Analysis of single-cell RNA-seq data (I)

Hao Wu
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

Ziyi Li
Department of Biostatistics
The University of Texas MD
Anderson Cancer Center

ENAR 2021 short course
March 2021

Course outline

- **8-9:15: Intro and data preprocessing.**
- 9:15-9:45: Lab: preprocessing and visualization.
- 10-11:15: Normalization, batch effect, imputation, DE, simulator.
- 11:15-12: Lab: Normalization, batch effect, imputation, DE, simulator
- 12-1: lunch break
- 1-2: Clustering and pseudotime construction
- 2-2:30: Lab: Clustering and pseudotime construction
- 2:45–3:30: Supervised cell typing & related single cell data sources
- 3:30-4: Lab: supervised cell typing.
- 4:15-5: scRNA-seq in cancer

Other useful resources

- https://github.com/theislab/single-cell-tutorial/
- https://scrnaseq-course.cog.sanger.ac.uk/website/index.html
- https://broadinstitute.github.io/2019_scWorkshop/

Outline for this session

- **Background**
  - Scientific motivation
  - Technology
  - UMI
- **Pre-processing:** Alignment, QC, GE quantification
- **Data visualization**: tSNE, UMAP

Background

- Most of the biological experiments are performed on “bulk” samples, which contains a large number of cells (millions).
- The “bulk” data measure the average signals (gene expression, TF binding, methylation, etc.) of many cells.
- The bulk measurement ignores the inter-cellular heterogeneities:
  - Different cell types.
  - Variation among the same cell type.

Single cell biology

- The study of individual cells.
- The cells are isolated from multi-cellular organism.
- Experiment is performed for each cell individually.
- Provides more detailed, higher resolution information.
- High-throughput experiments on single cell is possible.

Single cell sequencing

- Different types of sequencing at the single-cell level:
  - DNA-seq
  - ATAC-seq, ChIP-seq
  - BS-seq
  - RNA-seq
  - Multi-omics
- Very active research field in the past few years.

Basic experimental procedure

- Isolation of single cell. Techniques include
  - Laser-capture microdissection (LCM)
  - Fluorescence-activated cell sorting (FACS)
  - Microfluidics
- Open the cell and obtain DNA/mRNA/etc.
- PCR amplification to get enough materials.
- Perform sequencing.

Single cell RNA-seq (scRNA-seq)

- The most active in the single cell field.
- Scientific goals:
  - Composition of different cell types in complex tissues.
  - New/rare cell type discovery.
  - Gene expression, alternative splicing, allele specific expression at the level of individual cells.
  - Transcriptional dynamics (pseudotime construction).
  - Above can be investigated and compared spatially, temporally, or under different biological condition.

