ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Introduction into single cell RNA-seq

Konstantin Zaitsev

August 27th, 2021. Tomsk / My hotel room

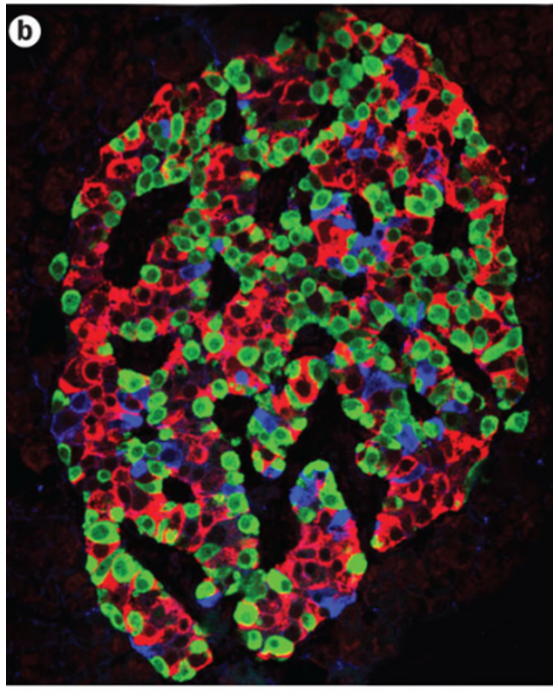
Single-cell RNA-seq (scRNA-seq)

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[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).

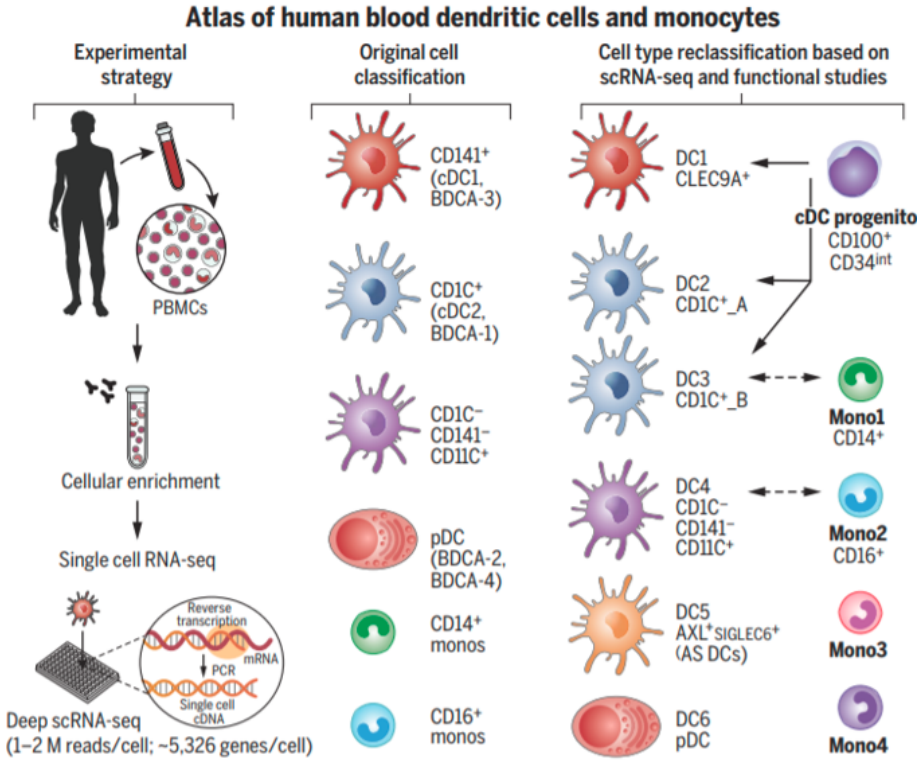
Published online 2 January 2014; doi:10.1038/nprot.2014.006

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 singl 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; sequencin 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.

Paper: https://www.nature.com/articles/nprot.2014.006

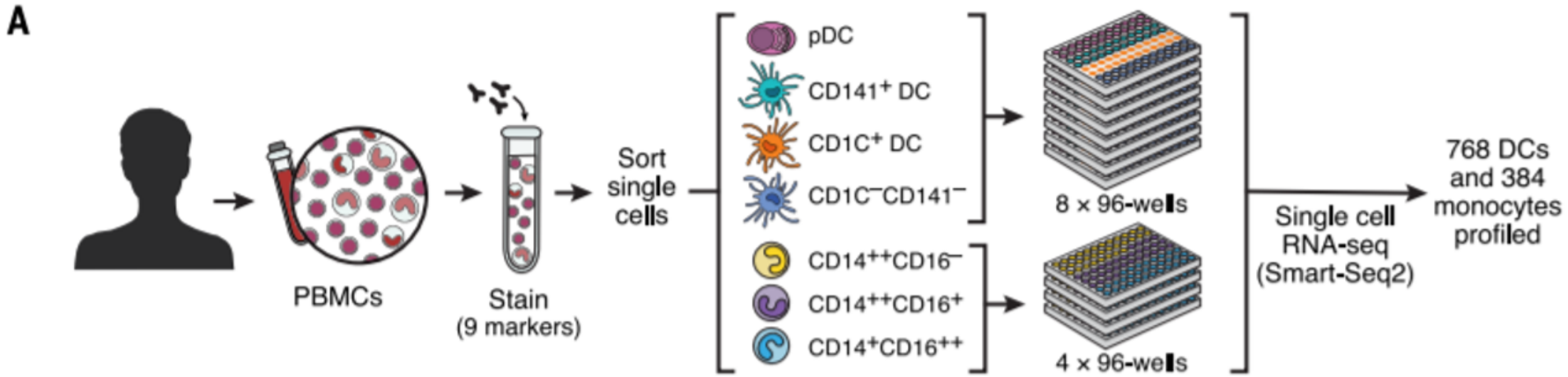
Single-cell RNA-seq of myeloid cells

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



Paper: https://science.sciencemag.org/content/356/6335/eaah4573

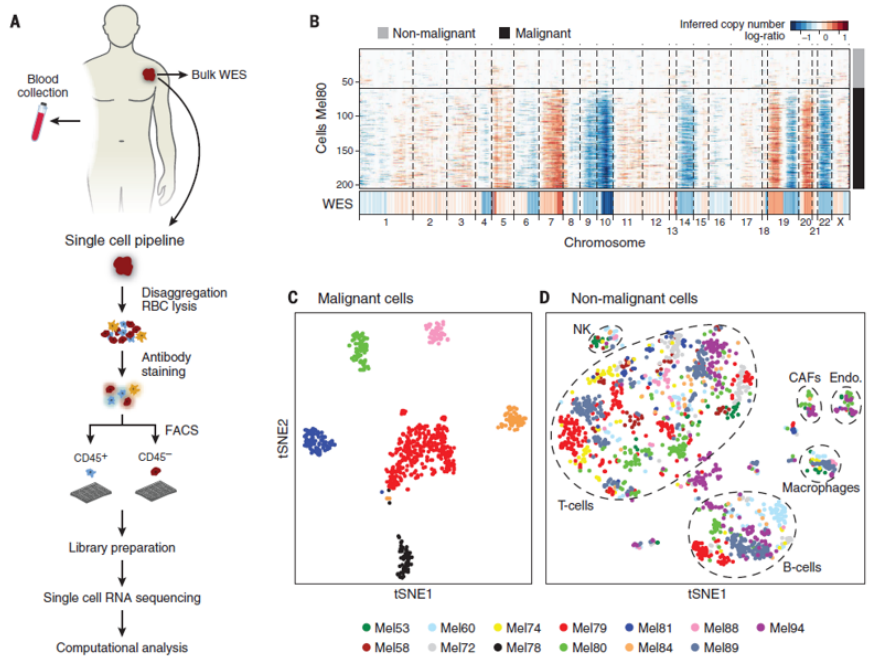
Single-cell RNA-seq of myeloid cells



Paper: https://science.sciencemag.org/content/356/6335/eaah4573

Single-cell RNA-seq of melanoma

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



Paper: https://science.sciencemag.org/content/352/6282/189.long

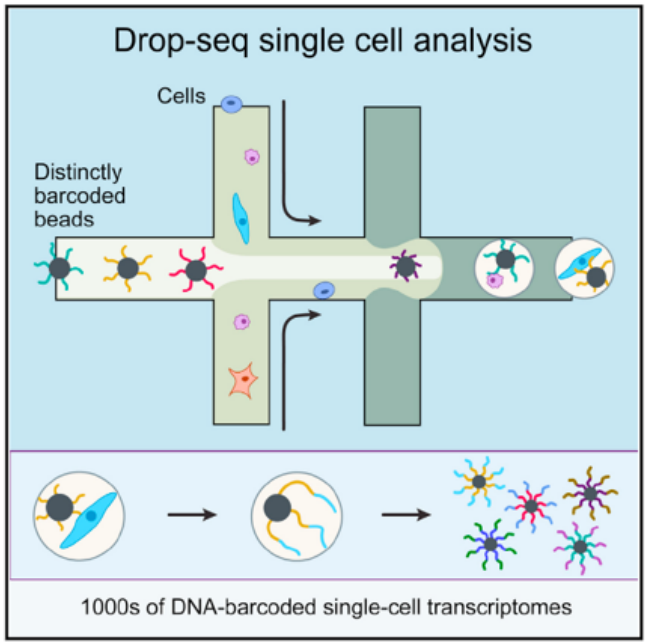
Drop-seq: Cell, 2015

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.

Paper: https://www.cell.com/abstract/S0092-8674(15)00549-8

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

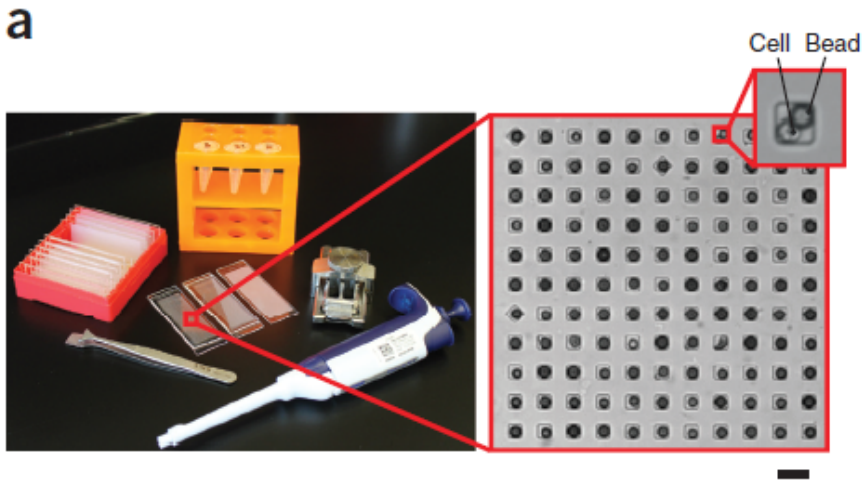
Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1],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

Paper: https://www.nature.com/articles/ncomms14049

Seq-Well: NMeth, 2017

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]



Paper: https://www.nature.com/articles/nmeth.4179

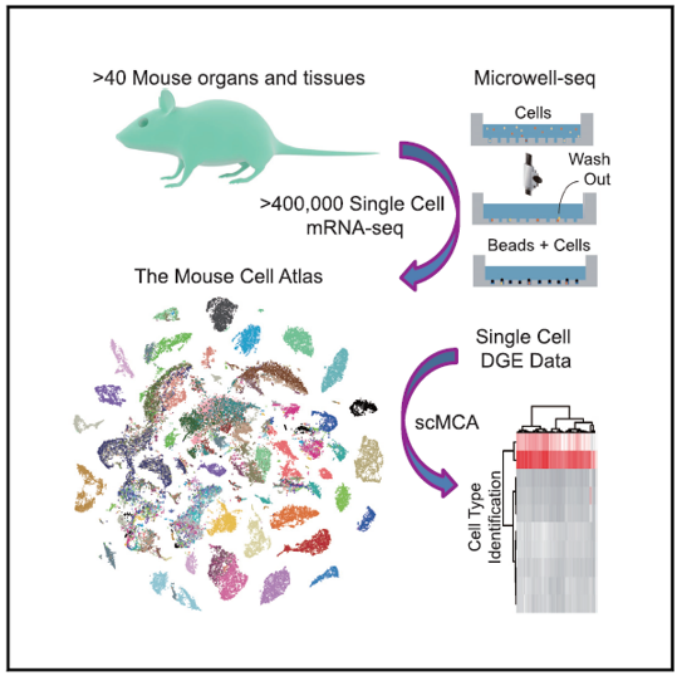
Microwell-Seq: Cell, 2018

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

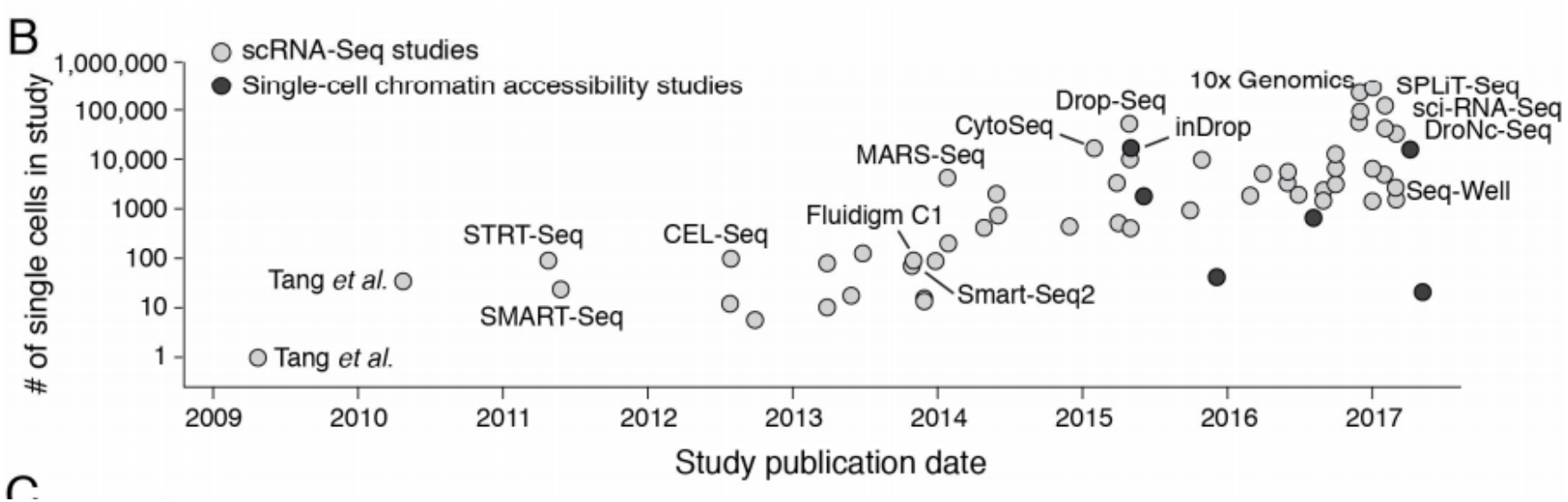
xhan@zju.edu.cn (X.H.), ggj@zju.edu.cn (G.G.)

