Introduction to single-cell RNA-seq

Konstantin Zaitsev
Washington University in Saint Louis / ITMO University

Bioinformatics summer school 2018

25.07.2018

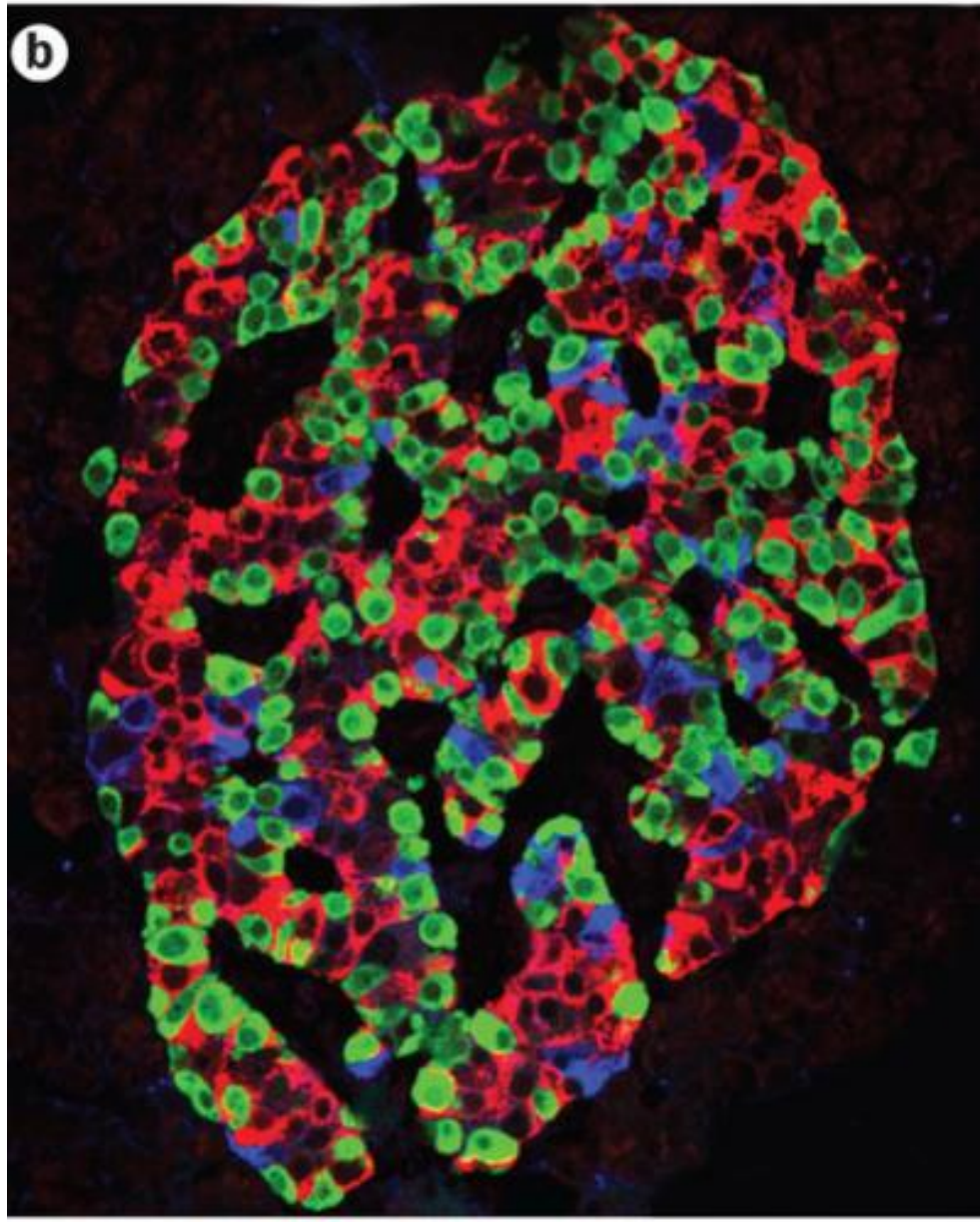
Mic check

RNA-seq шпаргалка

- RNA-seq – снимок того, что происходит внутри образца в данный момент
- В большинстве случаев мы не секвенируем РНК, мы секвенируем кДНК:
  - Избавление от рибосомальной РНК (poly-A selection, ribo zero kit)
  - Обратная транскрипция: РНК -> кДНК
  - Секвенирование

Single cell RNA-seq

- RNA-seq – «снимок» того, что происходит внутри образца
- Образец состоит из множества клеток
- Single cell RNA-seq –
  тысячи одновременных снимков клеток, находящихся в образце



Nature Reviews | Endocrinolog

Зачем single-cell RNA-seq

- Изучение клеточного разнообразия (в разных тканях, моделях)
- Описание новых клеточных подтипов
- Сравнение похожих клеточных подтипов
- Нахождение маркеров клеточных подтипов
- Отслеживание клеточной дифференциации

PROTOCOL

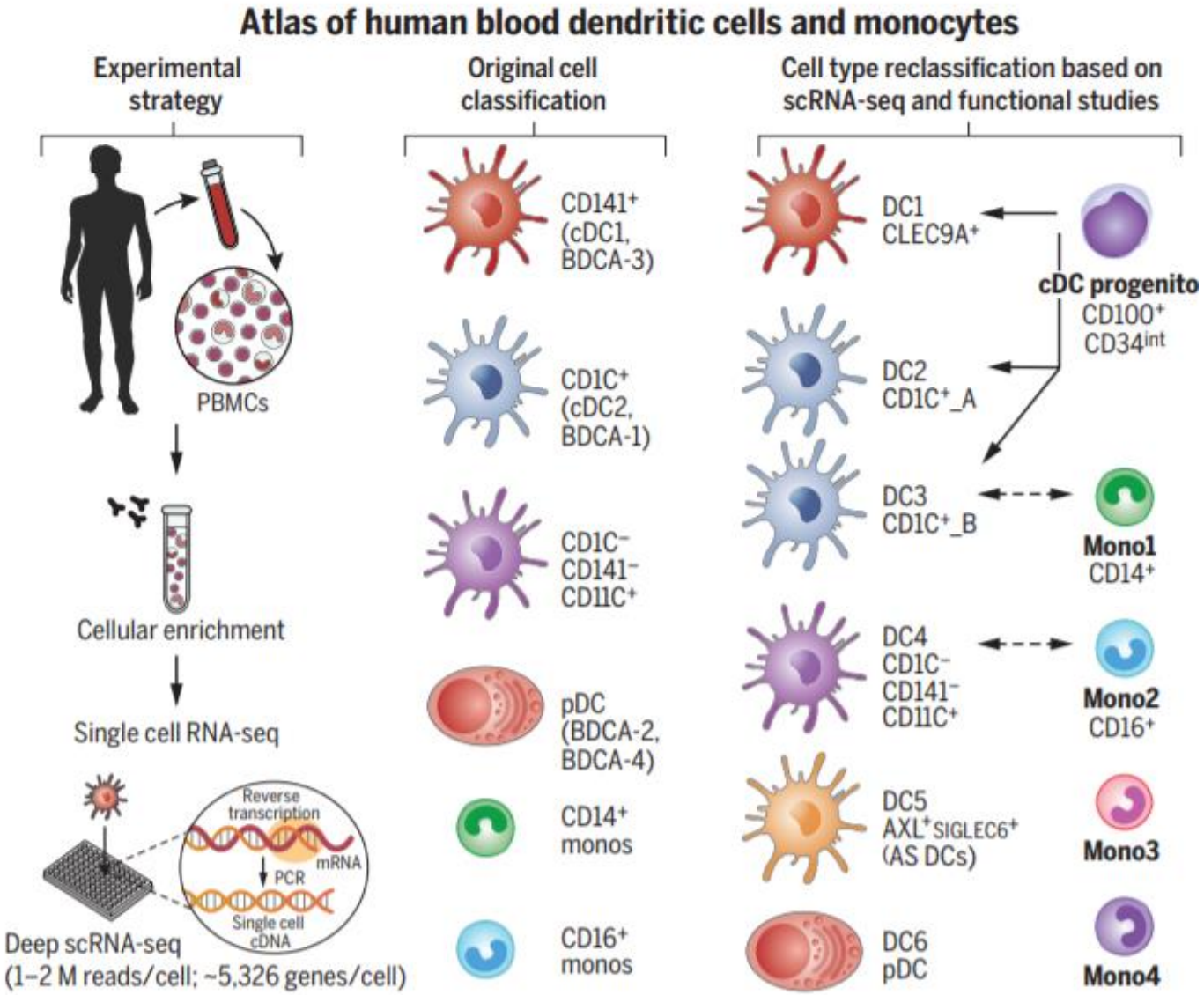
Full-length RNA-seq from single cells using Smart-seq2

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

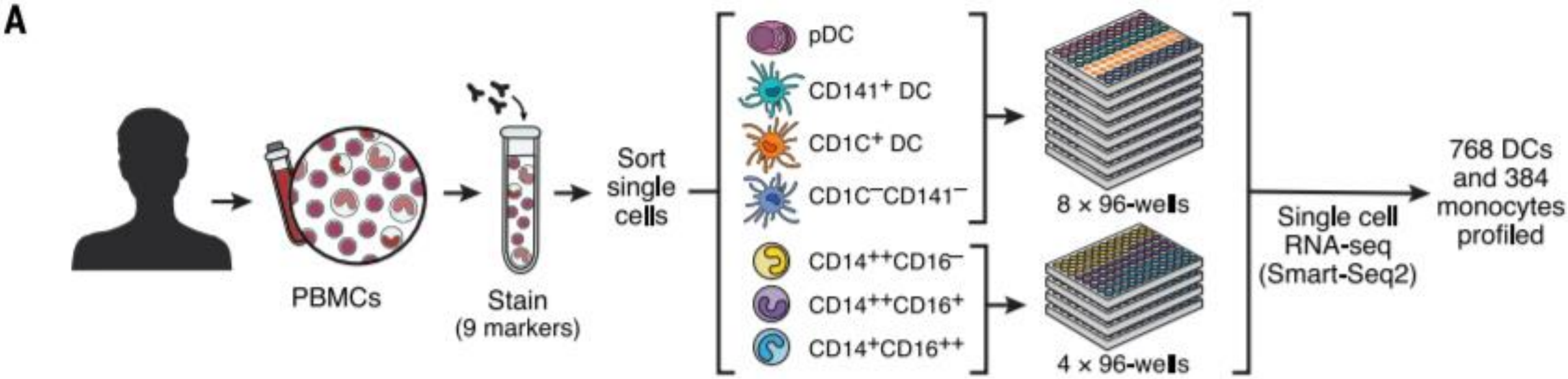
[1]Ludwig Institute for Cancer Research, Stockholm, Sweden. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.
Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

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

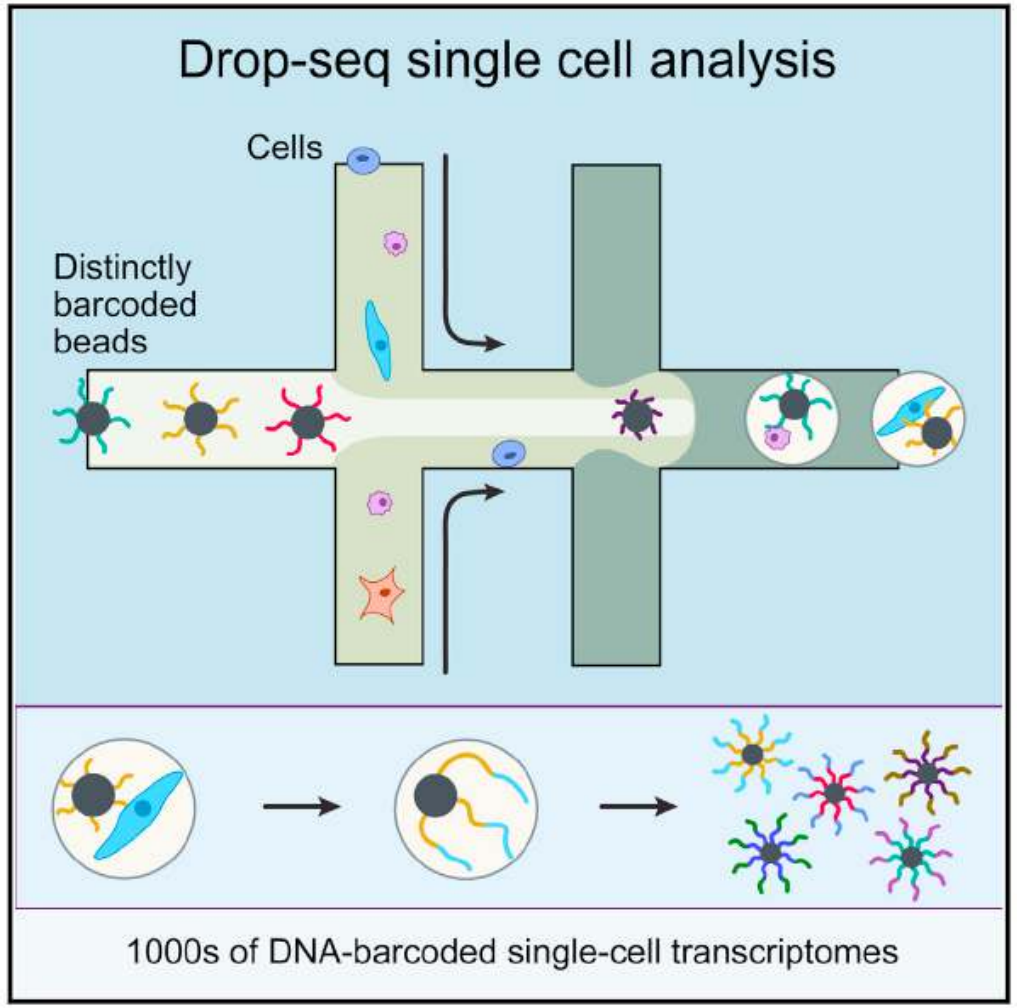

A
PBMCs
Stain
(9 markers)
Sort
single
cells
pDC
CD141+ DC
CD1C+ DC
CD1C−CD141−
CD14++CD16−
CD14++CD16+
CD14+CD16++
8 × 96-wells
4 × 96-wells
Single cell
RNA-seq
(Smart-Seq2)
768 DCs
and 384
monocytes
profiled