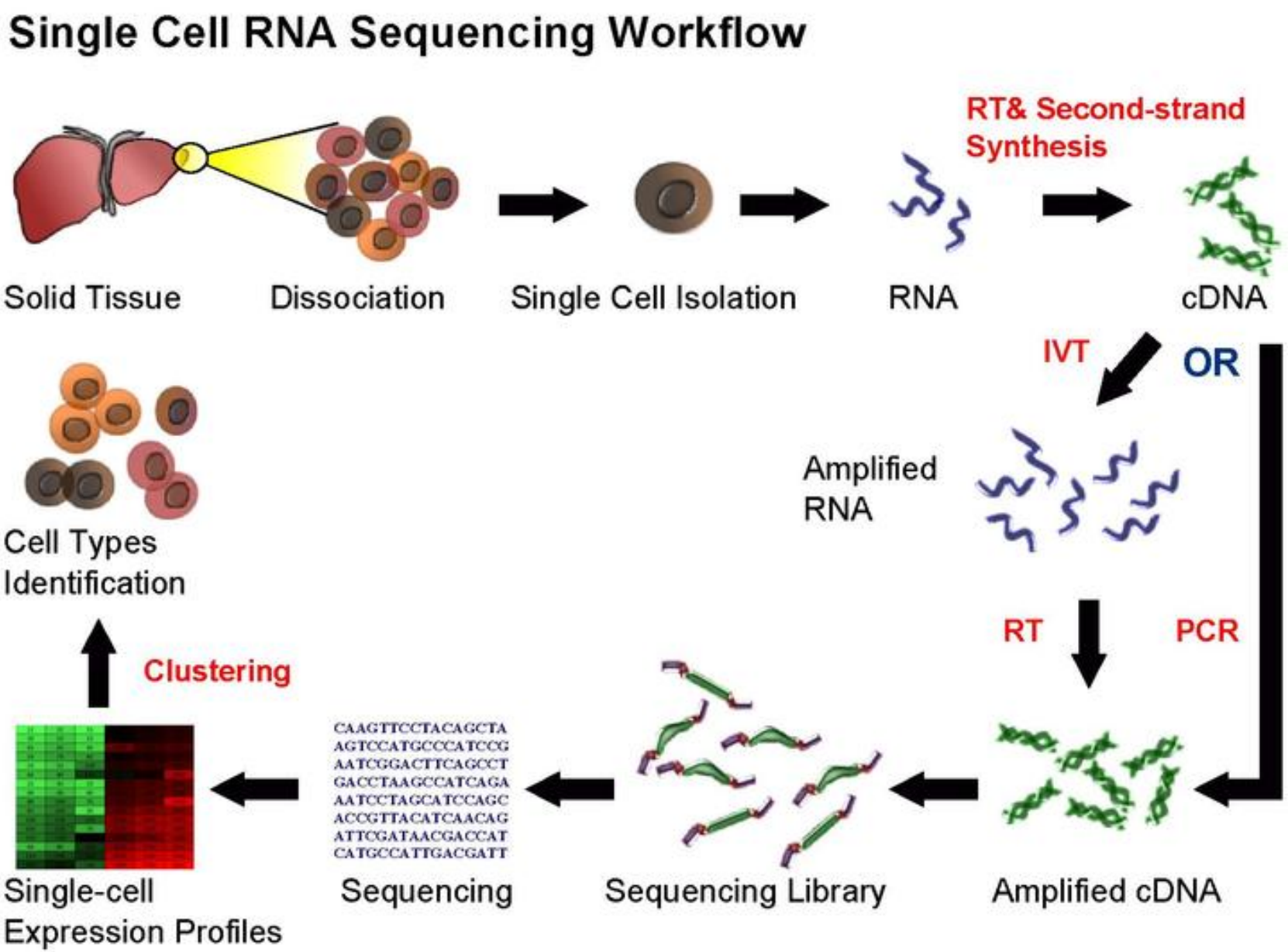


Figure source: Wikipedia

Technologies by cell capturing method

- **Plate-based methods**: Smart-Seq/Smart-Seq2, CEL-seq:
  - Sort cells into the wells on a multi-well plate.
  - Lower throughput (in terms of number of cells).
  - High sequencing depth
  - Can be combined with FACS for cell sorting.
  - Better at detecting low expression genes
  - Good for isoform analysis, allele specific expression