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.

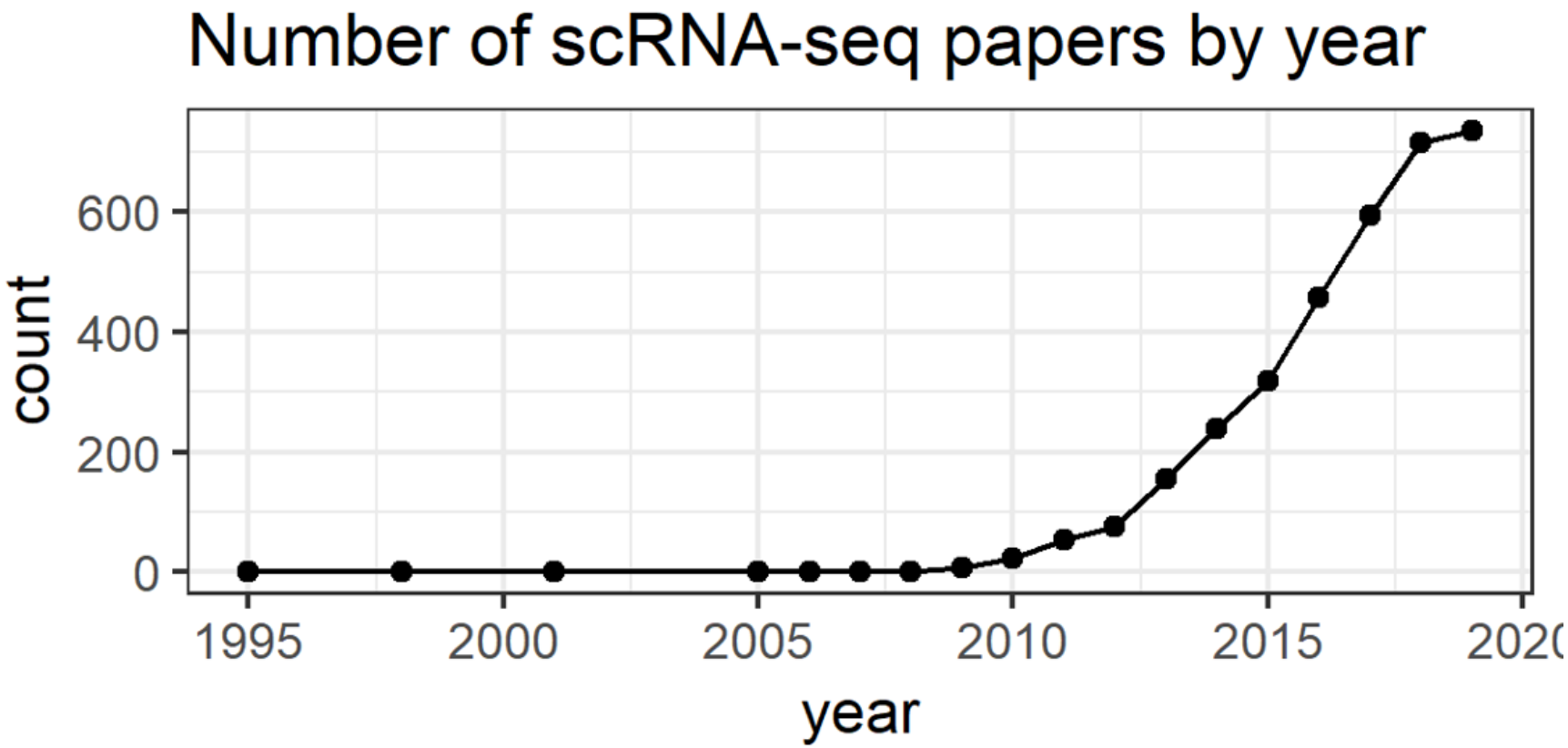
Paper: https://www.ncbi.nlm.nih.gov/pubmed/29474909

Growth of single-cell technologies



Human cell atlas white paper: https://arxiv.org/pdf/1810.05192.pdf

Growth of single-cell technologies



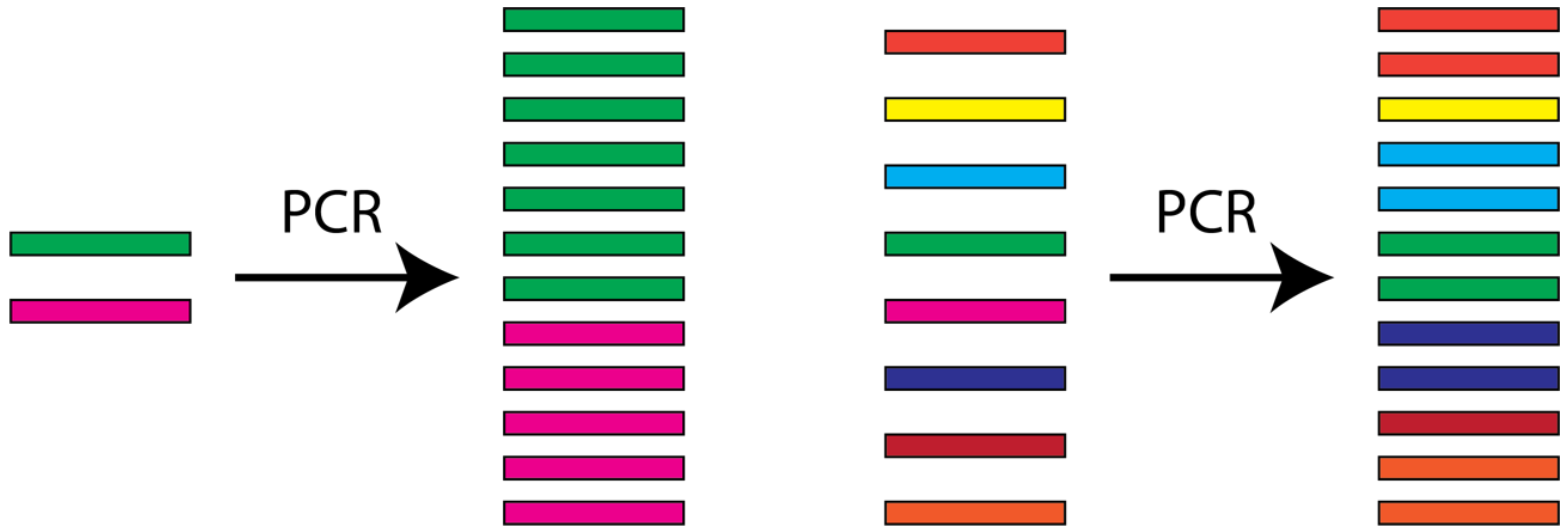
Based on pubmed results

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?