Resource

Cell

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

Graphical Abstract



Authors

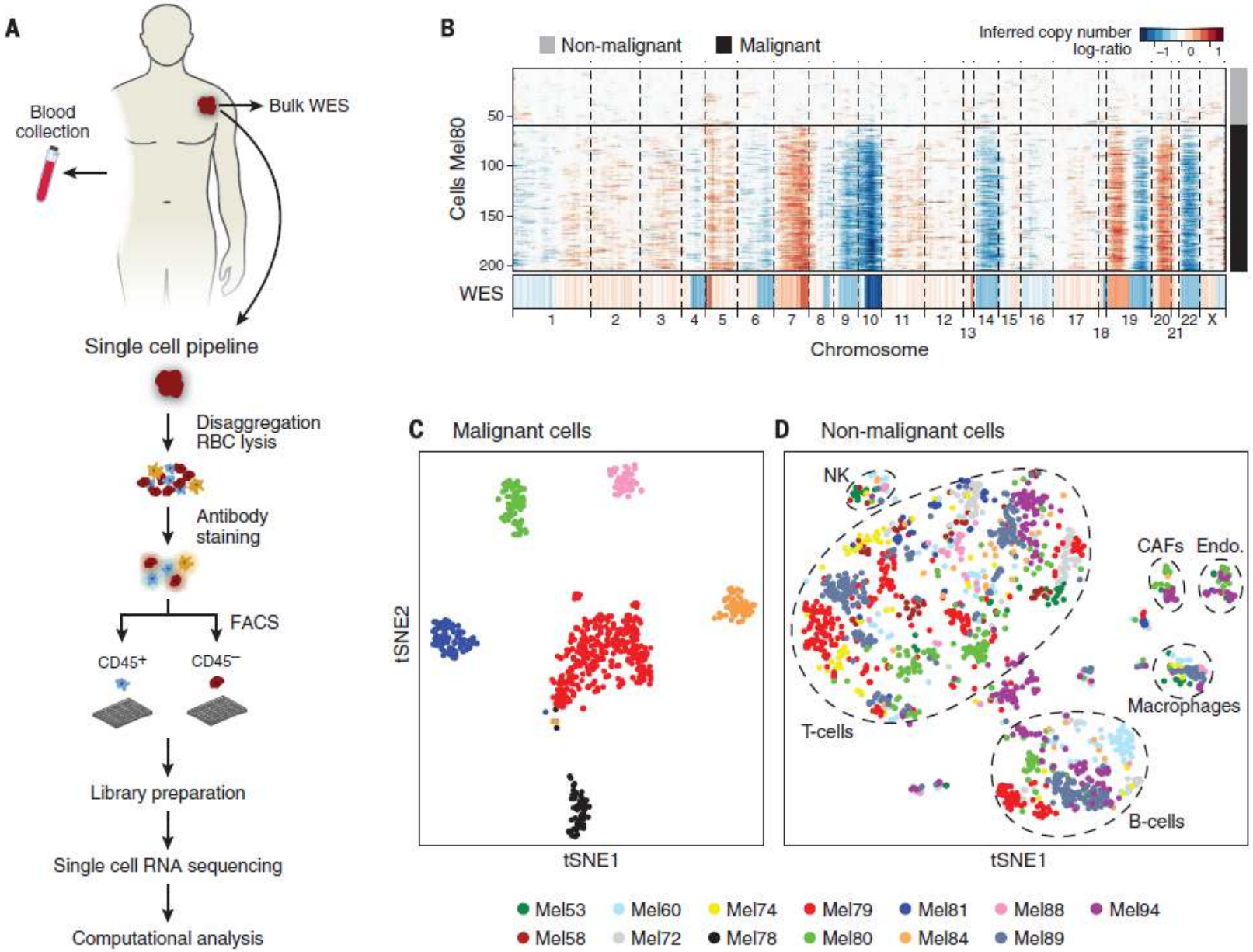
Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

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.



Tirosh, Izar et al, 2016
~ 4645 cells

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

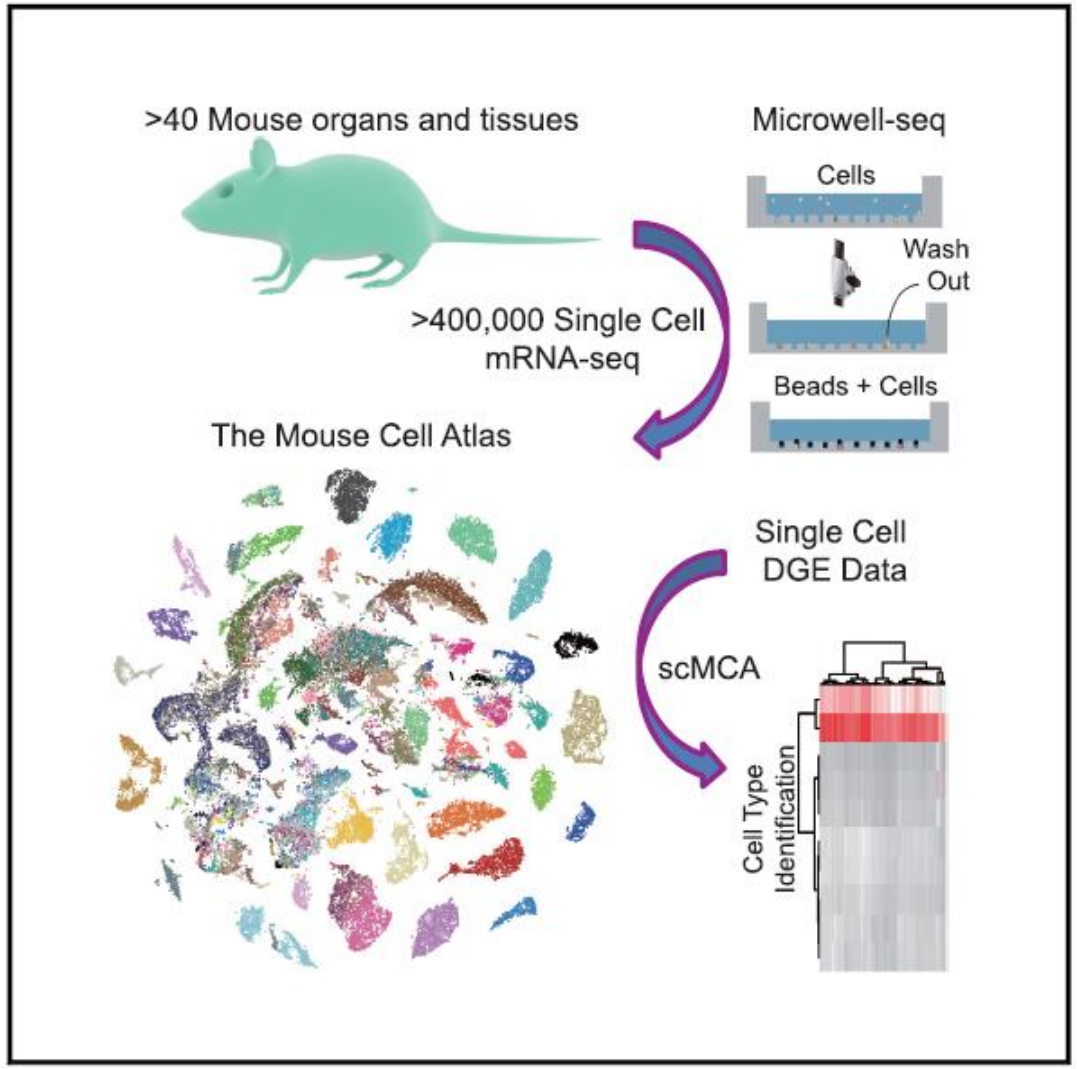
Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1], 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]

Resource

Cell

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.

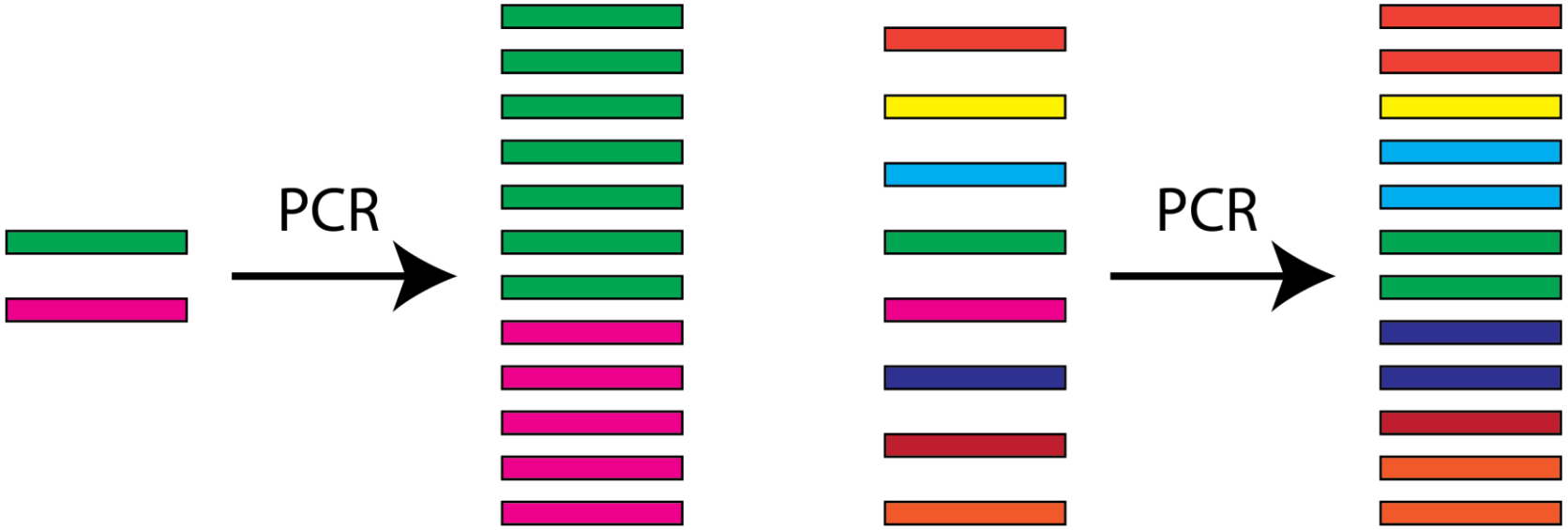
February 2018

Основные трудности

- **Как амплифицировать и секвенировать малое количество РНК (в среднем клетка млекопитающих содержит только 200 000 молекул мРНК)?**
- Как изолировать клетки друг от друга?
- Как работать с большим количеством клеток?

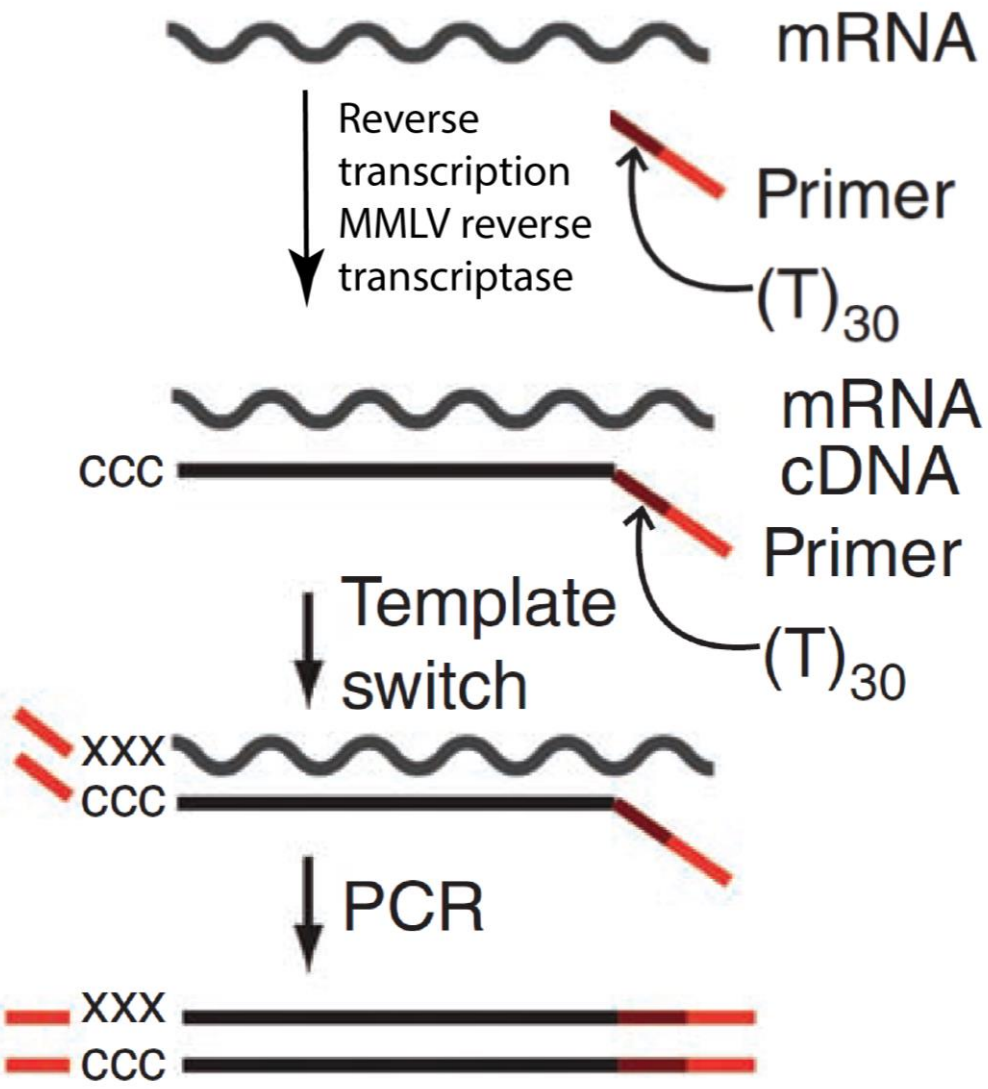
Малый размер библиотеки

- Малое количество мРНК дают ещё меньшее количество молекул кДНК
- кДНК амплифицируются ПЦР-ом
- Мы не хотим секвенировать много ПЦР дубликатов