Microwell plates

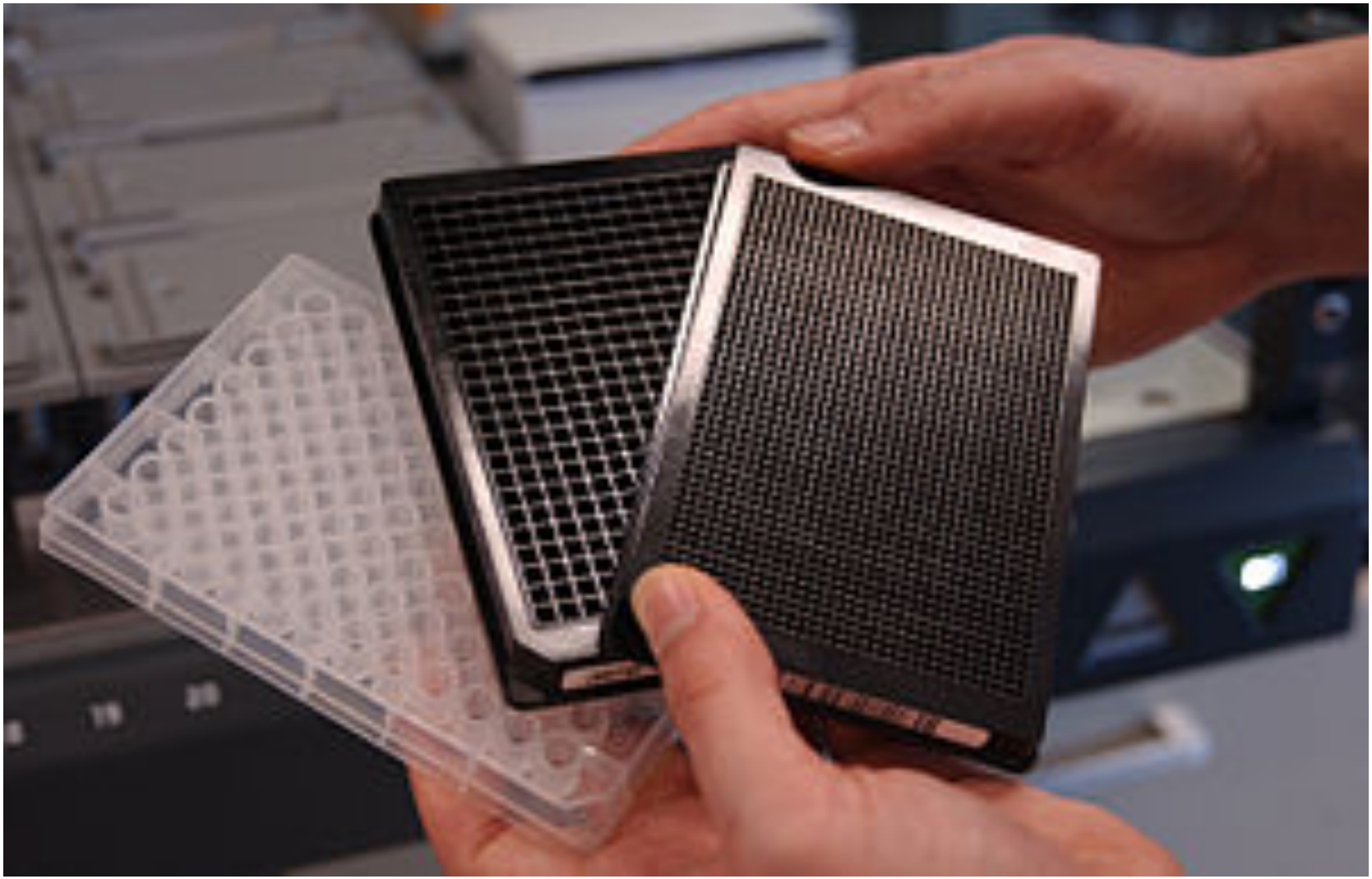
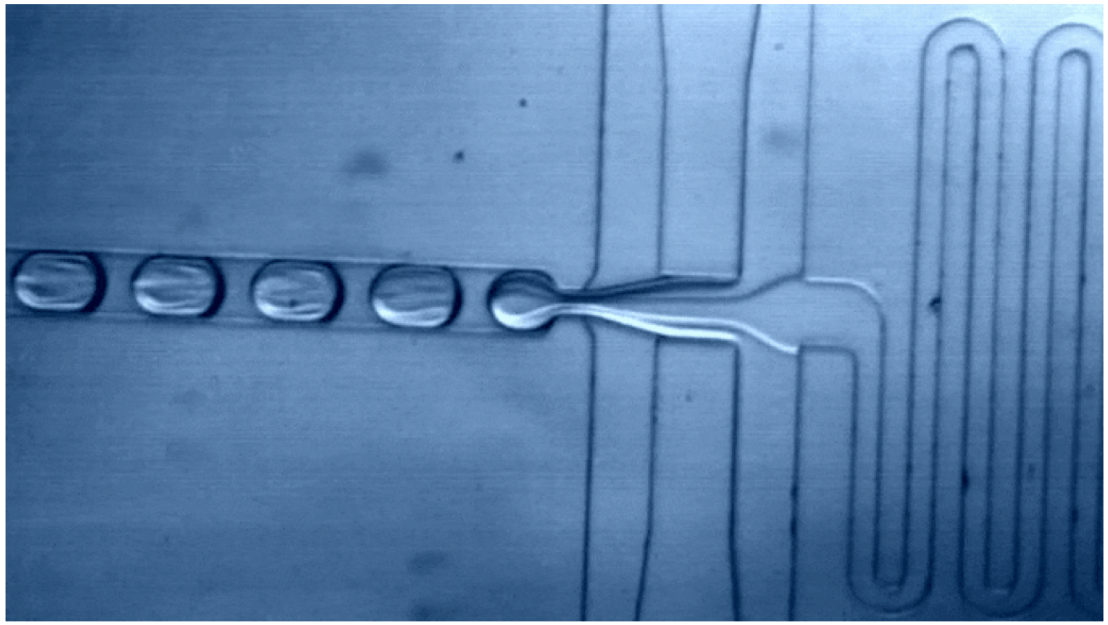
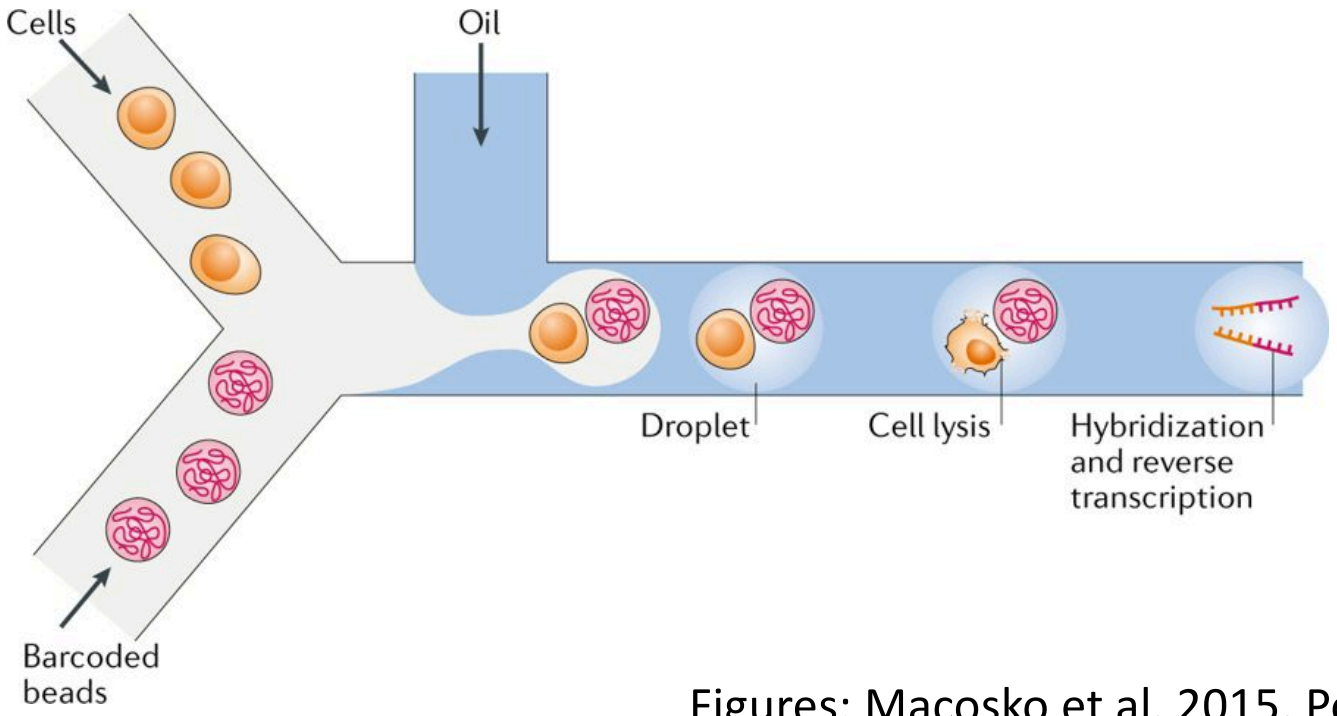