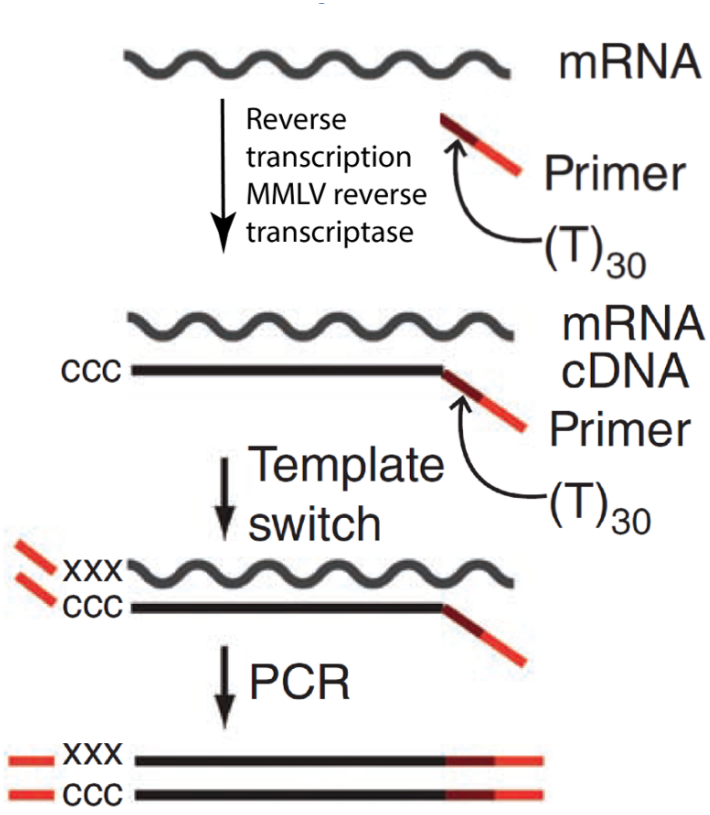
Low library complexity

- 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

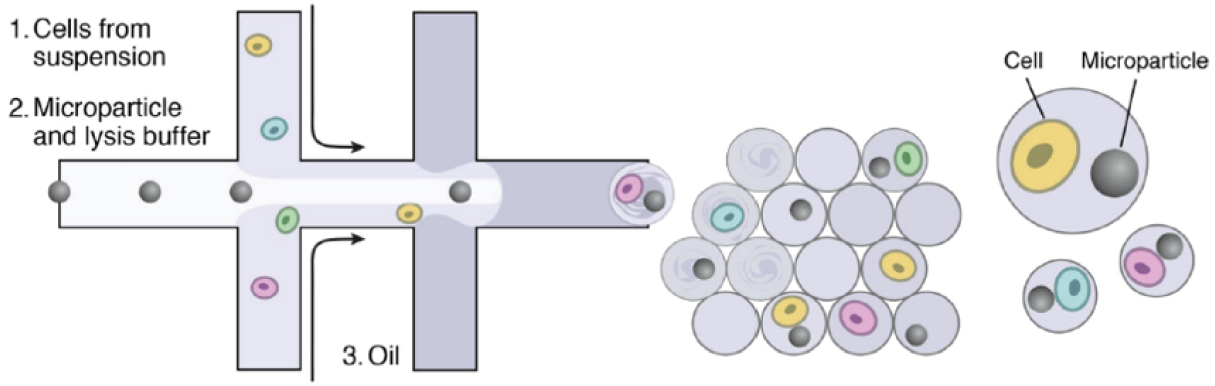


Template-switching PCR

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

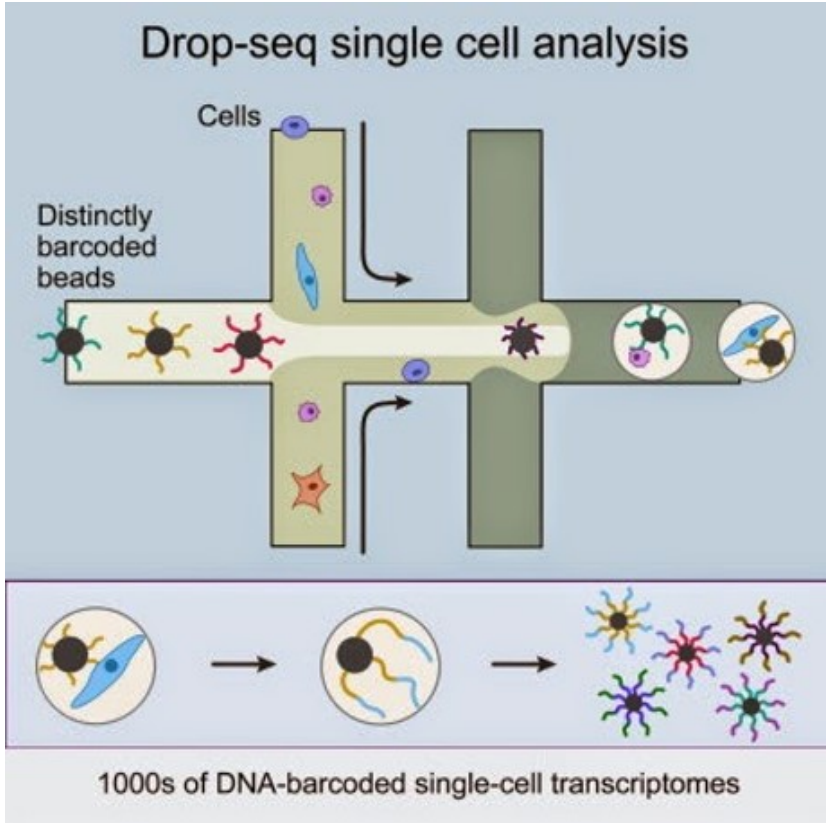


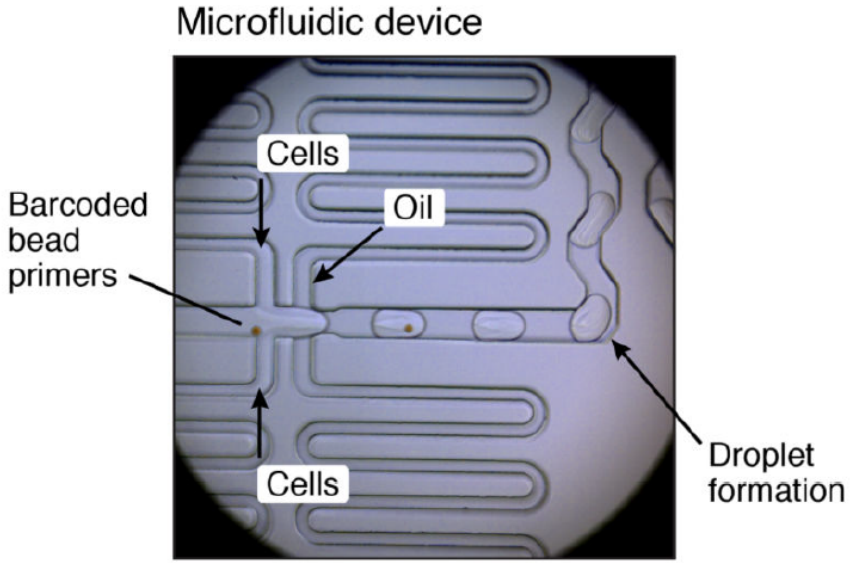
Drop-seq schematics



Paper: https://www.cell.com/abstract/S0092-8674(15)00549-8

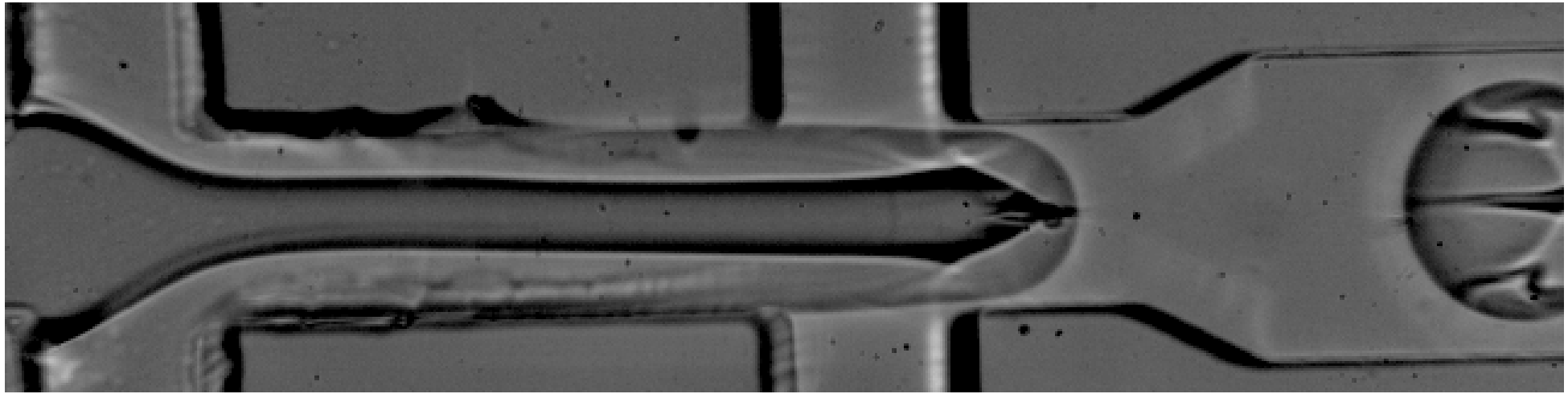
Drop-seq microfluidics



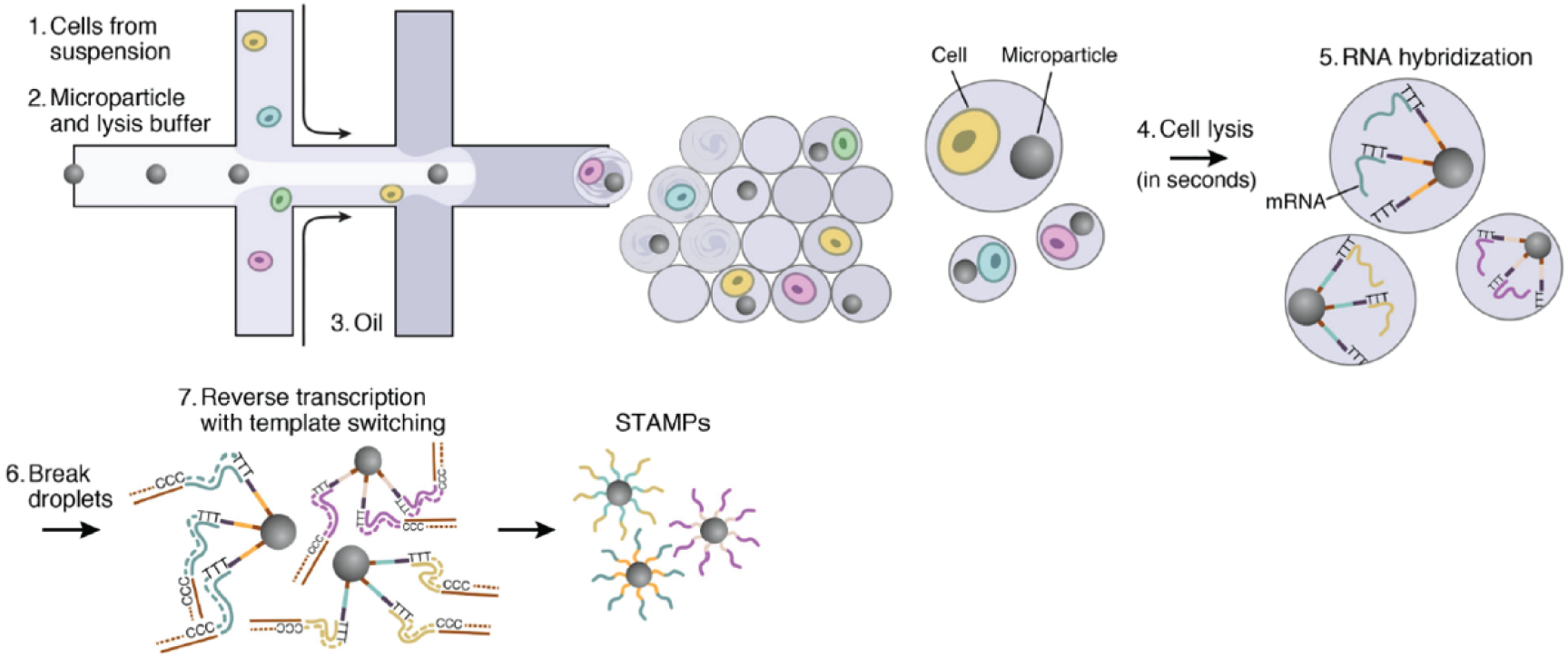


Paper: https://www.cell.com/abstract/S0092-8674(15)00549-8

Drop-seq microfluidics

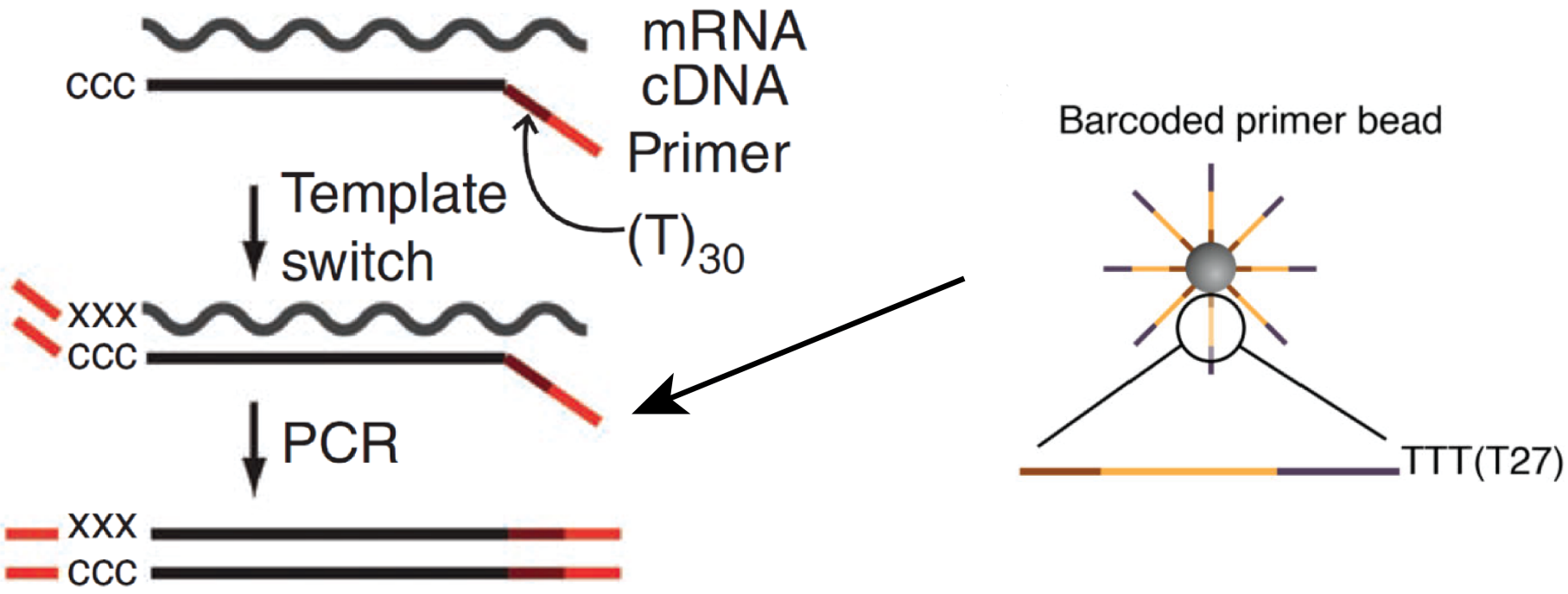


Drop-seq schematics

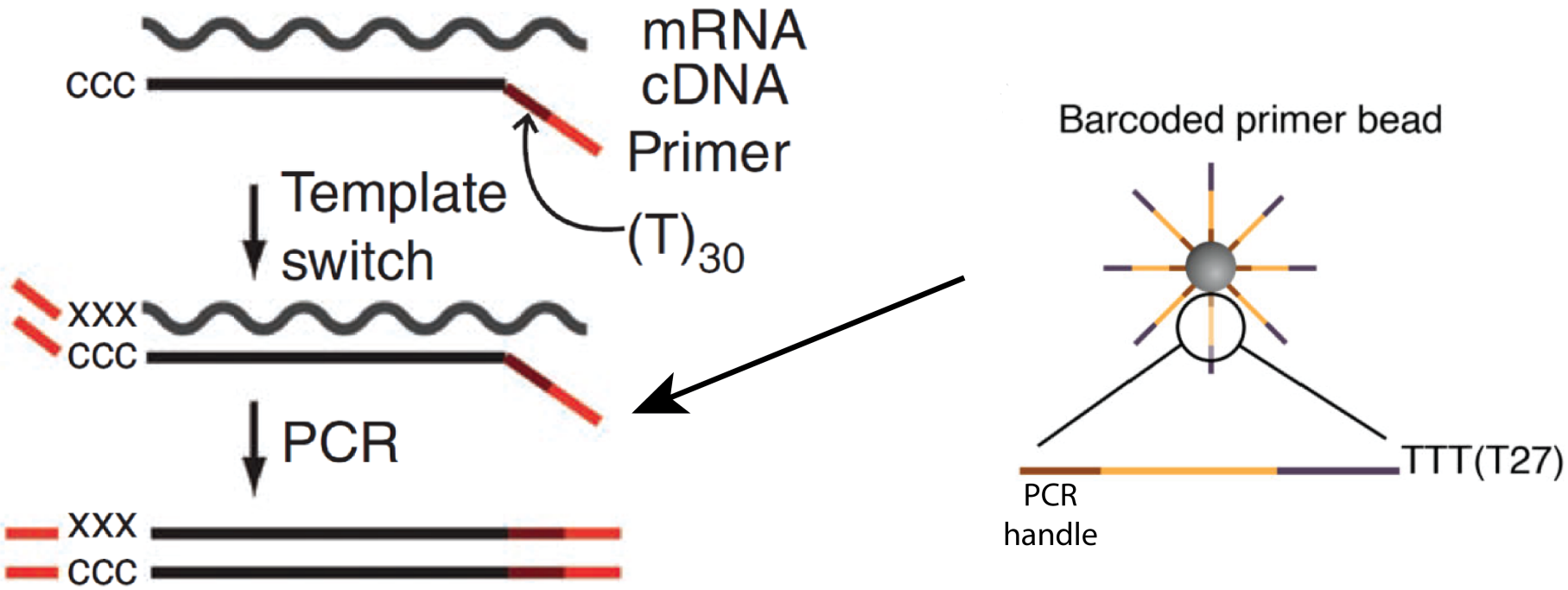


Paper: https://www.cell.com/abstract/S0092-8674(15)00549-8

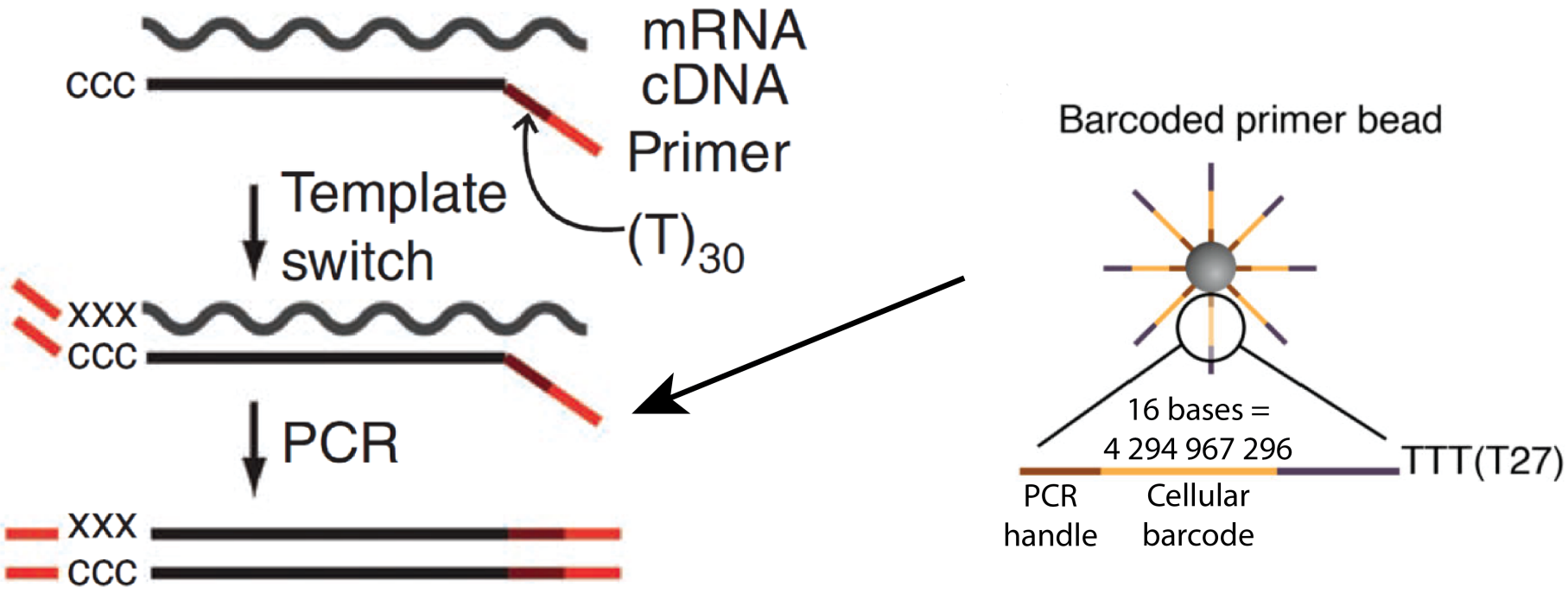
Barcoding


mRNA
cDNA
ccc
Primer
Template
switch
$(T)_{30}$
xxx
ccc
PCR
xxx
ccc
Barcoded primer bead
TTT(T27)