Template switching PCR for low input

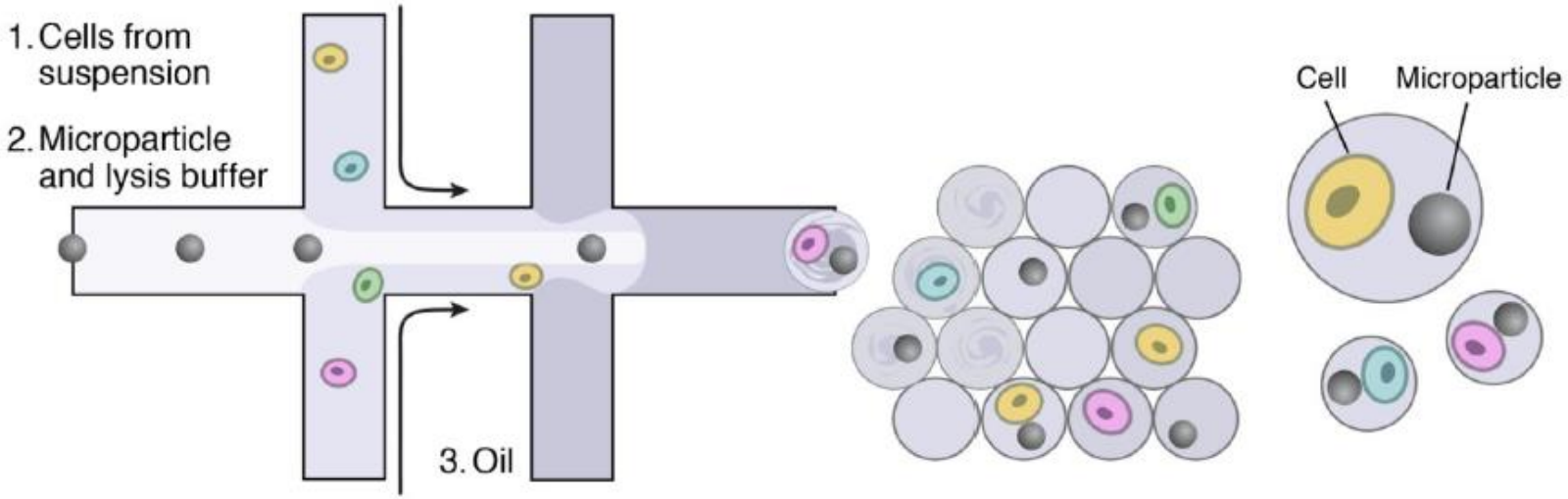
Switching Mechanism At the 5' end of RNA Template (SMART)



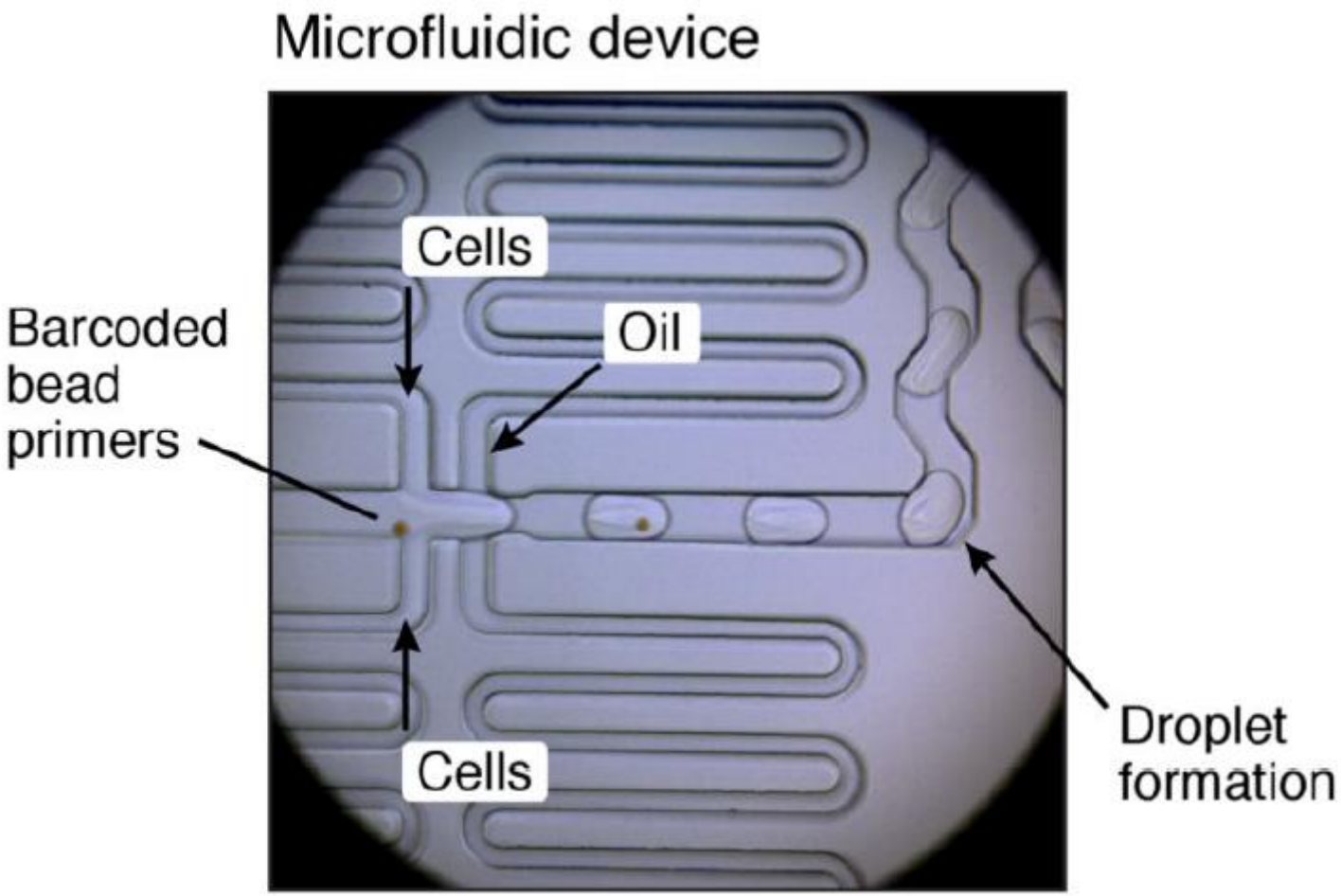
Основные трудности

- Как амплифицировать и секвенировать малое количество РНК (в среднем клетка млекопитающих содержит только 200 000 молекул мРНК)?
- **Как изолировать клетки друг от друга?**
- **Как работать с большим количеством клеток?**

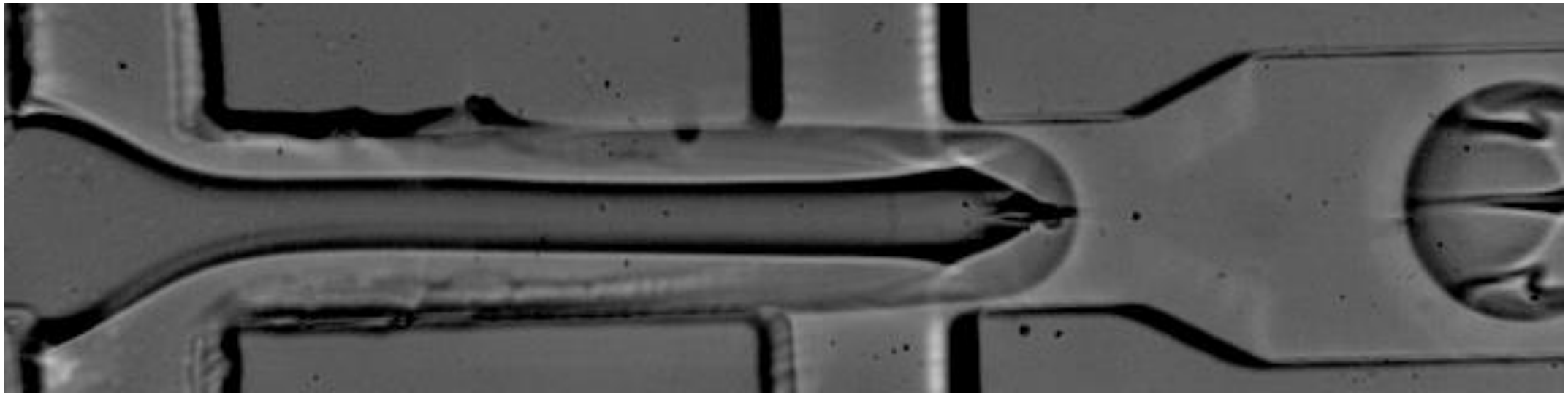
Drop-seq


1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle

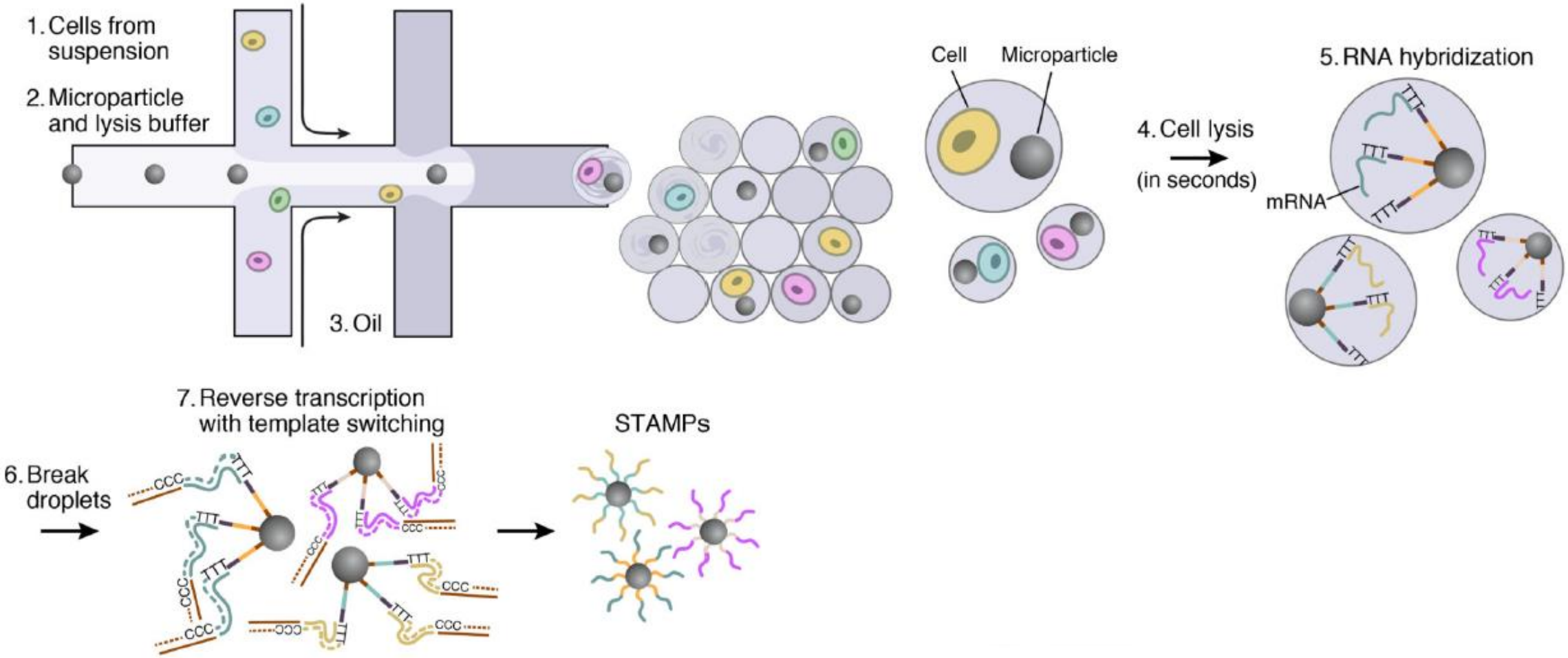
Как это выглядит в жизни?



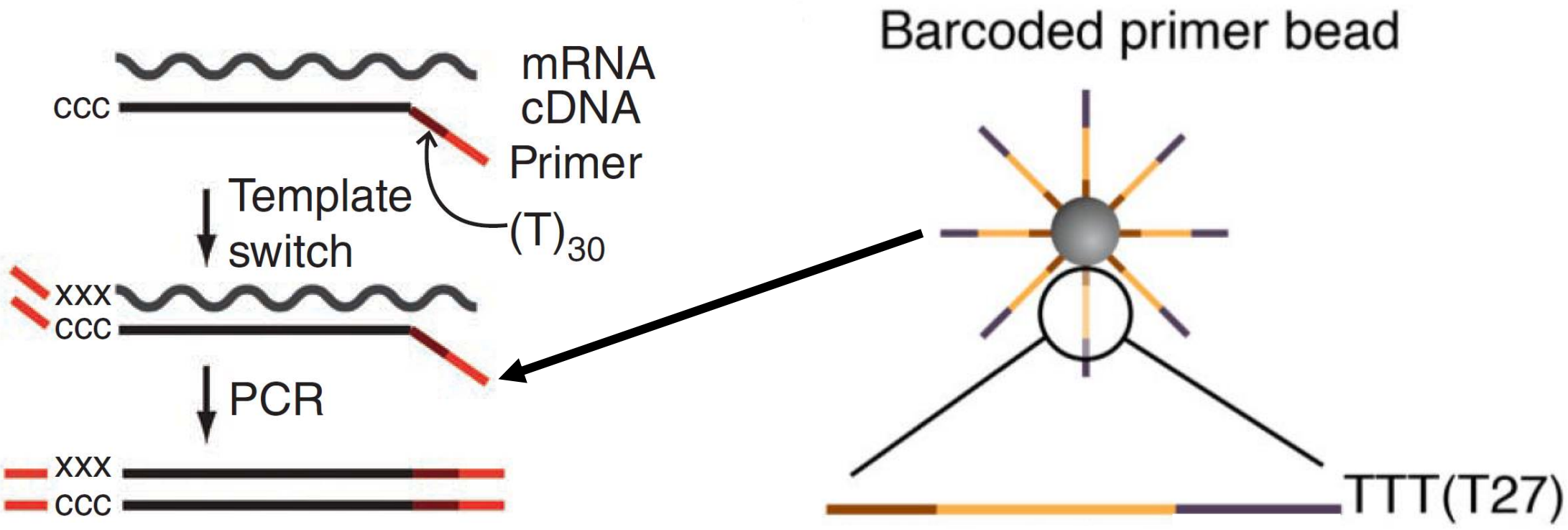
Как это выглядит в жизни?



