



ITMO UNIVERSITY

CT Lab  
ITMO UNIVERSITY

# Introduction into single cell RNA-seq

Konstantin Zaitsev

December 3<sup>rd</sup>, 2020. Tomsk / Saint-Petersburg

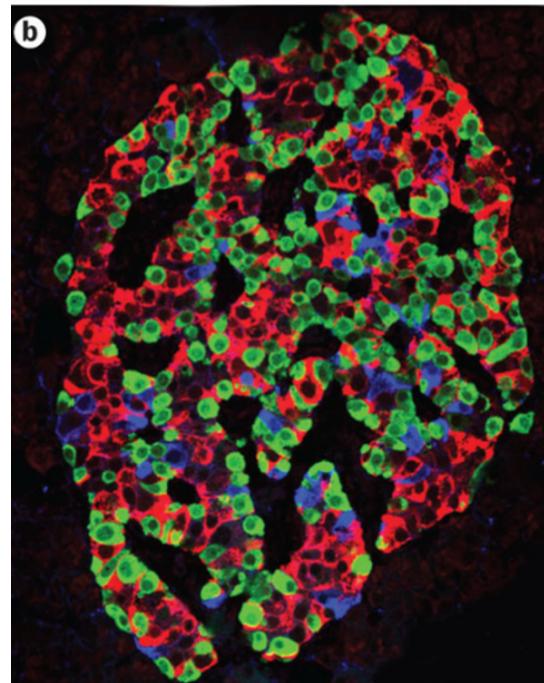
# Single-cell RNA-seq (scRNA-seq)

# Cell is the fundamental unit

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

Heterogeneous populations:

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

Homogeneous populations:

- Understanding heterogeneity
- Cellular states and cellular processes

Tracking of cell differentiation

# Smart-seq2

---

## PROTOCOL

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

Simone Picelli<sup>1</sup>, Omid R Faridani<sup>1</sup>, Åsa K Björklund<sup>1,2</sup>, Gösta Winberg<sup>1,2</sup>, Sven Sagasser<sup>1,2</sup> & Rickard Sandberg<sup>1,2</sup>

---

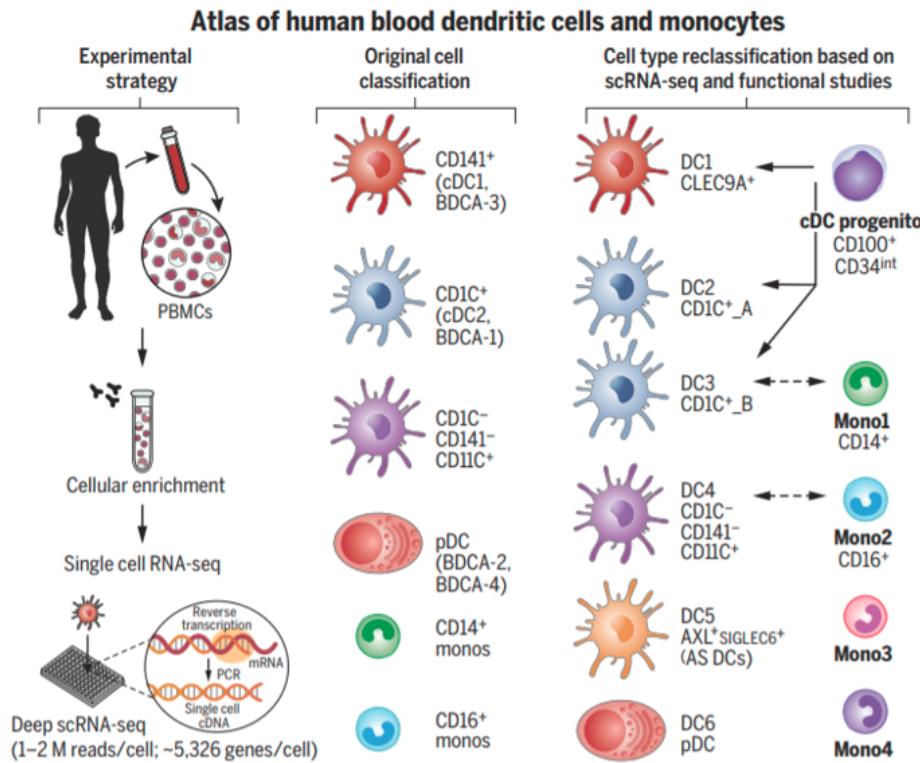
<sup>1</sup>Ludwig Institute for Cancer Research, Stockholm, Sweden. <sup>2</sup>Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.  
Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

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

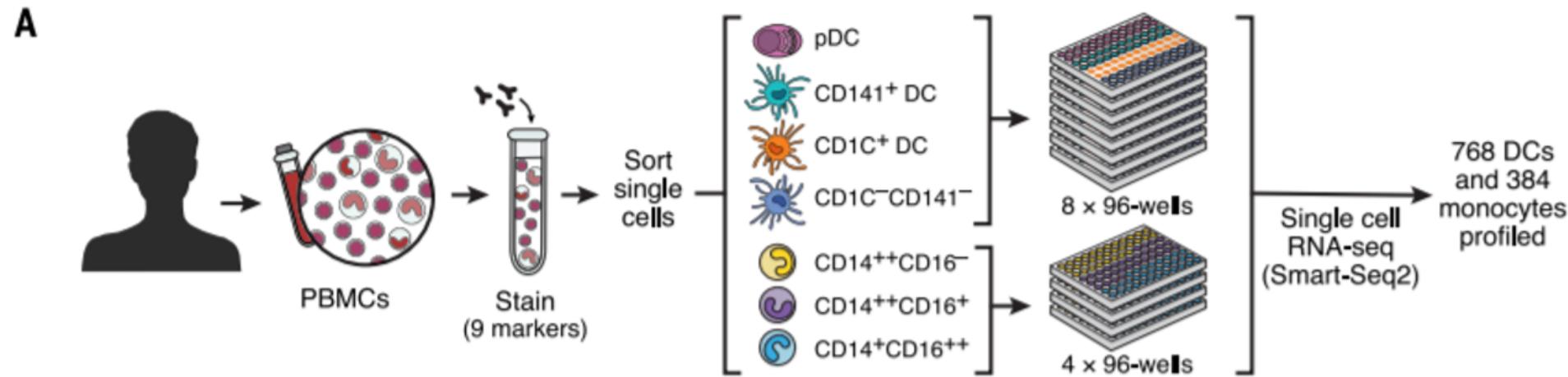
Emerging methods for the accurate quantification of gene expression in individual cells hold promise for revealing the extent, function and origins of cell-to-cell variability. Different high-throughput methods for single-cell RNA-seq have been introduced that vary in coverage, sensitivity and multiplexing 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<sup>-</sup>) RNA.

# Single-cell RNA-seq of myeloid cells

- Villani, Satija et al
- Science, 2017
- 1152 cells

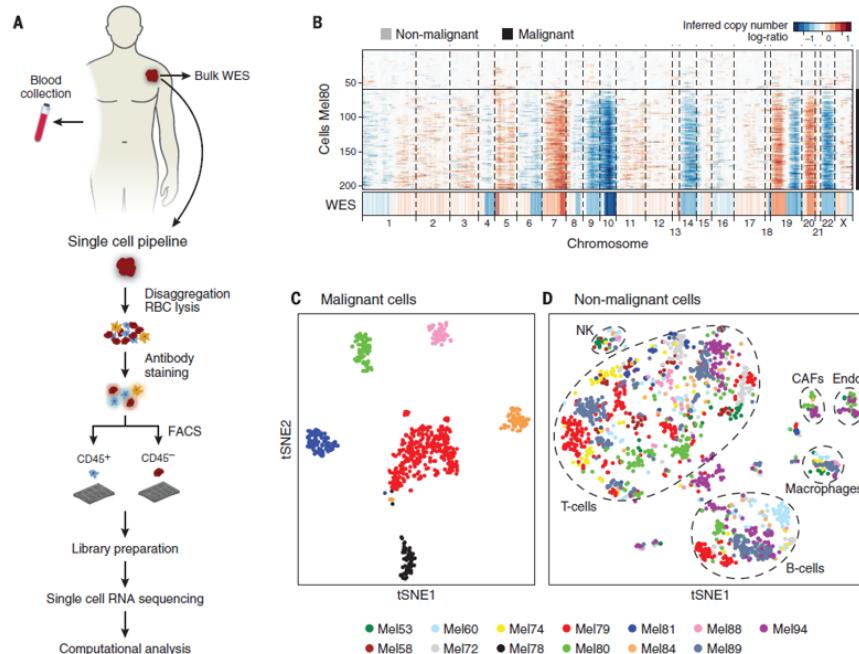


# Single-cell RNA-seq of myeloid cells



# Single-cell RNA-seq of melanoma

- Tirosh, Izar et al
- Science, 2016
- 4645 cells



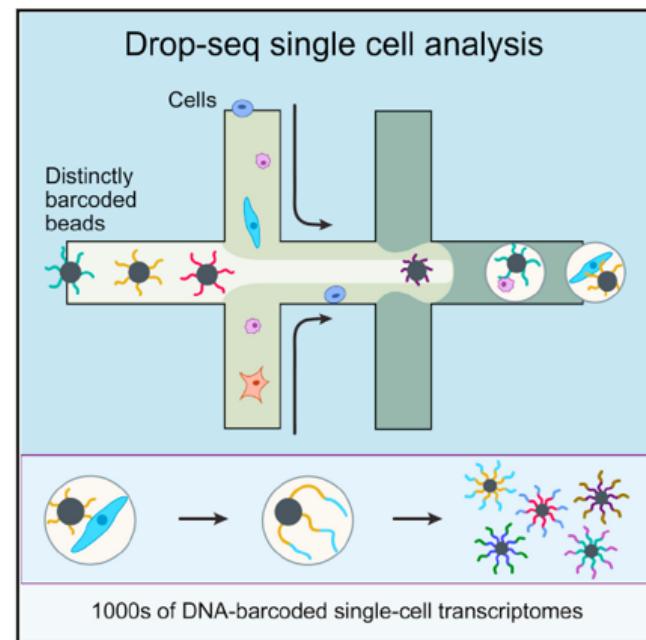
# Drop-seq: Cell, 2015

Cell

Resource

## 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.

# 10x chromium machine: NComms, 2017



## ARTICLE

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

DOI: 10.1038/ncomms14049

OPEN

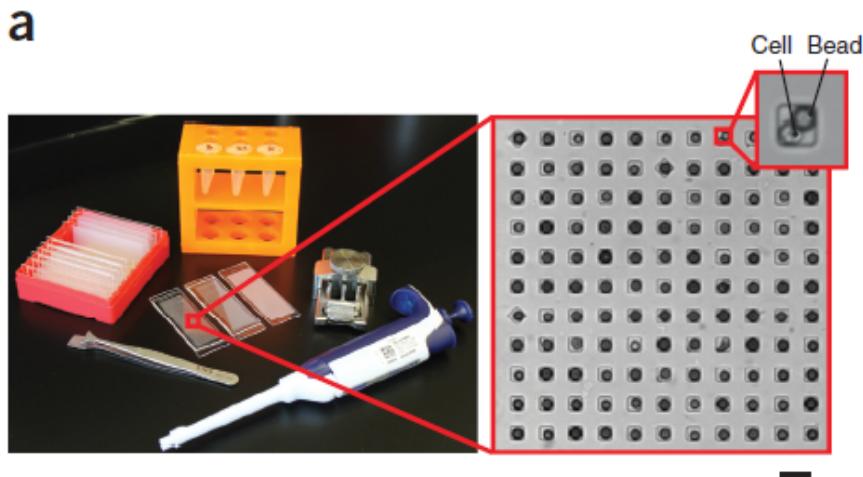
## Massively parallel digital transcriptional profiling of single cells

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

# Seq-Well: NMeth, 2017

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

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



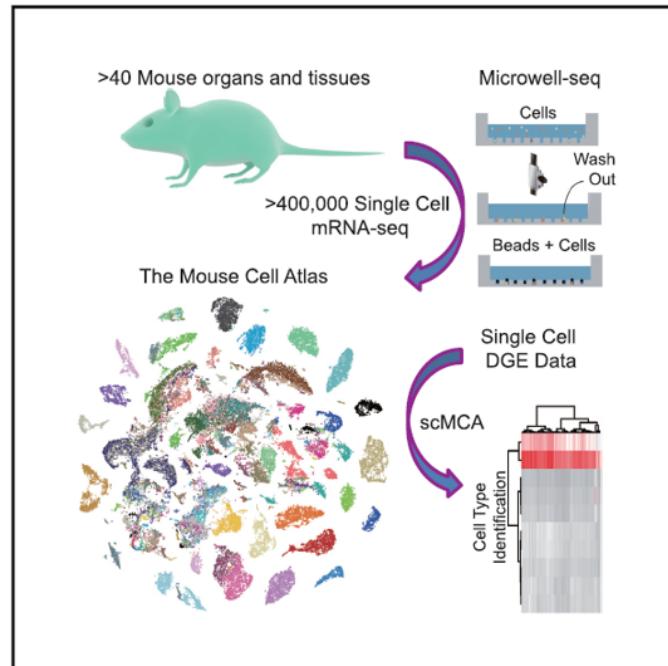
# Microwell-Seq: Cell, 2018

Cell

Resource

## Mapping the Mouse Cell Atlas by Microwell-Seq

### Graphical Abstract



### Authors

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

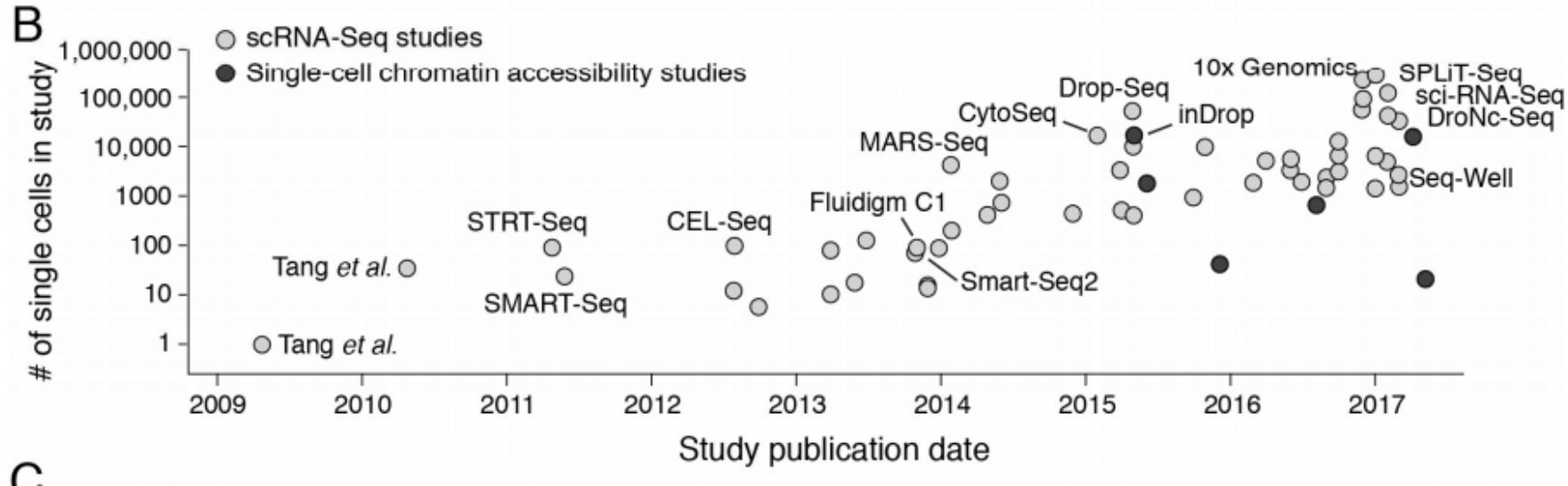
### Correspondence

xhan@zju.edu.cn (X.H.),  
gjj@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.