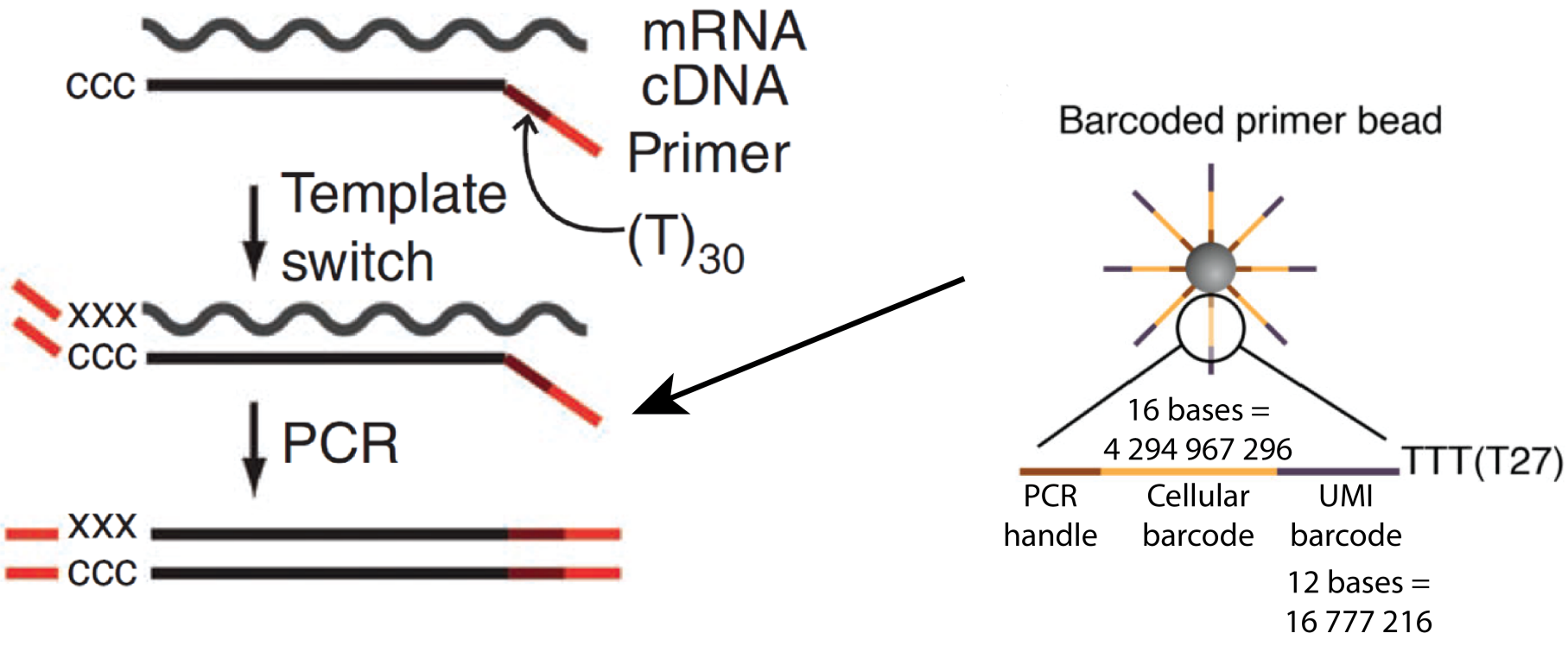
Barcoding


mRNA
cDNA
ccc
Primer
Template
switch
$(T)_{30}$
xxx
ccc
PCR
xxx
ccc
Barcoded primer bead
TTT(T27)
PCR
handle

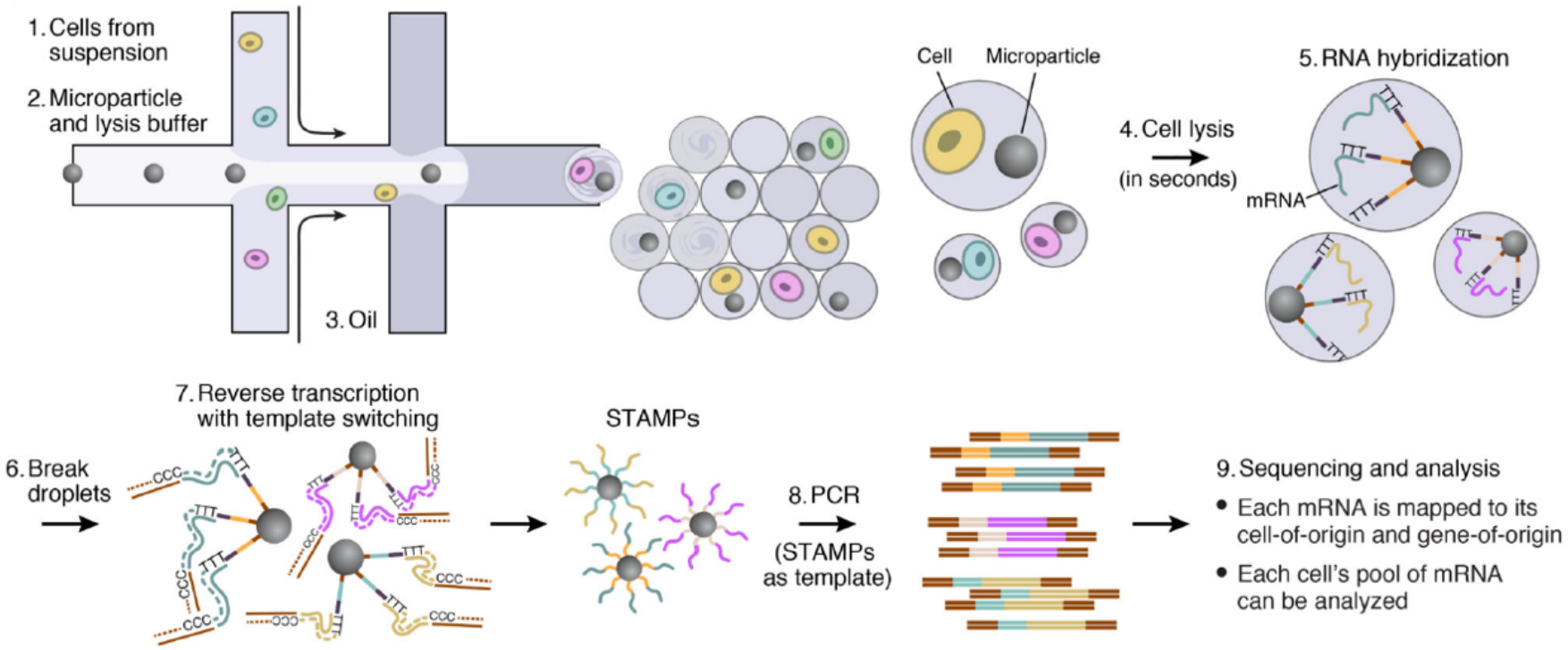
Barcoding


mRNA
cDNA
ccc
Primer
Template
switch
$(T)_{30}$
xxx
ccc
PCR
xxx
ccc
Barcoded primer bead
16 bases =
4 294 967 296
TTT(T27)
PCR
handle
Cellular
barcode

Barcoding

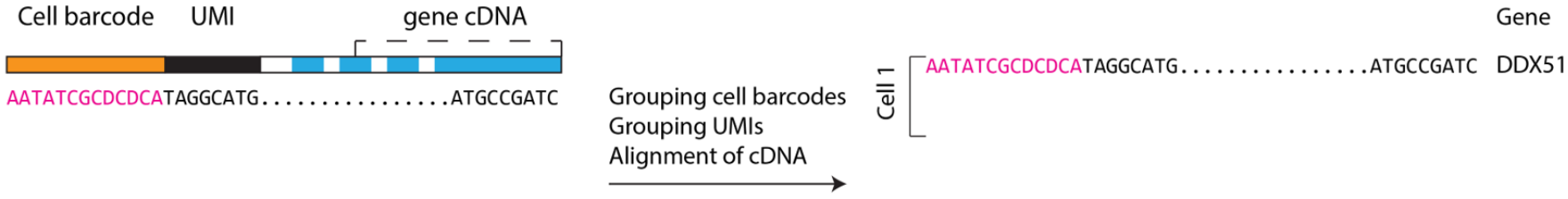

mRNA
cDNA
Primer
ccc
Template
switch
$(T)_{30}$
xxx
ccc
PCR
xxx
ccc
Barcoded primer bead
16 bases =
4 294 967 296
TTT(T27)
PCR
handle
Cellular
barcode
UMI
barcode
12 bases =
16 777 216

Drop-seq schematics

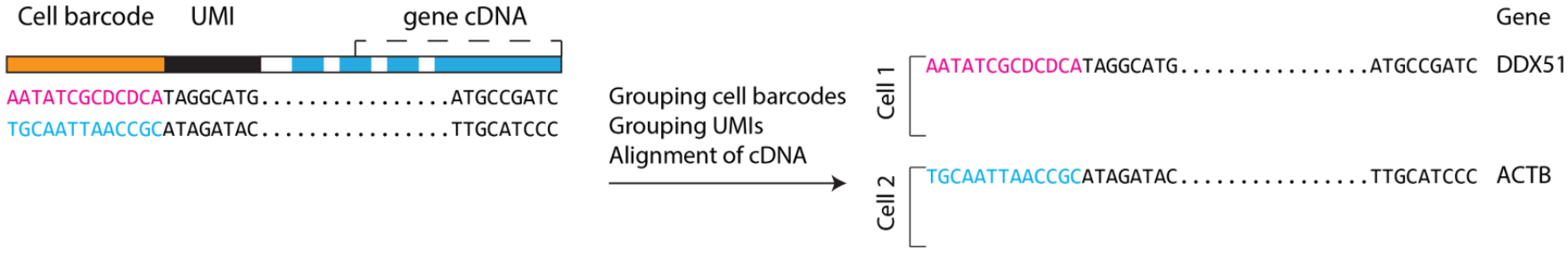


Paper: https://www.cell.com/abstract/S0092-8674(15)00549-8

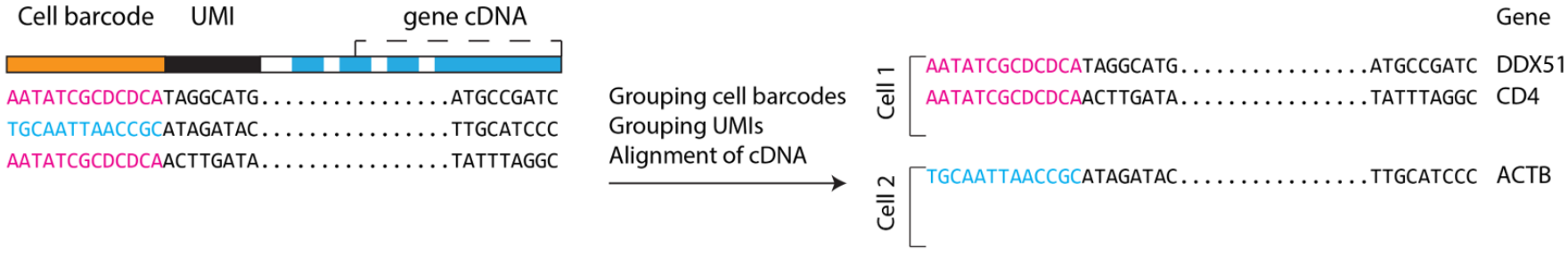
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC
Gene
DDX51

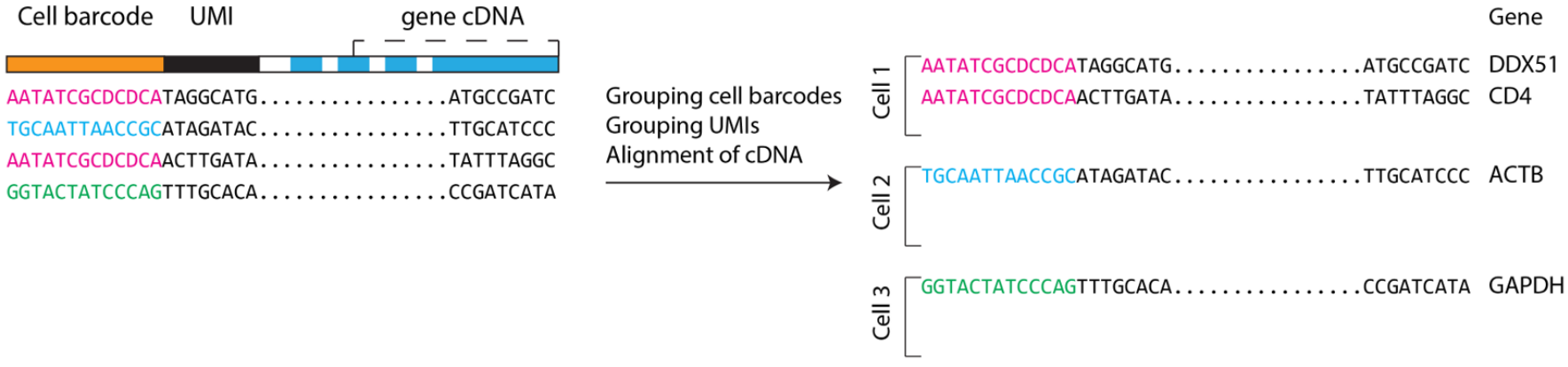
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Gene
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC
DDX51
Cell 2
TGCAATTAACCGCATAGATAC.....TTGCATCCC
ACTB