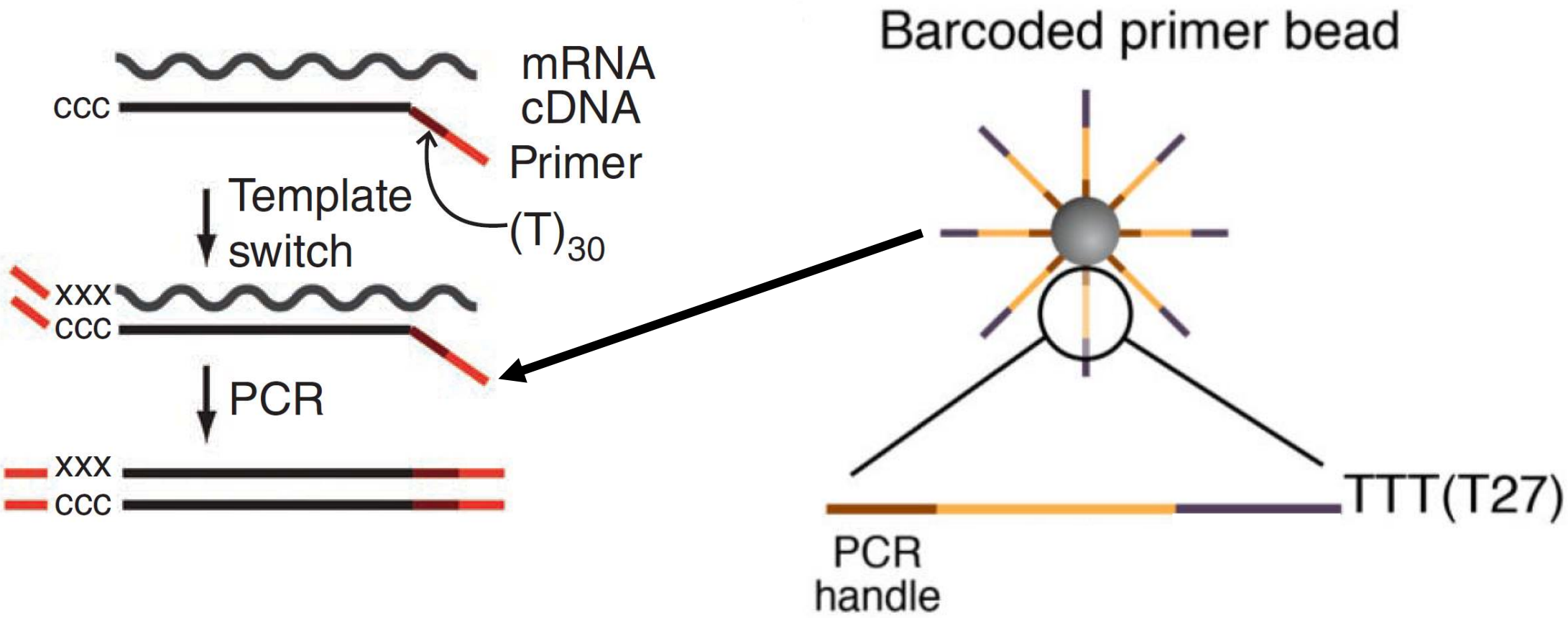
Drop-seq


1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
6. Break droplets
7. Reverse transcription with template switching
STAMPs

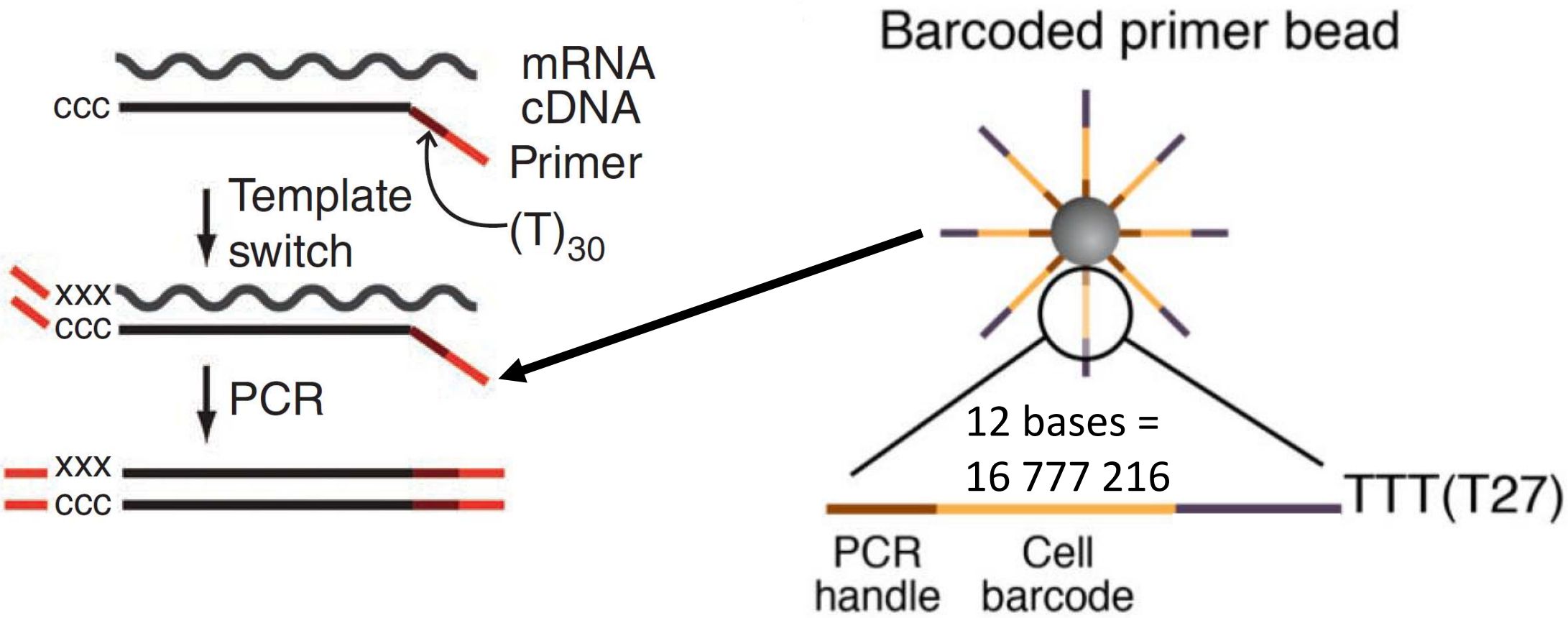
Баркодирование



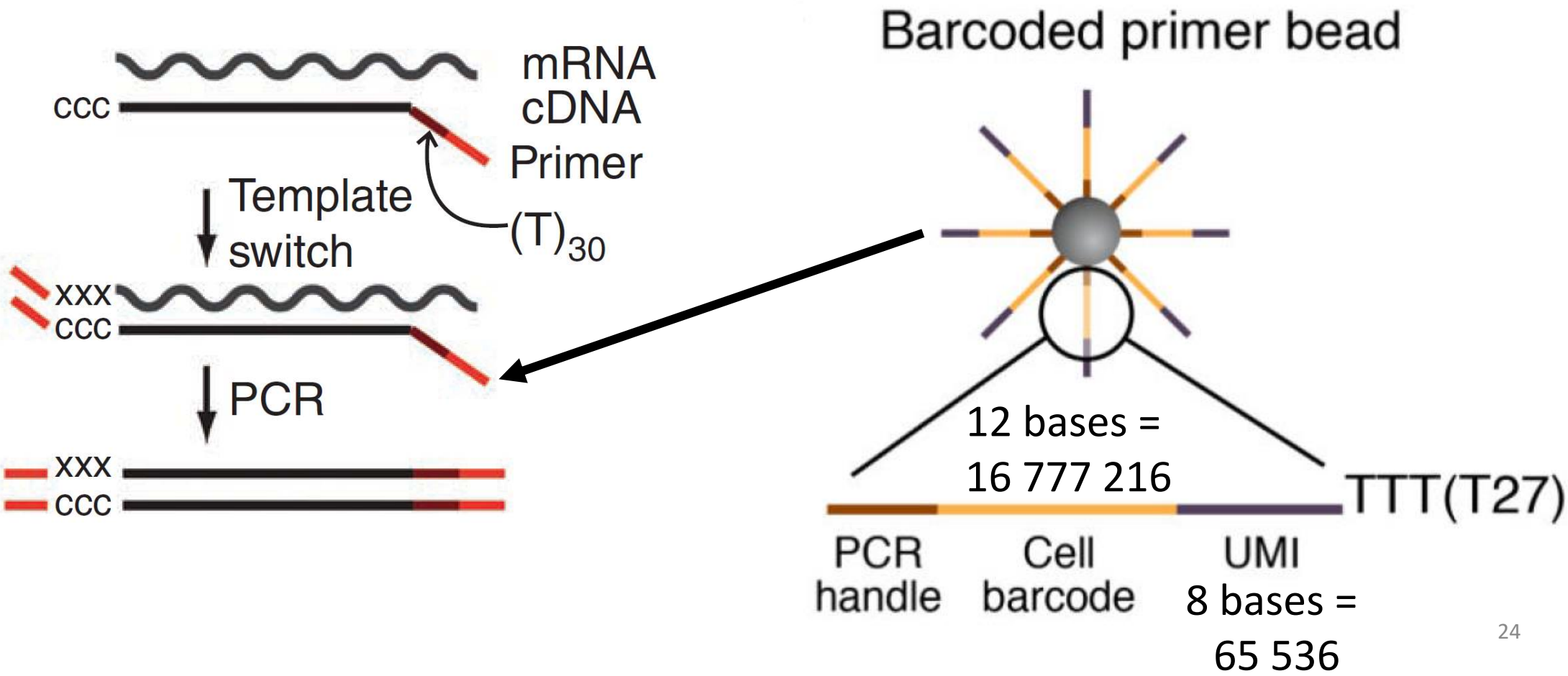
Баркодирование



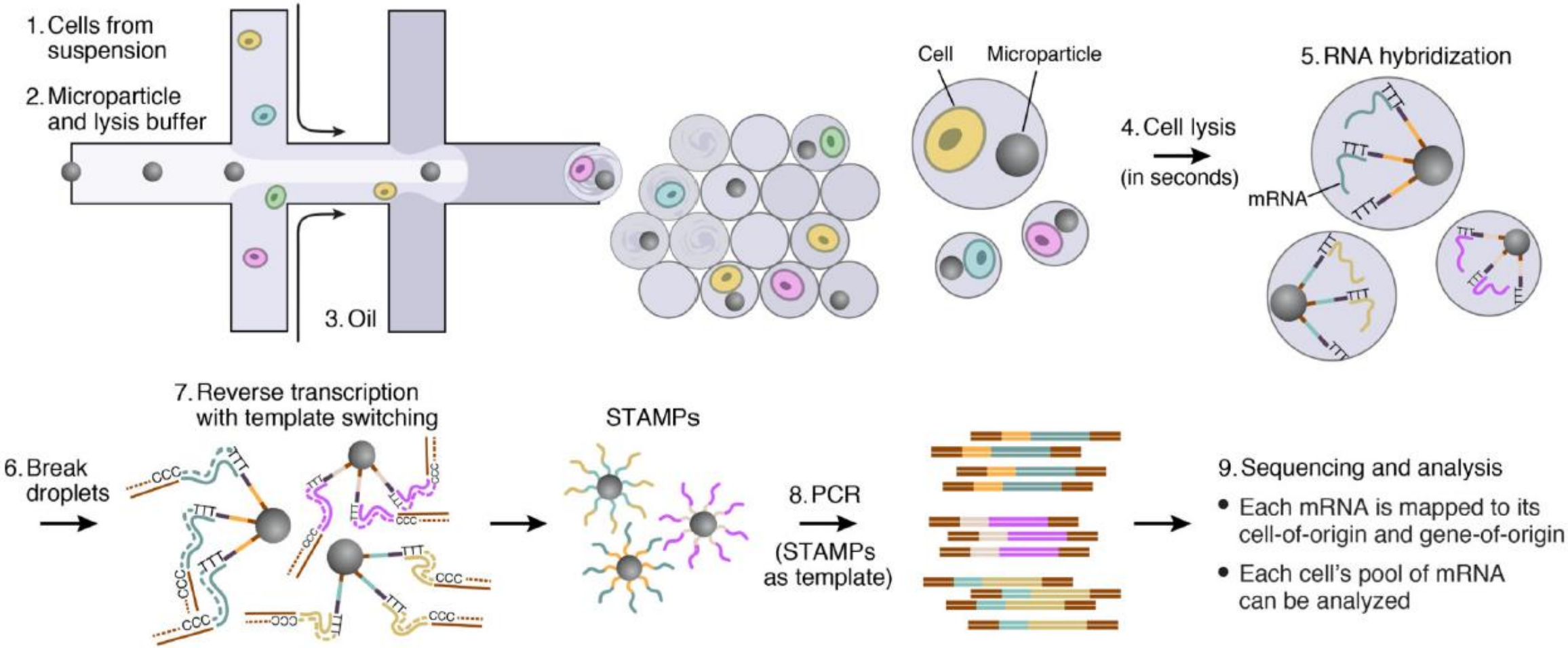
Баркодирование



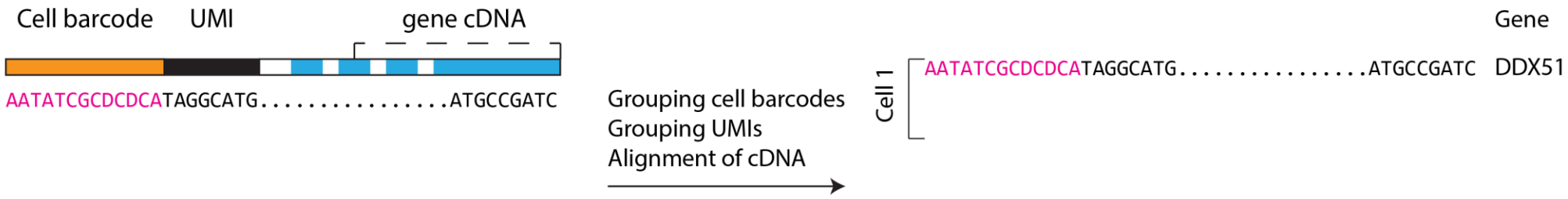
Баркодирование



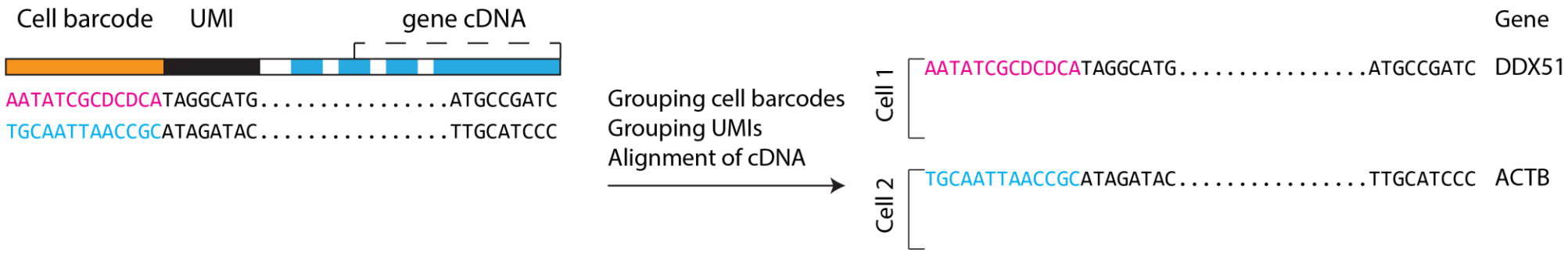
Drop-seq


1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
6. Break droplets
7. Reverse transcription with template switching
STAMPs
8. PCR
(STAMPs as template)
9. Sequencing and analysis
• Each mRNA is mapped to its cell-of-origin and gene-of-origin
• Each cell's pool of mRNA can be analyzed

