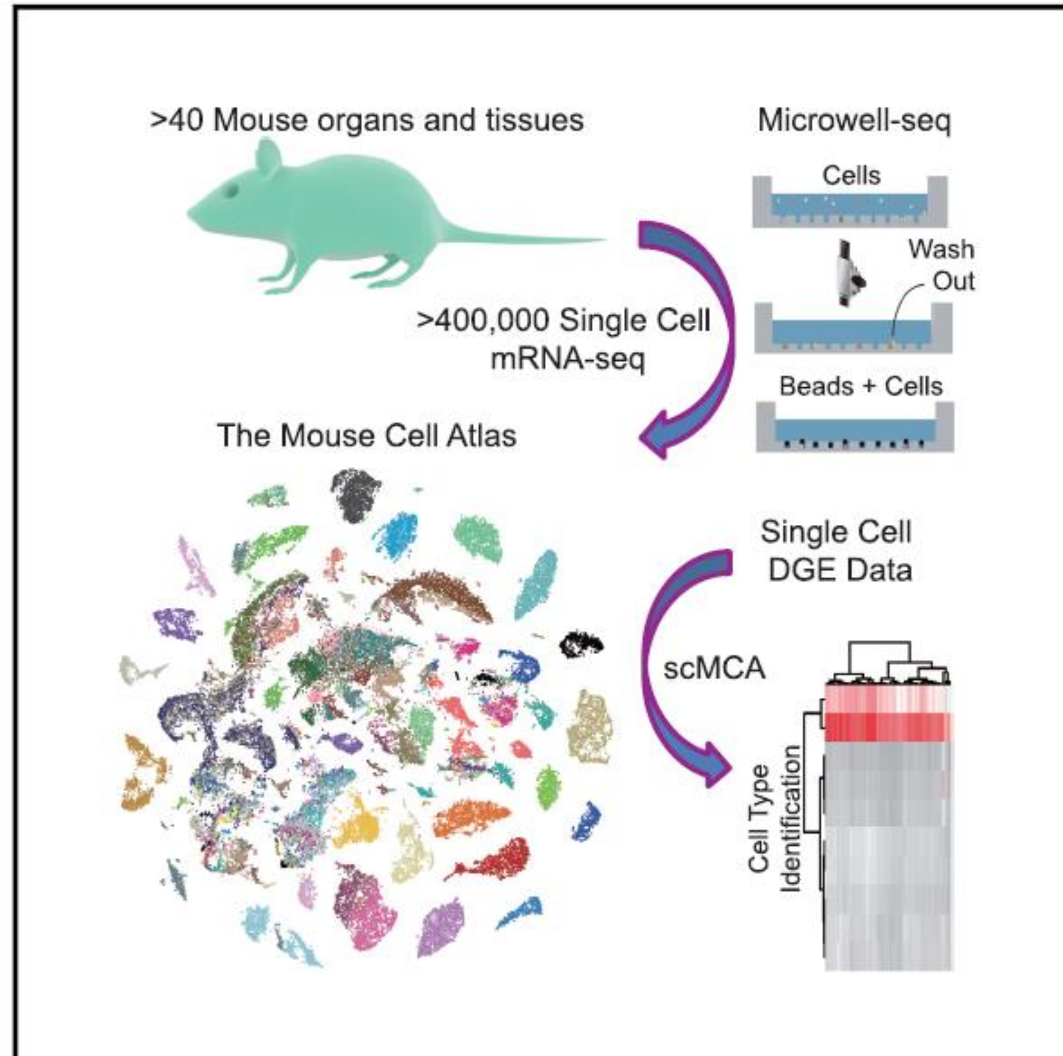
Cell

Resource

UNIVERSITY

# Mapping the Mouse Cell Atlas by Microwell-Seq

## Graphical Abstract




## Authors

Xiaoping Han, Renying Wang,
Yincong Zhou, ..., Guo-Cheng Yuan,
Ming Chen, Guoji Guo

## Correspondence

xhan@zju.edu.cn (X.H.),
ggj@zju.edu.cn (G.G.)

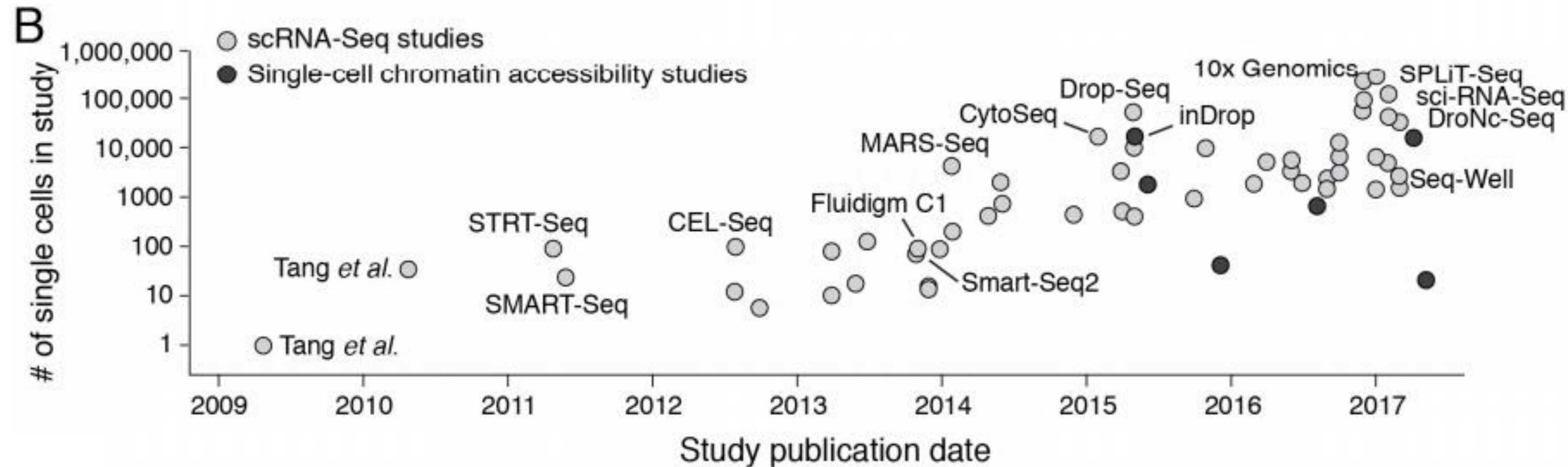
## In Brief

Development of Microwell-seq allows construction of a mouse cell atlas at the single-cell level with a high-throughput and low-cost platform.

February 2018

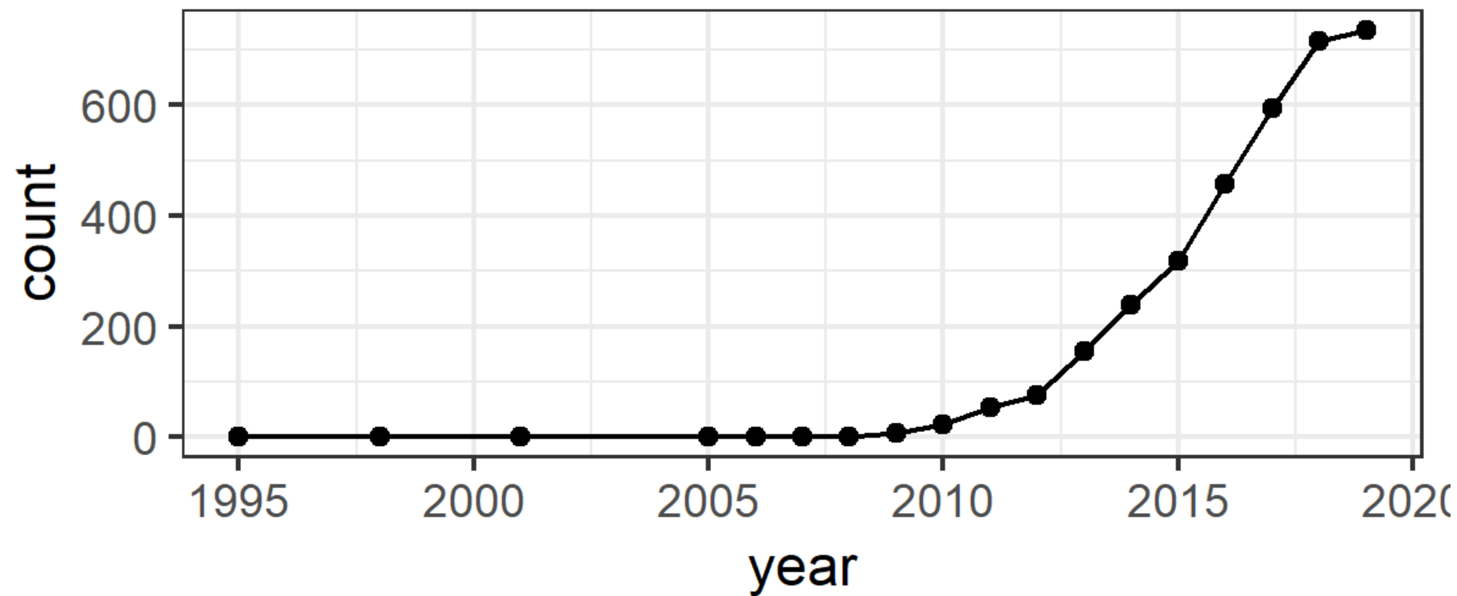
ITMO UNIVERSITY

# Growth of single cell technologies




* Taken from Human Cell Atlas white paper
https://www.humancellatlas.org/files/HCA_WhitePaper_18Oct2017-copyright.pdf

# Growth of single cell technologies



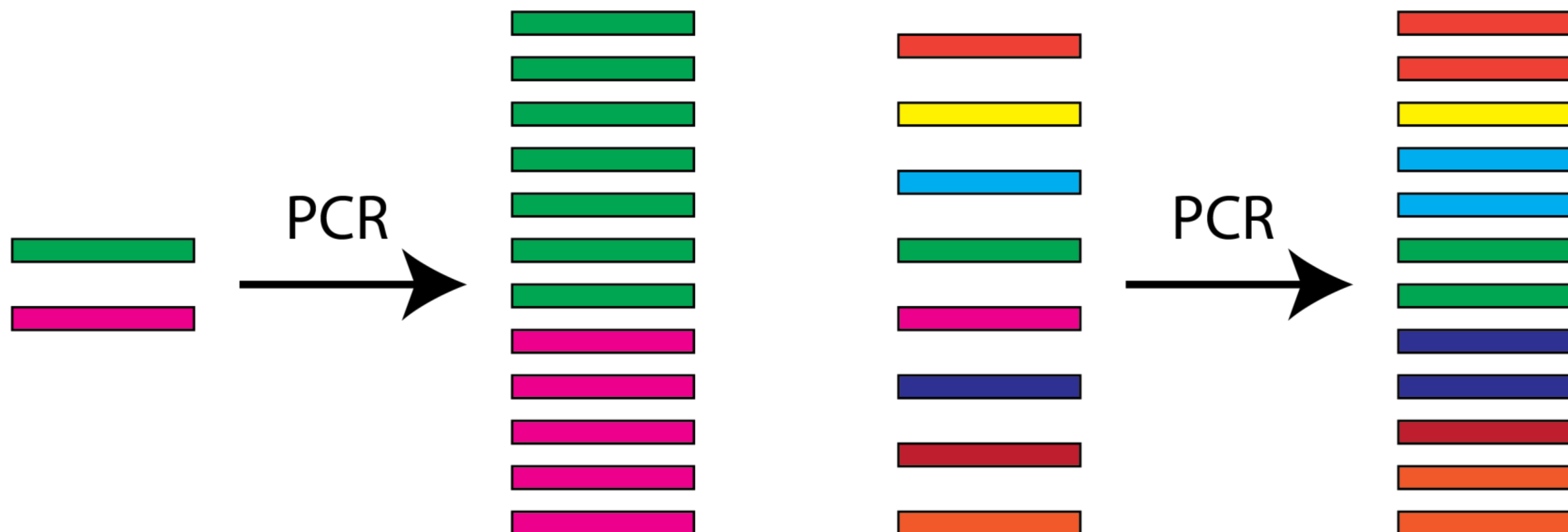

* Based on number of pubmed results

# Main challenges

- **How to amplify and sequence small number of RNA (typical mammalian cell has only 200 000 mRNA molecules)?**
- How to isolate cells?
- How to work with big number of cells?