Figure source: wikipedia

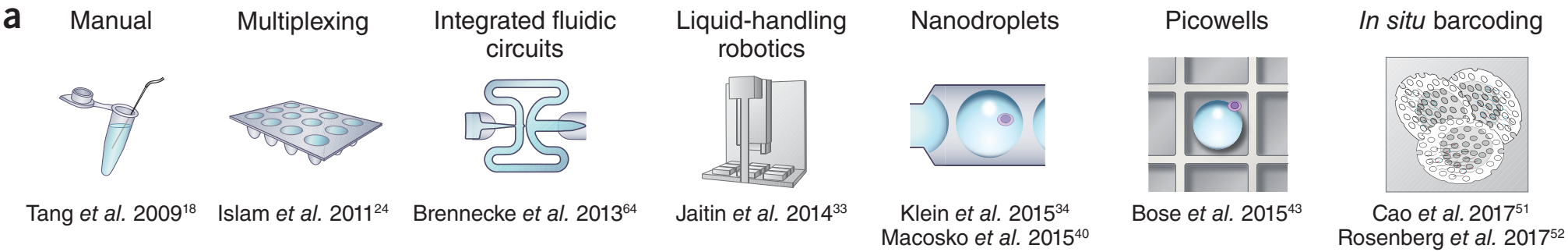
- **Droplet-based methods:** Drop-seq, inDrop, 10x genomics
  - Put each cell in a nanoliter droplet with a bead.
  - Each droplet is a reactor for PCR.
  - Each bead has a unique barcode, so all beads can be pooled and sequenced together.
  - Much higher throughput in terms of number of cells.
  - Lower sequencing depth.
  - Good for identifying cell subpopulations.

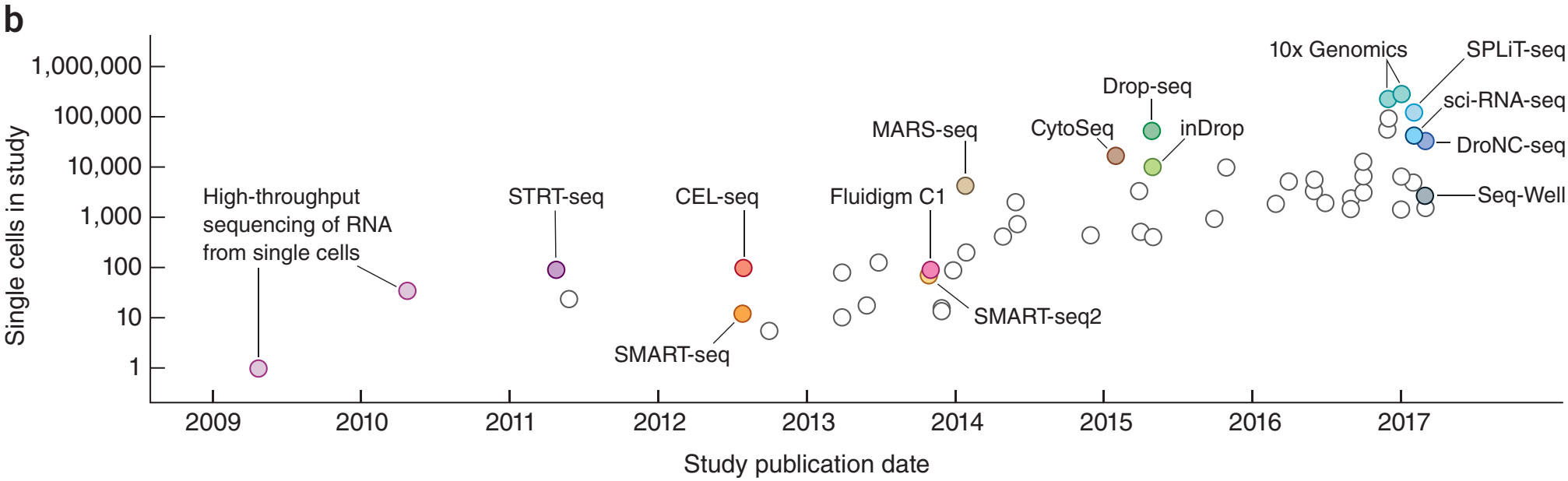




Figures: Macosko et al. 2015, Potter SS. 2018

Technologies over the years





Svensson et al. (2017) Nature Protocols

Unique molecule identifier (UMI)

NM 2014

Quantitative single-cell RNA-seq with unique molecular identifiers

Saiful Islam[1], Amit Zeisel[1], Simon Joost[2], Gioele La Manno[1], Pawel Zajac[1], Maria Kasper[2], Peter Lönnerberg[1] & Sten Linnarsson[1]