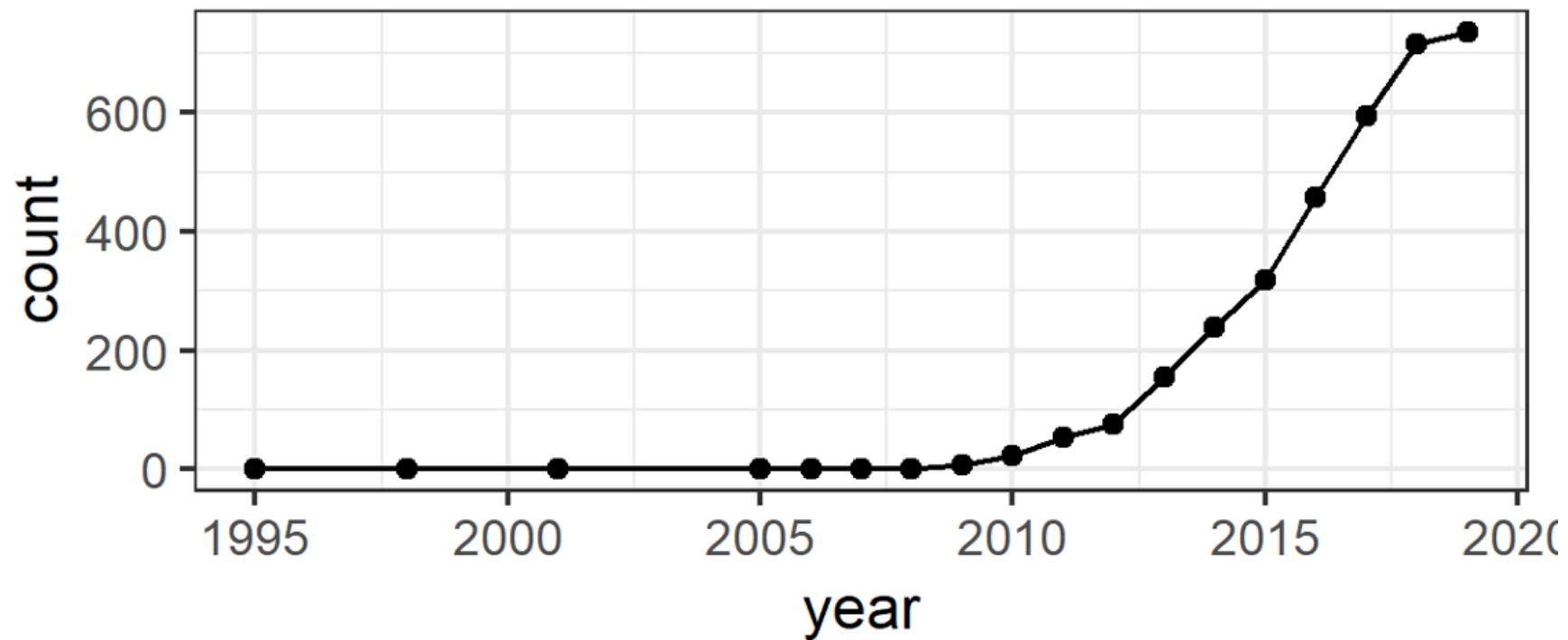
# Growth of single-cell technologies



C

# Growth of single-cell technologies

Number of scRNA-seq papers by year

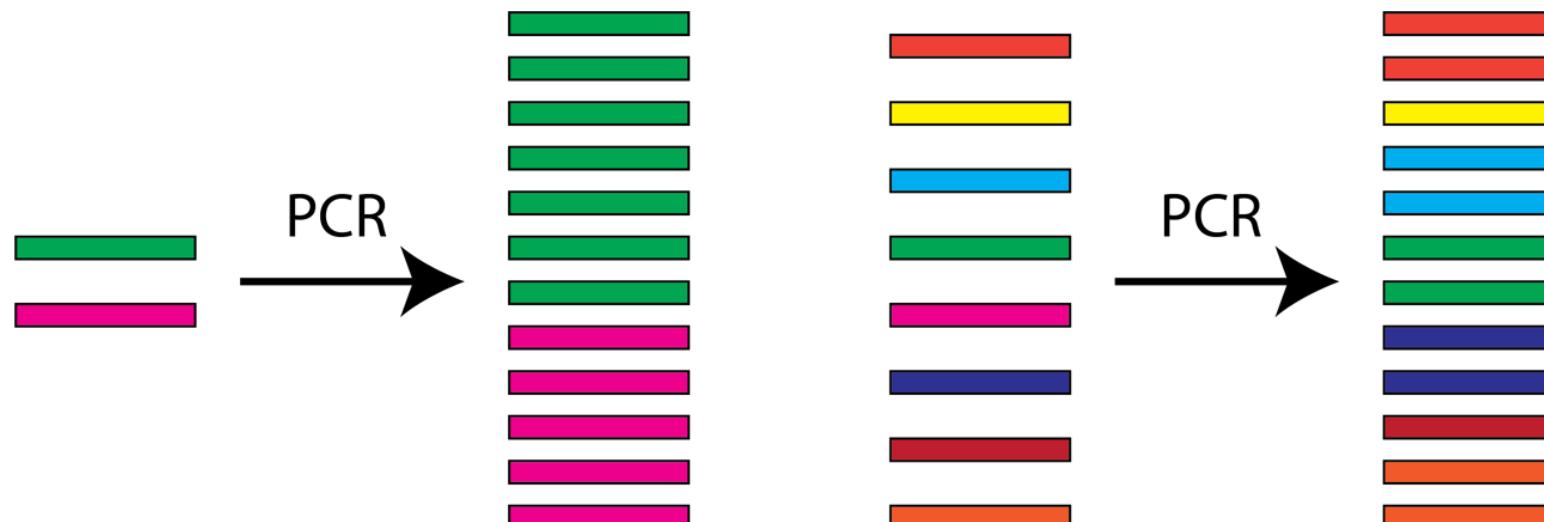


# Main challenges

- **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?

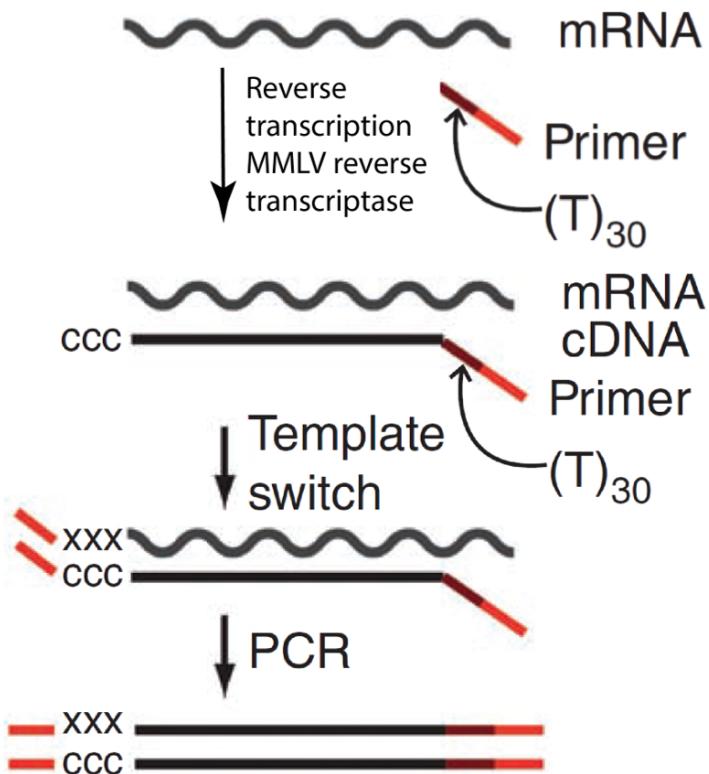
# Low library complexity

- 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

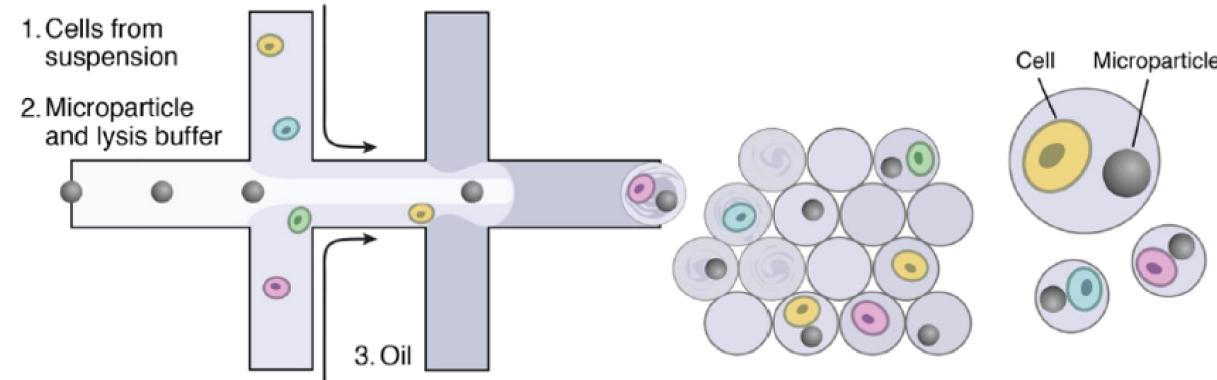


# Template-switching PCR

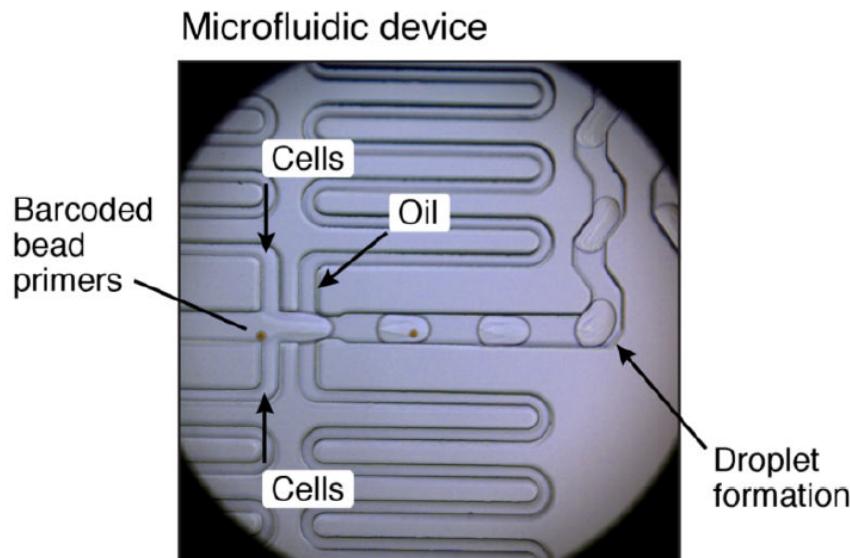
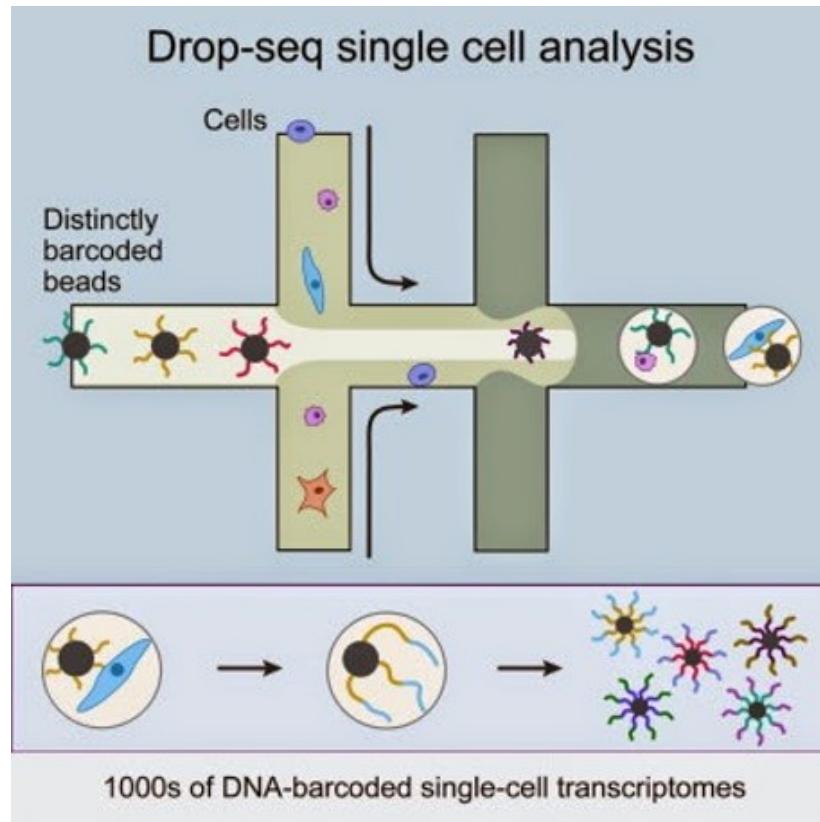
- Switching Mechanism At the 5' end of RNA Template (SMART)
- Robust for low input libraries