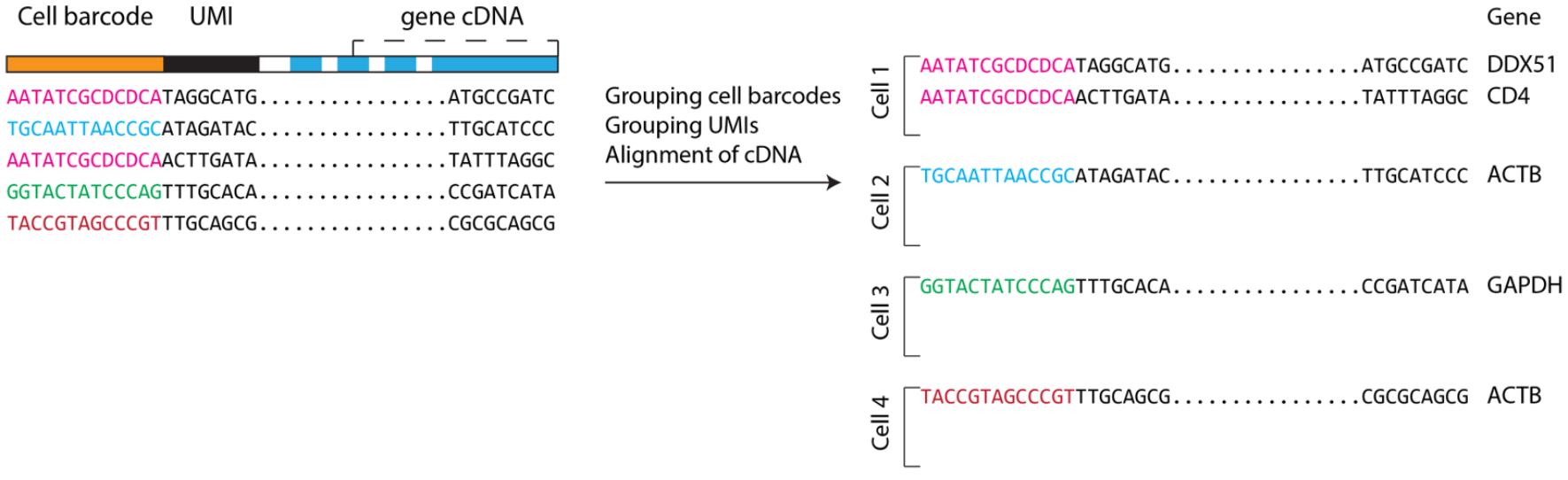
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
AATATCGCDCDCAACTTGATA.....TATTTAGGC
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Gene
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC DDX51
AATATCGCDCDCAACTTGATA.....TATTTAGGC CD4
Cell 2
TGCAATTAACCGCATAGATAC.....TTGCATCCC ACTB

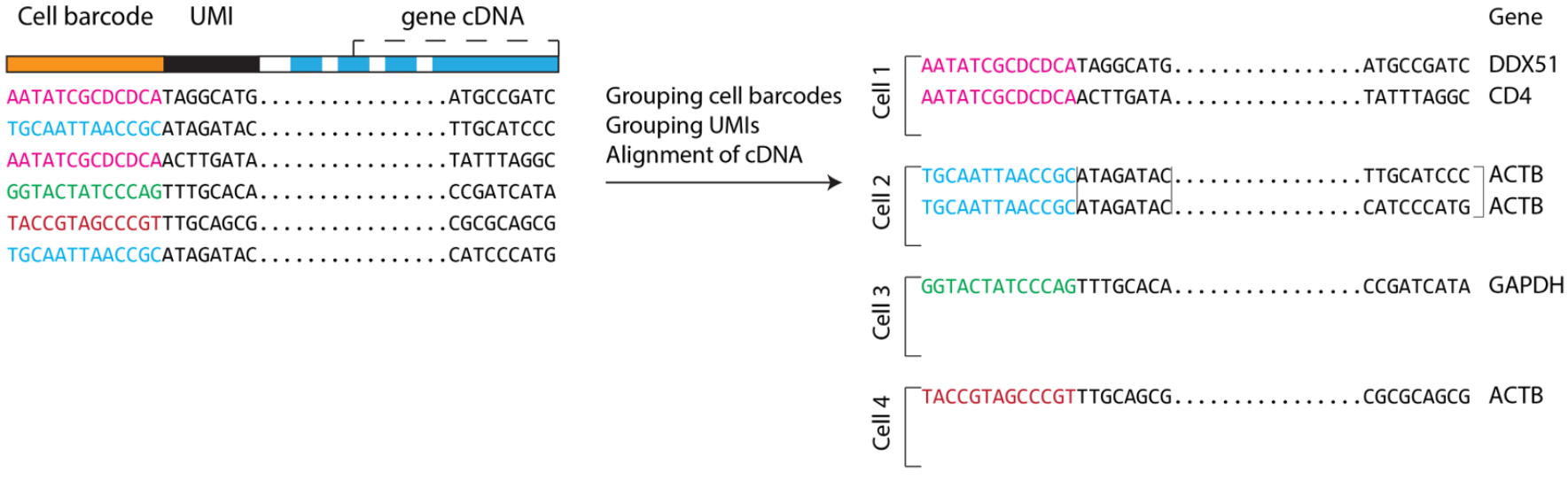
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
AATATCGCDCDCAACTTGATA.....TATTTAGGC
GGTACTATCCCAGTTTGCACA.....CCGATCATA
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Gene
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC DDX51
AATATCGCDCDCAACTTGATA.....TATTTAGGC CD4
Cell 2
TGCAATTAACCGCATAGATAC.....TTGCATCCC ACTB
Cell 3
GGTACTATCCCAGTTTGCACA.....CCGATCATA GAPDH

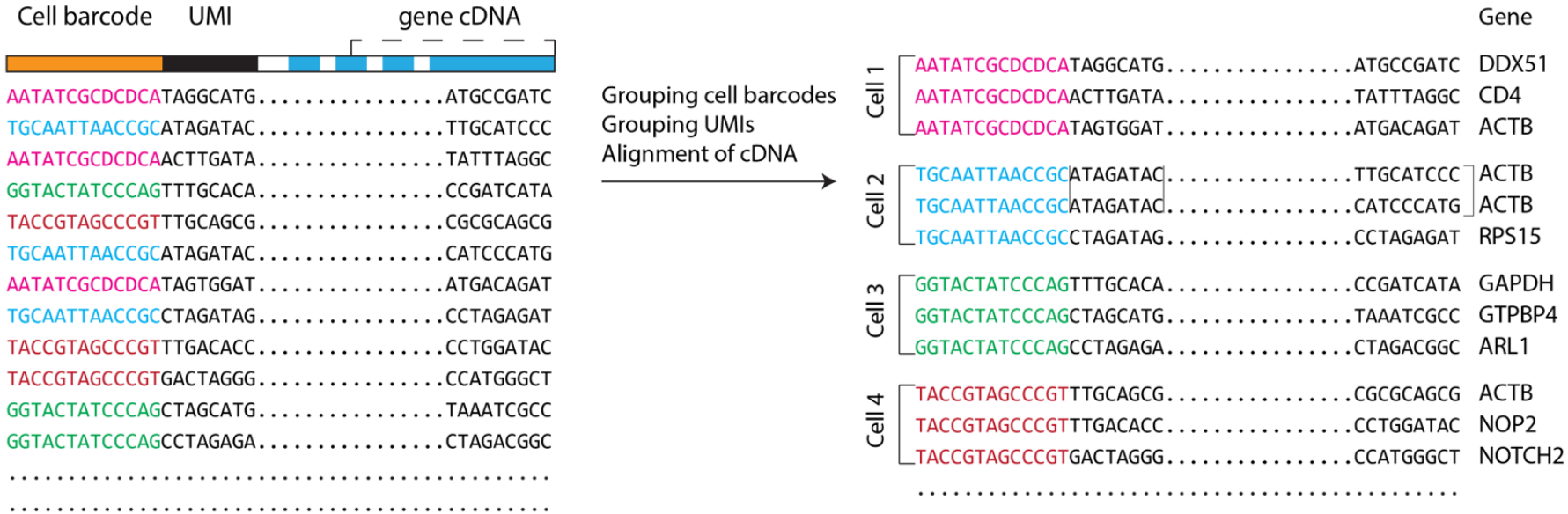
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
AATATCGCDCDCAACTTGATA.....TATTTAGGC
GGTACTATCCCAGTTTGCACA.....CCGATCATA
TACCGTAGCCCGTTTGCAGCG.....CGCGCAGCG
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Gene
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC DDX51
AATATCGCDCDCAACTTGATA.....TATTTAGGC CD4
Cell 2
TGCAATTAACCGCATAGATAC.....TTGCATCCC ACTB
Cell 3
GGTACTATCCCAGTTTGCACA.....CCGATCATA GAPDH
Cell 4
TACCGTAGCCCGTTTGCAGCG.....CGCGCAGCG ACTB

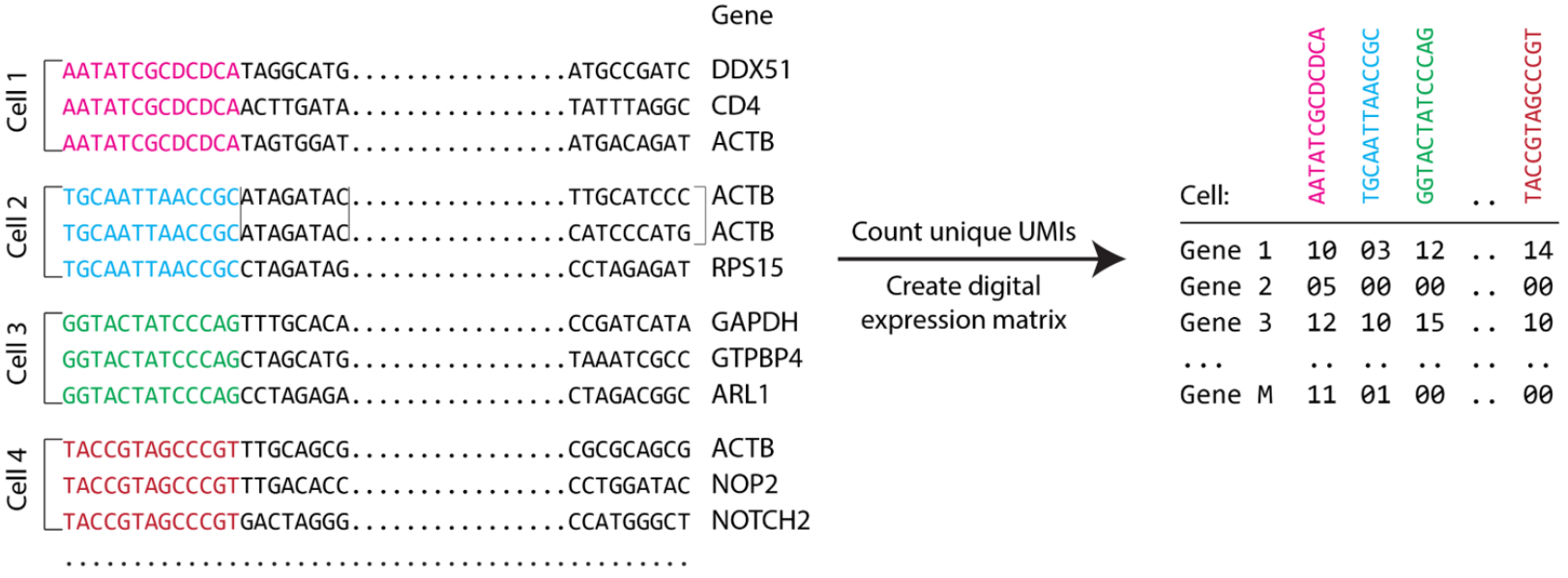
Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
AATATCGCDCDCAACTTGATA.....TATTTAGGC
GGTACTATCCCAGTTTGCACA.....CCGATCATA
TACCGTAGCCCGTTTGCAGCG.....CGCGCAGCG
TGCAATTAACCGCATAGATAC.....CATCCCATG
Grouping cell barcodes
Grouping UMIs
Alignment of cDNA
Gene
Cell 1
AATATCGCDCDCATAGGCATG.....ATGCCGATC DDX51
AATATCGCDCDCAACTTGATA.....TATTTAGGC CD4
Cell 2
TGCAATTAACCGCATAGATAC.....TTGCATCCC ACTB
TGCAATTAACCGCATAGATAC.....CATCCCATG ACTB
Cell 3
GGTACTATCCCAGTTTGCACA.....CCGATCATA GAPDH
Cell 4
TACCGTAGCCCGTTTGCAGCG.....CGCGCAGCG ACTB

Sequencing


Cell barcode
UMI
gene cDNA
AATATCGCDCDCATAGGCATG.....ATGCCGATC
TGCAATTAACCGCATAGATAC.....TTGCATCCC
AATATCGCDCDCAACTTGATA.....TATTTAGGC
GGTACTATCCCAGTTTGCACA.....CCGATCATA
TACCGTAGCCCGTTTGCAGCG.....CGCGCAGCG
TGCAATTAACCGCATAGATAC.....CATCCCATG
AATATCGCDCDCATAGTGGAT.....ATGACAGAT
TGCAATTAACCGCCTAGATAG.....CCTAGAGAT
TACCGTAGCCCGTTTGACACC.....CCTGGATAC
TACCGTAGCCCGTGACTAGGG.....CCATGGGCT
GGTACTATCCCAGCTAGCATG.....TAAATCGCC
GGTACTATCCCAGCCTAGAGA.....CTAGACGGC
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: AATATCGCDCDCDCA 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

All questions were addressed

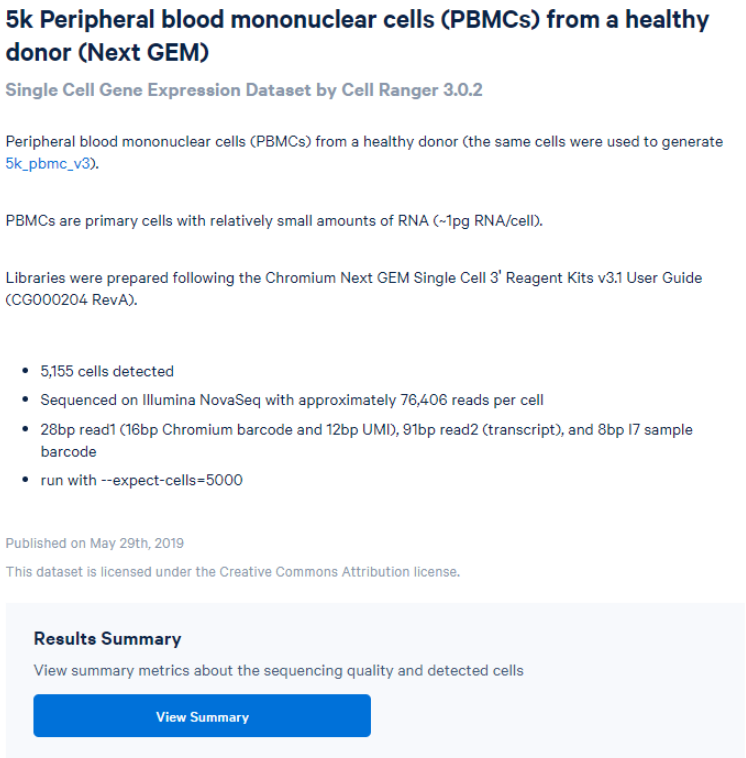
- Low input mRNA – template switching PCR
- Cell isolation – microfluidics
- Read identification – Cell barcodes
- Dealing with PCR duplicates – Cell/UMI barcodes

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:

10x genomics 5k cells, peripheral blood



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 29th, 2019

This dataset is licensed under the Creative Commons Attribution license.