UMI

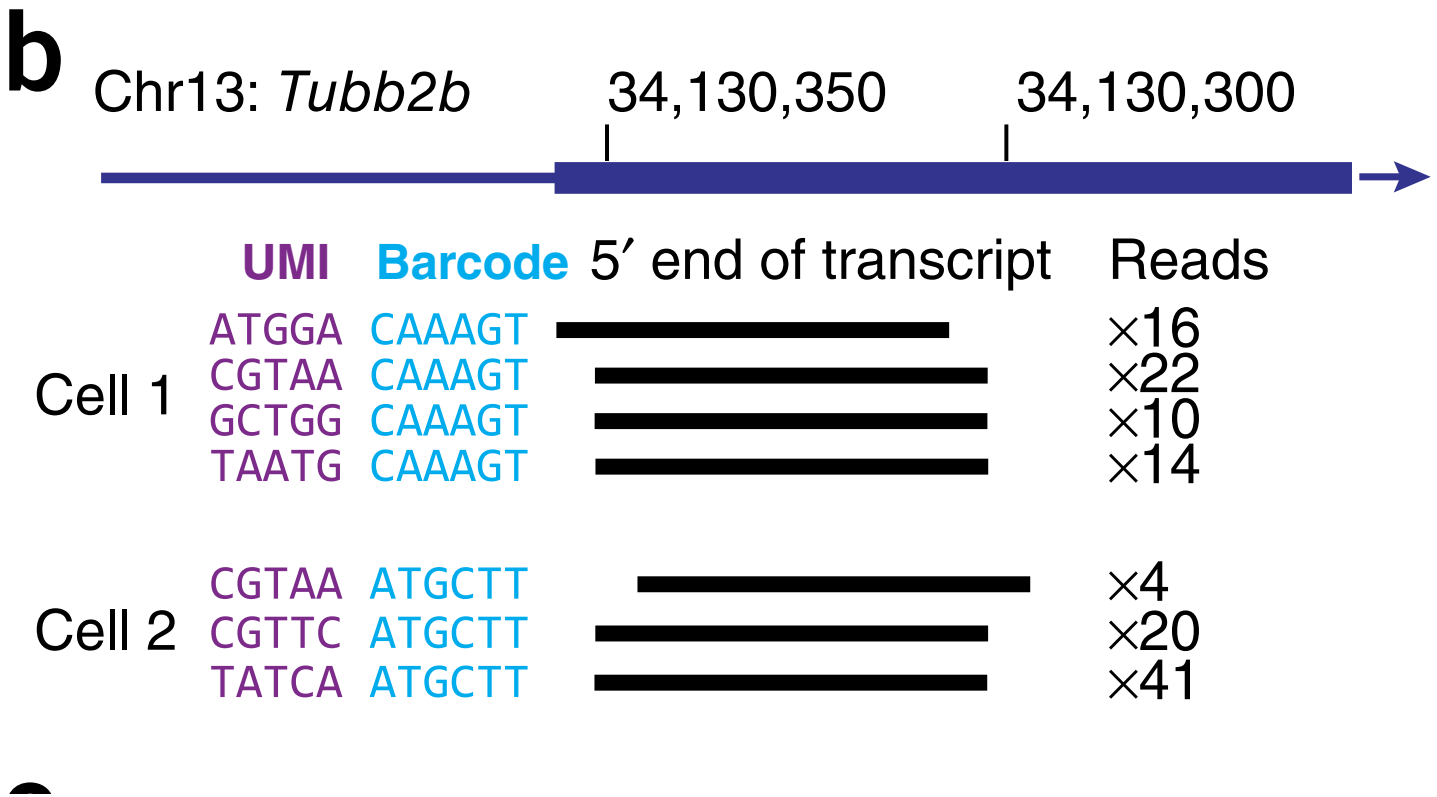
- PCR introduces nonlinear amplification bias
  - Factors influencing PCR: sequence content, chromatin structure, etc.
- UMIs are short sequence tag added to each unique mRNA molecular before PCR, for reducing PCR bias.
- Number of possible UMIs = 4^L, where L is the length of the UMI