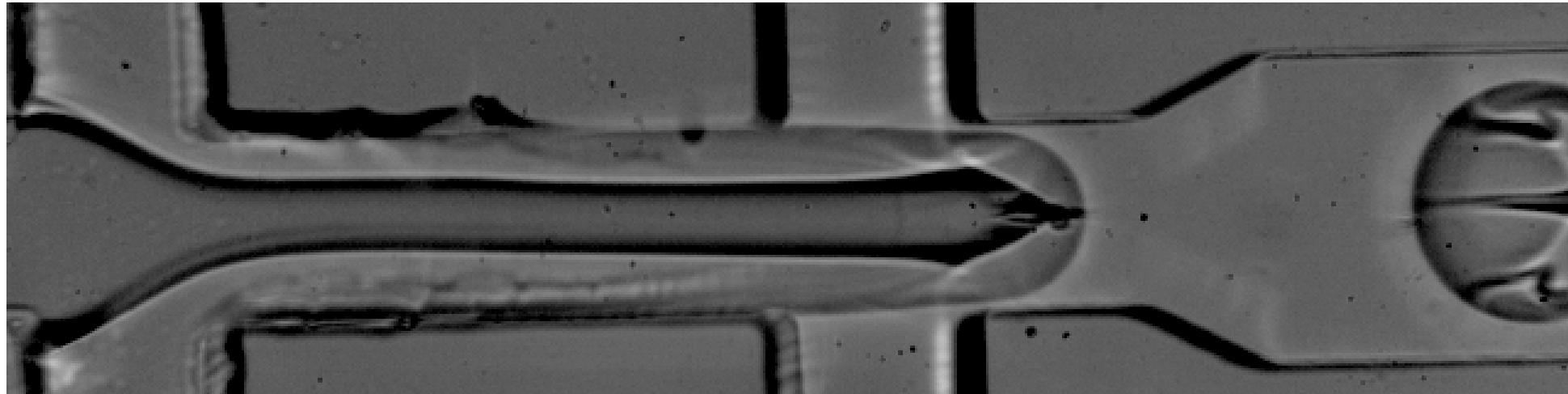
# Drop-seq schematics



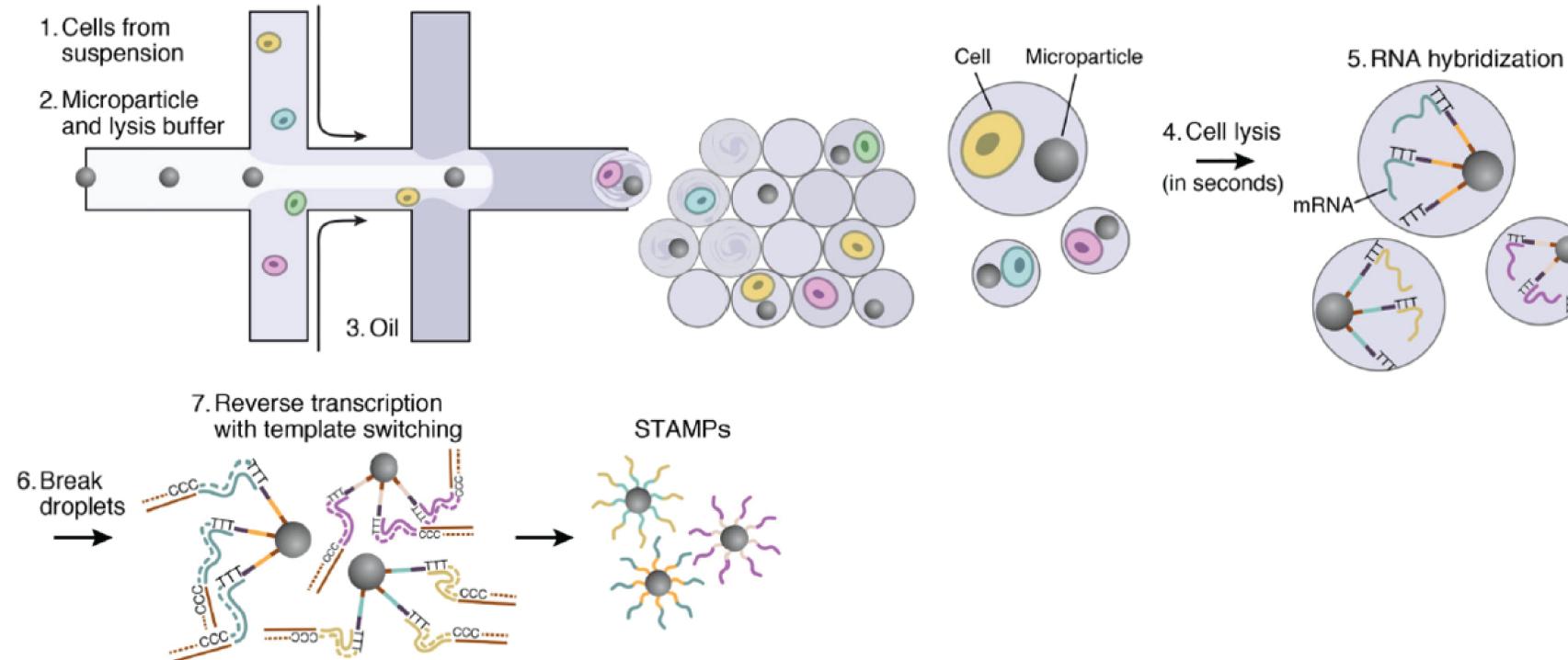
# Drop-seq microfluidics



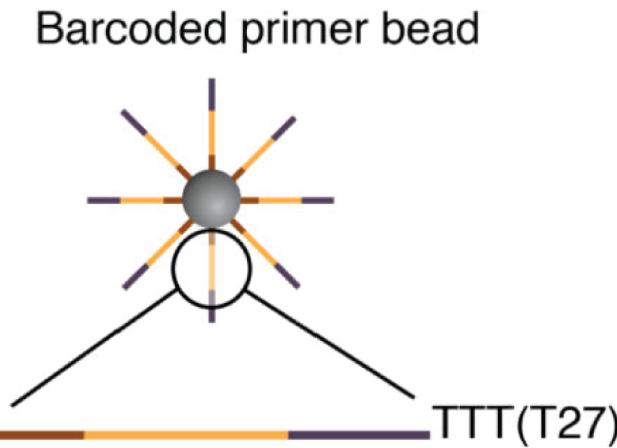
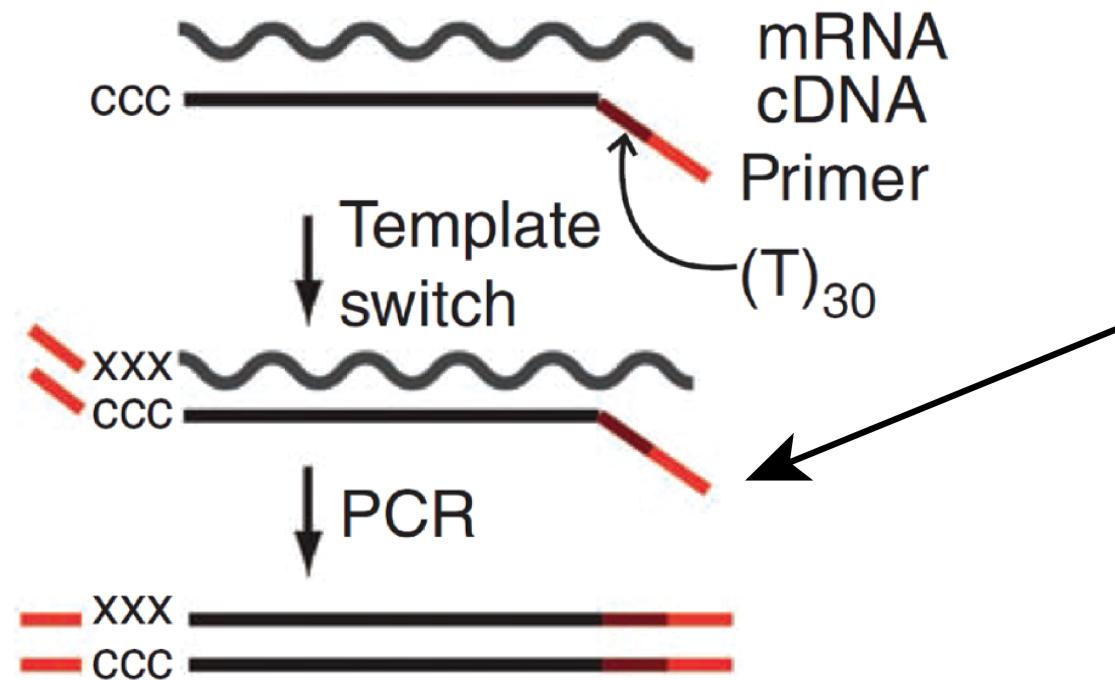
# Drop-seq microfluidics



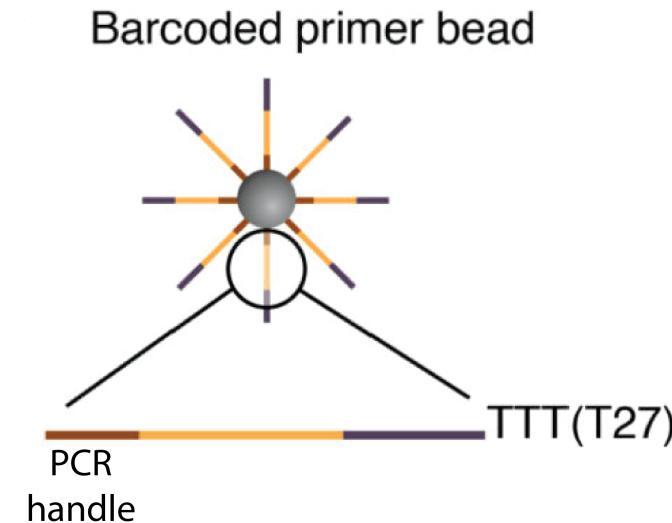
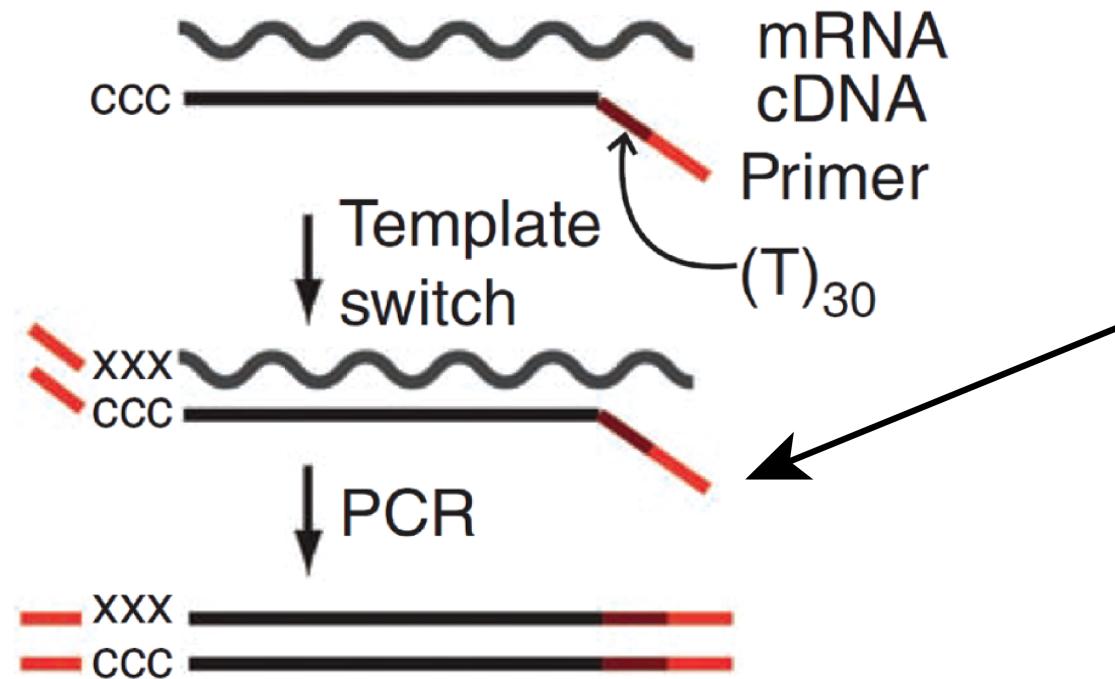
# Drop-seq schematics



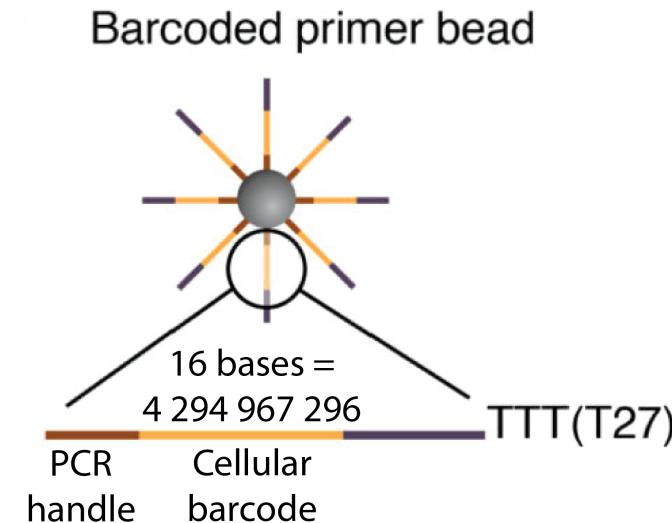
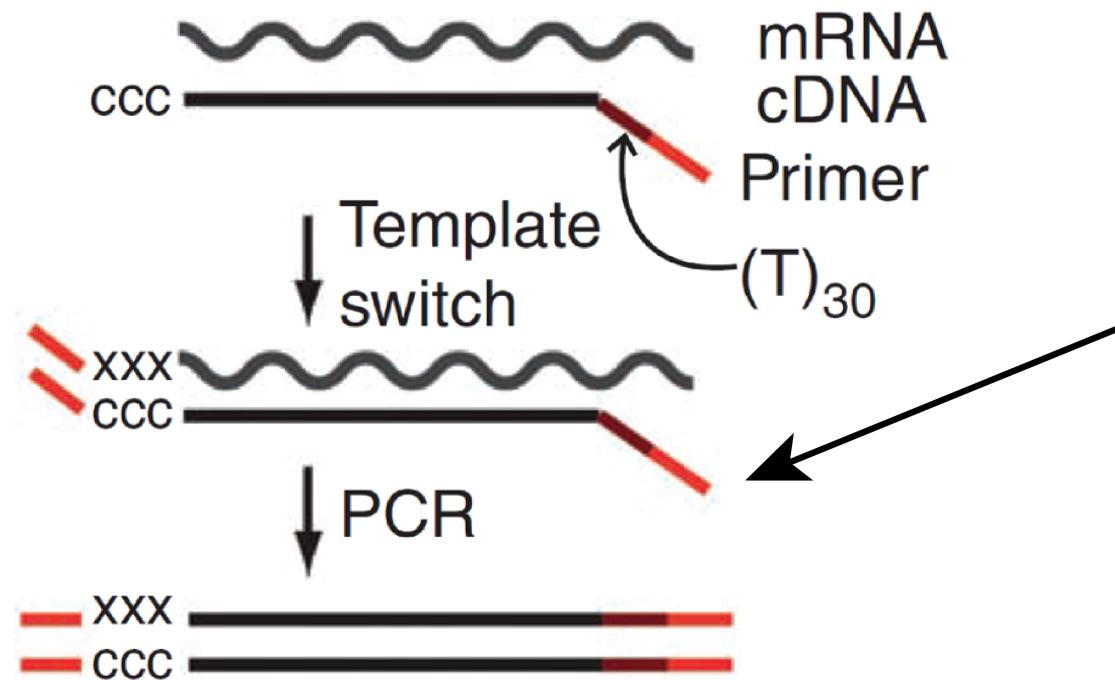
# Barcoding



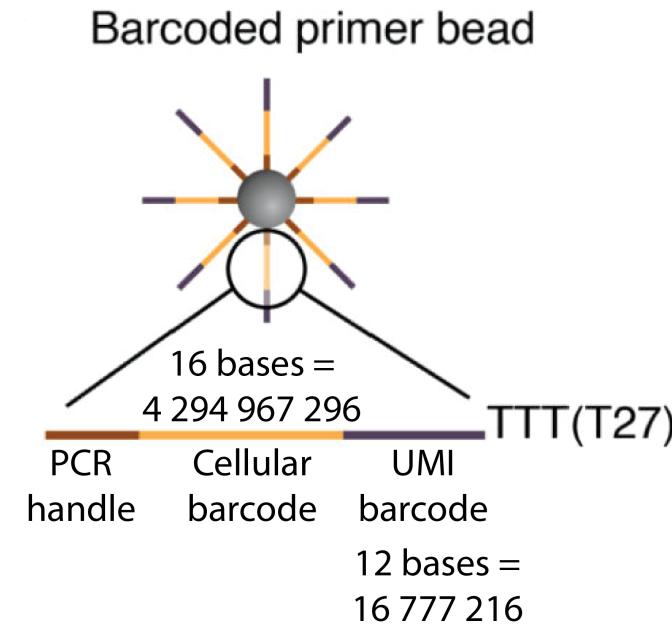
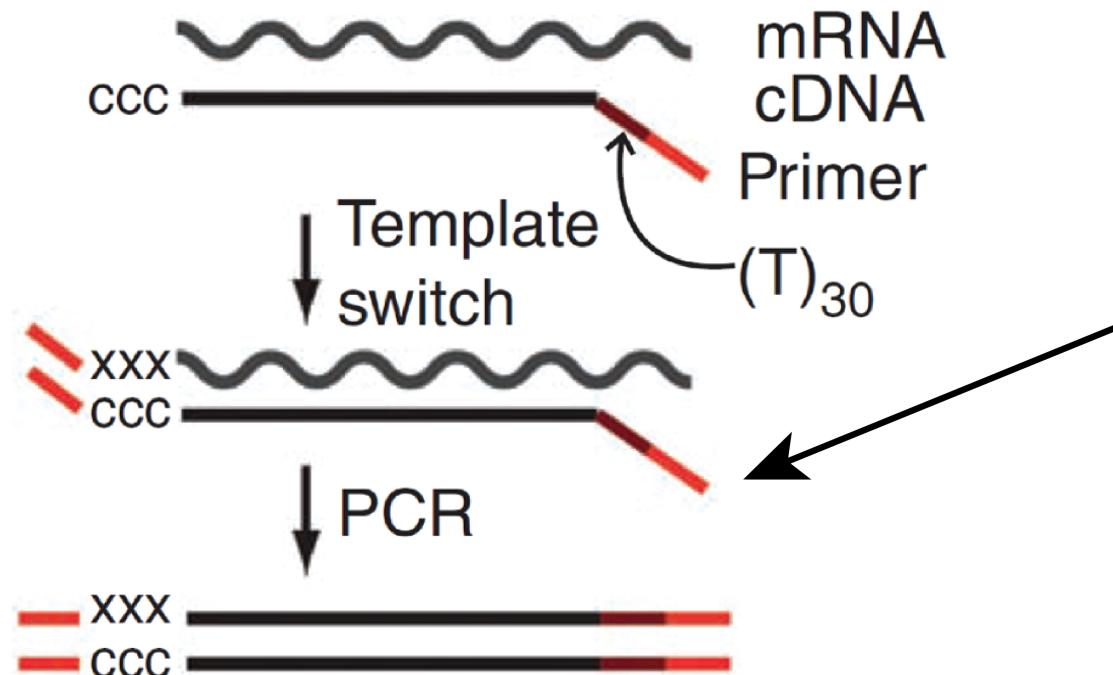
# Barcoding