Results Summary

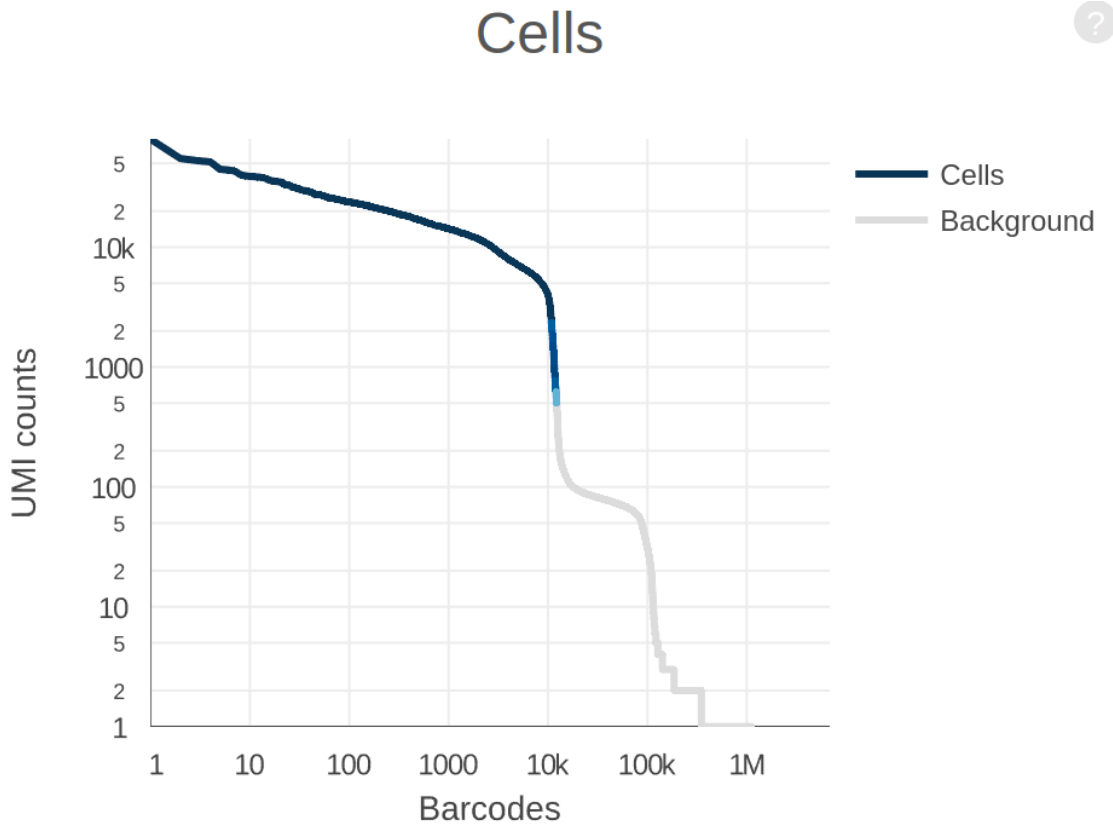
View summary metrics about the sequencing quality and detected cells

View Summary

Let's have a loot at the summary

- Summary for the same dataset

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:

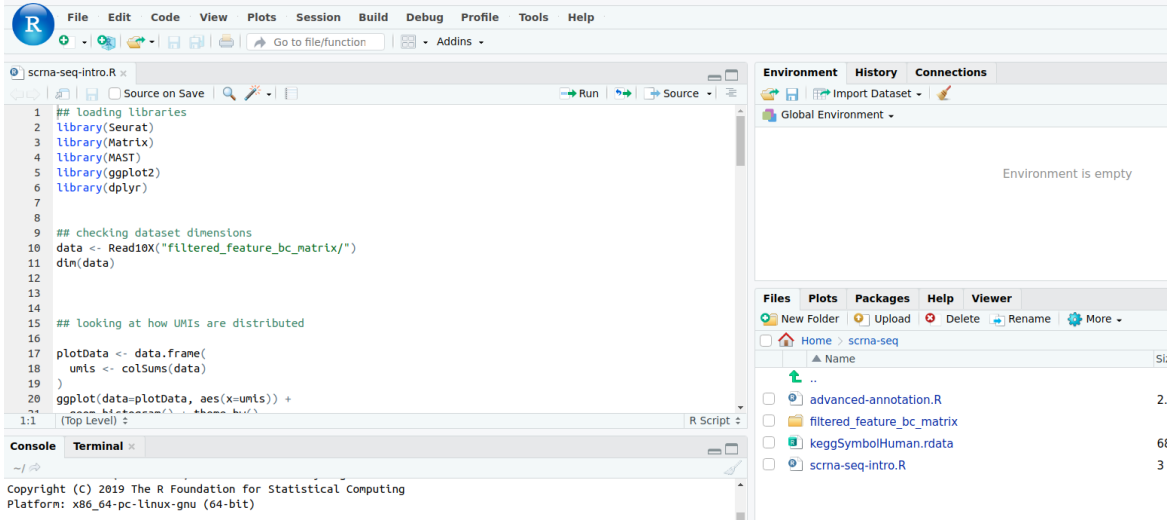
- We can set a hard threshold on UMI to filter noise from actual cells
- We can calculate noise signature and test every cell against this signature (emptyDrops, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y)

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

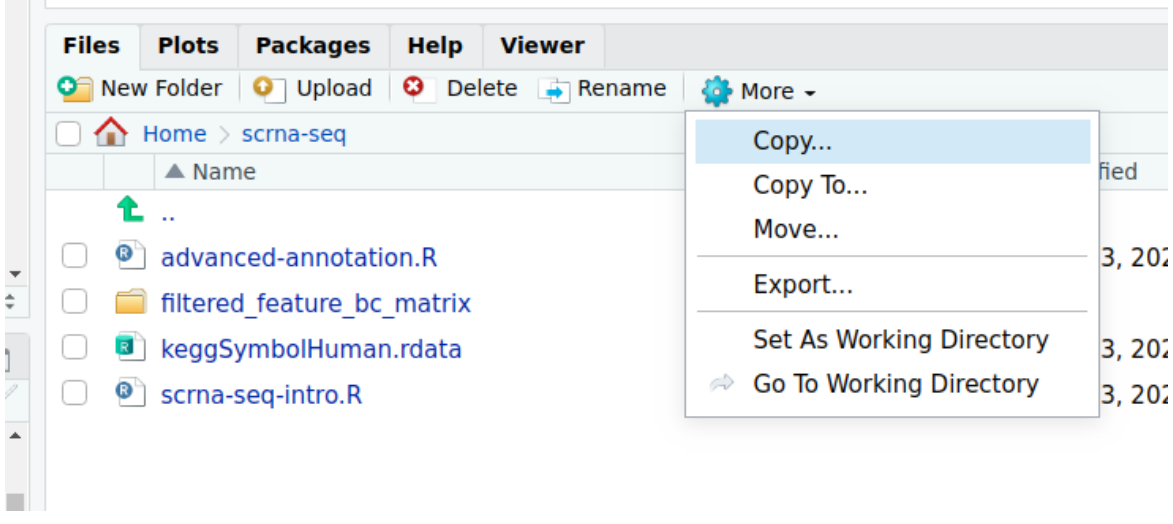
Our setup

- Your logins are the same as yesterday:
  - login: student
  - password: sysbiopass
- 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 librarires

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

Loading the data

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

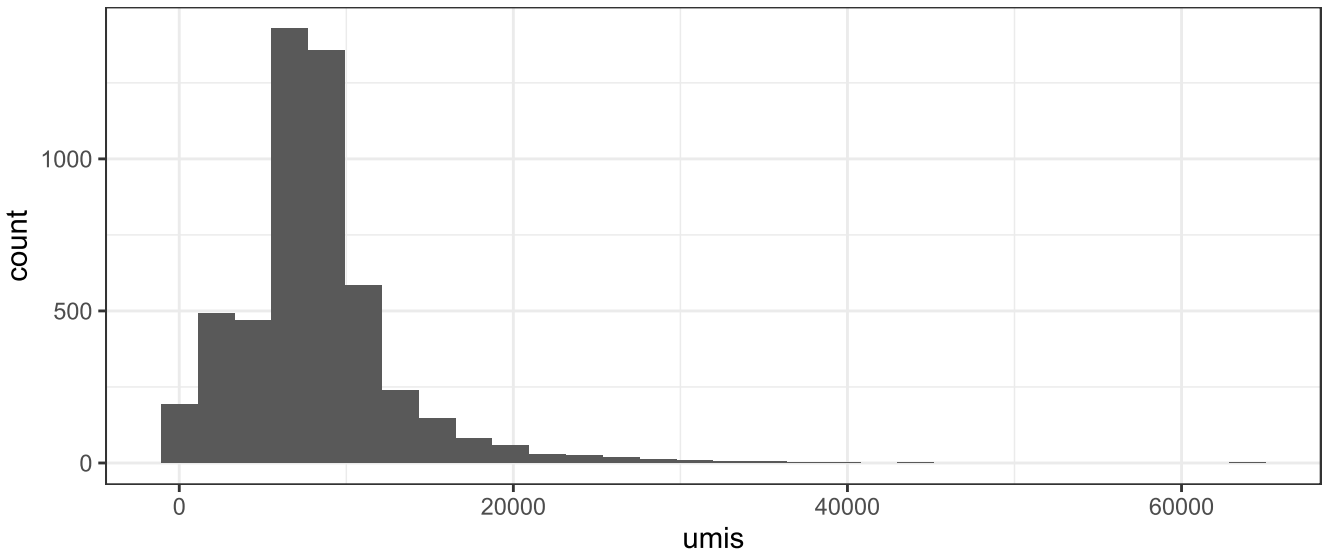
```
## [1] 33538  5155
```

Count matrix is large:

- 33538 genes
- 5155 cells

UMI distribution

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



Filtering (genes and barcodes)

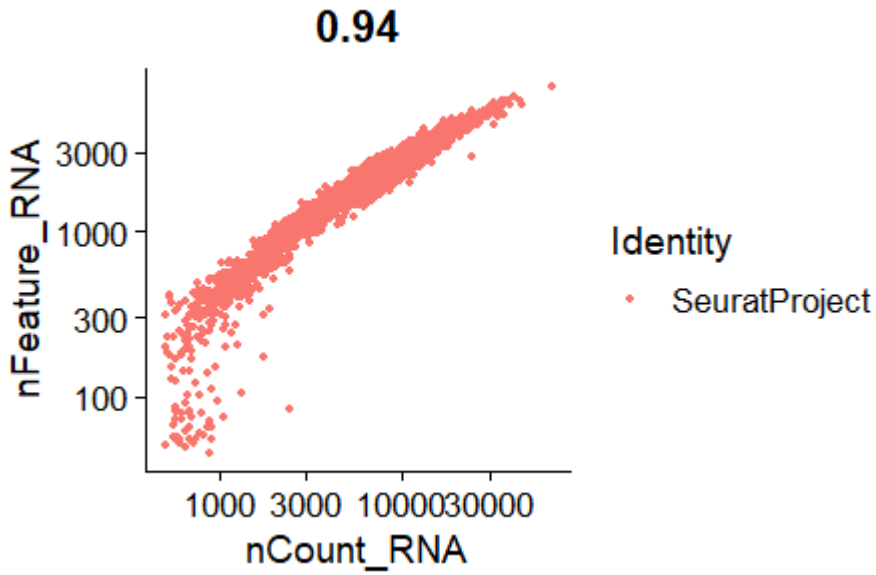
```
seurat <- CreateSeuratObject(data, min.cells = 10, min.features = 10)
dim(seurat)
```

```
## [1] 16022  5155
```

17527 of 33538 genes were filtered (detected in ≤ 10 cells)