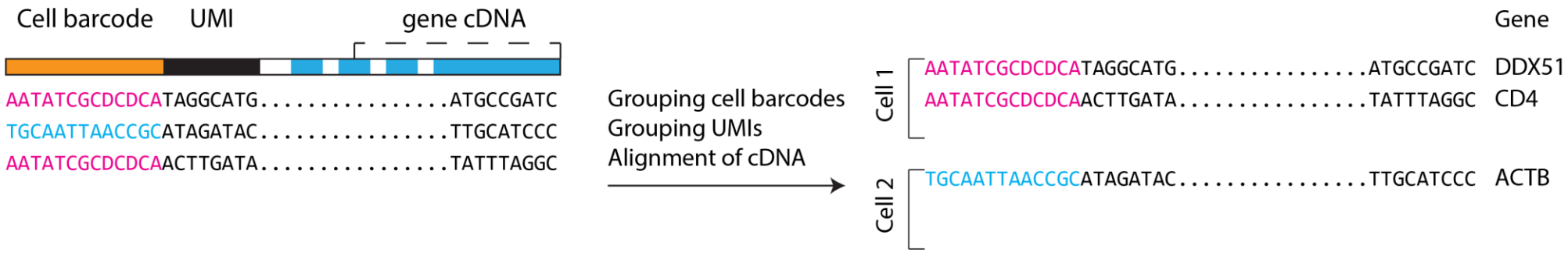
Секвенирование



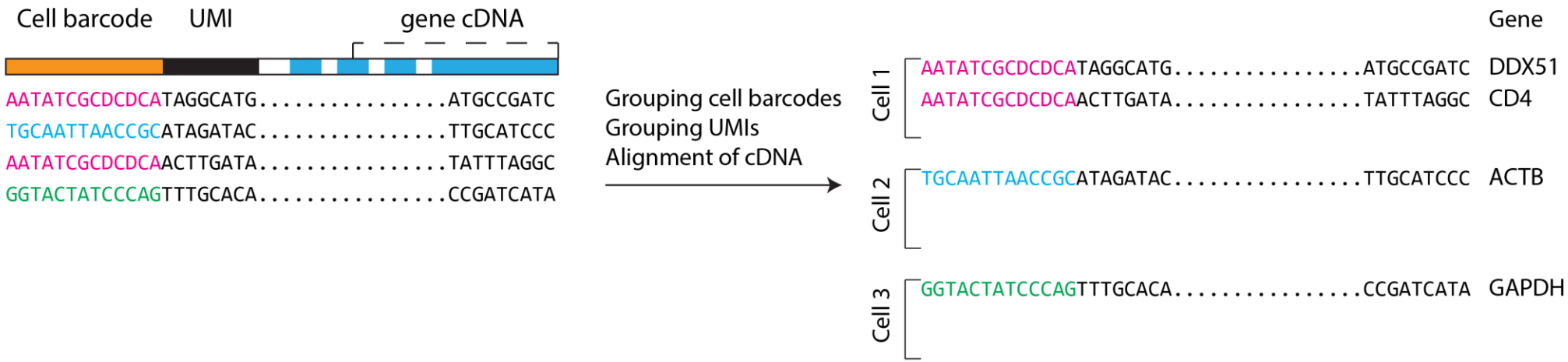
Секвенирование



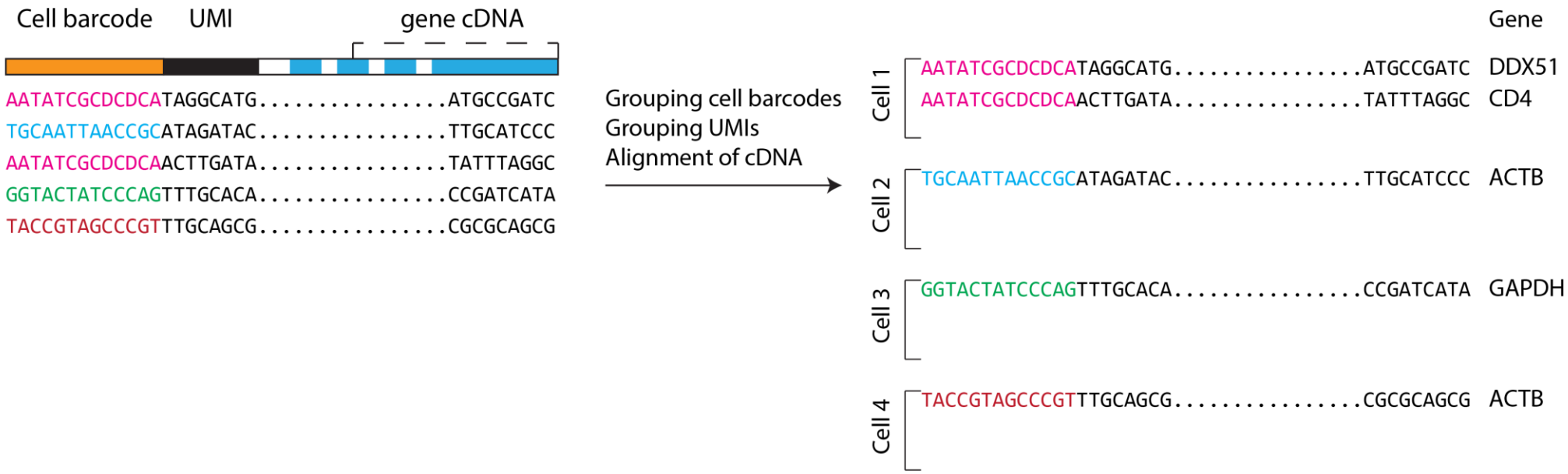
Секвенирование



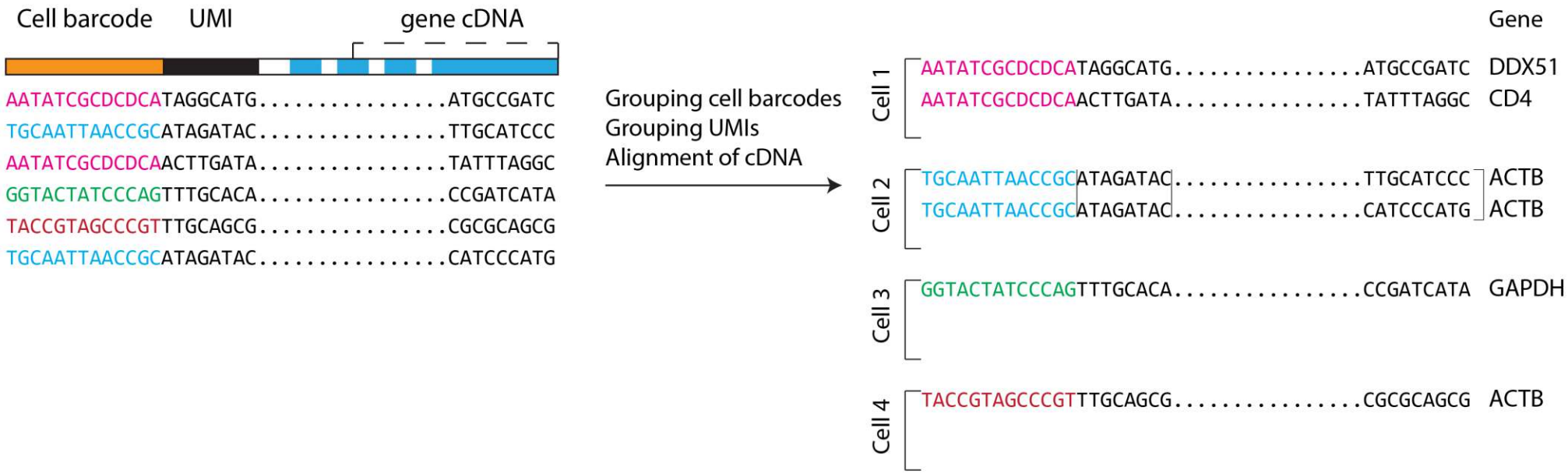
Секвенирование



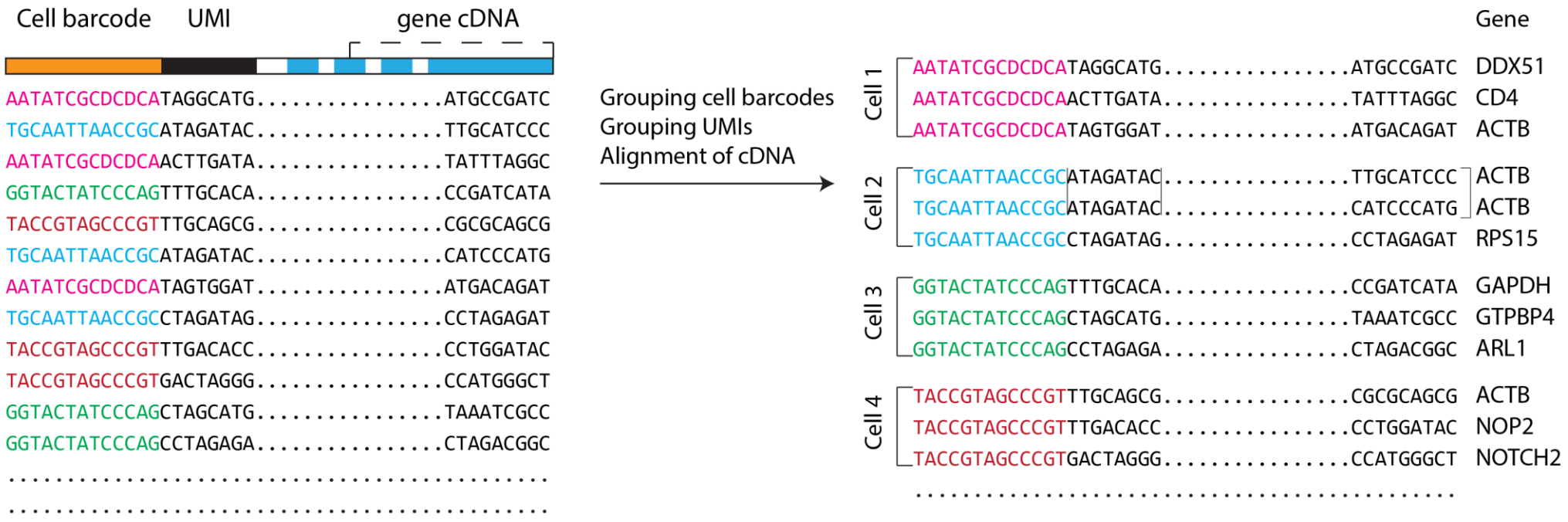
Секвенирование



Секвенирование



Секвенирование



Секвенирование

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

Матрица описывает экспрессию тысячи генов в тысячах клеток

Преодолевая трудности

- Малое количество РНК – template switching PCR
- Как изолировать клетки – microfluidics
- Как идентифицировать прочтения (риды) – Cell barcodes
- Как бороться с PCR-дупликатами – Cell/UMI barcodes