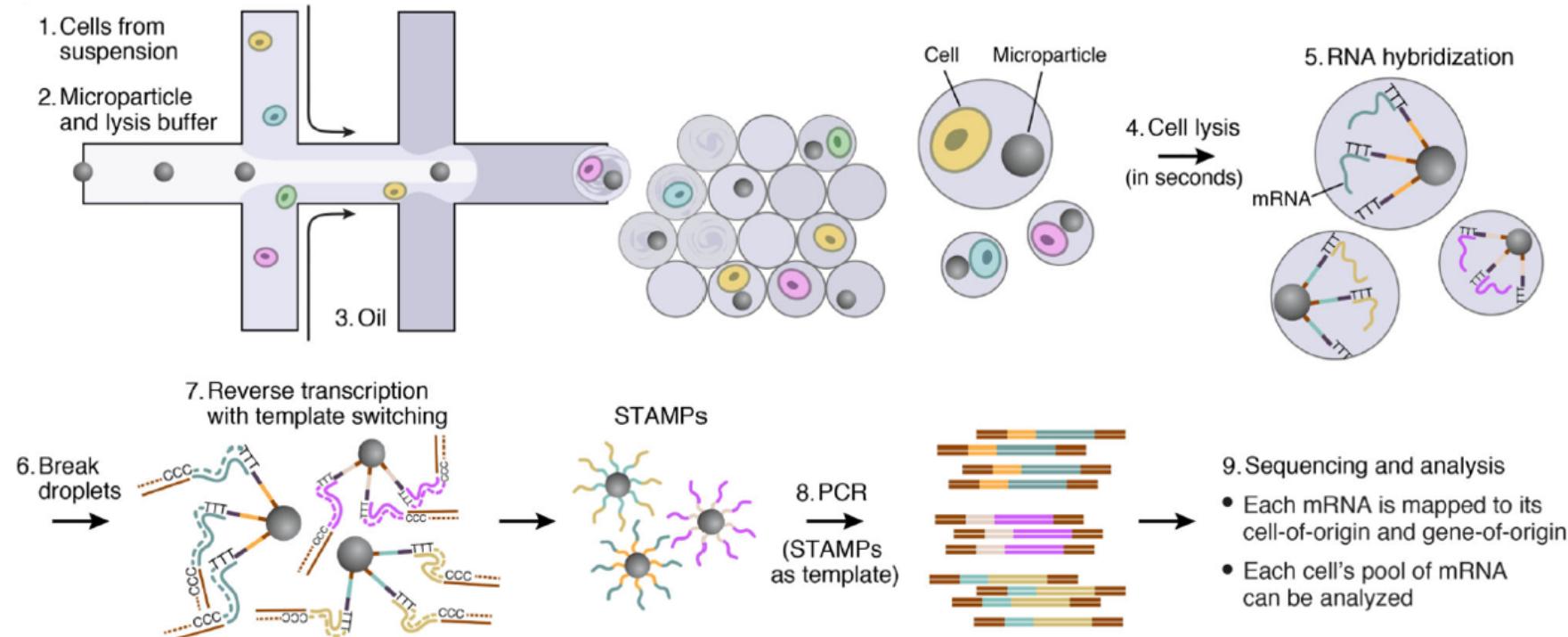
# Barcoding



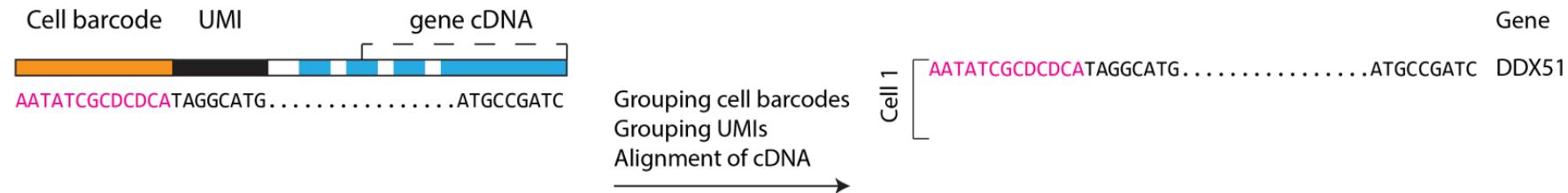
# Barcoding



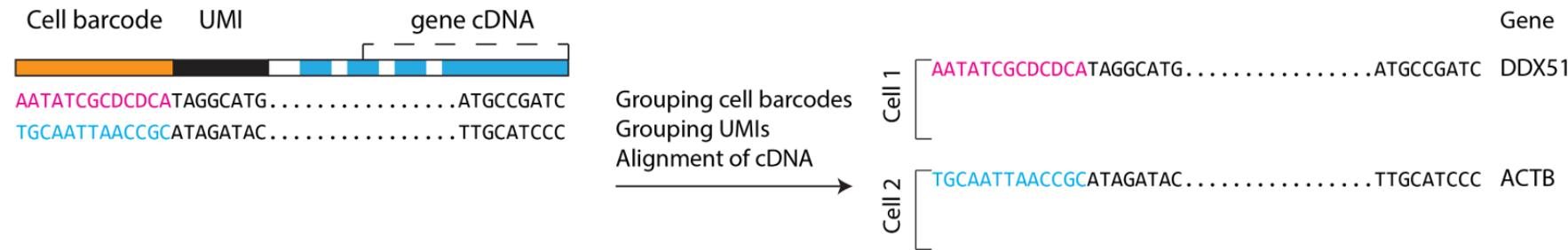
# Drop-seq schematics



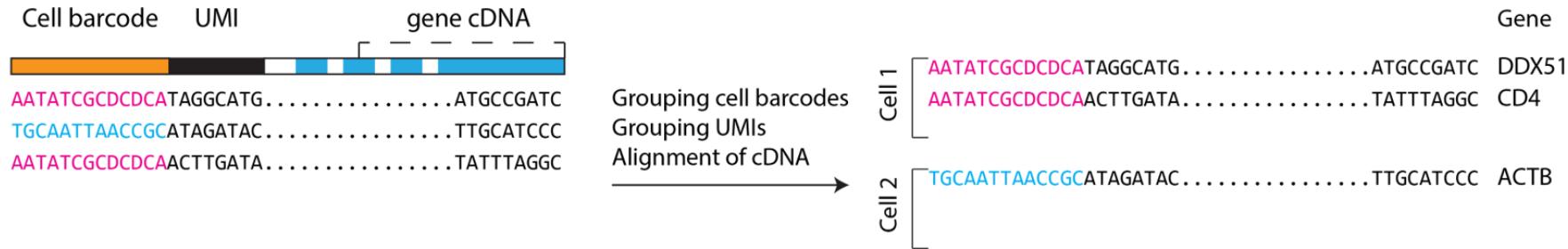
# Sequencing



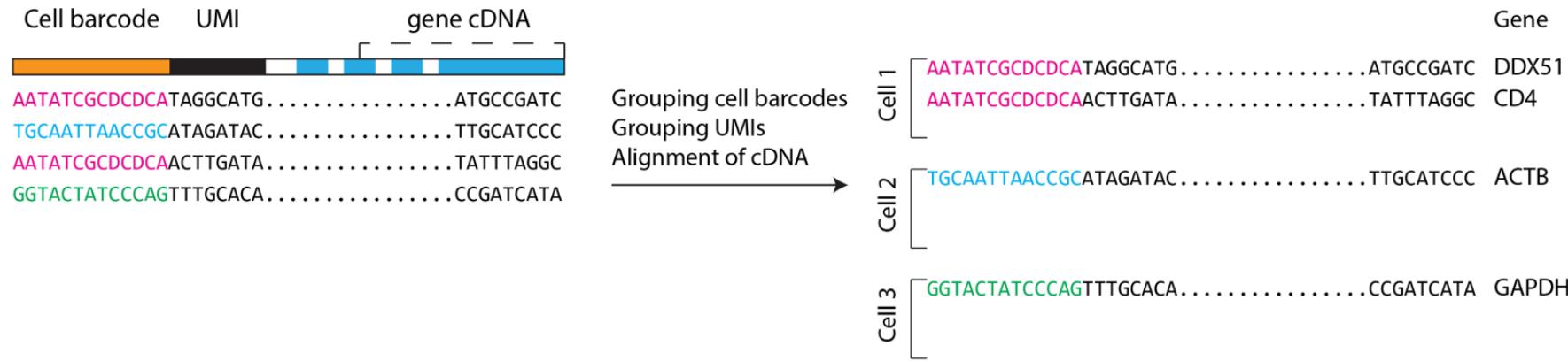
# Sequencing



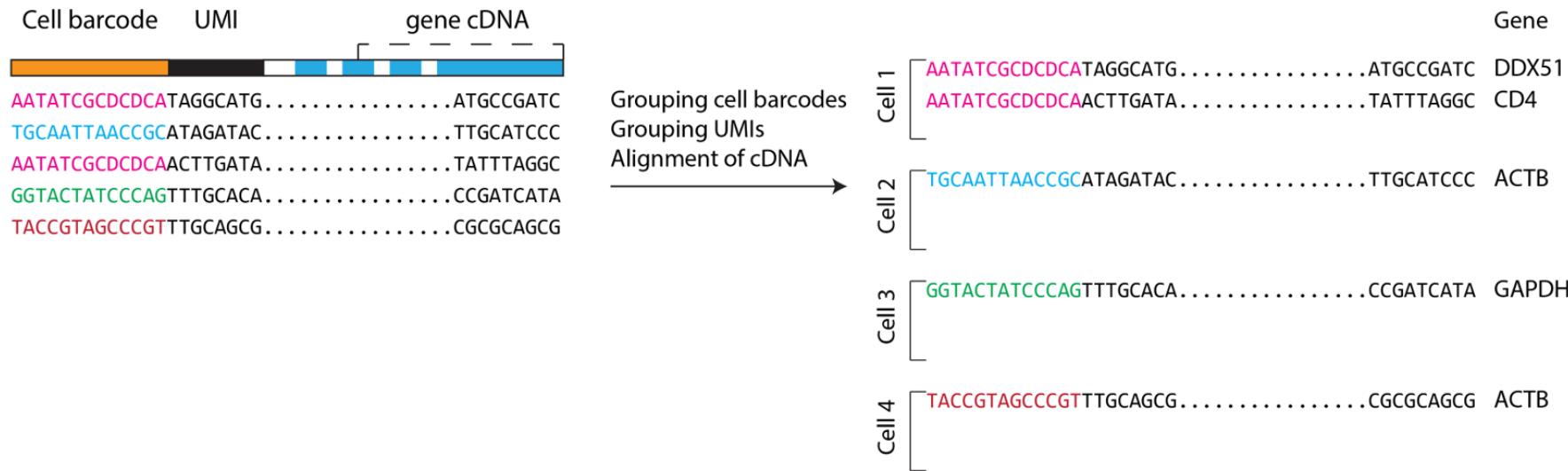
# Sequencing



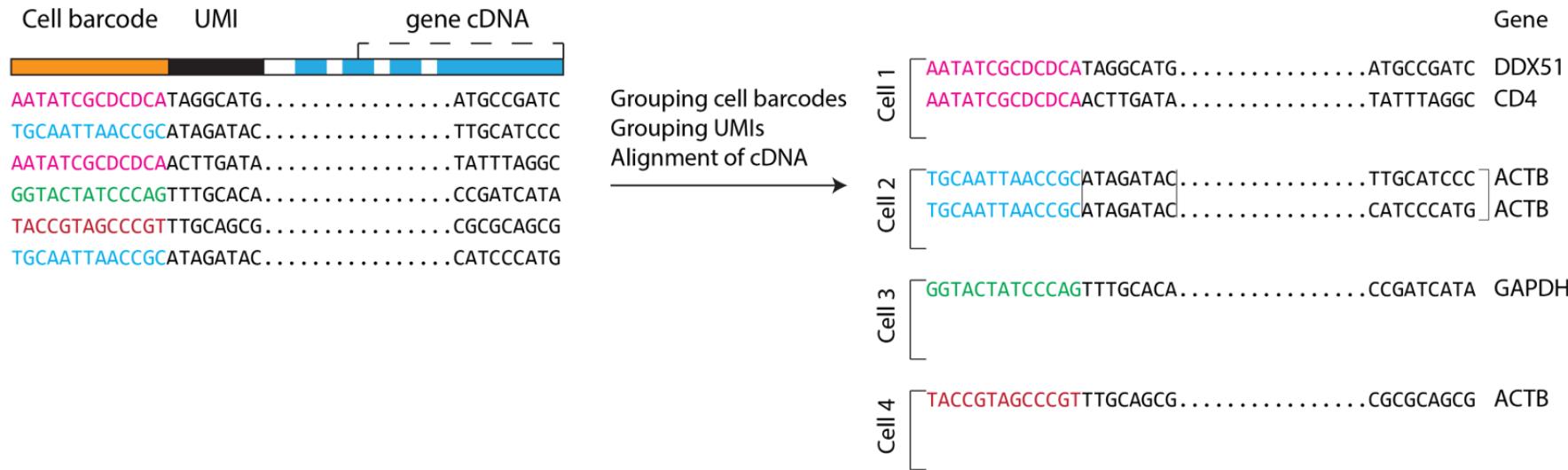
# Sequencing



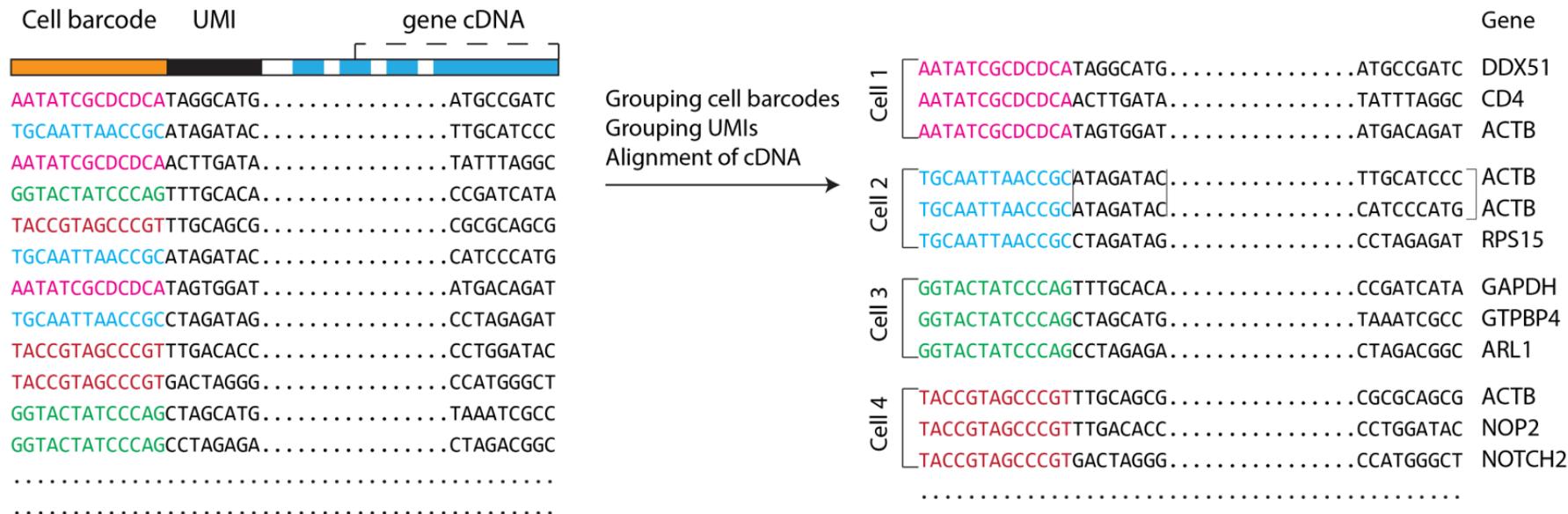
# Sequencing



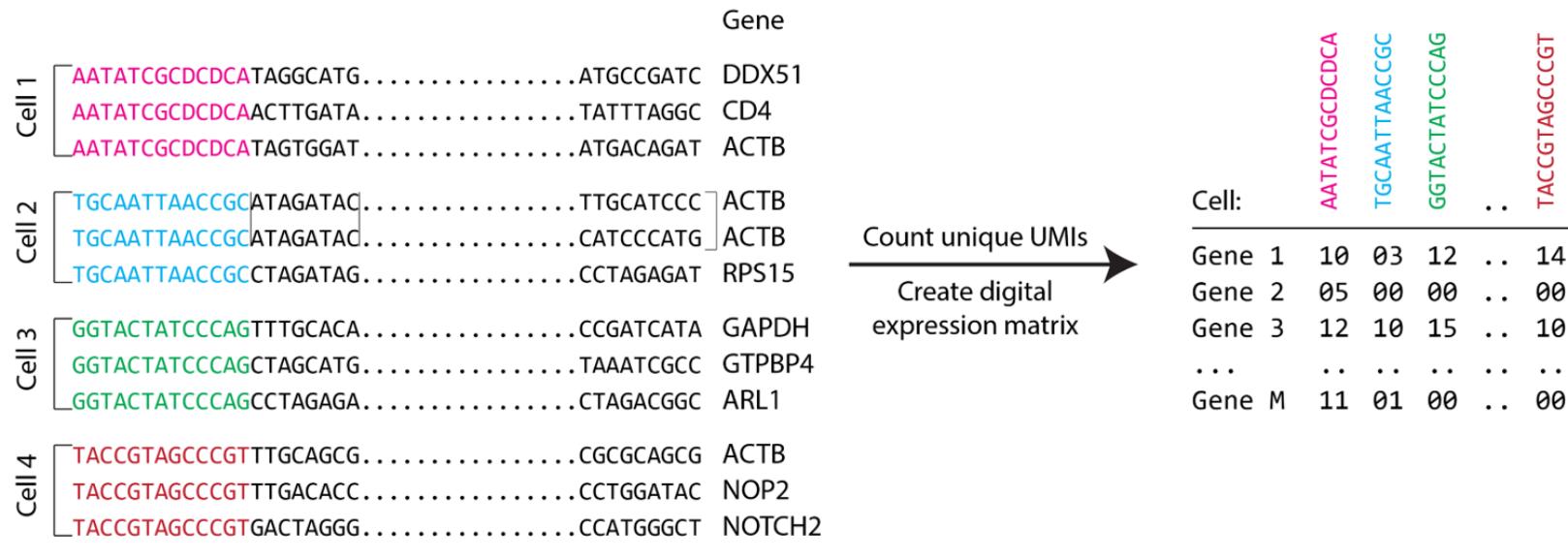
# Sequencing



# Sequencing



# Sequencing



# All questions were addressed

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

# All questions were addressed

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

# Dataset for today:

- [https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k\\_pbmc\\_v3\\_nextgem](https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_v3_nextgem)

## 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (Next GEM)

Single Cell Gene Expression Dataset by Cell Ranger 3.0.2

Peripheral blood mononuclear cells (PBMCs) from a healthy donor (the same cells were used to generate [5k\\_pbmc\\_v3](#)).

PBMCs are primary cells with relatively small amounts of RNA (~1pg RNA/cell).

Libraries were prepared following the Chromium Next GEM Single Cell 3' Reagent Kits v3.1 User Guide (CG000204 RevA).

- 5,155 cells detected
- Sequenced on Illumina NovaSeq with approximately 76,406 reads per cell
- 28bp read1 (16bp Chromium barcode and 12bp UMI), 91bp read2 (transcript), and 8bp I7 sample barcode
- run with --expect-cells=5000

Published on May 29, 2019