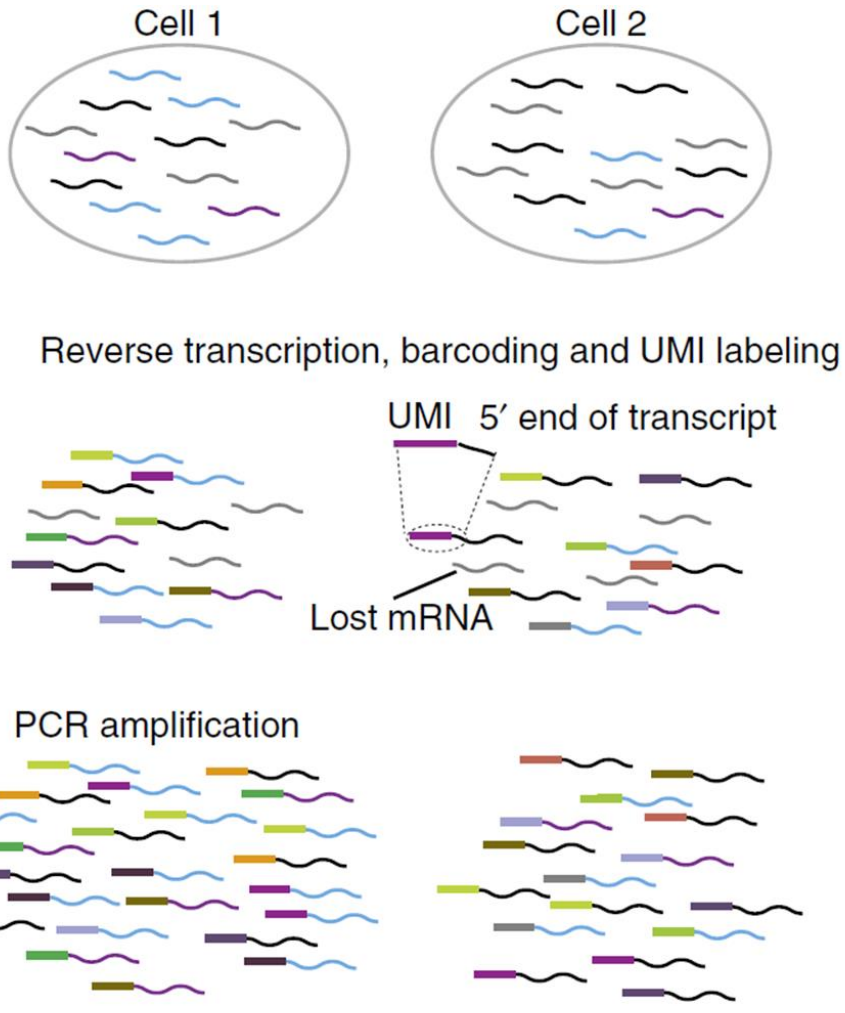
Multiplexing

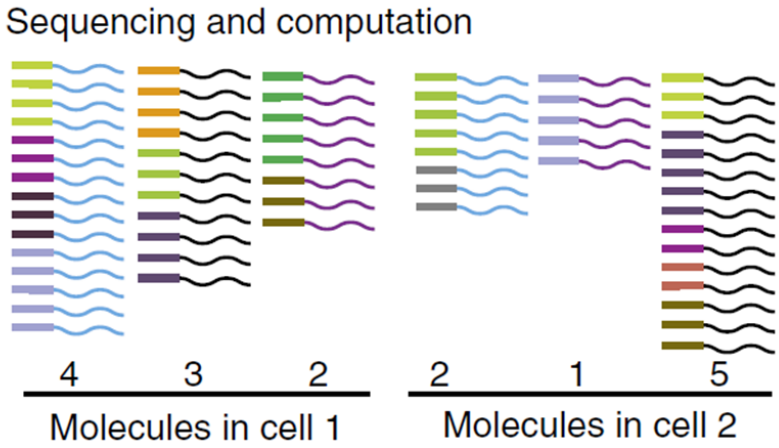
- Technology to pool many cells together for each sequencing lane.
- Each cell is uniquely marked by a **barcode** (short sequence tag).
- A combination of barcode and UMI can quantify unique transcripts in each cell.

UMI + barcode

- Assumption: number of identical mRNA is small (say, <100) for most genes.
- Use 5-bp UMI (can mark 1024 molecules)
  - When a transcript has, say, 20 mRNA molecules, the probability of two molecule having the same UMI is small.
- Use 6-bp barcode to identify cells (up to 4096 cells).
- Use molecule counts instead of read counts as gene expression measurements.

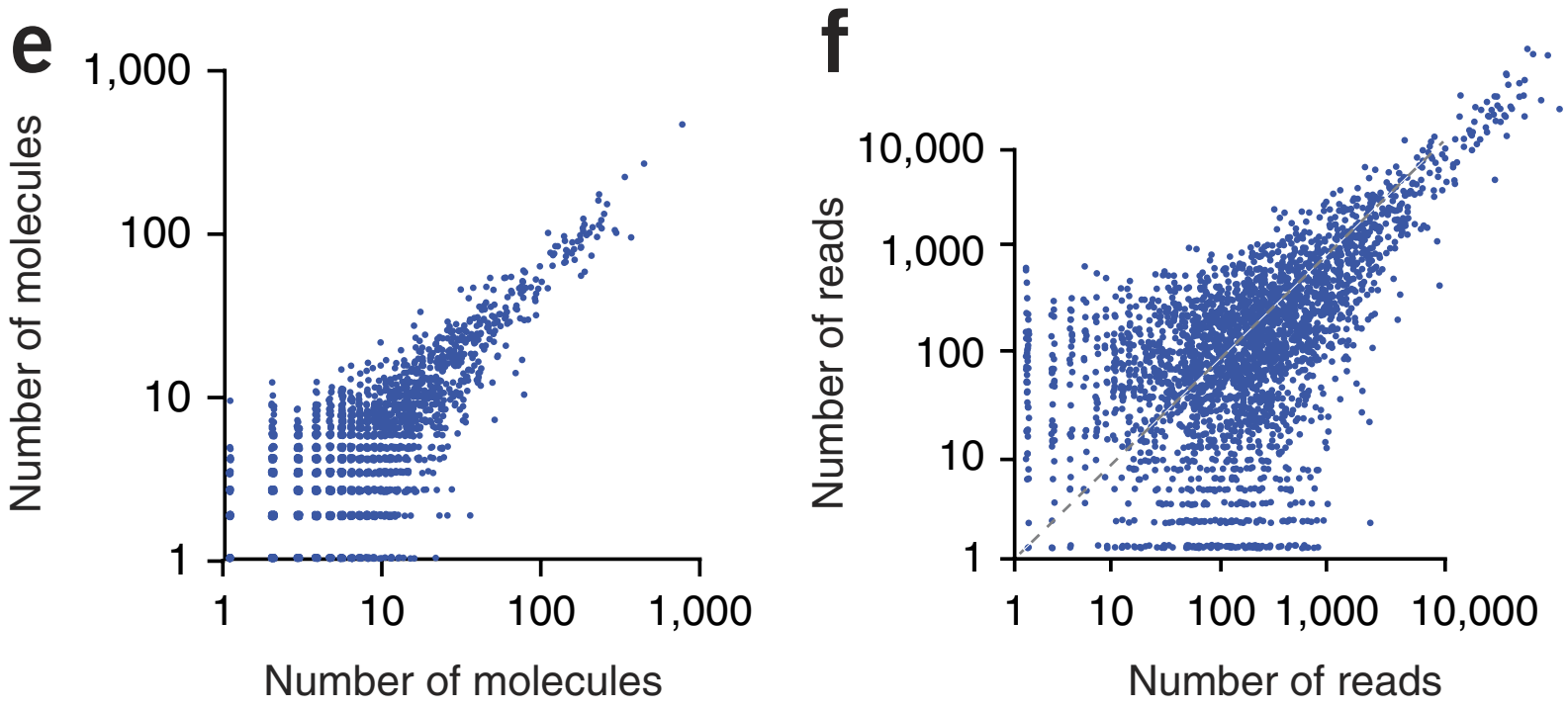

b
Chr13: Tubb2b
34,130,350
34,130,300
UMI
Barcode
5′ end of transcript
Reads
Cell 1
ATGGA CAAAGT
CGTAA CAAAGT
GCTGG CAAAGT
TAATG CAAAGT
×16
×22
×10
×14
Cell 2
CGTAA ATGCTT
CGTTC ATGCTT
TATCA ATGCTT
×4
×20
×41





Saiful Islam ••• Sten Linnarsson

UMI provides better measurements and reproducibility



Single nucleus RNA-seq (snRNA-seq)

- Profile gene expressions in nucleus, instead of the whole cell
  - Transcripts can be in cytoplasm and nucleus
- Useful when the cells are difficult to isolate
  - Frozen tissues
  - Highly connected cells such as neurons
- Analysis methods are similar.