Преодолевая трудности

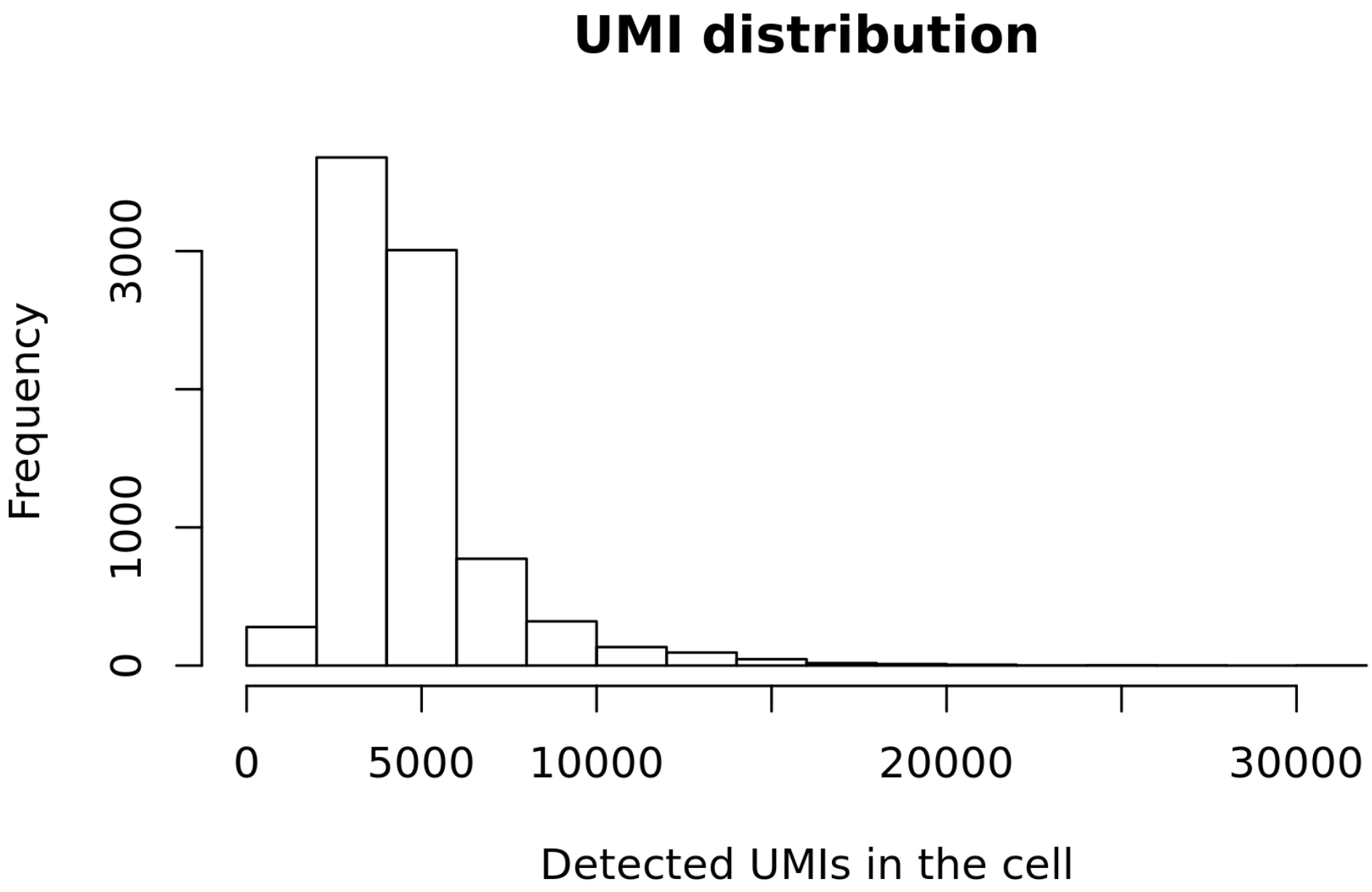
- Малое количество РНК – template switching PCR
- **Как изолировать клетки – microfluidics (может меняться)**
- **Баркодирование с помощью микрочастиц покрытых баркодами показало себя очень эффективным**

Есть ли вопросы?

- Лучше спросить сейчас вопросы по биологии, дальше будет часть про анализ

Базовые шаги в анализе single-cell RNA-seq

- **Фильтрация плохих баркодов**
- **Нормализация экспрессии**
- Визуализация (tSNE plots)
- Кластеризация
- Описание клеточных подтипов


UMI distribution
Frequency
0
1000
3000
0
5000
10000
20000
30000
Detected UMIs in the cell

Нужно нормализовать!

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

Нужно нормализовать!

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

$\longrightarrow$

AATATCGCDCDCA

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

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

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

$$\ldots \quad \ldots$$

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

Нужно нормализовать!

Cell:	TGCAATTAACCGC
Gene 1	03
Gene 2	00
Gene 3	10
...	..
Gene M	01

$\longrightarrow$

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

Нужно нормализовать!

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

→

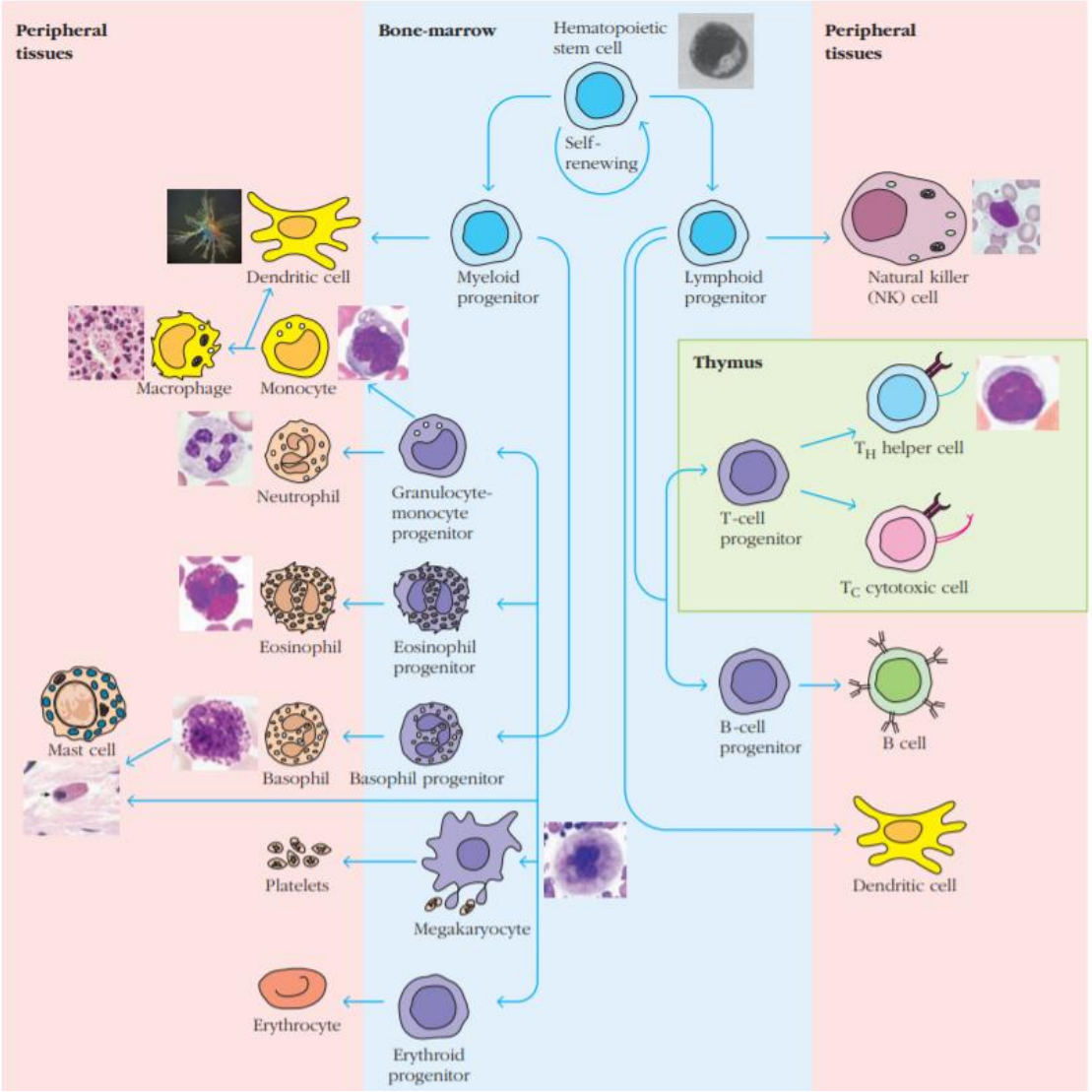
Cell:	AATATCGCDCDCA	TGCAATTAACCGC	GGTACTATCCCAG	..	TACCGTAGCCCGT
Gene 1	15	06	10	..	07
Gene 2	7.5	00	00	..	00
Gene 3	18	20	12.5	..	05
...	...	..	...	..	..
Gene M	16.5	02	00	..	00

Sum in every column is 10000

Базовые шаги в анализе single-cell RNA-seq

- Фильтрация плохих баркодов
- Нормализация экспрессии
- **Визуализация (tSNE plots)**
- Кластеризация
- Описание клеточных подтипов

Peripheral blood mononuclear cells mixture

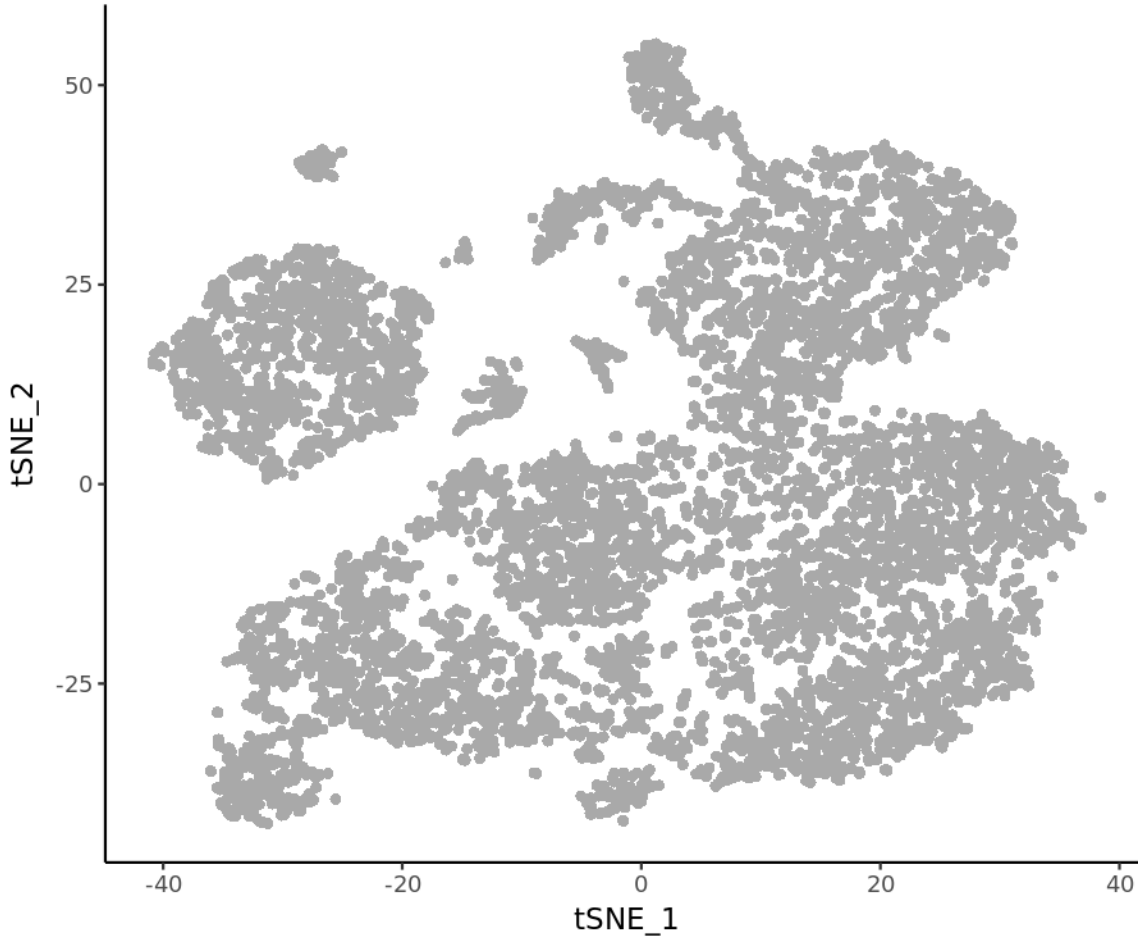


Kuby, Immunology, 7th edition

PBMC dataset: visualization by tSNE

tSNE (t-distributed stochastic neighbor embedding):
Нелинейный метод понижения размерности. Метод пытается поместить все точки (в нашем случае клетки) в 2 или 3мерное пространство, сохраняя расстояние* между точками в изначальном пространстве.

Мы подаём первые 10-20 главных компонент (из метода главных компонент) на вход алгоритма tSNE



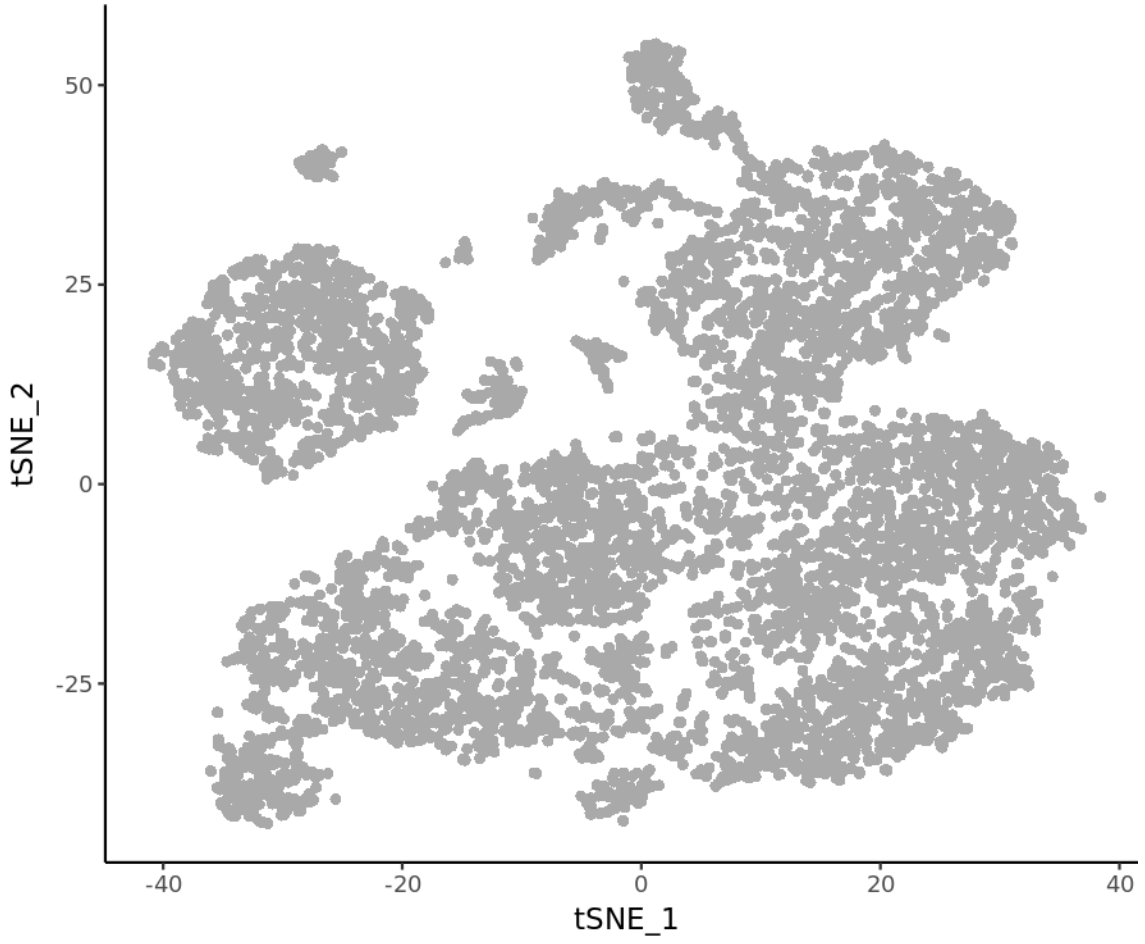
https://support.10xgenomics.com/single-cell-gene-expression/datasets/pbmc8k
http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