This dataset is licensed under the Creative Commons Attribution license.

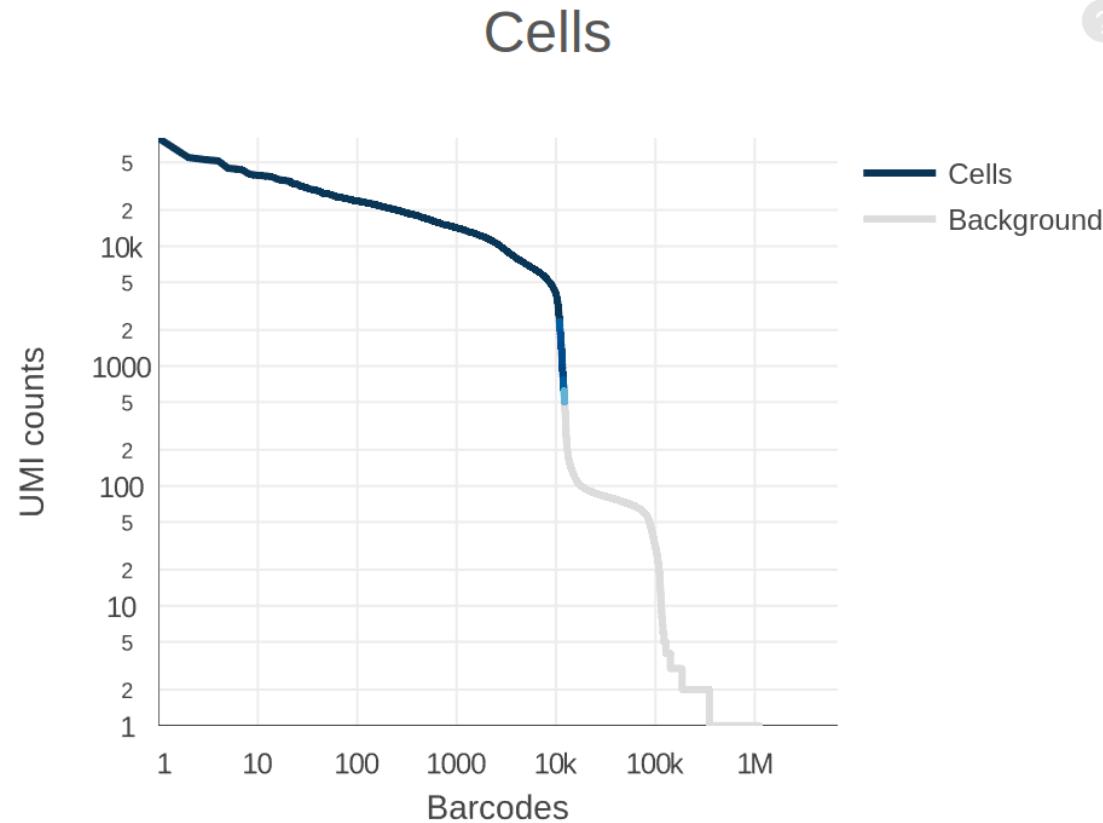
[View Summary](#)

[Show batch download instructions](#)

# Let's have a loot at the summary

- [http://cf.10xgenomics.com/samples/cell-exp/3.0.2/5k\\_pbmc\\_v3\\_nextgem/5k\\_pbmc\\_v3\\_nextgem\\_web\\_summary.html](http://cf.10xgenomics.com/samples/cell-exp/3.0.2/5k_pbmc_v3_nextgem/5k_pbmc_v3_nextgem_web_summary.html)

# Understaing what's noise



# Understaing what's noise

There is cell-free RNA in the cellular suspension that will be captured in empty droplets (with beads), and we must distinguish cells from empty droplets:

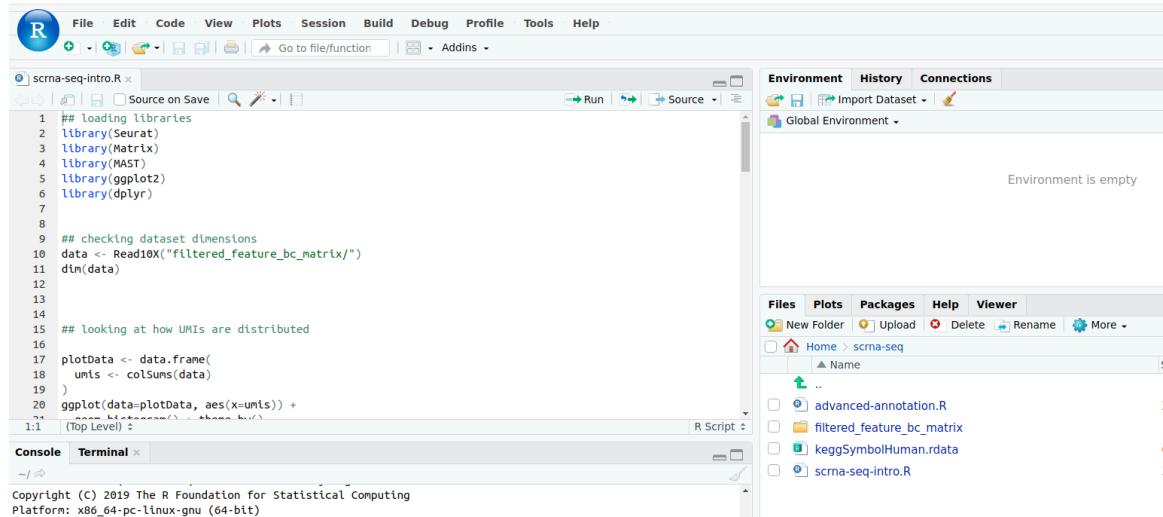
- We can set a hard threshold on UMI to filter noise from actual cells
- We can calculate noise signature and test every cell against this signature (emptyDrops, <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y>)

# Basic steps to analysis of scRNA-seq

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

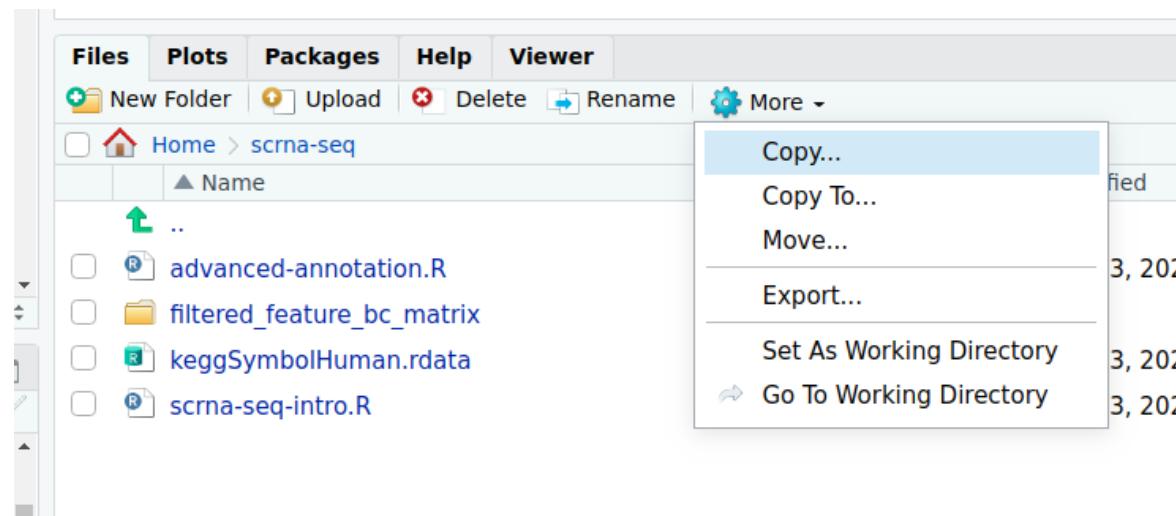
# Our setup

- Address is the same <https://ctlab.itmo.ru/rstudio-sbNN/>
- Folder scrna-seq



# Our setup

- Open the folder
- Set this as a working directory
- Open file scRNA-seq-intro.R



# Our setup

- You will see chunks of code appearing on my slides
- Follow along!
- Select the code and press Ctrl + Enter to run the chunk

# Loading all the libraries

```
library(Seurat)
library(Matrix)
library(MAST)
library(ggplot2)
library(dplyr)
```

# Loading the data

```
data <- Read10X("filtered_feature_bc_matrix/")
dim(data)

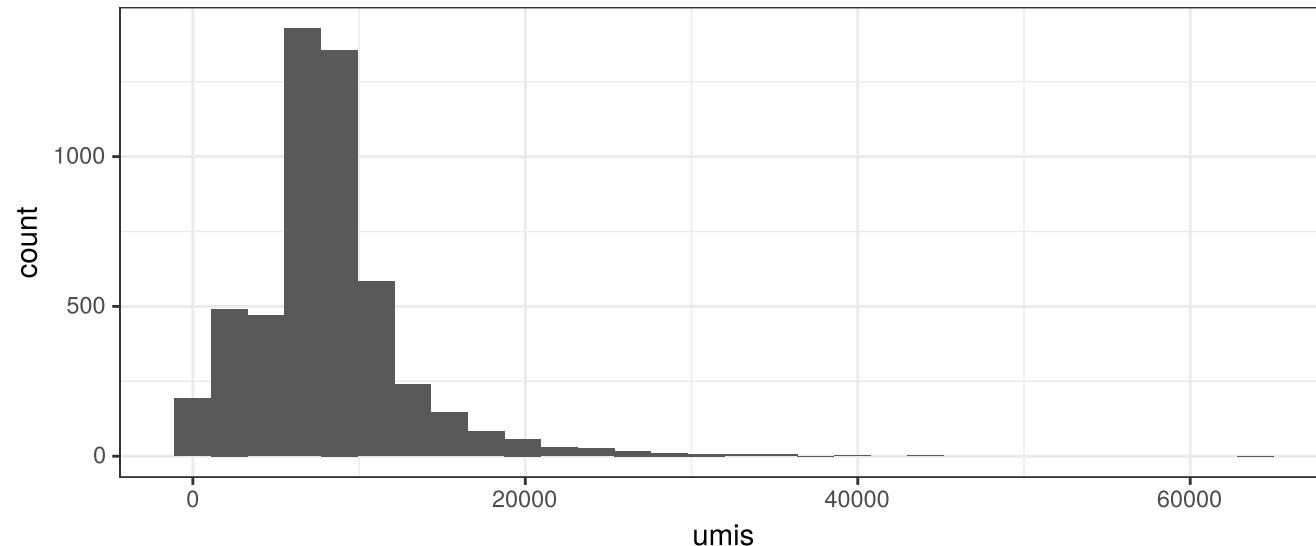
## [1] 33538 5155
```