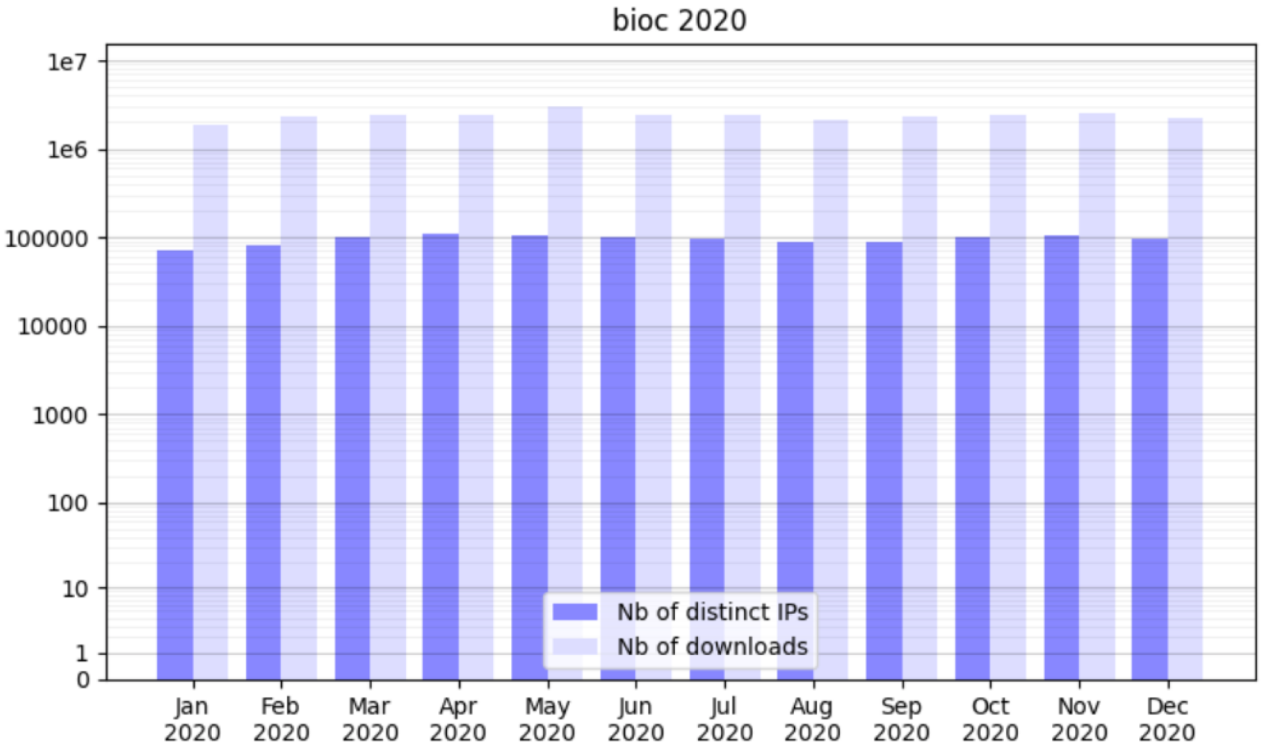
Multi-omics single cell assays

- CITE-seq (**C**ellular **I**ndexing of **T**ranscriptomes and **E**pitopes by **Seq**uencing)
  - Jointly profile transcriptome and proteome.
- scNMT-seq (single-cell **N**ucleosome, **M**ethylation and **T**ranscription sequencing)
  - Jointly profile chromatin accessibility, DNA methylation, and transcription

Brief introduction to Bioconductor

- A collection of R packages
- The *de facto* language for genomic data analysis.

2020



Month	Nb of distinct IPs	Nb of downloads
Jan/2020	71347	1863031
Feb/2020	82959	2327549
Mar/2020	100156	2437796
Apr/2020	109245	2445530
May/2020	107201	3059277
Jun/2020	99529	2406886
Jul/2020	96776	2409421
Aug/2020	86995	2119491
Sep/2020	90269	2280596
Oct/2020	100693	2427302
Nov/2020	103036	2572492
Dec/2020	95193	2225497
2020	**816065**	**28574868**

bioc_2020_stats.tab

Functionalities

- "*Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development.*"
- Provides close to 2000 packages for:
  - microarrays.
  - second generation sequencing.
  - other high-throughput assays.
  - annotation.
- Most of the packages are contributed.

Bioconductor installation

- Use **`BiocManager::install()`**.
- Basic installation: installing default (core) packages:

```
if (!requireNamespace("BiocManager"))
    install.packages("BiocManager")
BiocManager::install()
```

- Installing a specific package:

```
BiocManager::install("limma")
```

Data processing

- Preprocessing
  - QC
  - Alignment
  - Expression quantification
- Normalization
- Batch effect correction
- Imputation

scRNA-seq data preprocessing

- QC:
  - FastQC is a popular tool for checking a single sample.
  - MultiQC: create a single report with interactive plots for multiple QC reports.
- Read trimming:
  - cutadapt (with a wrapper Trim Galore!)
- Bioconductor package “scater” provides useful and easy-to-use functions for QC and data visualization.