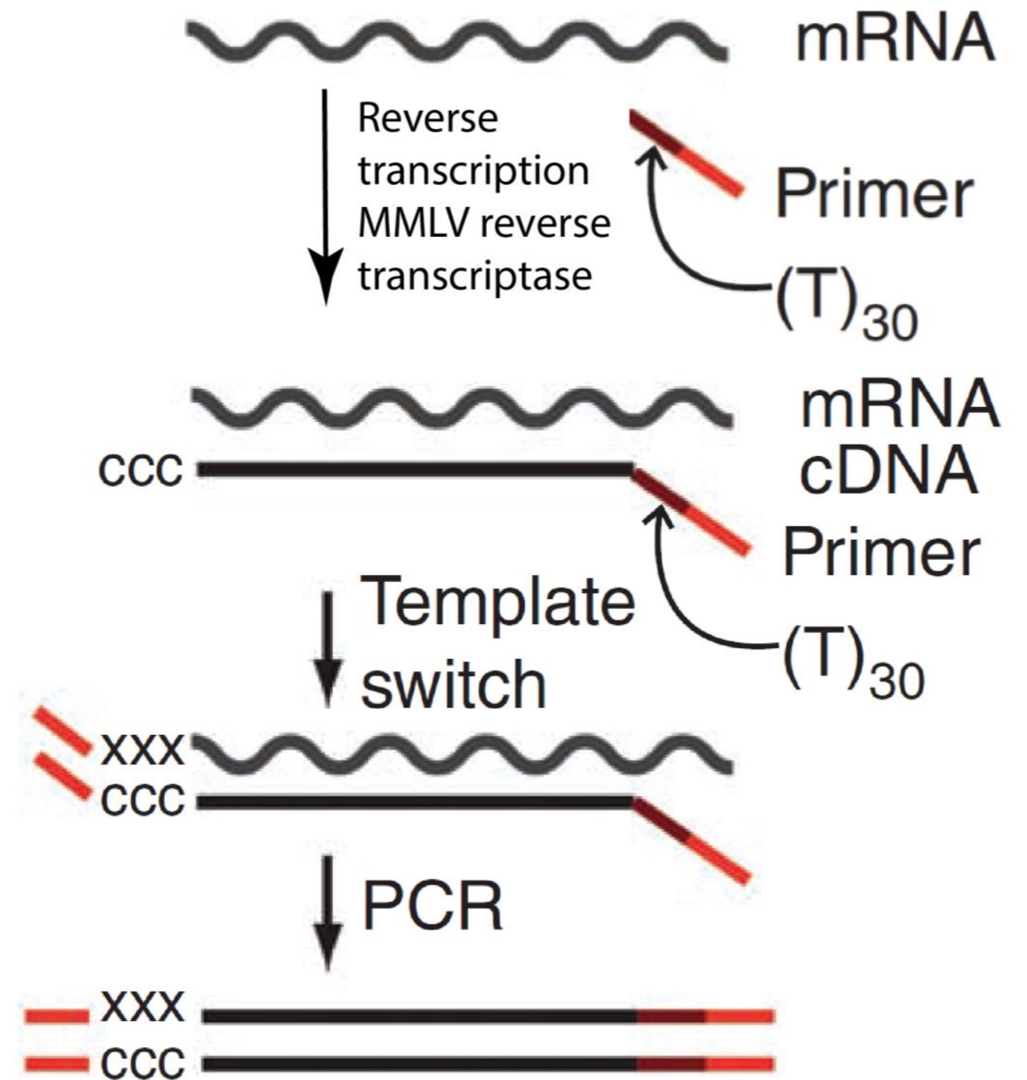
# Low library complexity

- Small numbers of mRNA molecules yield low complexity cDNA library
- cDNA molecules to be amplified by PCR
- We don’t want to sequence tons of PCR duplicates



# Template switching PCR for low input

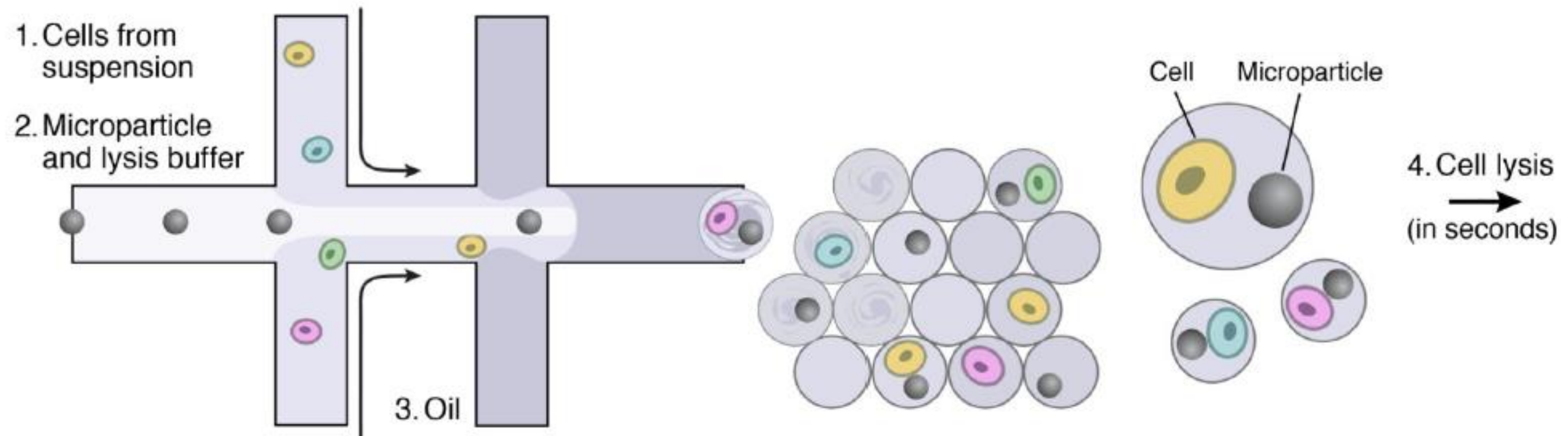
# Switching Mechanism At the 5' end of RNA Template (SMART)



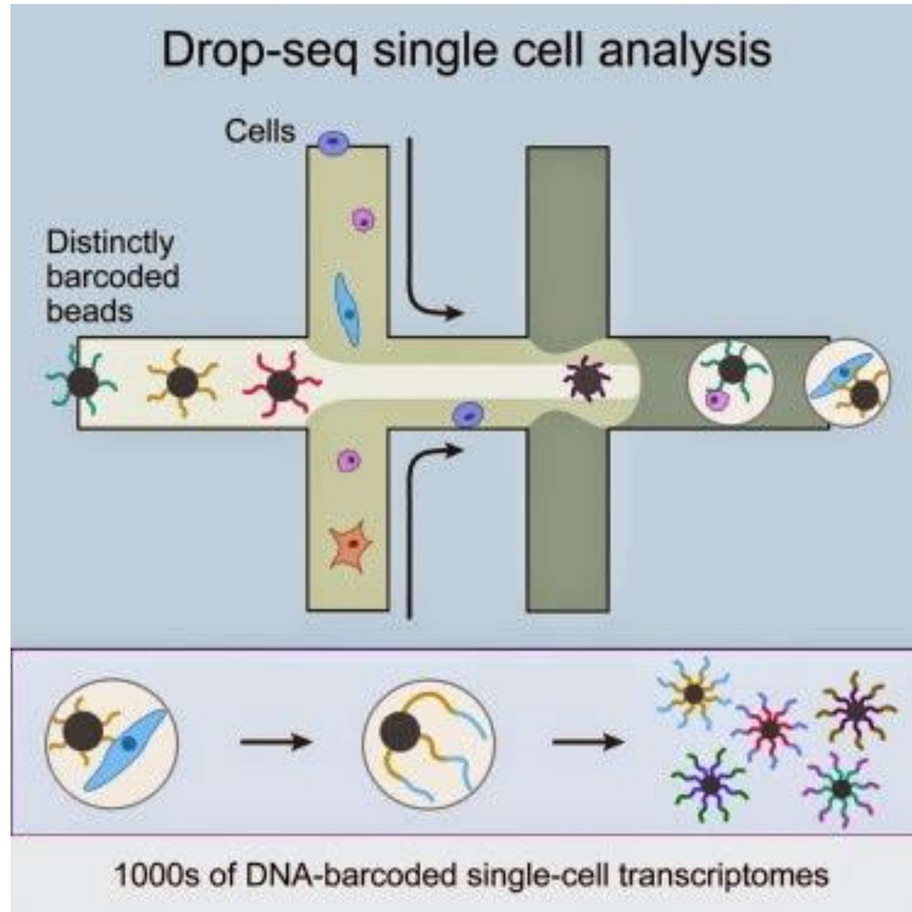
# Main challenges

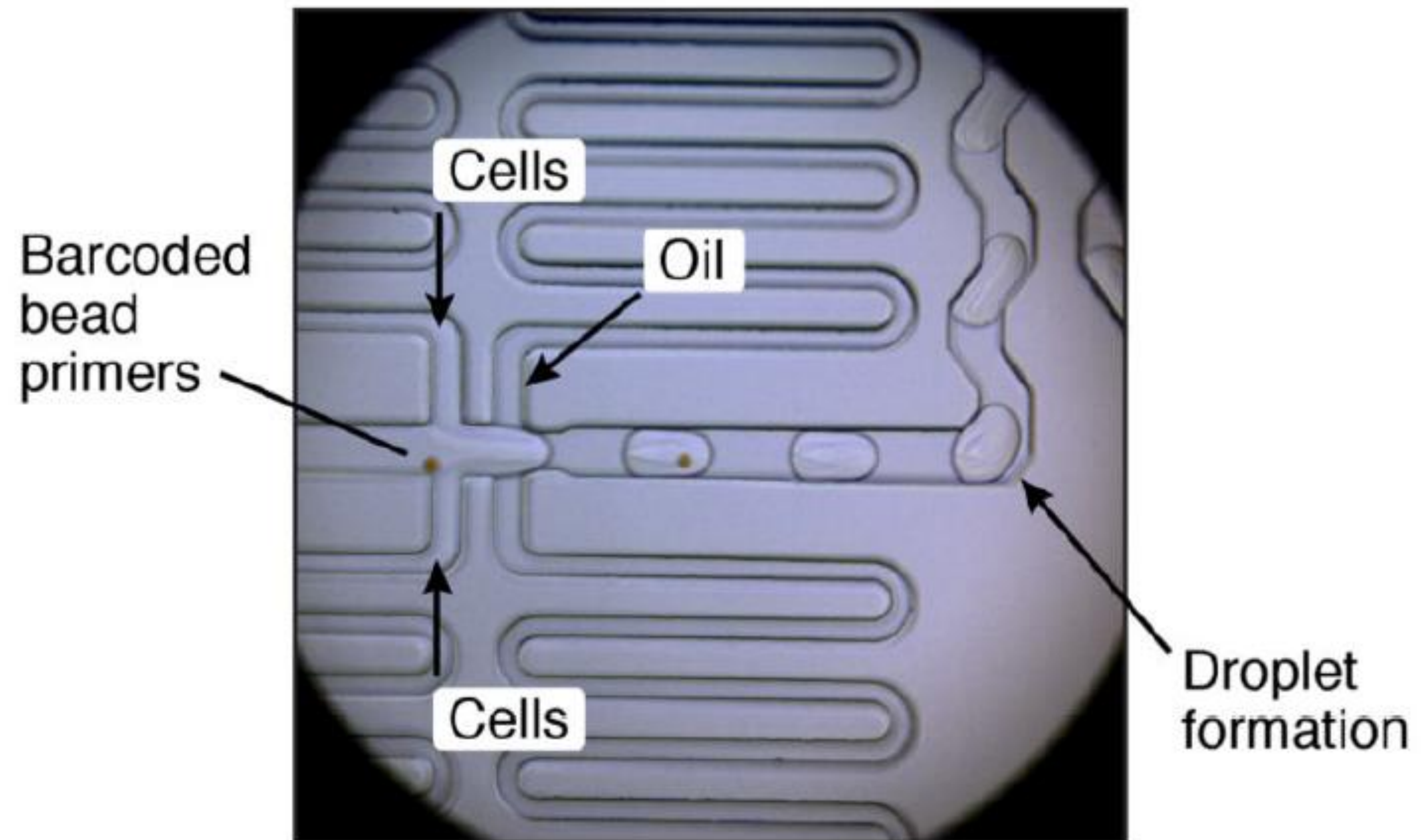
- How to amplify and sequence small number of RNA (typical mammalian cell has only 200 000 mRNA molecules)?
- **How to isolate cells?**
- **How to work with big number of cells?**