Count matrix is large:

- 33538 genes
- 11769 cells

# UMI distribution

```
plotData <- data.frame(  
  umis <- colSums(data))  
)  
ggplot(data=plotData, aes(x=umis)) +  
  geom_histogram() + theme_bw()
```



# Filtering (genes and barcodes)

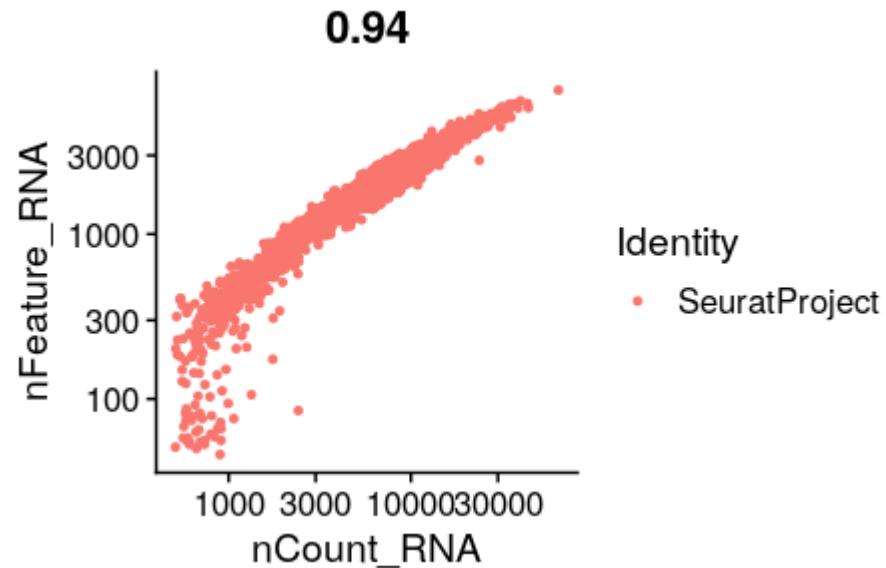
```
seurat <- CreateSeuratObject(data, min.cells = 10, min.features = 10)
dim(seurat)

## [1] 16022 5155
```

17527 of 33538 genes were filtered (detected in  $\leq 10$  cells)

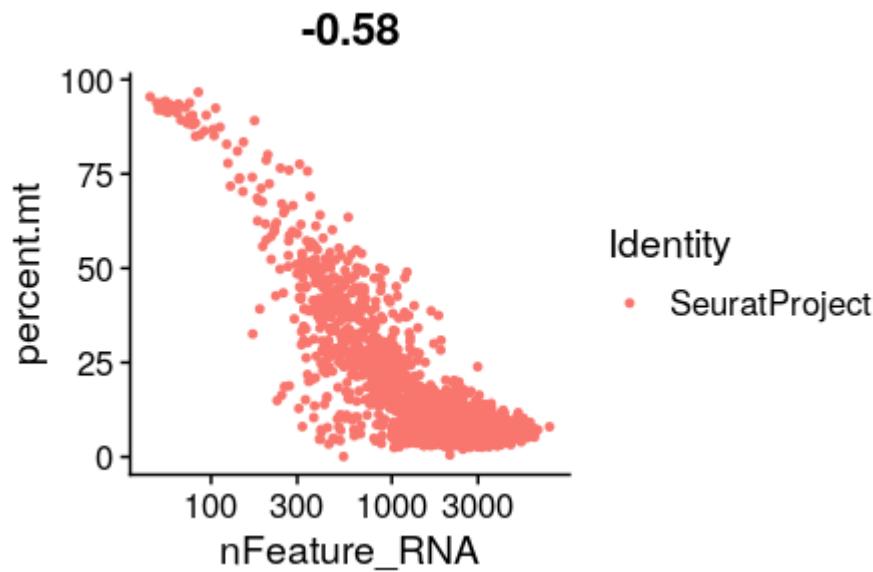
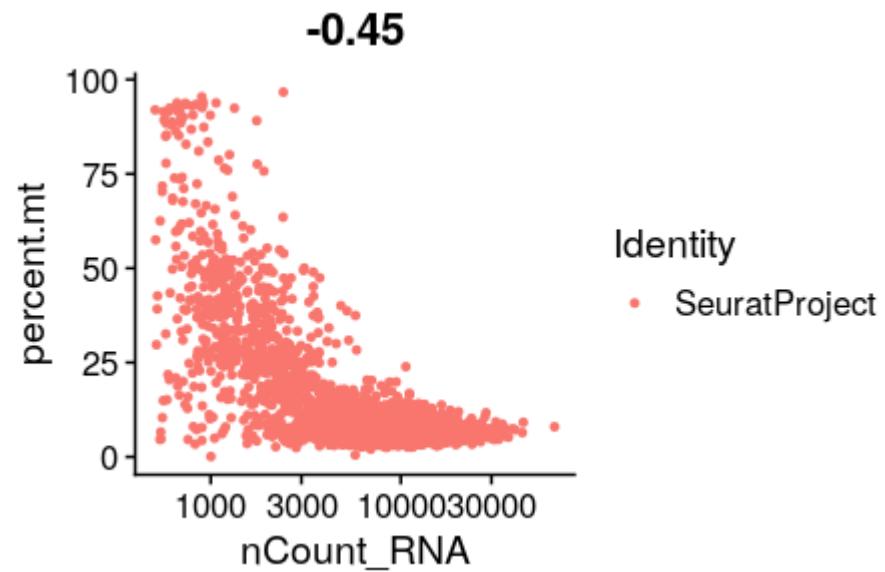
# Filtering (genes and barcodes)

```
FeatureScatter(seurat, "nCount_RNA", "nFeature_RNA") + scale_x_log10() + scale_y_log10()
```



# Filtering (genes and barcodes)

```
seurat[["percent.mt"]] <- PercentageFeatureSet(seurat, pattern = "^MT-")
FeatureScatter(seurat, "nCount_RNA", "percent.mt") + scale_x_log10()
FeatureScatter(seurat, "nFeature_RNA", "percent.mt") + scale_x_log10()
```



# Filtering (genes and barcodes)

```
seurat <- subset(seurat, subset = nFeature_RNA > 300 & percent.mt < 25)
dim(seurat)

## [1] 16022 4714
```

# Normalization (old way)

```
## seurat <- NormalizeData(seurat, normalization.method = "LogNormalize", scale.factor =  
## seurat <- FindVariableFeatures(seurat, selection.method = "vst", nfeatures = 2000)  
## seurat <- ScaleData(seurat)
```

- Scaling expression to 10 000 UMIs (instead of million in RPM)
- Finding features with high variance (features above mean/variance trend)
- Scaling data (for PCA and so on)

# Normalization (better way)

```
seurat <- SCTransform(seurat, vars.to.regress = "percent.mt", verbose = FALSE)
```

- SCTransform does all of these things in one command

# SCTransform:

## Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression

Christoph Hafemeister<sup>1</sup> & Rahul Satija<sup>1,2</sup>

<sup>1</sup>New York Genome Center, New York City, NY, USA

<sup>2</sup>Center for Genomics and Systems Biology, New York University, New York City, NY, USA

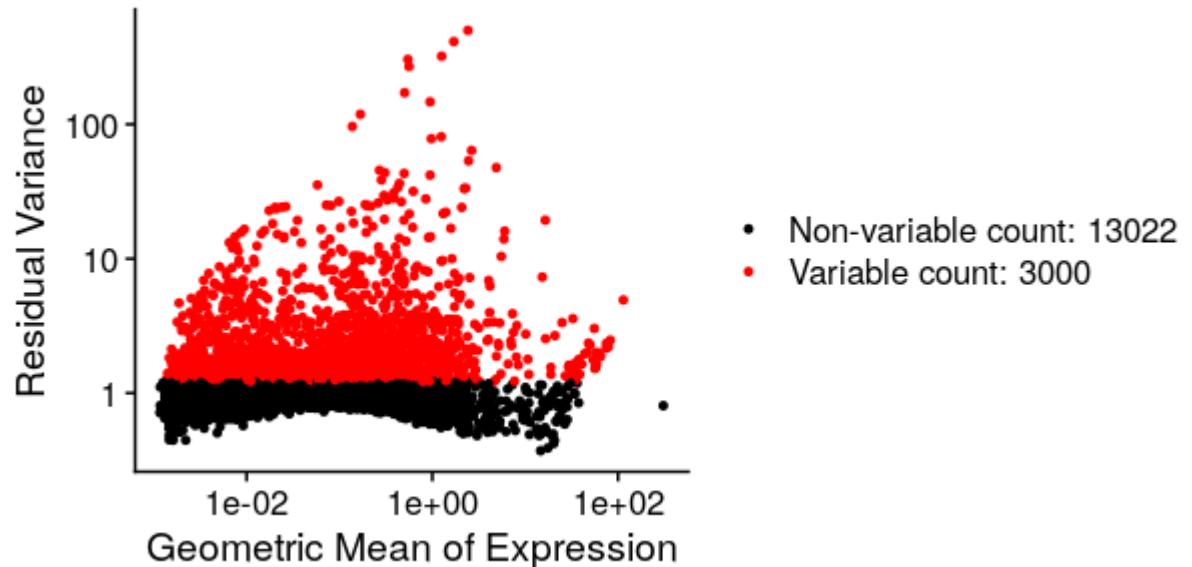
Correspondence to: chafemeister@nygenome.org, rsatija@nygenome.org

### Abstract

Single-cell RNA-seq (scRNA-seq) data exhibits significant cell-to-cell variation due to technical factors, including the number of molecules detected in each cell, which can confound biological heterogeneity with technical effects. To address this, we present a modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiments. We propose that the Pearson residuals from 'regularized negative binomial regression', where cellular sequencing depth is utilized as a covariate in a generalized linear model, successfully remove the influence of technical characteristics from downstream analyses while preserving biological heterogeneity. Importantly, we show that an unconstrained negative binomial model may overfit scRNA-seq data, and overcome this by pooling information across genes with similar abundances to obtain stable parameter estimates. Our procedure omits the need for heuristic steps including pseudocount addition or log-transformation, and improves common downstream analytical tasks such as variable gene selection, dimensional reduction, and differential expression. Our approach can be applied to any UMI-based scRNA-seq dataset and is freely available as part of the R package `sctransform`, with a direct interface to our single-cell toolkit `Seurat`.

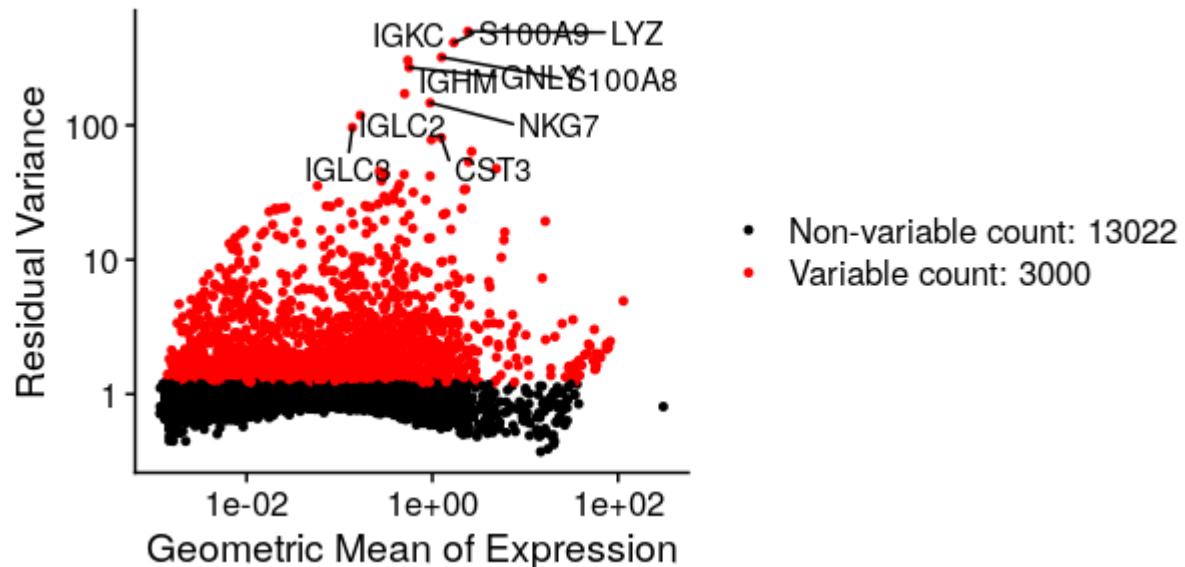
# Variable genes

```
VariableFeaturePlot(seurat) + scale_y_log10()
```



# Variable genes

```
top10_variable_genes <- head(VariableFeatures(seurat), 10)
VariableFeaturePlot(seurat) %>%
  LabelPoints(points = top10_variable_genes, repel = TRUE) +
  scale_y_log10()
```



# Basic steps to analysis of scRNA-seq

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

# High-dimensionality of scRNA-seq

Initially matrix is very large in size, this causes different kind of issues:

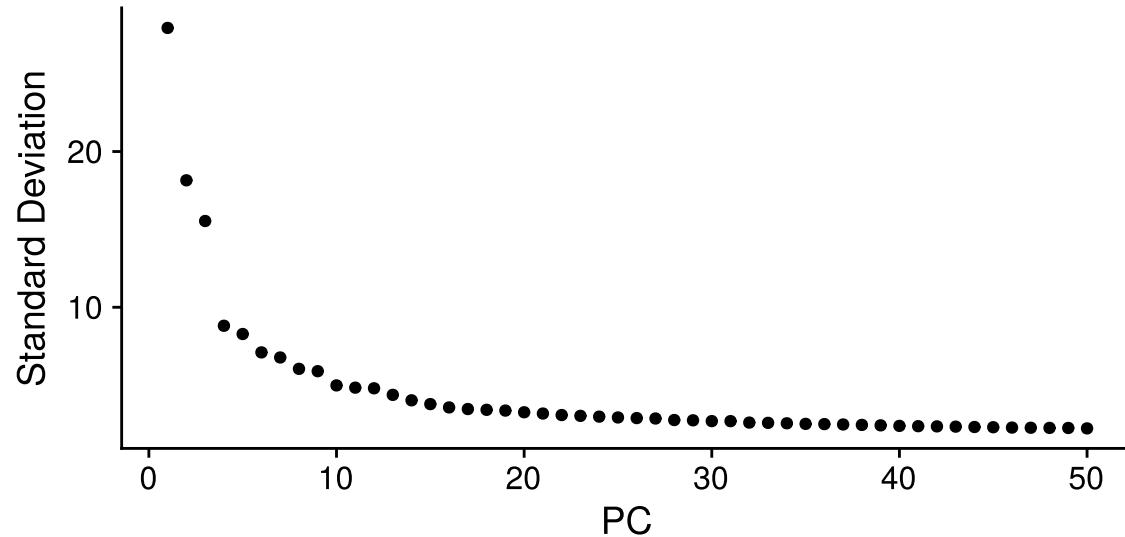
- Some algorithms are just slow when performed in this high-dimensionality data
- Curse of dimensionality

We usually take several steps to reduce dims before creating 2d clustered map of our dataset