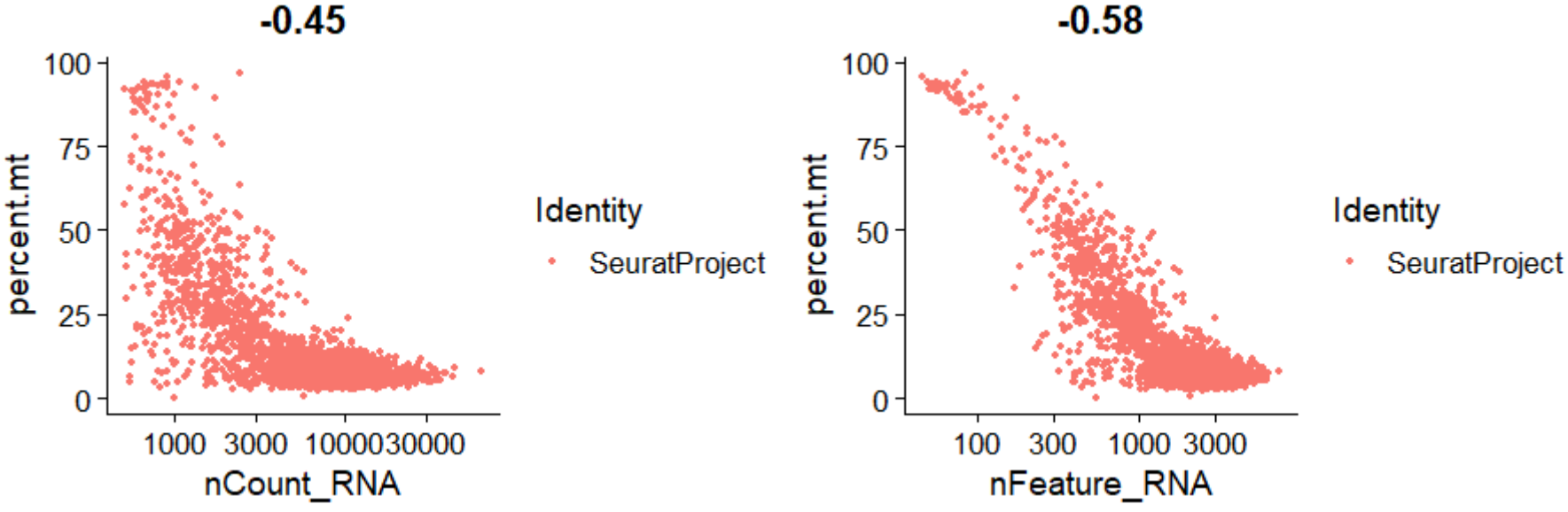
Filtering (genes and barcodes)

```
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[1] & Rahul Satija[1,2]

[1]New York Genome Center, New York City, NY, USA
[2]Center for Genomics and Systems Biology, New York University, New York City, NY, USA
Correspondence to: chafemeister@nygenome.org, rsatija@nygenome.org

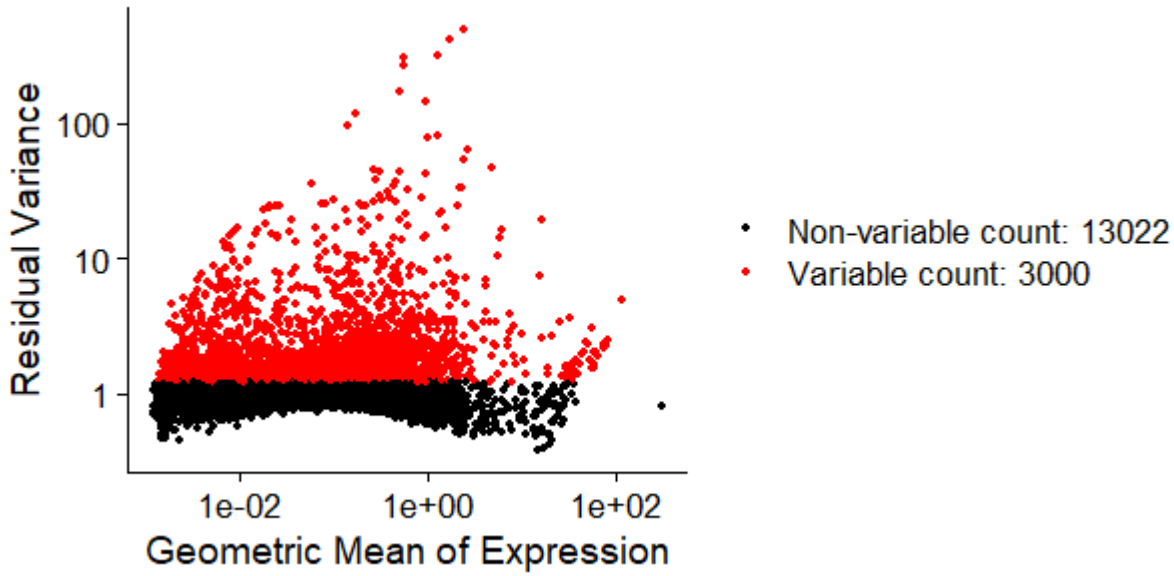
Abstract

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`.

Paper: https://www.biorxiv.org/content/biorxiv/early/2019/03/18/576827.full.pdf

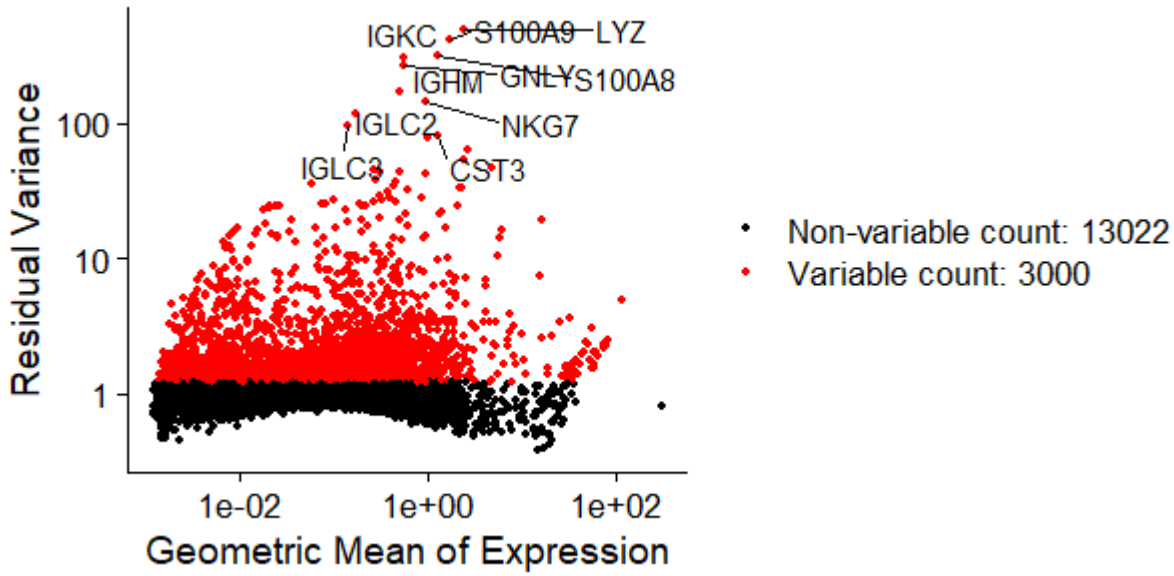
Variable genes

```
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:

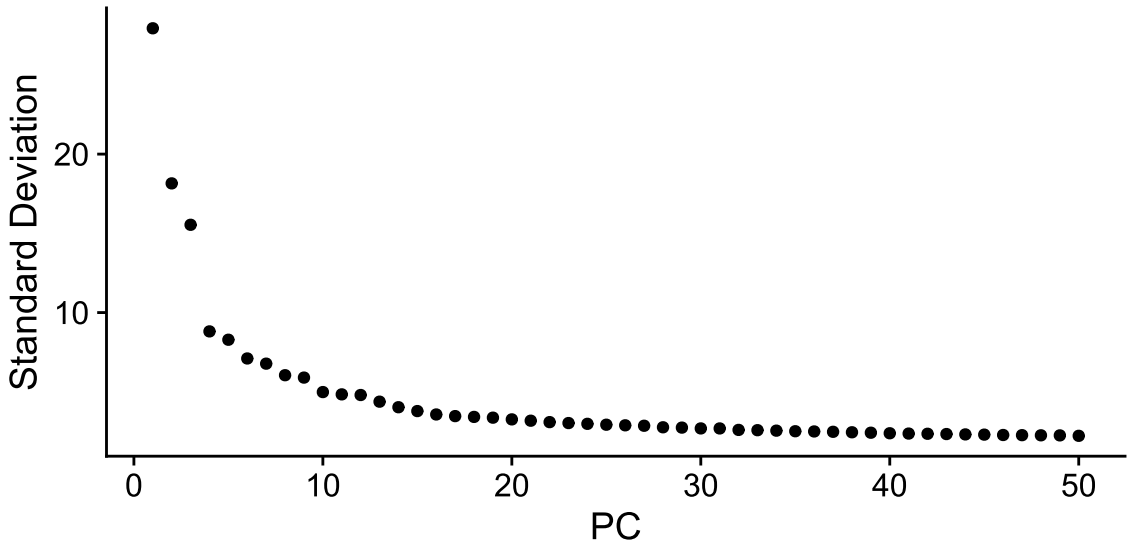
- Some algorithms are just slow when performed in this high-dimensionality data
- Curse of dimensionality

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:

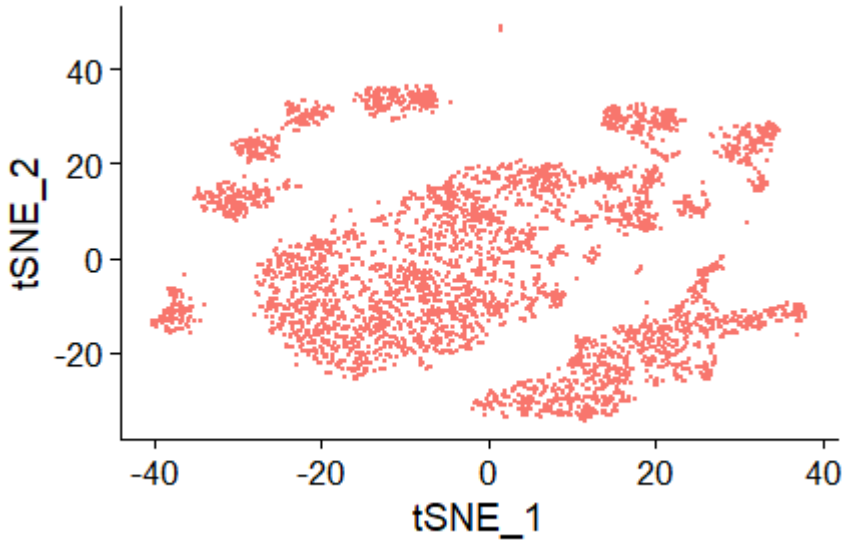
- Dots that are close to each other are cells that are transcriptionally similar to each other
- Dots that are far from each other are cells that are trasncriptionally different from each other