# Drop-seq schematics

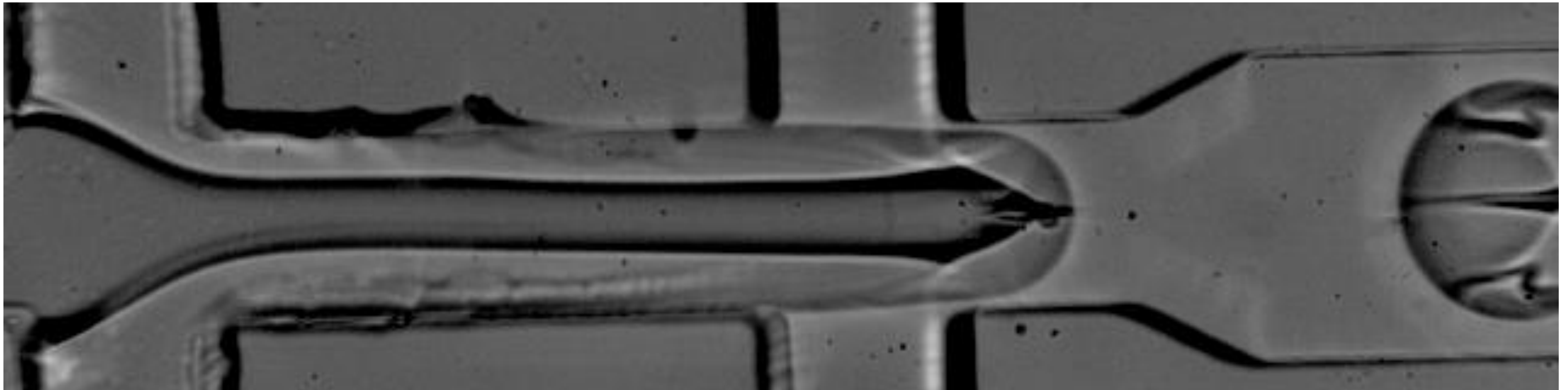


# How does it look like?

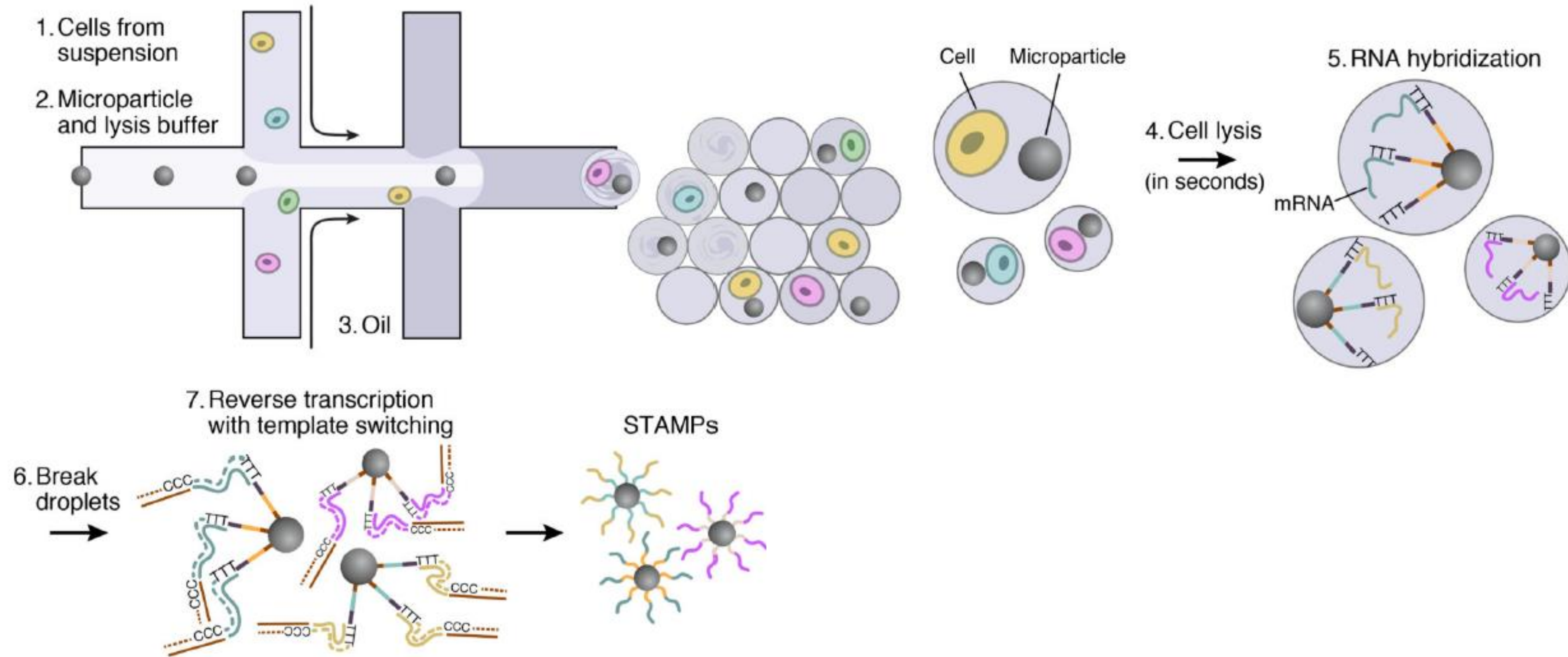


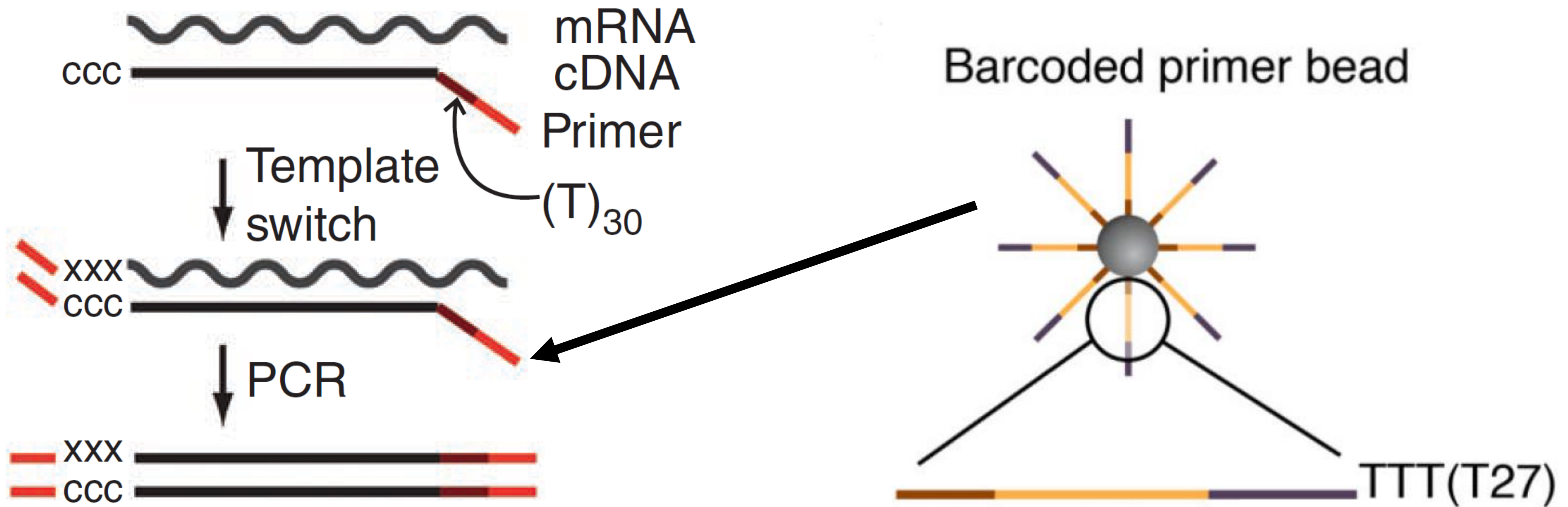
# How does it look like



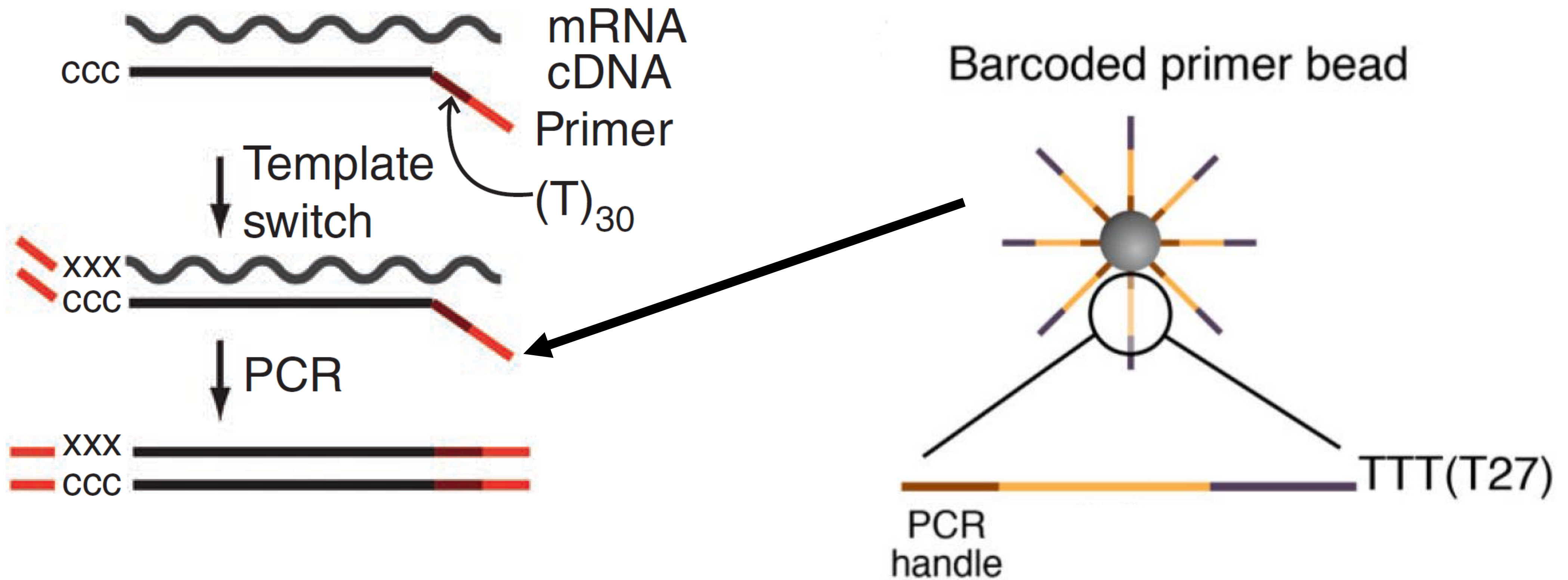
# Drop-seq schematics



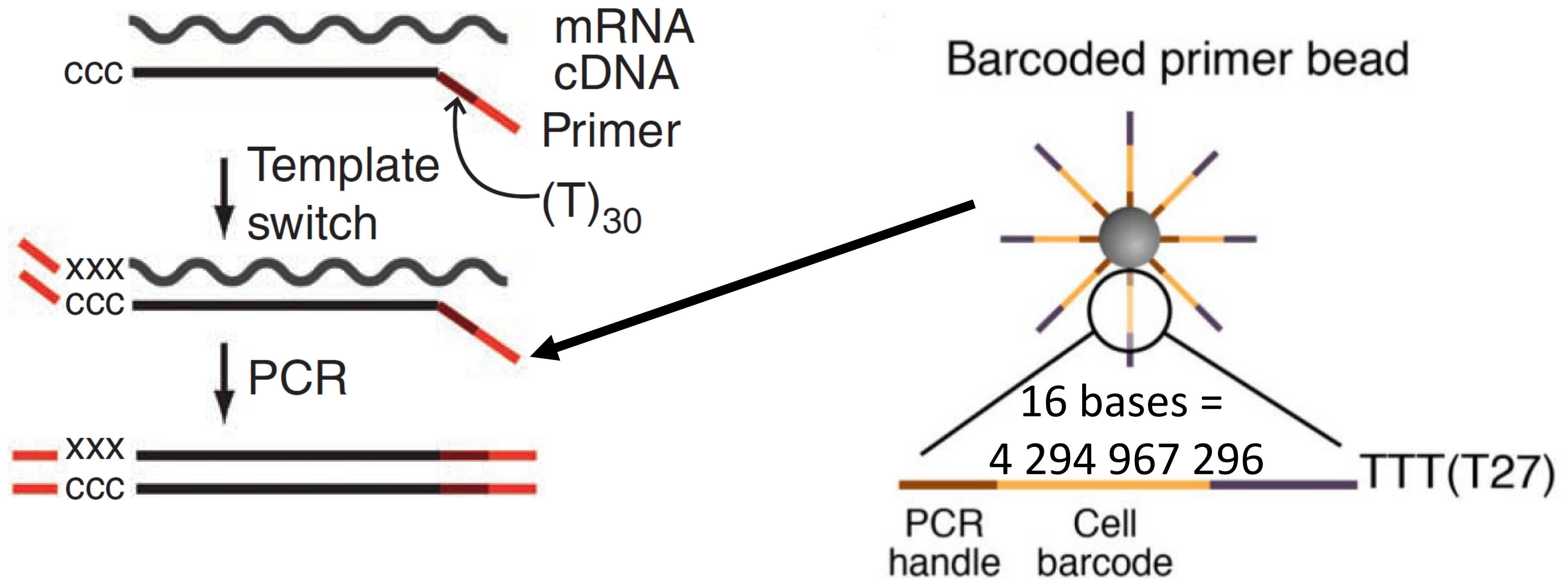
# Barcoding



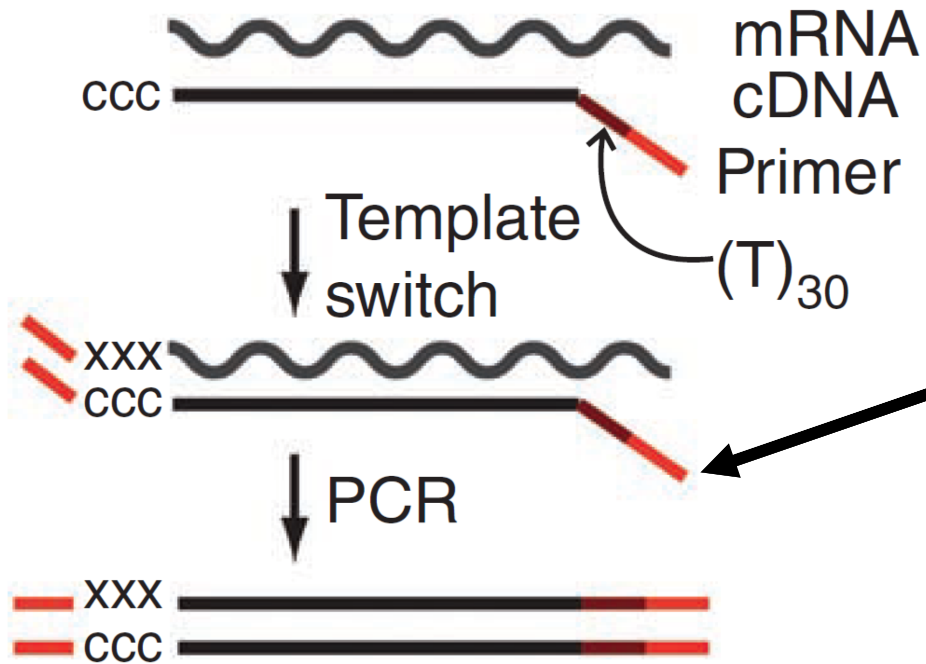
# Barcoding

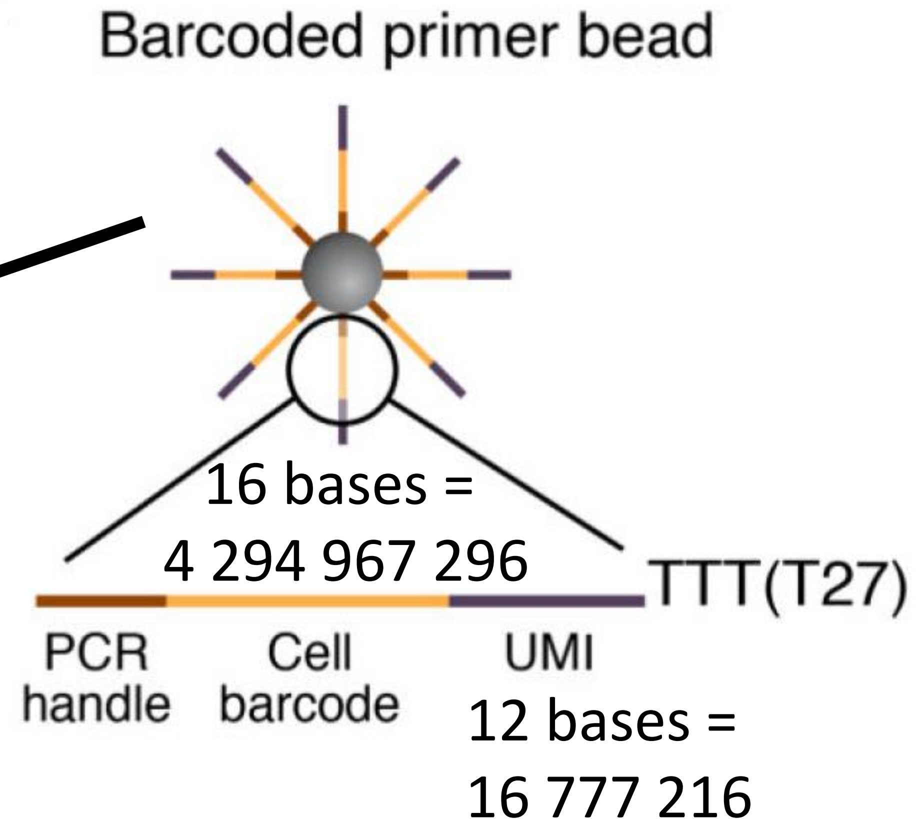


# Barcoding

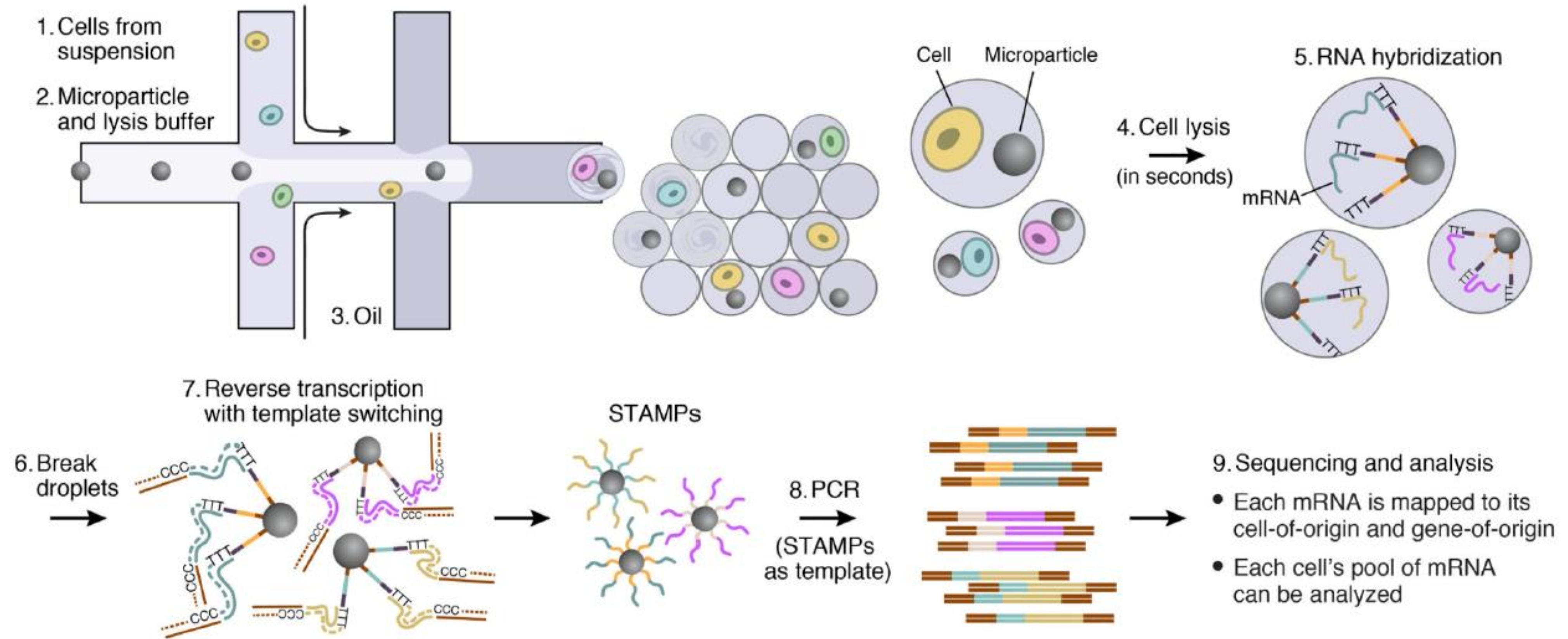


# Barcoding

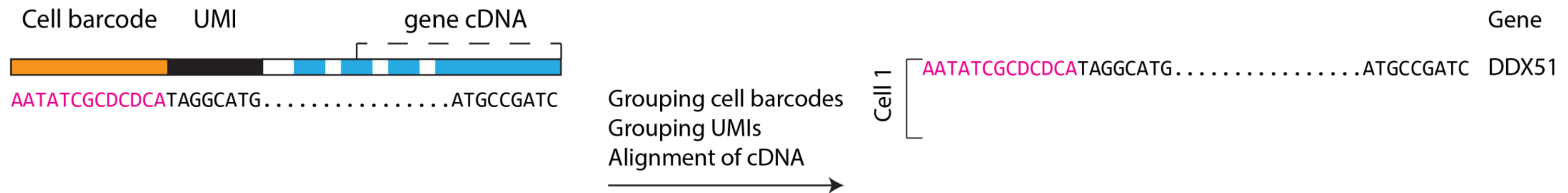


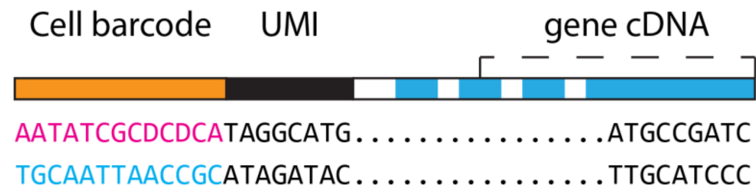
# Drop-seq schematics



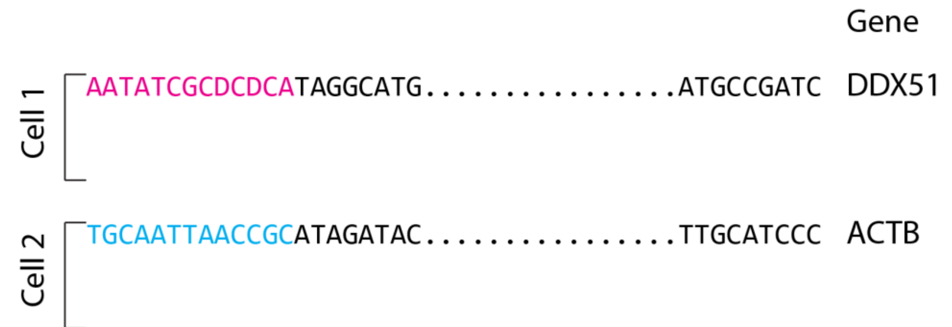
# Sequencing



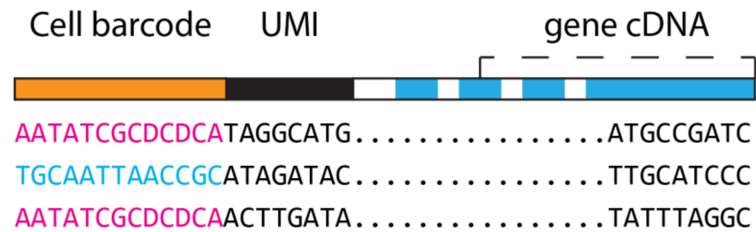
# Sequencing



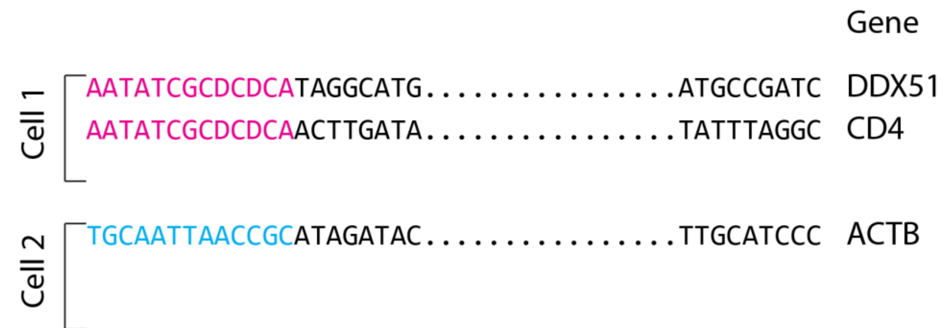

Grouping cell barcodes
Grouping UMIs
Alignment of cDNA



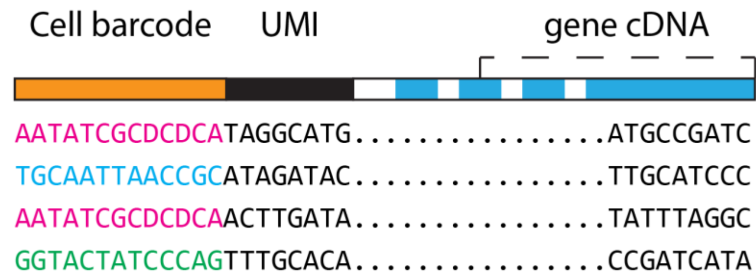
# Sequencing




Grouping cell barcodes
Grouping UMIs
Alignment of cDNA



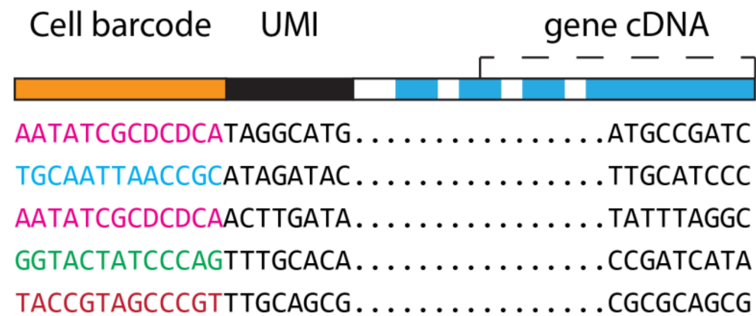
# Sequencing




Grouping cell barcodes
Grouping UMIs
Alignment of cDNA

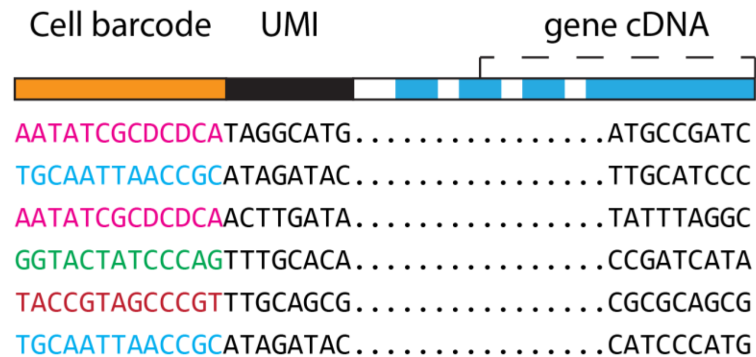
ITMO UNIVERSITY

# Sequencing



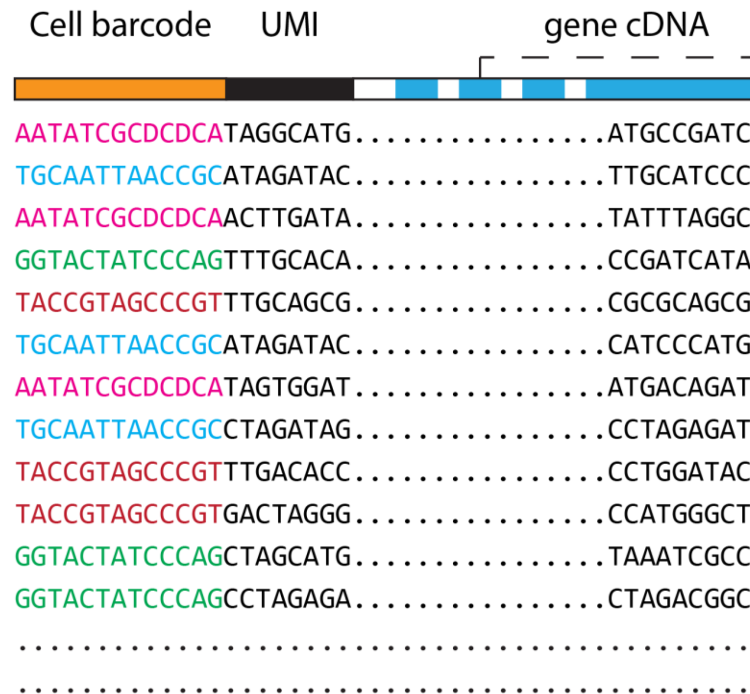

Grouping cell barcodes
Grouping UMIs
Alignment of cDNA

# Sequencing




Grouping cell barcodes
Grouping UMIs
Alignment of cDNA

# Sequencing




Grouping cell barcodes
Grouping UMIs
Alignment of cDNA