PBMC dataset: visualization by tSNE

tSNE:

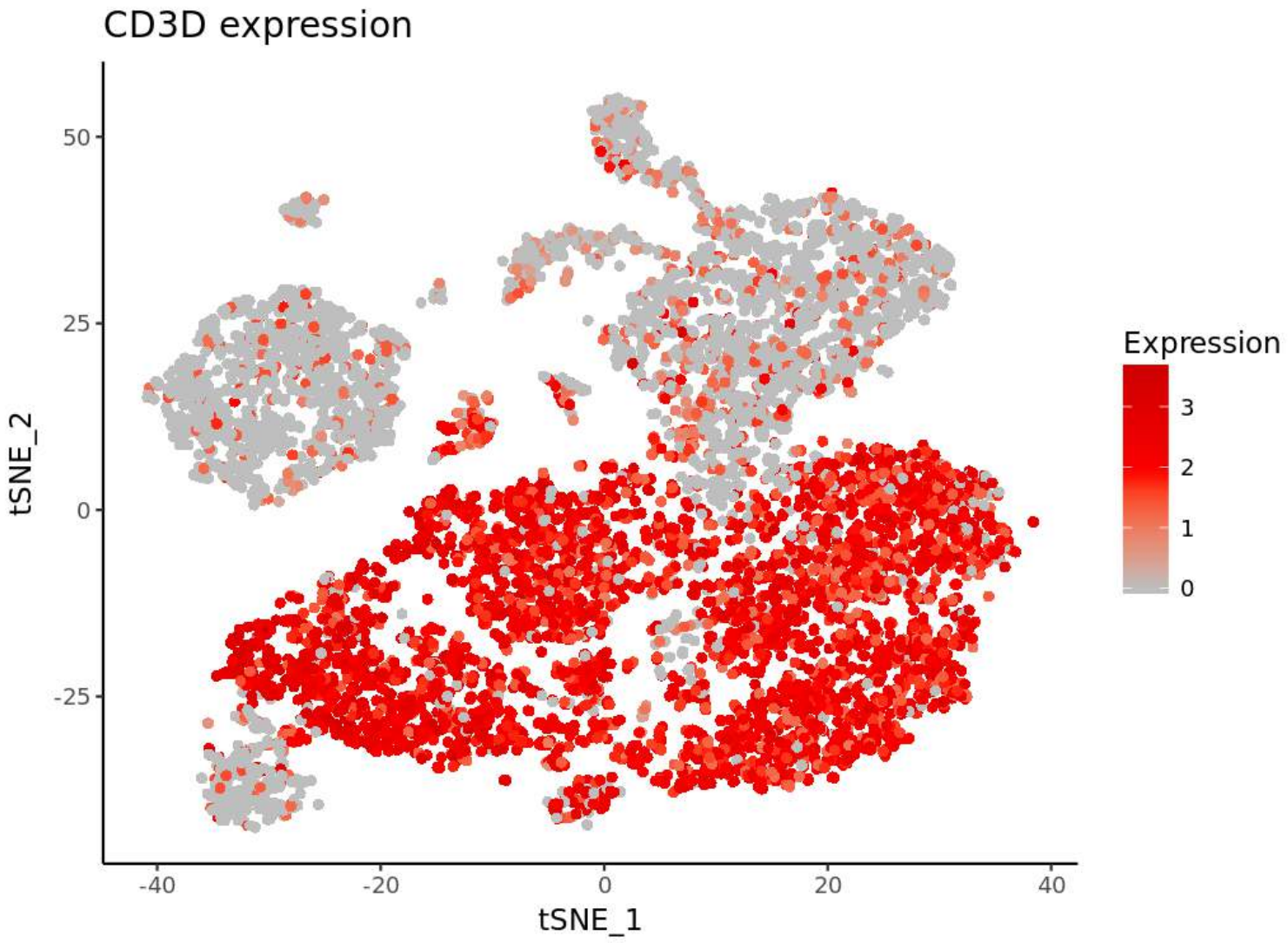
Клетки (точки), которые ближе друг к другу с оригинальном многомерном пространстве, будут ближе друг к другу в tSNE пространстве

tSNE – нелинейная трансформация, и координаты точек в tSNE-пространстве не имеют интерпретируемого биологического значения



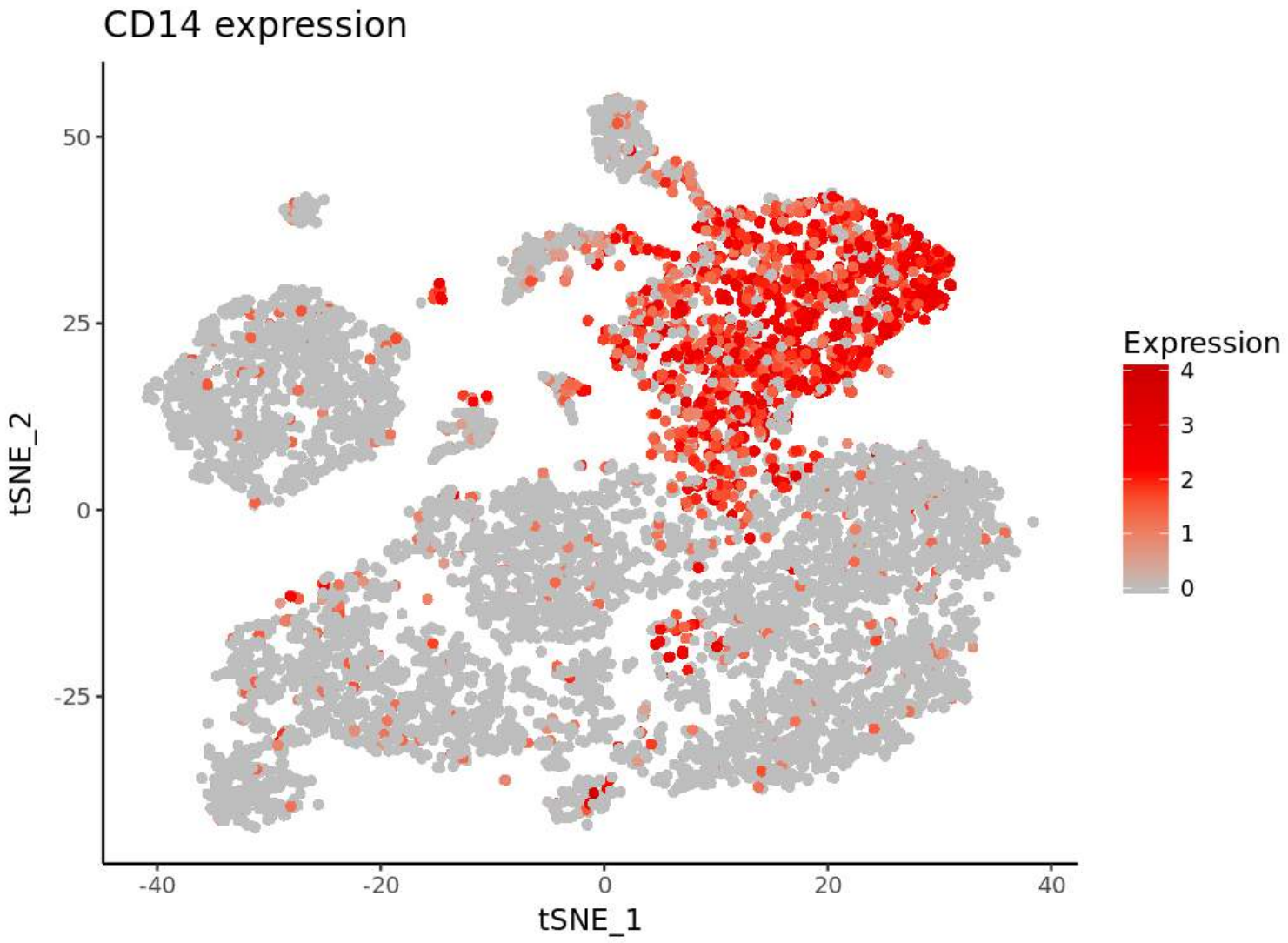
https://support.10xgenomics.com/single-cell-gene-expression/datasets/pbmc8k
http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

PBMC dataset: check for the known markers CD3D – T cell marker



PBMC dataset: check for the known markers CD79A – B cell marker

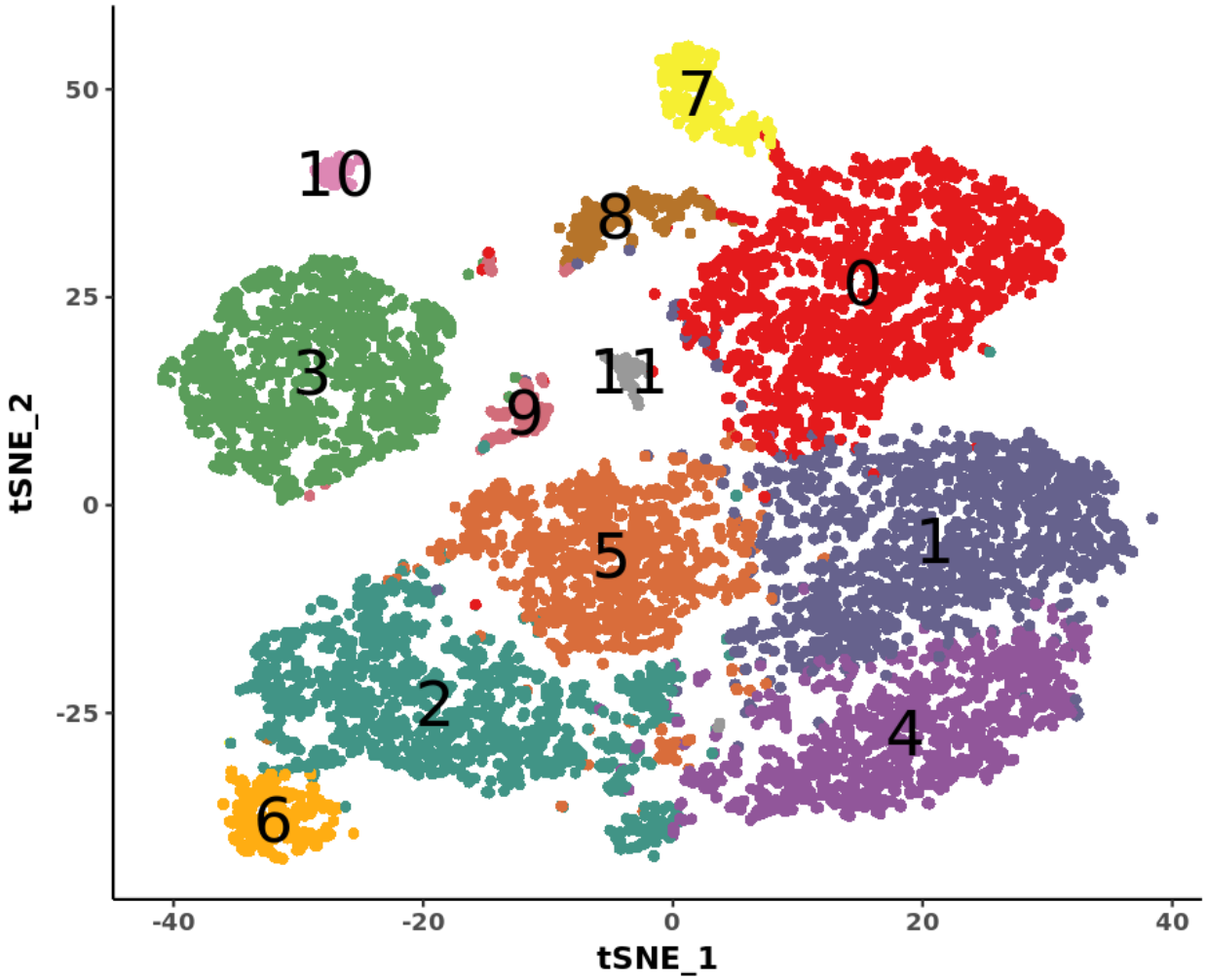
PBMC dataset: check for the known markers CD14 – monocyte marker



PBMC dataset: clustering and annotation

Clustering:

- **Graph-based clustering**
- K-means



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

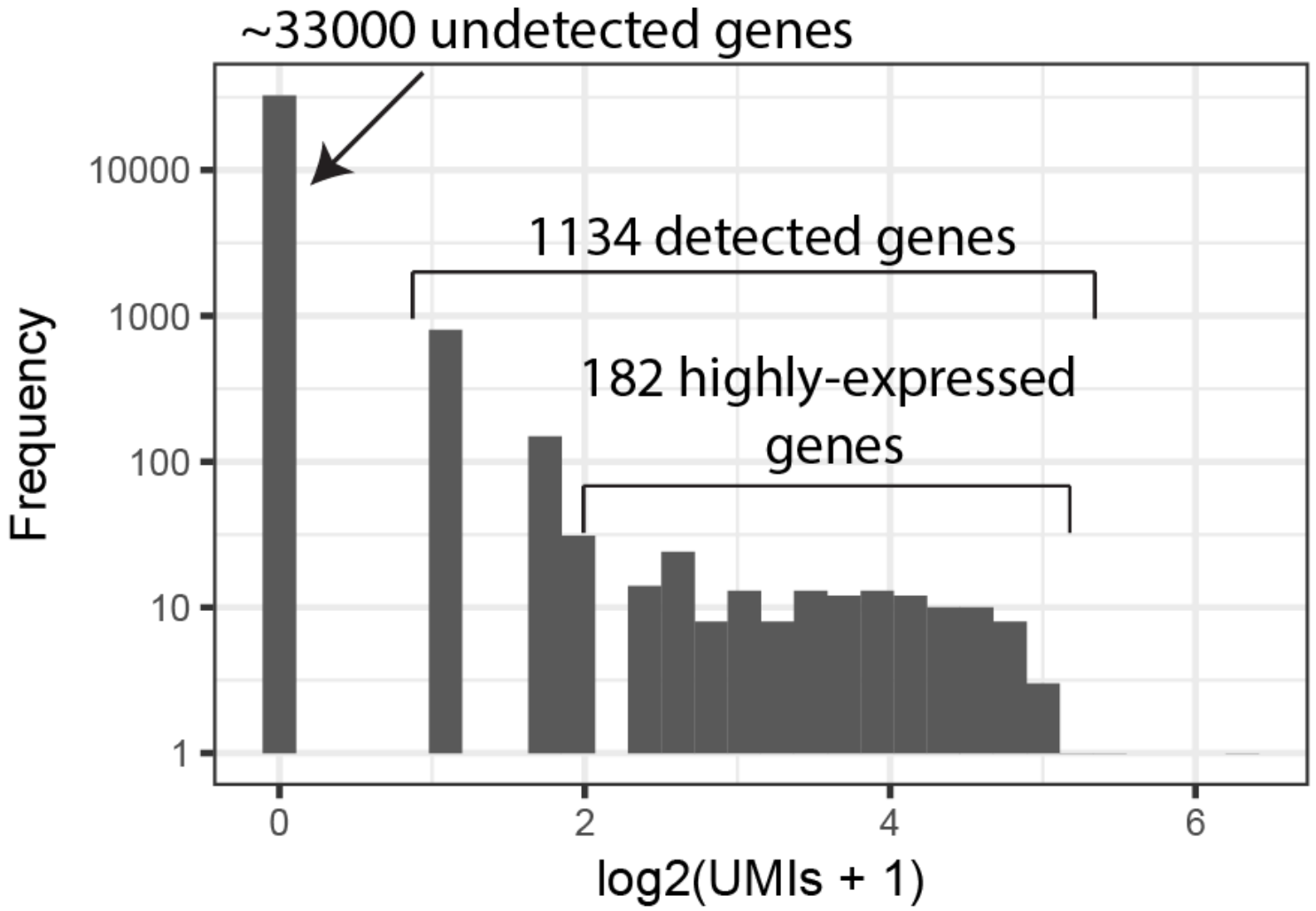
Supplementary information: Supplementary data are available at *Bioinformatics* online.