- Keeping only variable genes (since those introduce variance to the dataset)
- PCA will reduce dimensionality to 20-30 first components

# PCA

```
seurat <- RunPCA(seurat, verbose = FALSE)  
ElbowPlot(seurat, ndims = 50)
```



# Basic steps to analysis of scRNA-seq

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

# Visualization

Both tSNE and UMAP will put our data after PCA into a 2D plane:

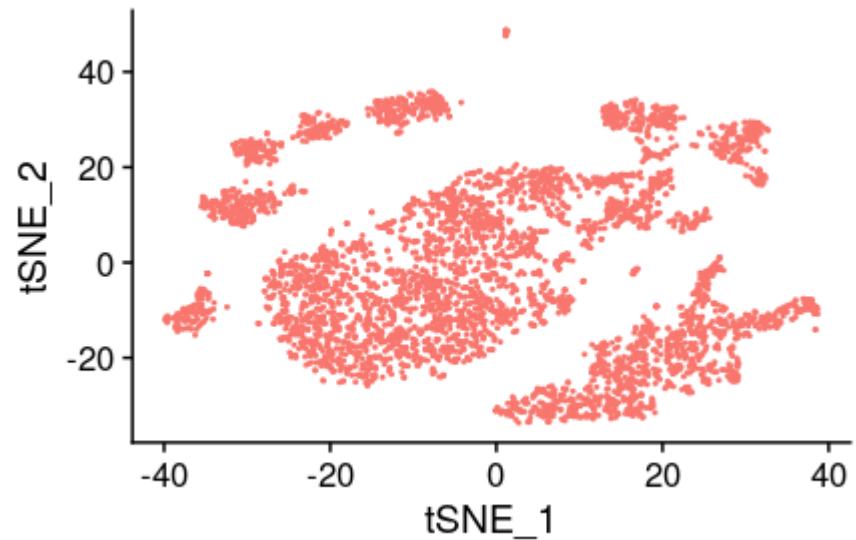
- Dots that are close to each other are cells that are transcriptionally similar to each other
- Dots that are far from each other are cells that are transcriptionally different from each other

Lets look at dimensionality of each cell:

- Variable genes only: 30k+ -> 2-3k of variable genes
- PCA: 2-3k of variable genes -> 20-30 principal components
- tSNE or UMAP: 20-30 PCs -> 2d or 3d plots

# tSNE

```
seurat <- RunTSNE(seurat, dims=1:20)
DimPlot(seurat, reduction = "tsne") + NoLegend()
```



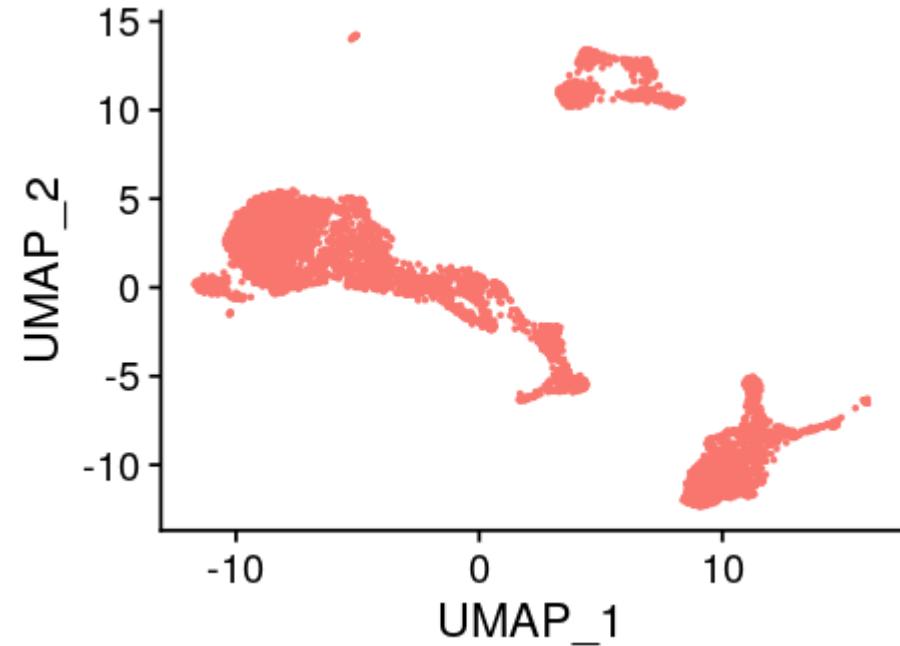
# tSNE

tSNE - "t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets"

- We can calculate distances in original space, and then calculate conditional probabilities  $p_{i|j}$  that point  $i$  would choose point  $j$  as a neighbor.  $p_{i|j}$  are proportional to "distances" from  $i$  to all other points (actually probability density around point  $i$ , but it doesn't matter here).
- Once all  $p_{i,j}$  are calculated in original space we try to find such 2d/3d space that would have similar probabilities

# UMAP

```
seurat <- RunUMAP(seurat, dims=1:20)
DimPlot(seurat, reduction = "umap") + NoLegend()
```



# UMAP

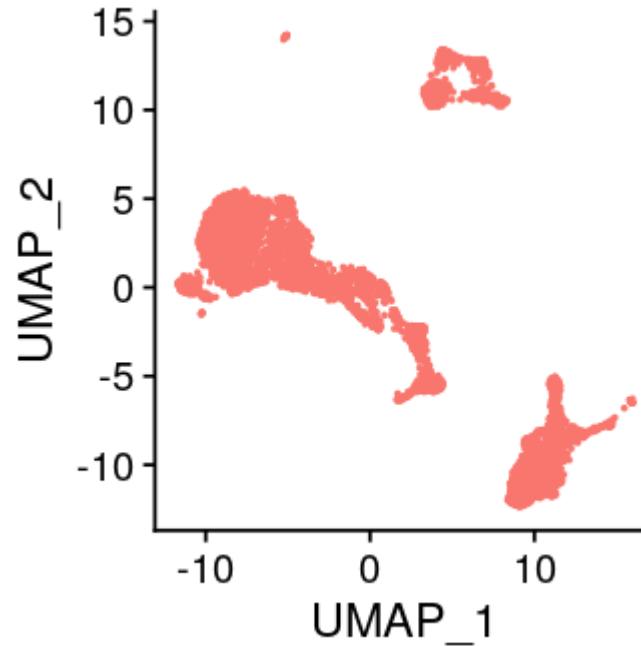
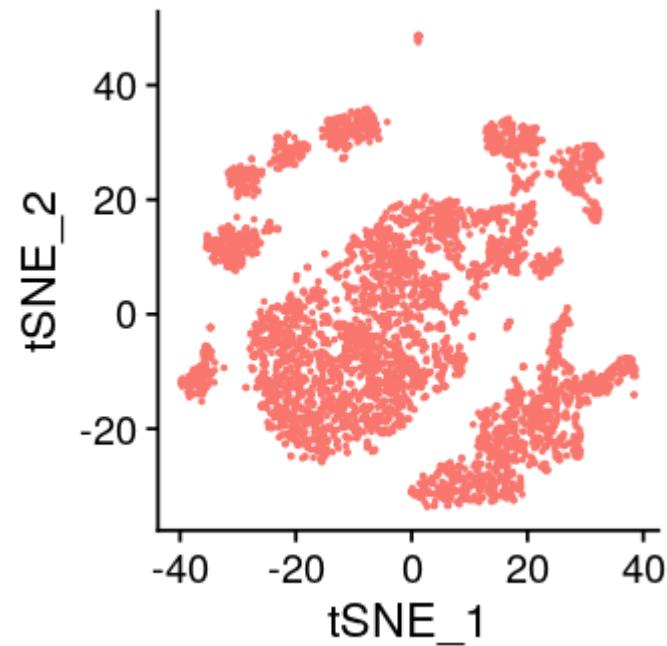
UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that applies to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning.

# UMAP

<https://pair-code.github.io/understanding-umap/>

# Comparing two

```
DimPlot(seurat, reduction = "tsne") + NoLegend()  
DimPlot(seurat, reduction = "umap") + NoLegend()
```



# Basic steps to analysis of scRNA-seq

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

# Clustering and annotation

## Clustering:

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

## Annotation:

- First, check known markers
- For each cluster, perform differential expression: cluster against all others
- Top DE genes expected to be highly distinctive marker genes

# Clustering

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

---

## 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 homogenous 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>.

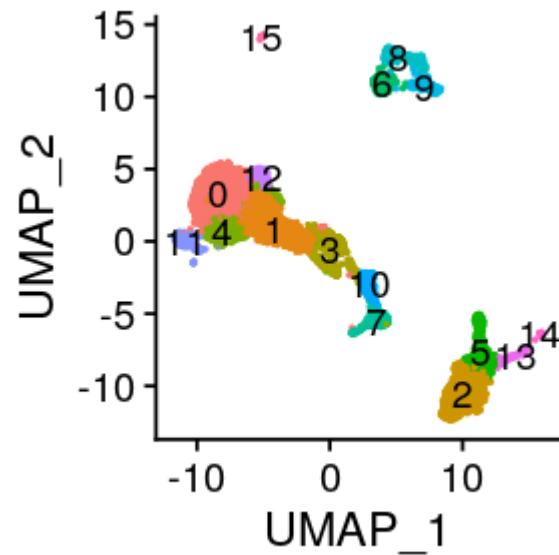
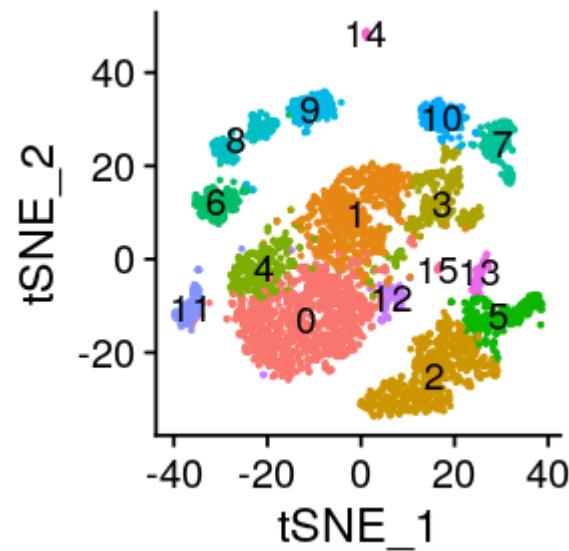
**Contact:** [zcsu@uncc.edu](mailto:zcsu@uncc.edu)

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

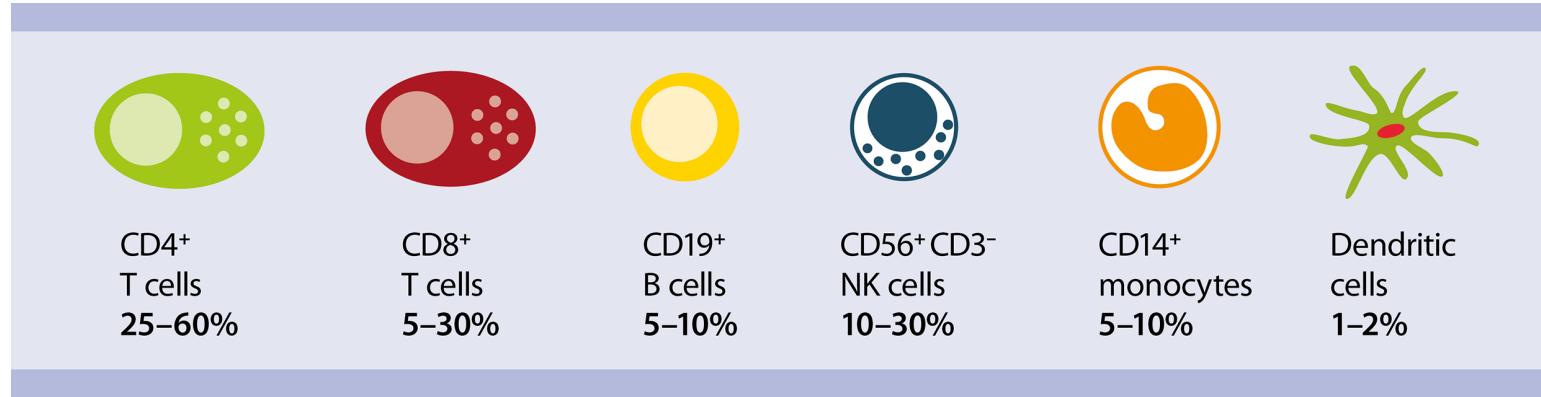
---

# Clustering

```
seurat <- FindNeighbors(seurat, dims = 1:20, verbose = FALSE)
seurat <- FindClusters(seurat, resolution=0.6, verbose = FALSE)
DimPlot(seurat, reduction = "tsne", label = TRUE) + NoLegend()
DimPlot(seurat, reduction = "umap", label = TRUE) + NoLegend()
```



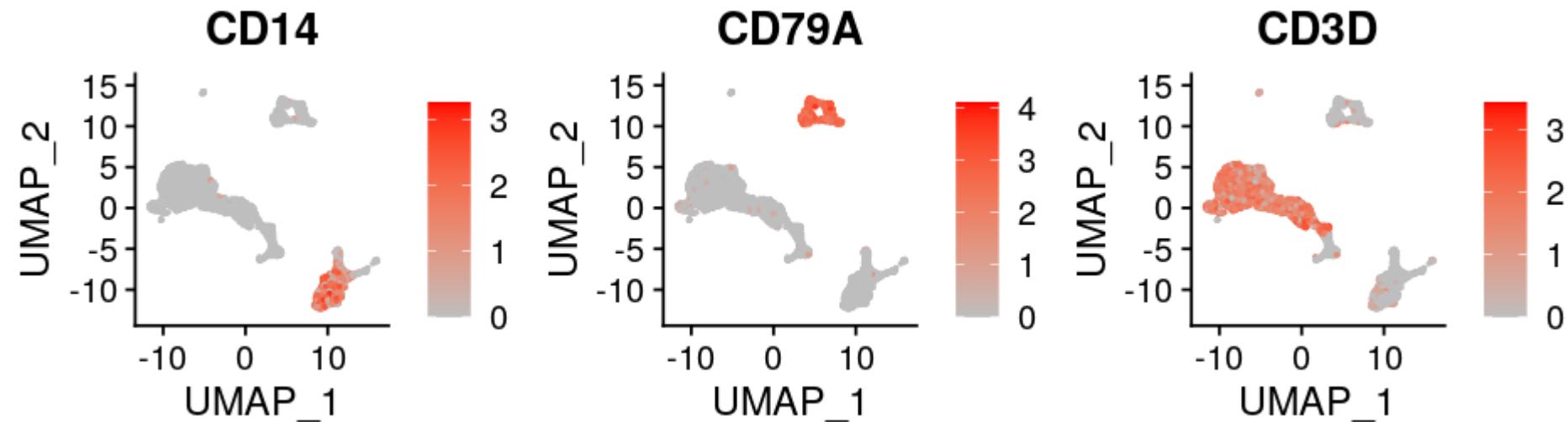
# PBMC



# Annotation

Known markers: CD14, CD79A, CD3D are known markers of Monocytes, B cells and T cells respectively

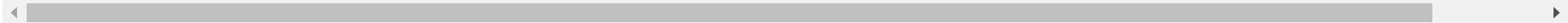
```
FeaturePlot(seurat, c("CD14", "CD79A", "CD3D"), cols=c("grey", "red"), reduction="umap", ncol=3)
```



# Annotation

We can run DE to identify markers automatically using MAST test

```
# max cells per ident is only seed to speed up the whole thing
allMarkers <- FindAllMarkers(seurat, max.cells.per.ident = 100, test.use = "MAST", only.p
goodMarkers <- allMarkers %>% group_by(cluster) %>% top_n(n = 1, wt = avg_logFC) %>% pull
goodMarkers
```