| | | Gene |
|---|---|---|
| Cell 1 | AATATCGCDCDCATAGGCATG...............ATGCCGATC | DDX51 |
| | AATATCGCDCDCAACTTGATA...............TATTTAGGC | CD4 |
| | AATATCGCDCDCATAGTGGAT...............ATGACAGAT | ACTB |
| Cell 2 | TGCAATTAACCGCATAGATAC...............TTGCATCCC | ACTB |
| | TGCAATTAACCGCATAGATAC...............CATCCCATG | ACTB |
| | TGCAATTAACCGCCTAGATAG...............CCTAGAGAT | RPS15 |
| Cell 3 | GGTACTATCCCAGTTTGCACA...............CCGATCATA | GAPDH |
| | GGTACTATCCCAGCTAGCATG...............TAAATCGCC | GTPBP4 |
| | GGTACTATCCCAGCCTAGAGA...............CTAGACGGC | ARL1 |
| Cell 4 | TACCGTAGCCCGTTTGCAGCG...............CGCGCAGCG | ACTB |
| | TACCGTAGCCCGTTTGACACC...............CCTGGATAC | NOP2 |
| | TACCGTAGCCCGTGACTAGGG...............CCATGGGCT | NOTCH2 |

# Sequencing

Gene

Cell 1
AATATCGCDCDCATAGGCATG................ATGCCGATC DDX51
AATATCGCDCDCAACTTGATA................TATTTAGGC CD4
AATATCGCDCDCATAGTGGAT................ATGACAGAT ACTB

Cell 2
TGCAATTAACCGCATAGATAC................TTGCATCCC ACTB
TGCAATTAACCGCATAGATAC................CATCCCATG ACTB
TGCAATTAACCGCCTAGATAG................CCTAGAGAT RPS15

Cell 3
GGTACTATCCCAGTTTGCACA................CCGATCATA GAPDH
GGTACTATCCCAGCTAGCATG................TAAATCGCC GTPBP4
GGTACTATCCCAGCCTAGAGA................CTAGACGGC ARL1

Cell 4
TACCGTAGCCCGTTTGCAGCG................CGCGCAGCG ACTB