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()
```



Paper: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

tSNE

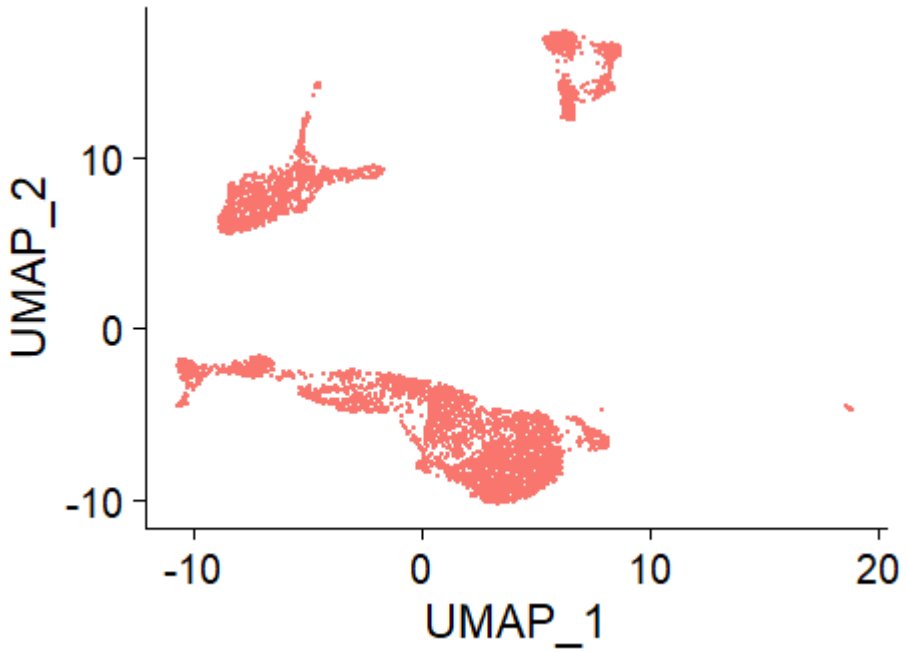
tSNE - "t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets"

- 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

Paper: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

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.

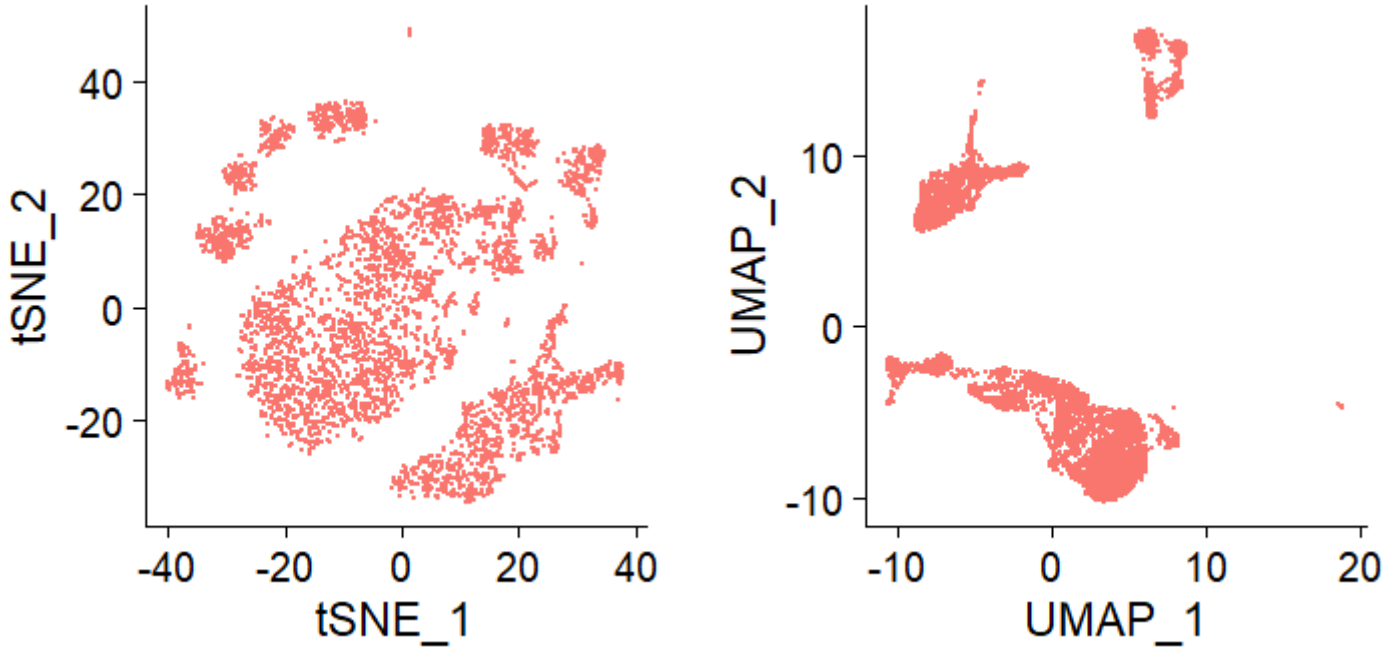
Paper: https://arxiv.org/pdf/1802.03426.pdf

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:

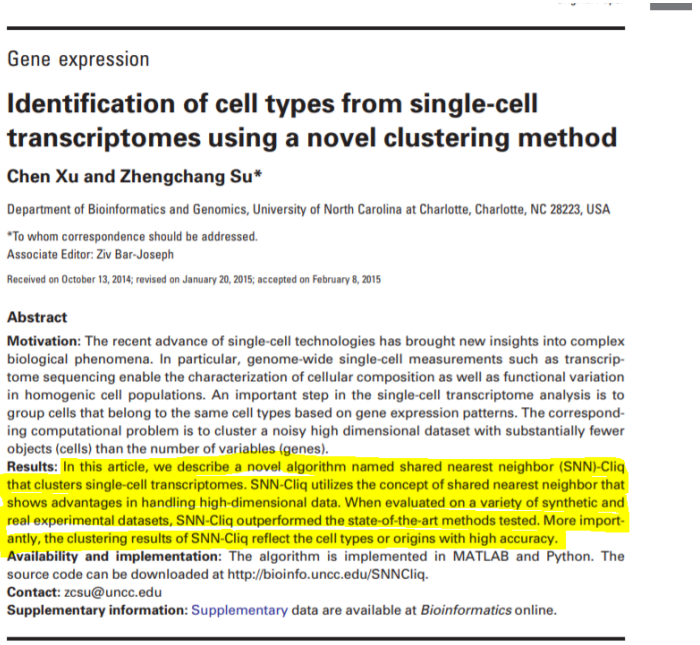
- Graph-based clustering (preferred)
- K-means

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 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

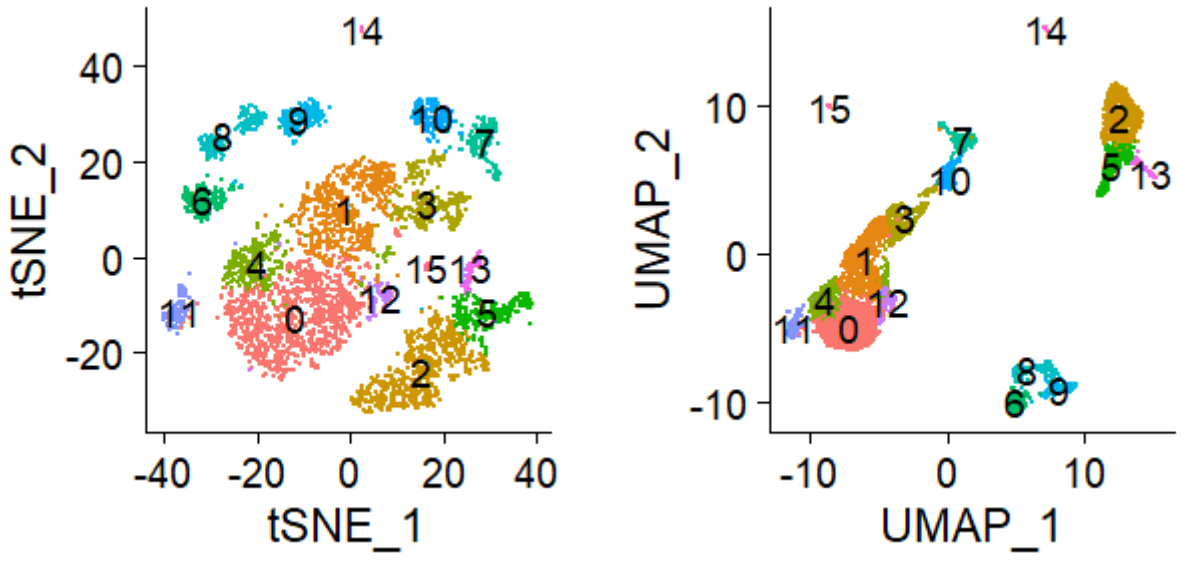
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Paper: https://academic.oup.com/bioinformatics/article/31/12/1974/214505

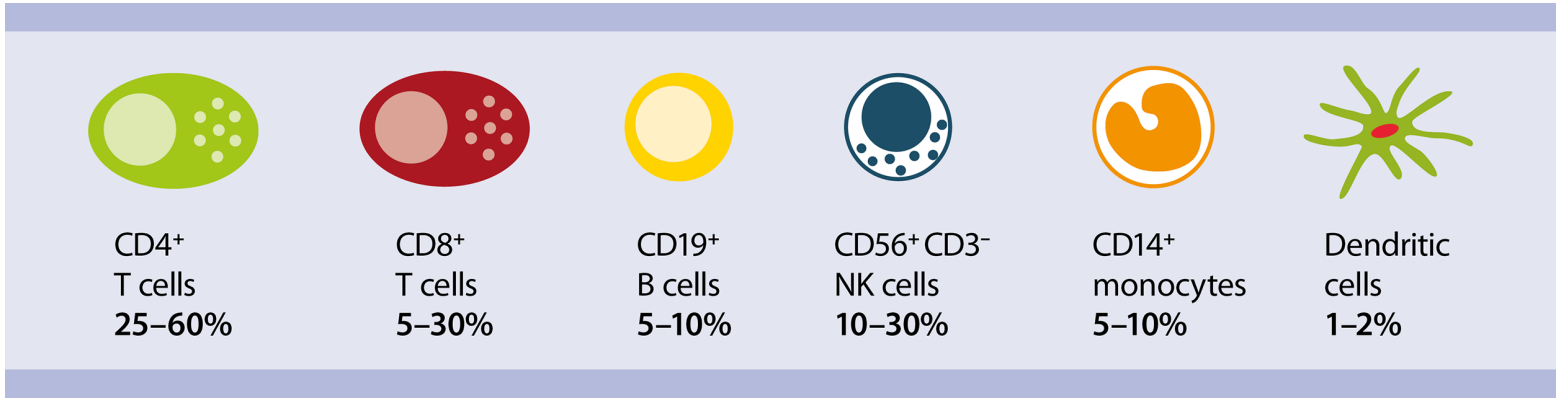
Clustering

```
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()
```



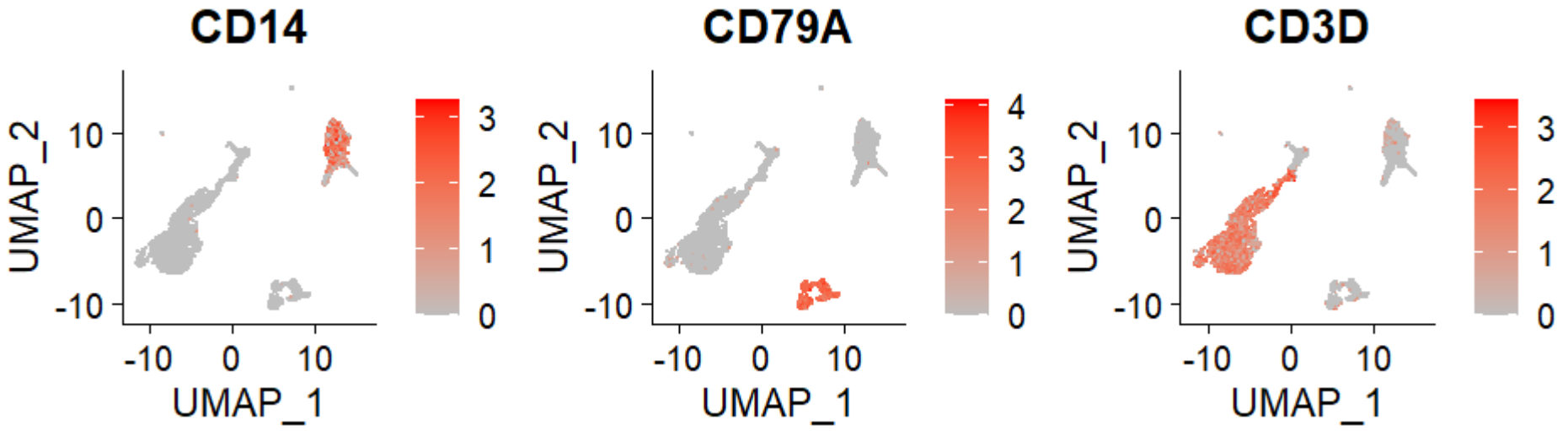
PBMC


CD4+
T cells
25–60%
CD8+
T cells
5–30%
CD19+
B cells
5–10%
CD56+ CD3−
NK cells
10–30%
CD14+
monocytes
5–10%
Dendritic
cells
1–2%

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

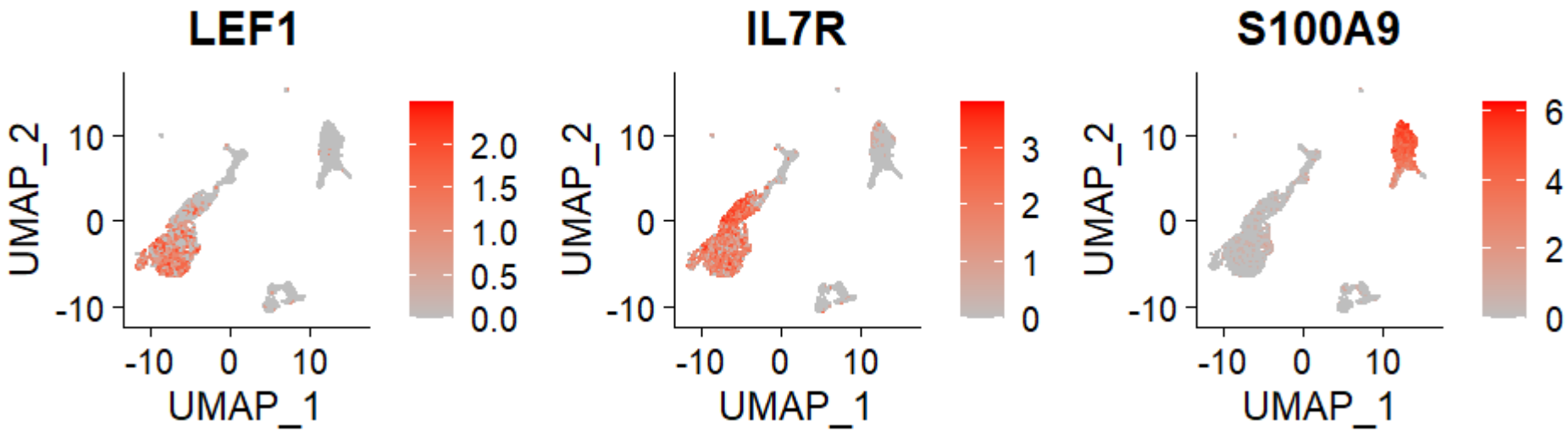
```
# max cells per ident is only seed to speed up the whole thing
allMarkers <- FindAllMarkers(seurat, max.cells.per.ident = 100, test.use = "MAST", only.p
goodMarkers <- allMarkers %>% group_by(cluster) %>% top_n(n = 1, wt = avg_logFC) %>% pull
goodMarkers
```

```
##  [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

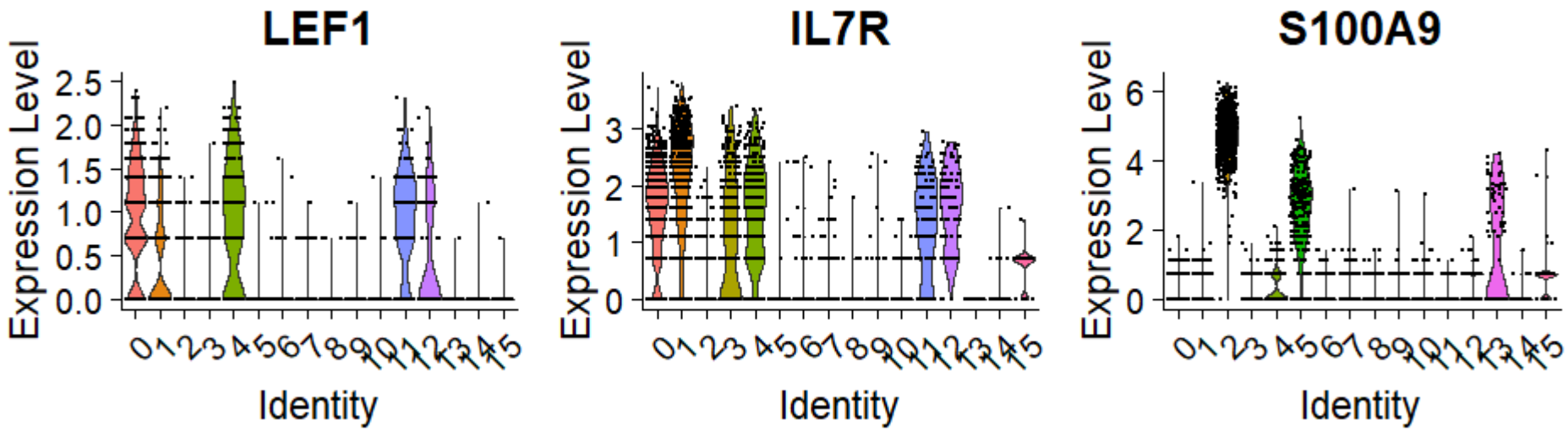
```
FeaturePlot(seurat, goodMarkers[1:3], cols=c("grey", "red"), reduction="umap", ncol=3)
```



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 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

ITMOre than a UNIVERSITY

Latest things:

- 5' vs 3' sequencing
- Single-cell immune profiling
- Single-cell ATAC-seq (with and without gene expression) - multiomics data
- Spatial sequencing