- Вместо того, чтобы пытаться находить кластеры по расстояниям, мы для каждой клетки найдем её «соседей»
- Клетки, у которых много соседей скорее всего живут в одном «районе»
- Алгоритм пытается находить «районы» клеток в датасете

Making sense

- Когда кластеры найдены и есть возможность проверить экспрессию генов, нужно идентифицировать клеточные подтипы: сказать кто, есть кто
- Если вы биоинформатик и вы анализируете чужие данные, в этот момент вы обязательно должны выйти из домика и пойти поговорить с биологом, которые эти данные сделал
- Если вы биолог, который делает single-cell RNA-seq, то скорее всего, никто лучше вас не знает все маркеры и все клеточные подтипы в ваших данных.

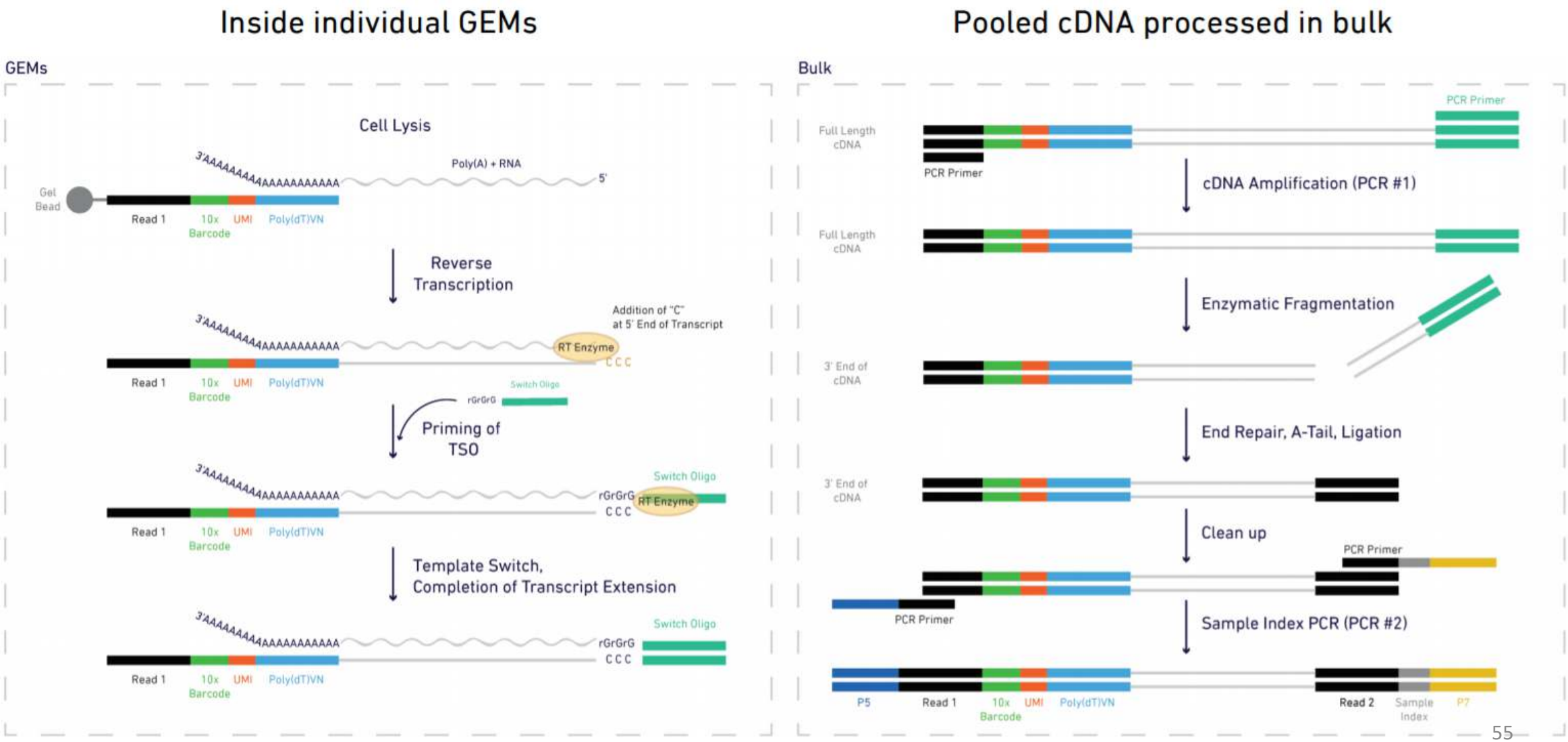
Gene expression in the cell



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



https://assets.ctfassets.net/an68im79xiti/4fIy9tr6qQuCWamIii0iEa/40658acce7a6756e38537584897840e3/CG000108_AssayConfiguration_SC3v2.pdf

Brief Communication

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

BRIEF COMMUNICATIONS

nature biotechnology

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]

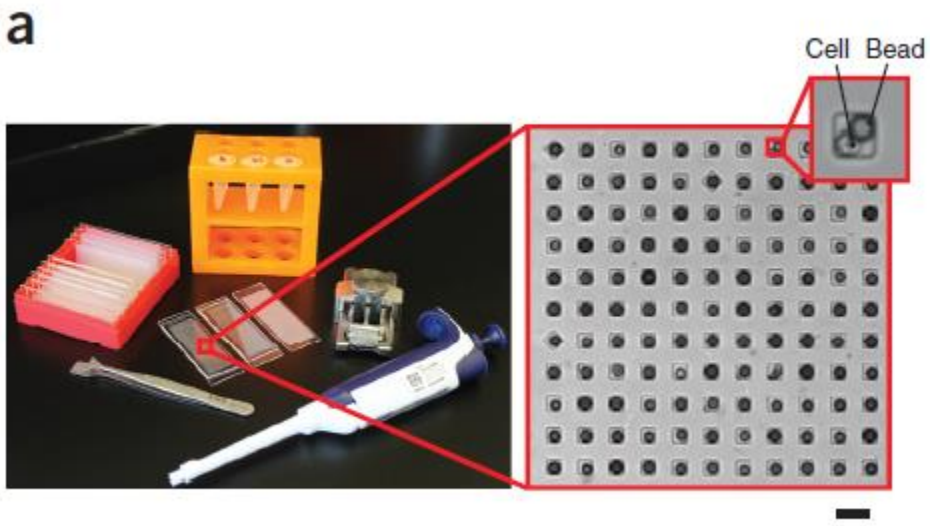
rights reserved.

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

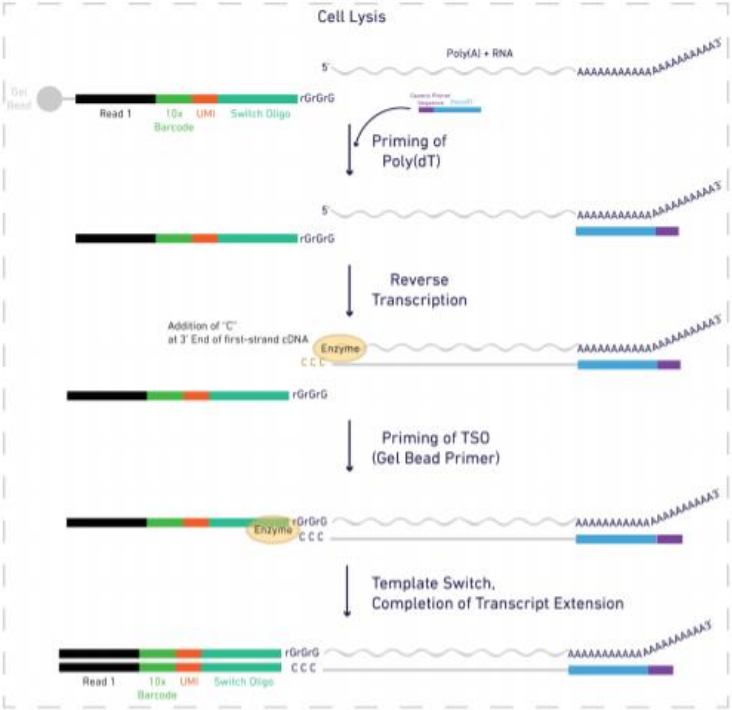
Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

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]

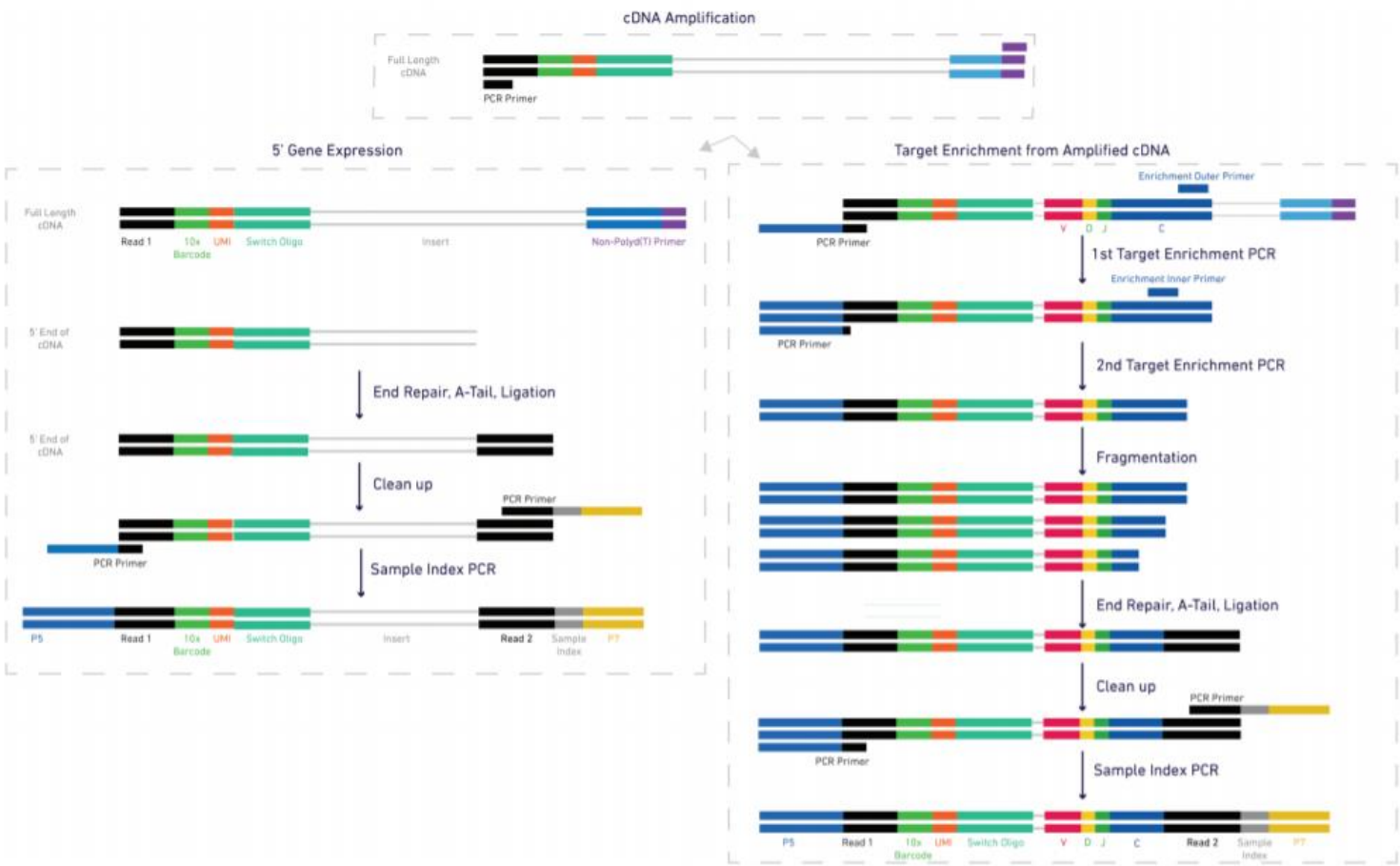


5’ scRNA-seq

Inside individual GEMs



Pooled cDNA processed in bulk



https://assets.ctfassets.net/an68im79xiti/j7qXxnyhAAiIWcSWA6g0M/46d70e508f7a3019a5cef2d64f203a3d/CG000109_AssayConfiguration_VDJ_RevD.pdf