TACCGTAGCCCGTTTGACACC................CCTGGATAC NOP2
TACCGTAGCCCGTGACTAGGG................CCATGGGCT NOTCH2
........................................

Count unique UMIs

Create digital expression matrix

| Cell: | AATATCGCDCDCA | TGCAATTAACCGC | GGTACTATCCCAG | .. | TACCGTAGCCCGT |
|---|---|---|---|---|---|
| Gene 1 | 10 | 03 | 12 | .. | 14 |
| Gene 2 | 05 | 00 | 00 | .. | 00 |
| Gene 3 | 12 | 10 | 15 | .. | 10 |
| ... | .. | .. | .. | .. | .. |
| Gene M | 11 | 01 | 00 | .. | 00 |

Matrix describing expression of thousands of genes in thousands of cells

# All questions were addressed and answered

- Low input mRNA – template switching PCR
- Cell isolation – microfluidics
- Read identification – Cell barcodes
- Dealing with PCR duplicates – Cell/UMI barcodes

# All questions were addressed and answered

- Low input mRNA – template switching PCR
- **Cell isolation – microfluidics (might be different)**
- **Barcoding with beads proved to be very effective**

# Dataset for today

- https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3

## 10k PBMCs from a Healthy Donor (v3 chemistry)

Single Cell Gene Expression Dataset by Cell Ranger 3.0.0

Peripheral blood mononuclear cells (PBMCs) from a healthy donor (the same cells were used to generate pbmc_1k_v2, pbmc_10k_v3). PBMCs are primary cells with relatively small amounts of RNA (~1pg RNA/cell).