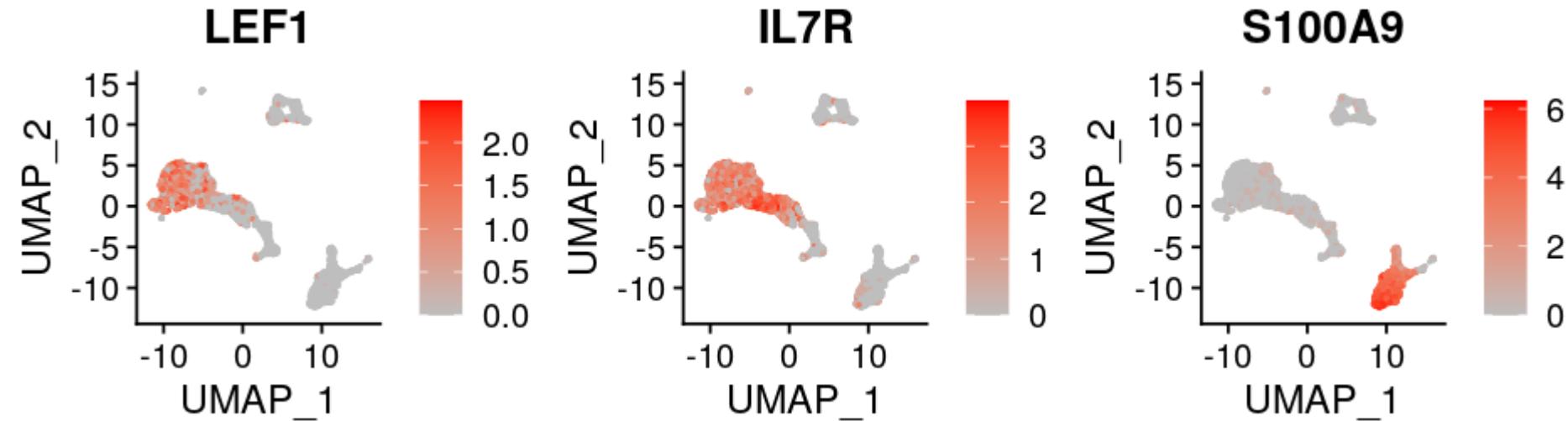


```
## [1] "LEF1"      "IL7R"      "S100A9"     "CCL5"      "CCR7"      "AIF1"
## [7] "IGKC"      "GNLY"      "IGLC2"      "IGKC"      "GNLY"      "CD8B"
## [13] "INTS6"     "HLA-DQA1"  "TCF4"      "PPBP"
```

# Ways to show expression

We usually either show expression **on top of reduction plot**, or show a violin plot for expression

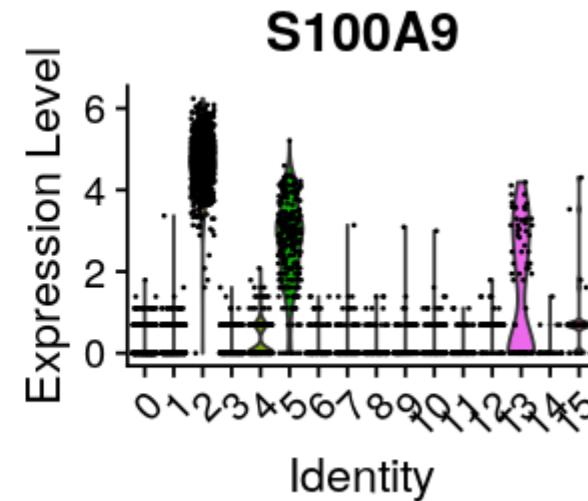
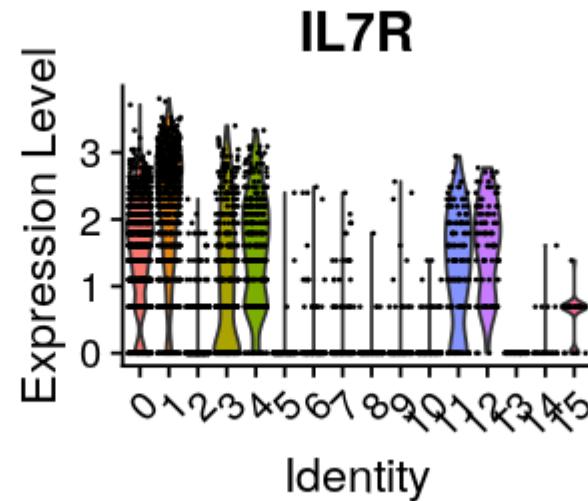
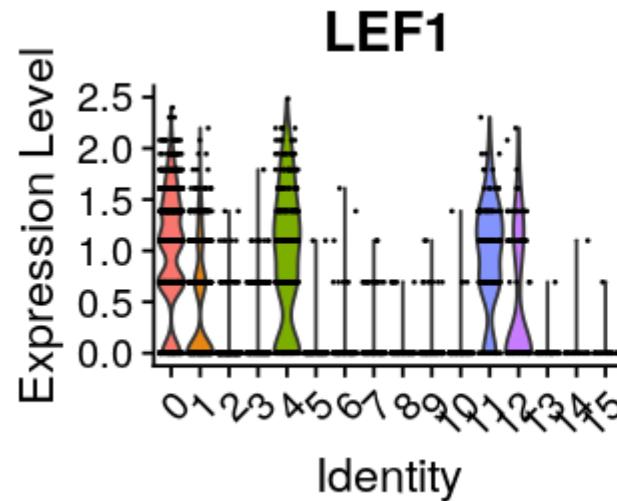
```
FeaturePlot(seurat, goodMarkers[1:3], cols=c("grey", "red"), reduction="umap", ncol=3)
```



# Ways to show expression

We usually either show expression of top of reduction plot, or show a **violin plot** for expression

```
VlnPlot(seurat, goodMarkers[1:3], pt.size = 0.1)
```



# Save the Seurat object for later

```
saveRDS(seurat, file="blood_seurat.rds")
```

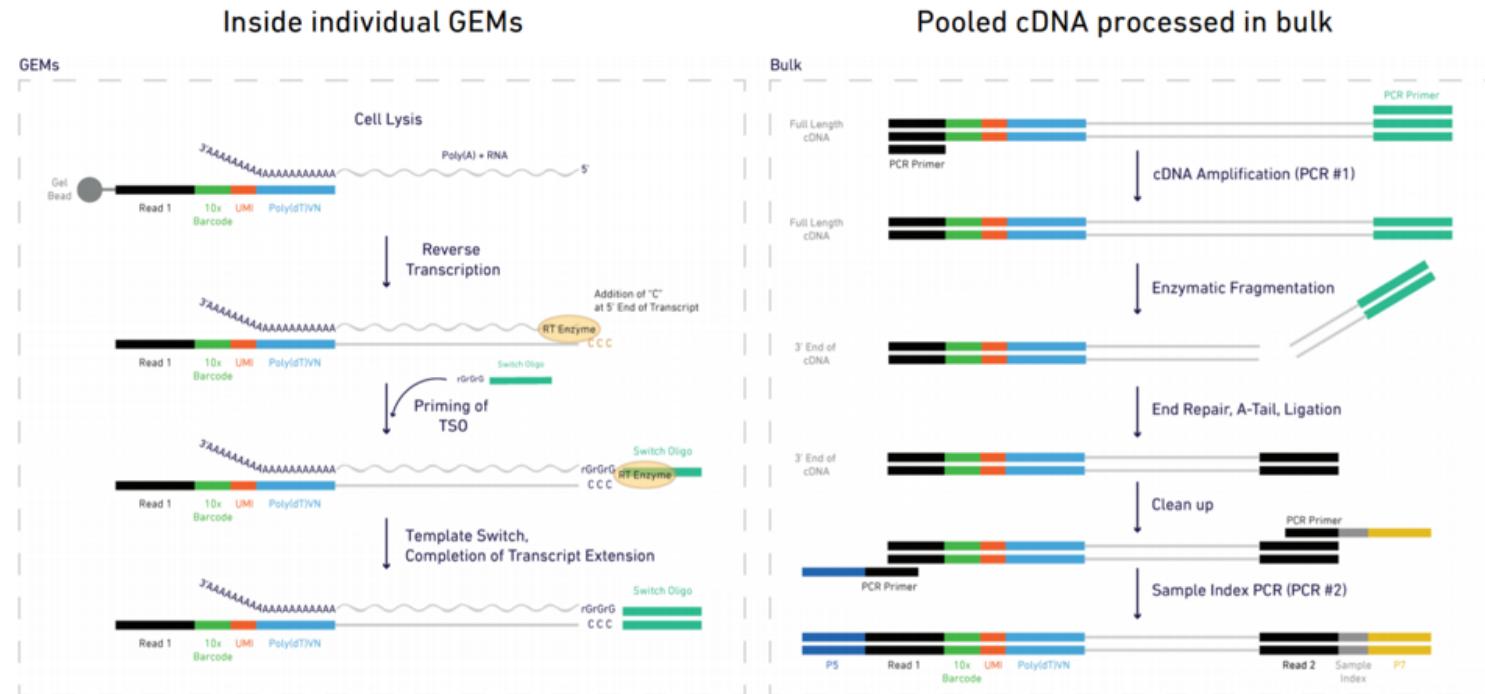
# Communication is important

- 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 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

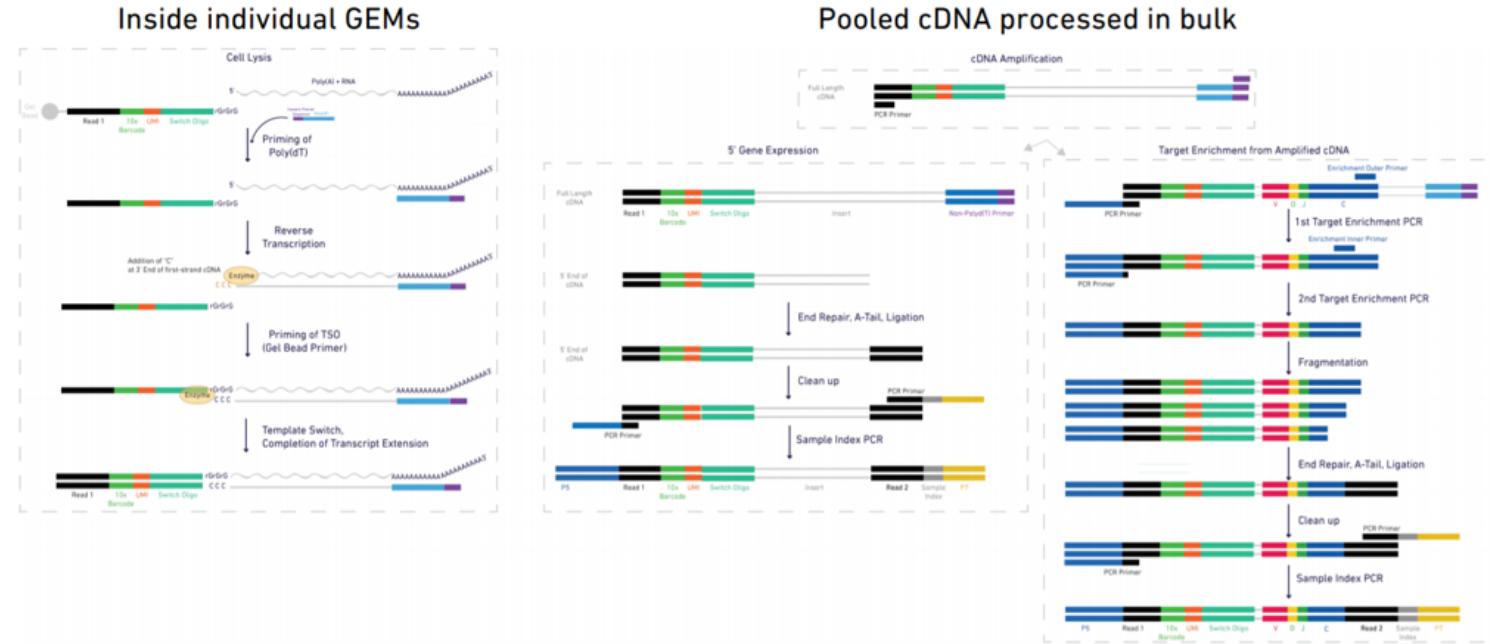
# Latest things: 5'

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

# 3' sequencing



# 5' sequencing



# Surface protein expression

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

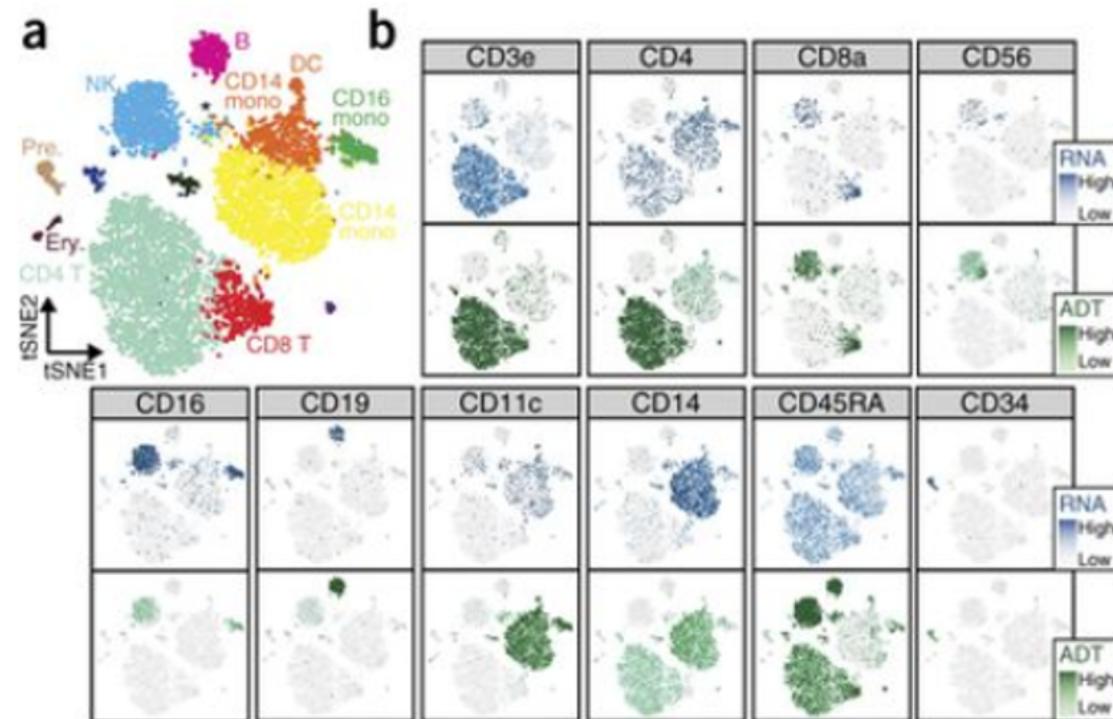
Vanessa M Peterson<sup>1,5</sup>, Kelvin Xi Zhang<sup>2,5</sup>,  
Namit Kumar<sup>1</sup>, Jerelyn Wong<sup>3</sup>, Lixia Li<sup>1</sup>, Douglas  
C Wilson<sup>3</sup>, Renee Moore<sup>4</sup>, Terrill K McClanahan<sup>3</sup>,  
Svetlana Sadekova<sup>3</sup> & Joel A Klappanbach<sup>1</sup>

rights reserved.

the standard 10x Genomics single-cell (sc)RNA-seq platform<sup>3</sup>, 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

# Surface protein expression



# Questions?