- 11,769 cells detected
- Sequenced on Illumina NovaSeq with approximately 54,000 reads per cell
- 28bp read1 (16bp Chromium barcode and 12bp UMI), 91bp read2 (transcript), and 8bp I7 sample barcode
- run with --expect-cells=10000

*Published on November 19, 2018*

*This dataset is licensed under the Creative Commons Attribution license.*

View Summary

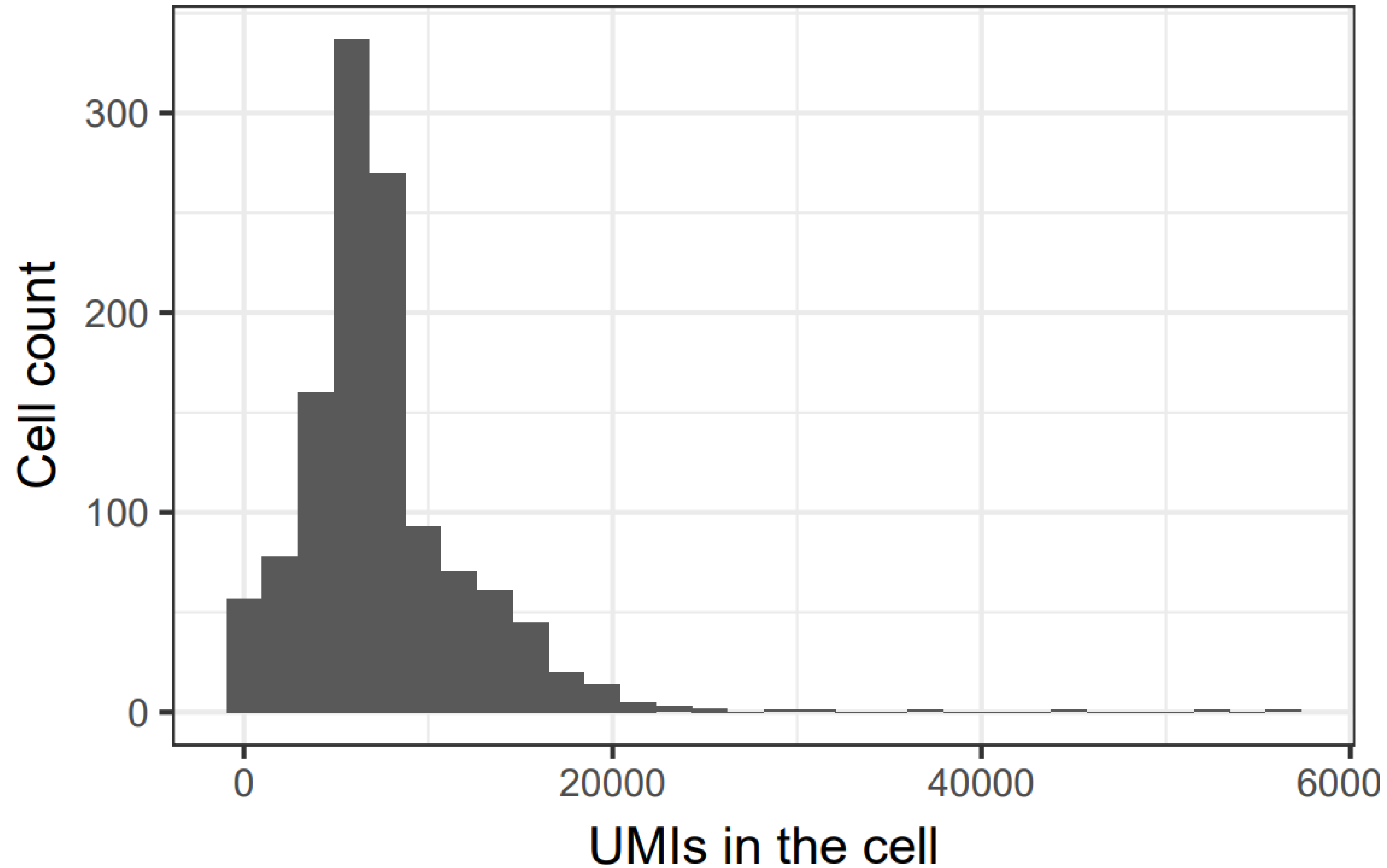
# Dataset summary

- http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_10k_v3/pbmc_10k_v3_web_summary.html

# Basic steps to analysis of scRNA-seq

- **Filtering out “bad” barcodes**
- **Normalizing expression levels: (scaling and log2 normalizing)**
- Visualization (tSNE plots)
- Clustering
- Cellular subset annotation

# UMI distribution



# We have to normalize it

| Cell: | AATATCGCDCDCA | TGCAATTAACCGC | GGTACTATCCCAG | .. | TACCGTAGCCCGT |
|---|---|---|---|---|---|
| Gene 1 | 10 | 03 | 12 | .. | 14 |
| Gene 2 | 05 | 00 | 00 | .. | 00 |
| Gene 3 | 12 | 10 | 15 | .. | 10 |
| ... | .. | .. | .. | .. | .. |
| Gene M | 11 | 01 | 00 | .. | 00 |

# We have to normalize it

| Cell: | AATATCGCDCDCA |
|---|---|
| Gene 1 | 10 |
| Gene 2 | 05 |
| Gene 3 | 12 |
| ... | .. |
| Gene M | 11 |

→

AATATCGCDCDCA

Gene 1 $\frac{10}{10 + 5 + 12 + \ldots + 11} \cdot 10000$

Gene 2 $\frac{5}{10 + 5 + 12 + \ldots + 11} \cdot 10000$

Gene 3 $\frac{12}{10 + 5 + 12 + \ldots + 11} \cdot 10000$

... ...

Gene M $\frac{11}{10 + 5 + 12 + \ldots + 11} \cdot 10000$

# We have to normalize it

Cell: TGCAATTAACCGC

| Gene | Count |
| --- | --- |
| Gene 1 | 03 |
| Gene 2 | 00 |
| Gene 3 | 10 |
| ... | .. |
| Gene M | 01 |

→

TGCAATTAACCGC

Gene 1 $\frac{3}{3 + 0 + 10 + \ldots + 1} \cdot 10000$

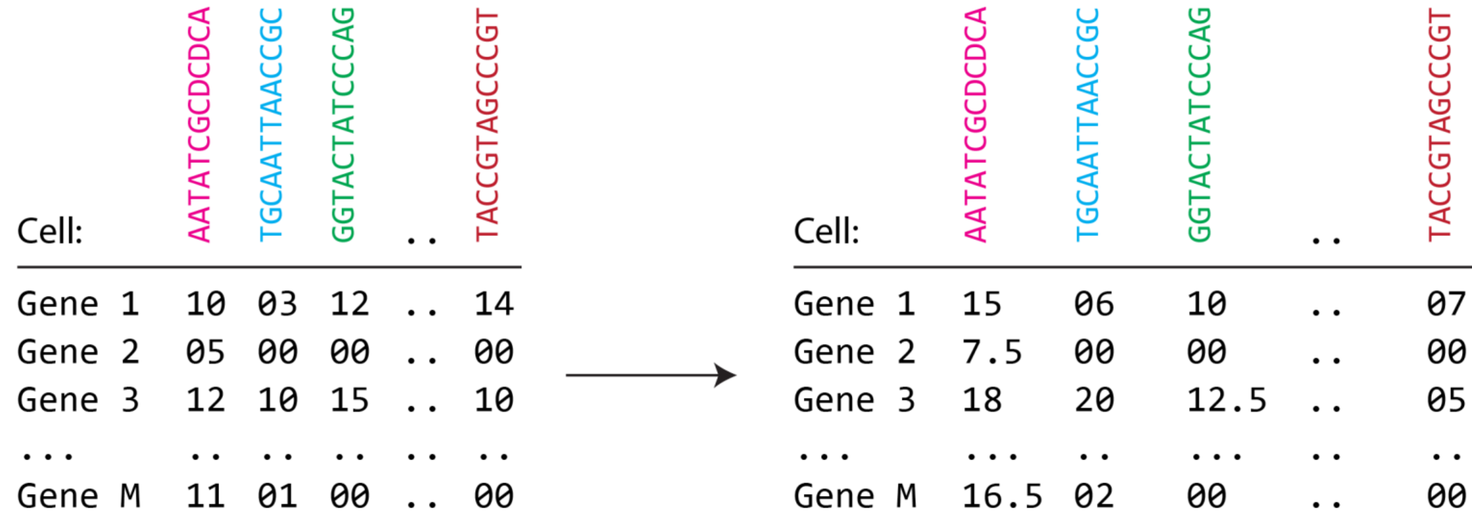
Gene 2 $\frac{0}{3 + 0 + 10 + \ldots + 1} \cdot 10000$

Gene 3 $\frac{10}{3 + 0 + 10 + \ldots + 1} \cdot 10000$

... ...

Gene M $\frac{1}{3 + 0 + 10 + \ldots + 1} \cdot 10000$

# We have to normalize it

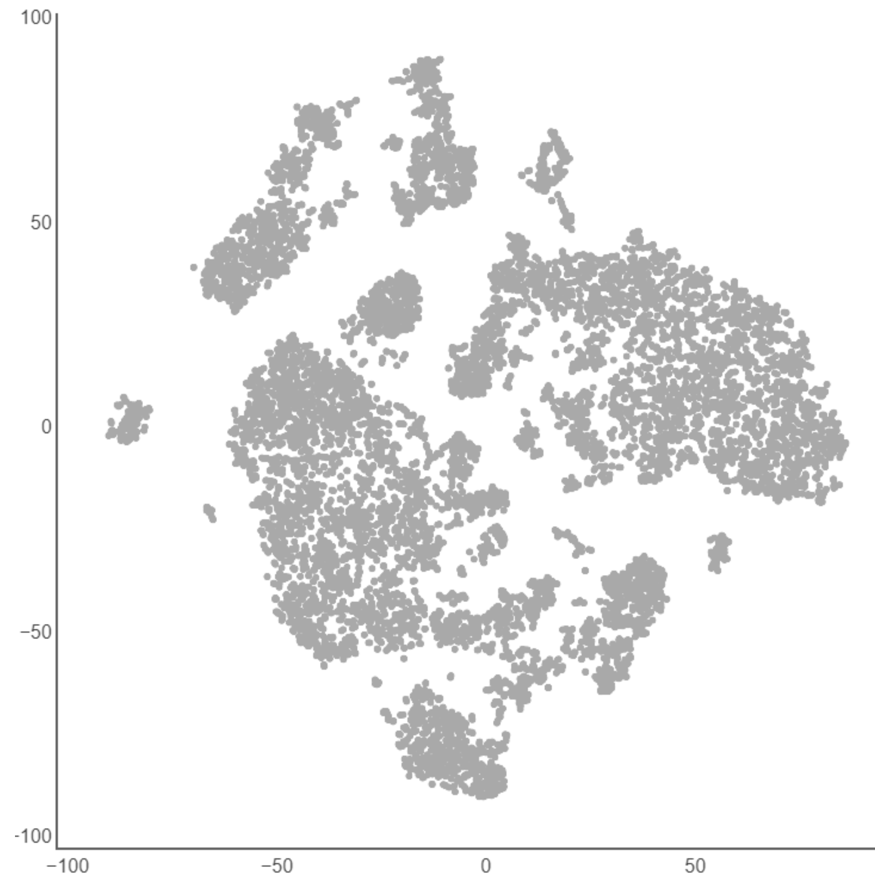


# Basic steps to analysis of scRNA-seq

- Filtering out “bad” barcodes
- Normalizing expression levels
- **Visualization (tSNE plots)**
- **Clustering**
- Cellular subset annotation

# PBMC dataset: visualization by tSNE

**tSNE** (t-distributed stochastic neighbor embedding):
Non-linear dimensionality reduction technique that aims to put data in 2 or 3 dimensional space and save “distance” between each two dots.




https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3
http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

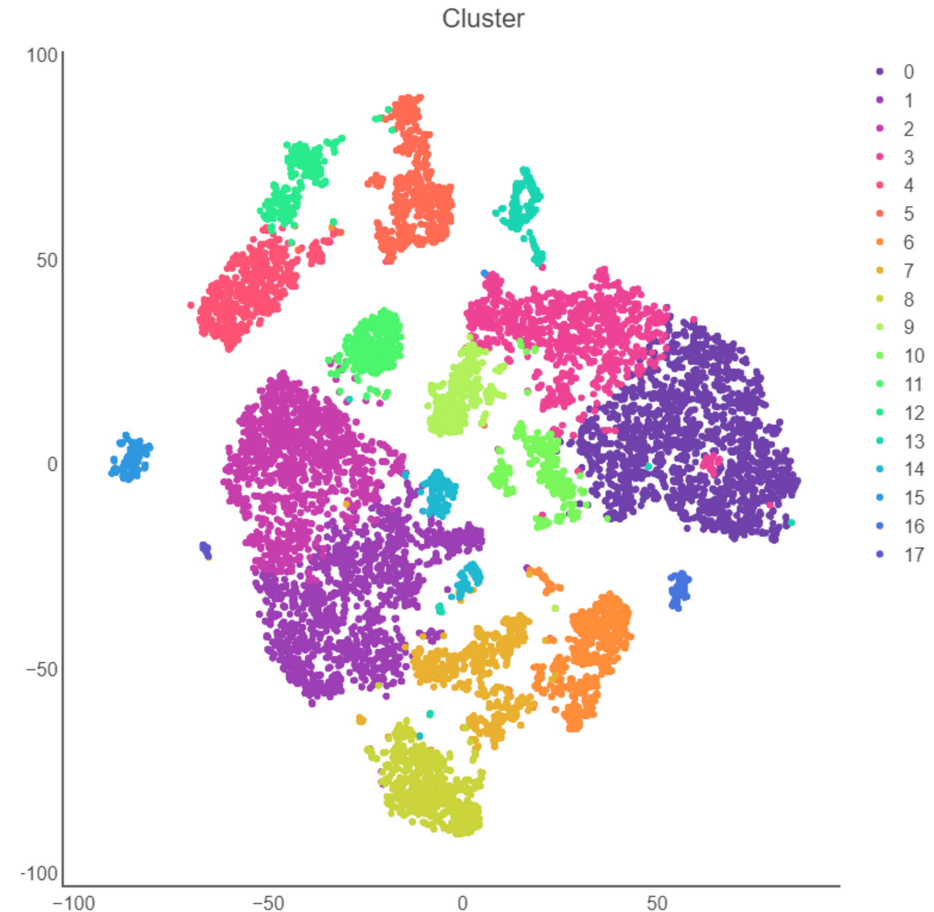
# PBMC dataset: clustering and annotation

Clustering:

- Graph-based clustering (preferred);
- K-means.

Annotation:

- For each cluster, perform differential expression: cluster against all others;
- Top DE genes exprected to be highly distinctive marker genes.




https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3

Gene expression

# Identification of cell types from single-cell transcriptomes using a novel clustering method

**Chen Xu and Zhengchang Su***

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on October 13, 2014; revised on January 20, 2015; accepted on February 8, 2015

## Abstract

**Motivation:** The recent advance of single-cell technologies has brought new insights into complex biological phenomena. In particular, genome-wide single-cell measurements such as transcriptome sequencing enable the characterization of cellular composition as well as functional variation in homogenic cell populations. An important step in the single-cell transcriptome analysis is to group cells that belong to the same cell types based on gene expression patterns. The corresponding computational problem is to cluster a noisy high dimensional dataset with substantially fewer objects (cells) than the number of variables (genes).
**Results:** In this article, we describe a novel algorithm named shared nearest neighbor (SNN)-Cliq that clusters single-cell transcriptomes. SNN-Cliq utilizes the concept of shared nearest neighbor that shows advantages in handling high-dimensional data. When evaluated on a variety of synthetic and real experimental datasets, SNN-Cliq outperformed the state-of-the-art methods tested. More importantly, the clustering results of SNN-Cliq reflect the cell types or origins with high accuracy.
**Availability and implementation:** The algorithm is implemented in MATLAB and Python. The source code can be downloaded at http://bioinfo.uncc.edu/SNNCliq.
**Contact:** zcsu@uncc.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.



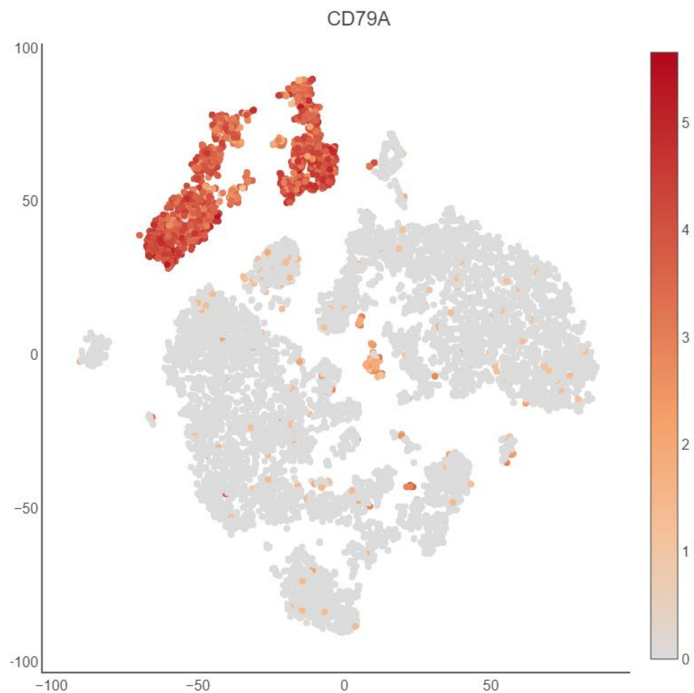
- Instead of defining clusters based on distance we first find “neighbors” for each of the cells
- Cells that have a lot of neighbors in common, most likely “live in the same neighborhood”
- Algorithm is trying to find such neighborhoods

# Basic steps to analysis of scRNA-seq

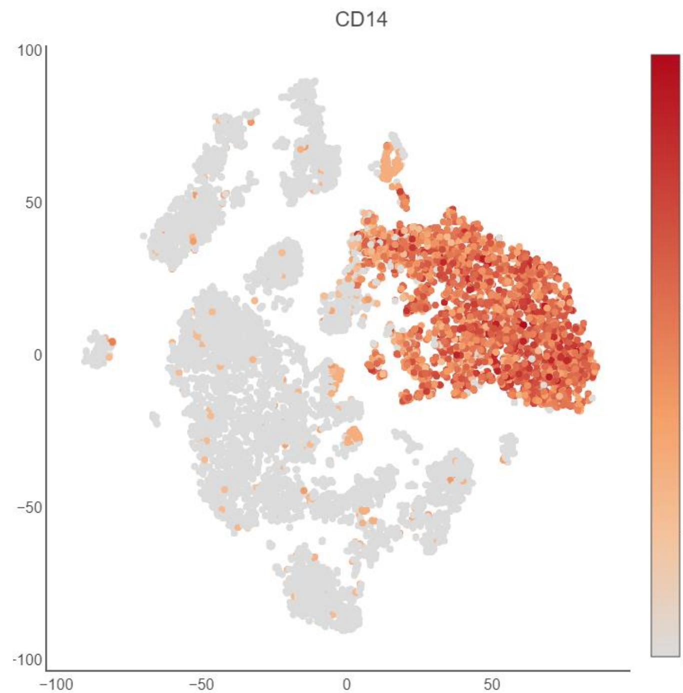
- Filtering out “bad” barcodes
- Normalizing expression levels
- Visualization (tSNE plots)
- Clustering
- **Cellular subset annotation**

# PBMC dataset: check for the known marker genes

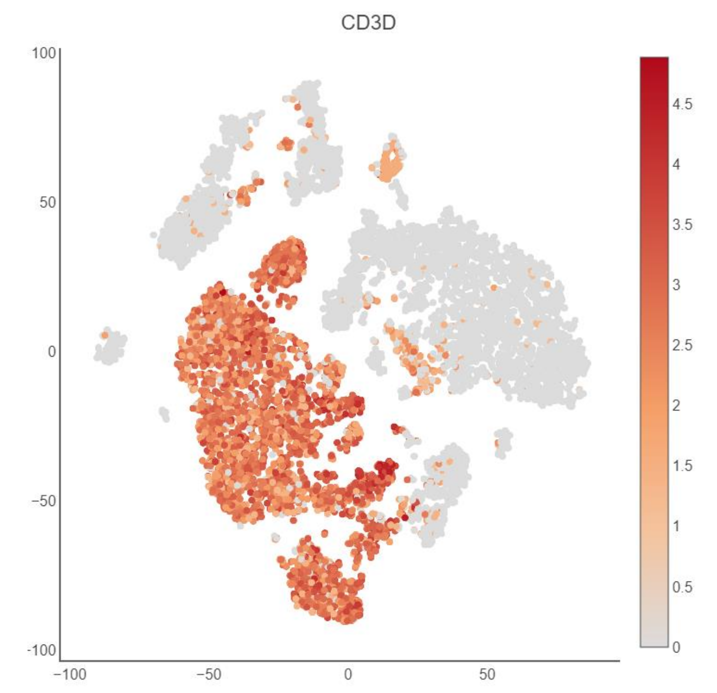
CD79 expression: B-cell marker




CD14 expression:
CD14+ Monocyte marker




CD3E expression:
T-cell marker




https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_1k_v3

# Communication is important

- When clusters are found we want to identify which cell subsets are presented, to “annotate” them
- If you are a bioinformatician and you are a single-cell RNA-seq dataset that have been designed\done by you, this is a perfect to time to go and talk to a biologist who performed\designed the experiment
- If you are a biologist who designed\performed single-cell RNA-seq experiment, chances are, you know all cellular subsets and markers better than almost anyone else
- This is where you communicate and try to make sense of the data

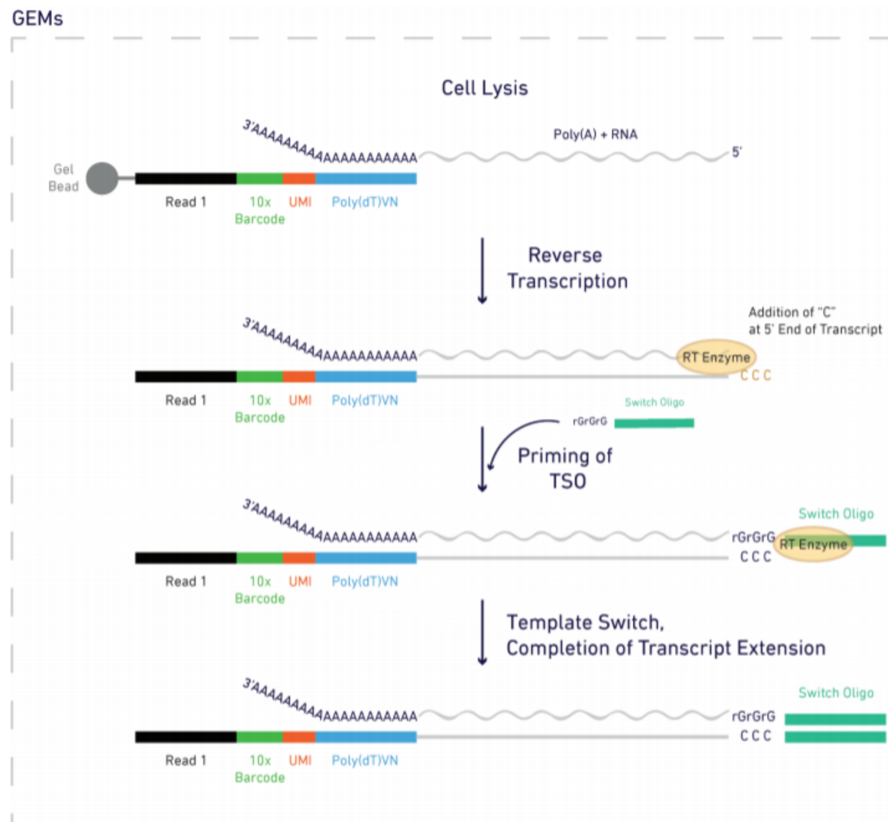
# Latest things

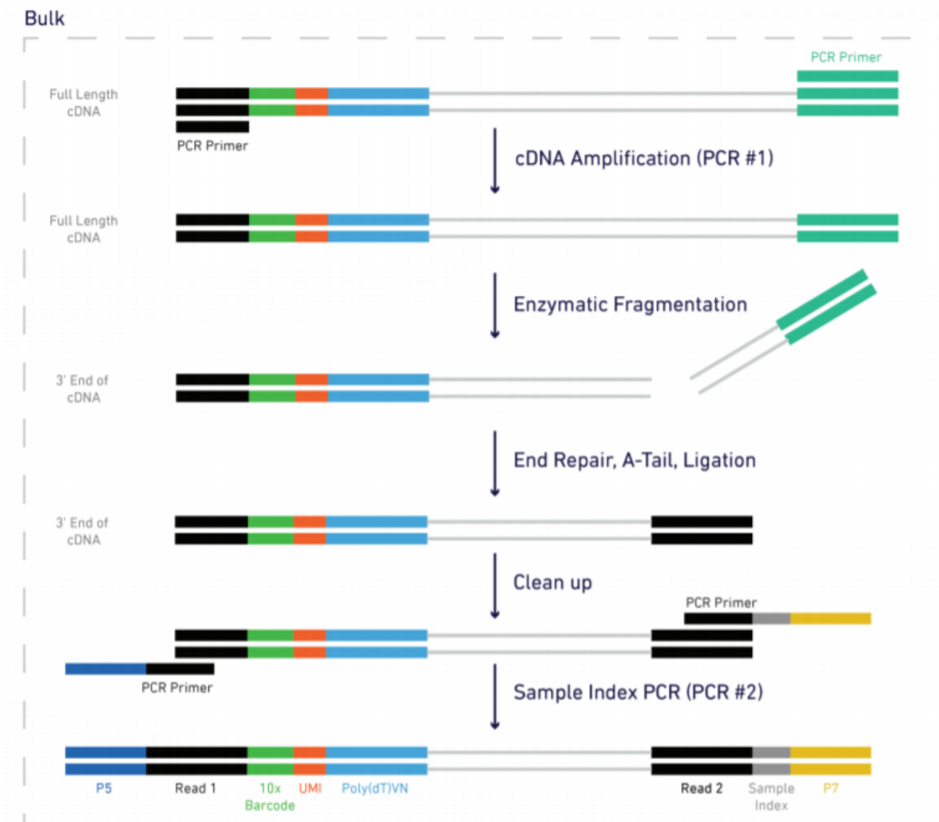
- 5’ scRNA-seq is different from 3’ scRNA-seq
- 5’ scRNA-seq allows us to get TCR and IG repertoires of the cells

# 3’ scRNA-seq

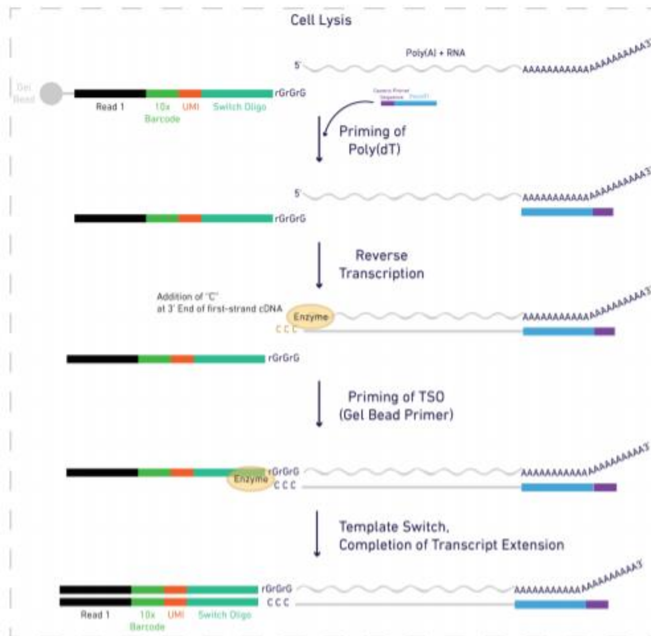
Inside individual GEMs

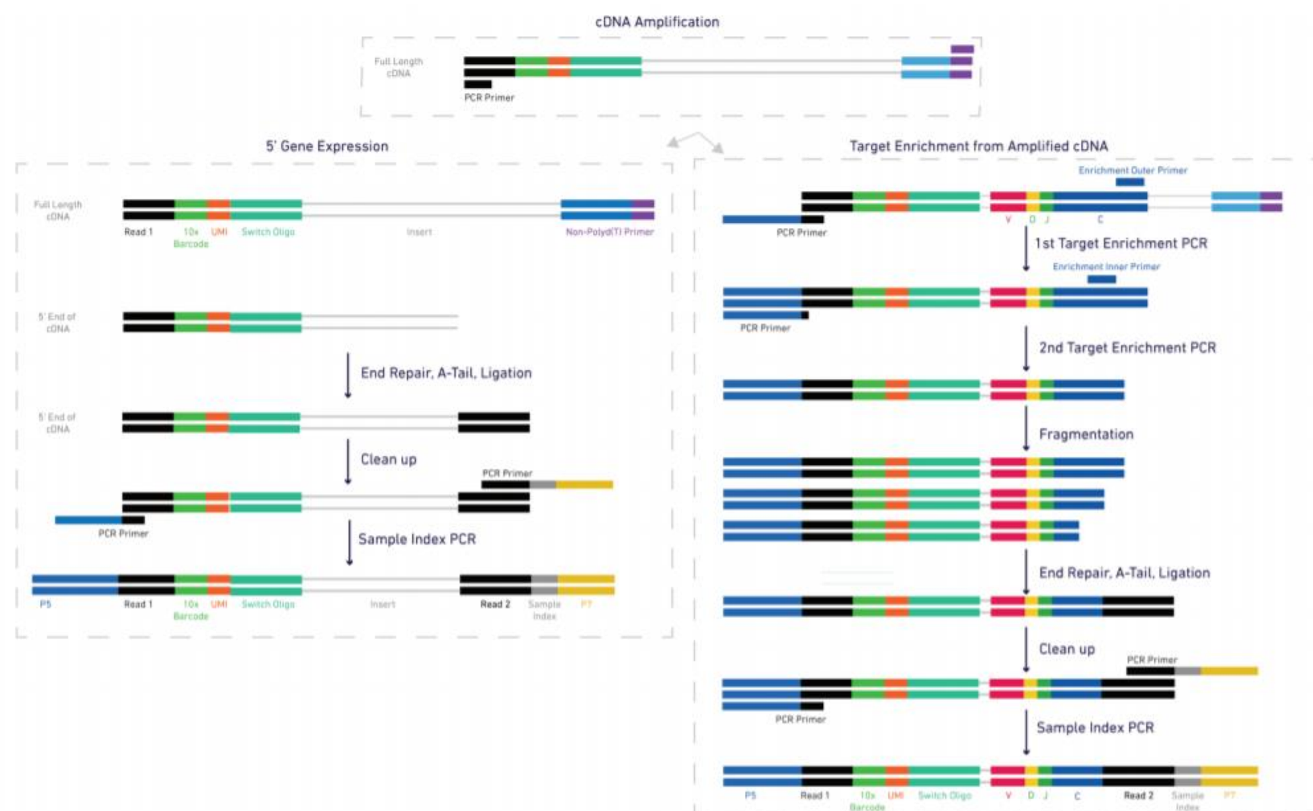
Pooled cDNA processed in bulk



# 5’ scRNA-seq



ITMO UNIVERSITY

# Simultaneous epitope and transcriptome measurement in single cells

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija & Peter Smibert

E

nature biotechnology

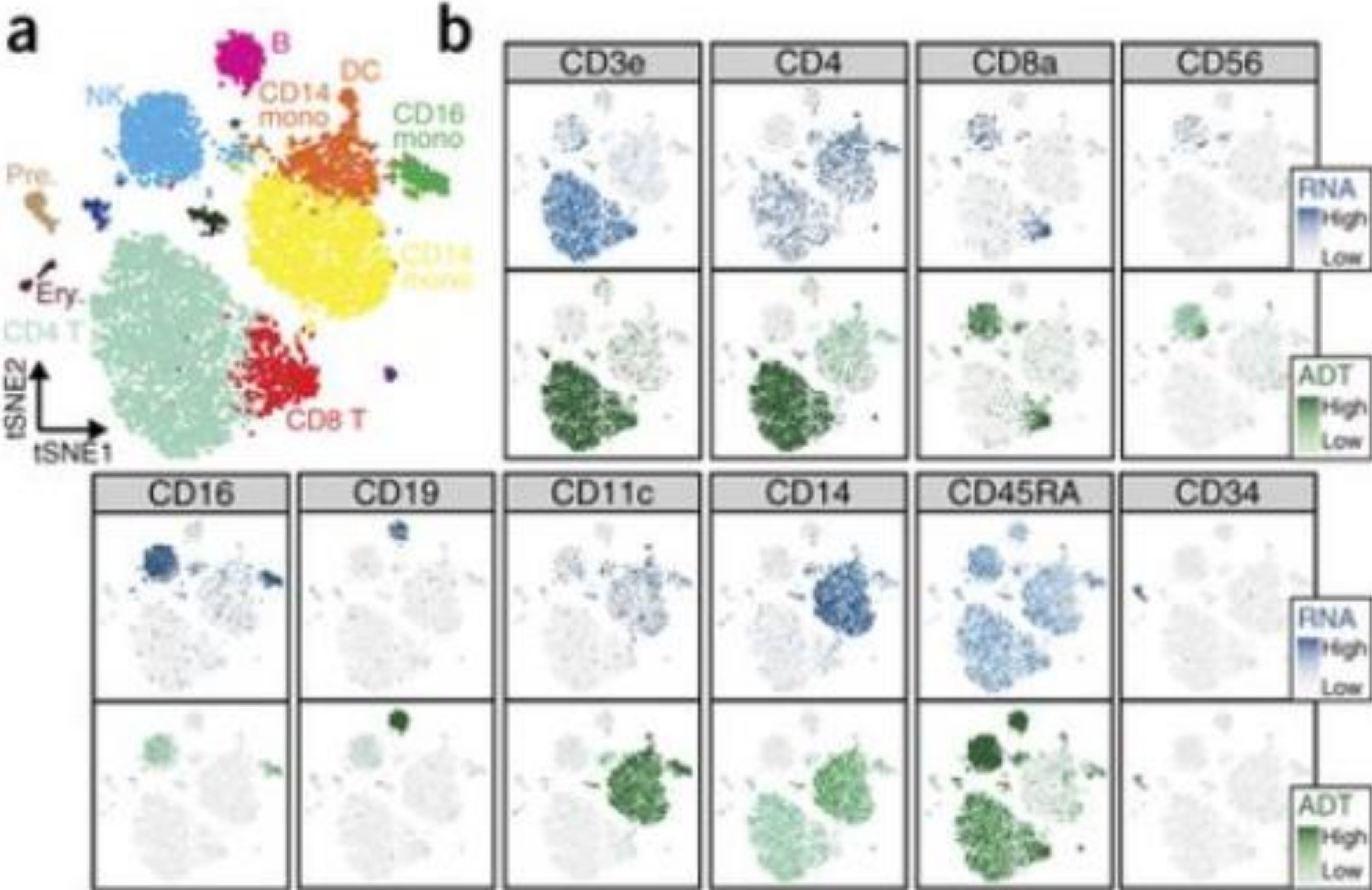
rights reserved.

# Multiplexed quantification of proteins and transcripts in single cells

Vanessa M Peterson[1,5], Kelvin Xi Zhang[2,5], Namit Kumar[1], Jerelyn Wong[3], Lixia Li[1], Douglas C Wilson[3], Renee Moore[4], Terrill K McClanahan[3], Svetlana Sadekova[3] & Joel A Klappenbach[1]

the standard 10x Genomics single-cell (sc)RNA-seq platform[3], which is a droplet-based system designed for 3′ digital counting of mRNA in thousands of single cells.

REAP-seq leverages the DNA polymerase activity of the reverse transcriptase to simultaneously extend the primed AbB with the poly(dT) cell barcode and synthesize complementary DNA from mRNA in the same reaction. Exonuclease I is then used to degrade any excess unbound single-stranded oligonucleotides from the protein double-stranded (ds) DNA (~155 bp) products to prevent crosstalk between AbBs and cell barcodes from different cells (**Supplementary Fig. 4**). Dextran sulfate was added to AbB labeling buffer to reduce non-specific binding of negatively charged DNA barcodes to the cell surface and isotype controls




* taken single
cells", https://www.nature.com/articles/nmeth.4380/figures/3