Alignment and quantification

- Alignment
  - Bulk RNA-seq alignment software (Tophat, STAR, HISAT, etc.) can be used.
  - Some commercial software, such as CellRanger for 10x genomics data.
- Quantification (to obtain count matrix from aligned reads)
  - Most alignment software provide such functionality.

Some data characteristics

- Data is very sparse (many zeros), especially for Drop-seq data.
- Number of transcripts detected is much lower compared to bulk RNA-seq under the same sequencing depth.

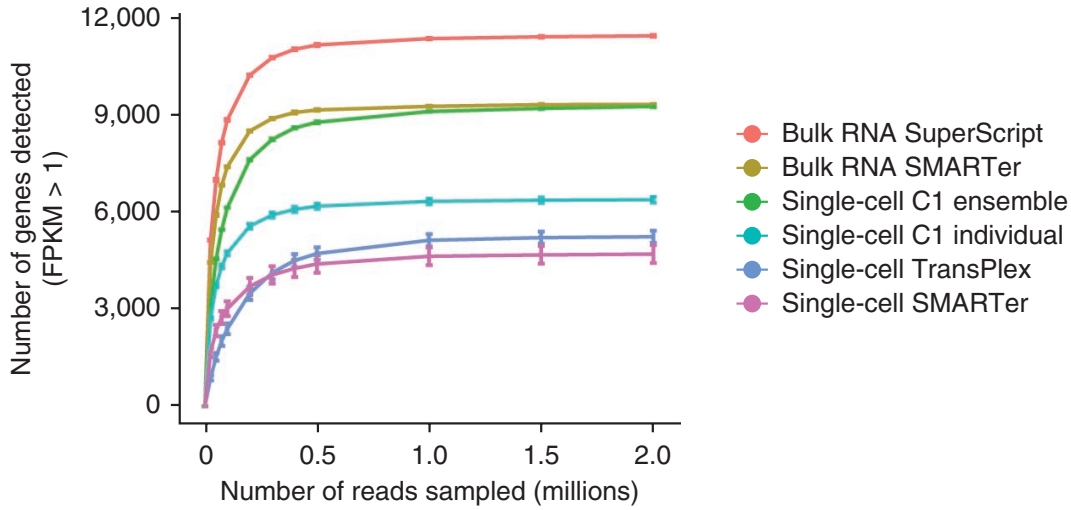
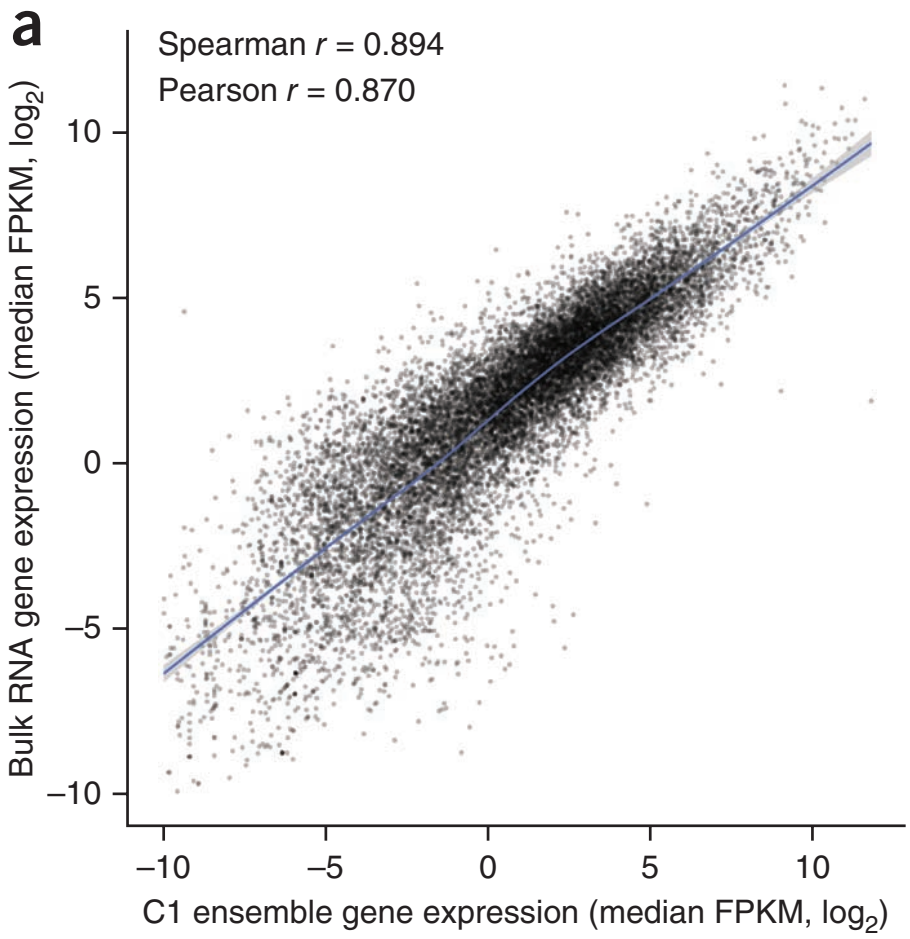


Figure 5 | Saturation curves for the different sample preparation methods. Each point on the curve was generated by randomly selecting a number of raw reads from each sample library and then using the same alignment pipeline to call genes with mean FPKM >1. Each point represents four replicate subsamplings. Error bars, standard error.

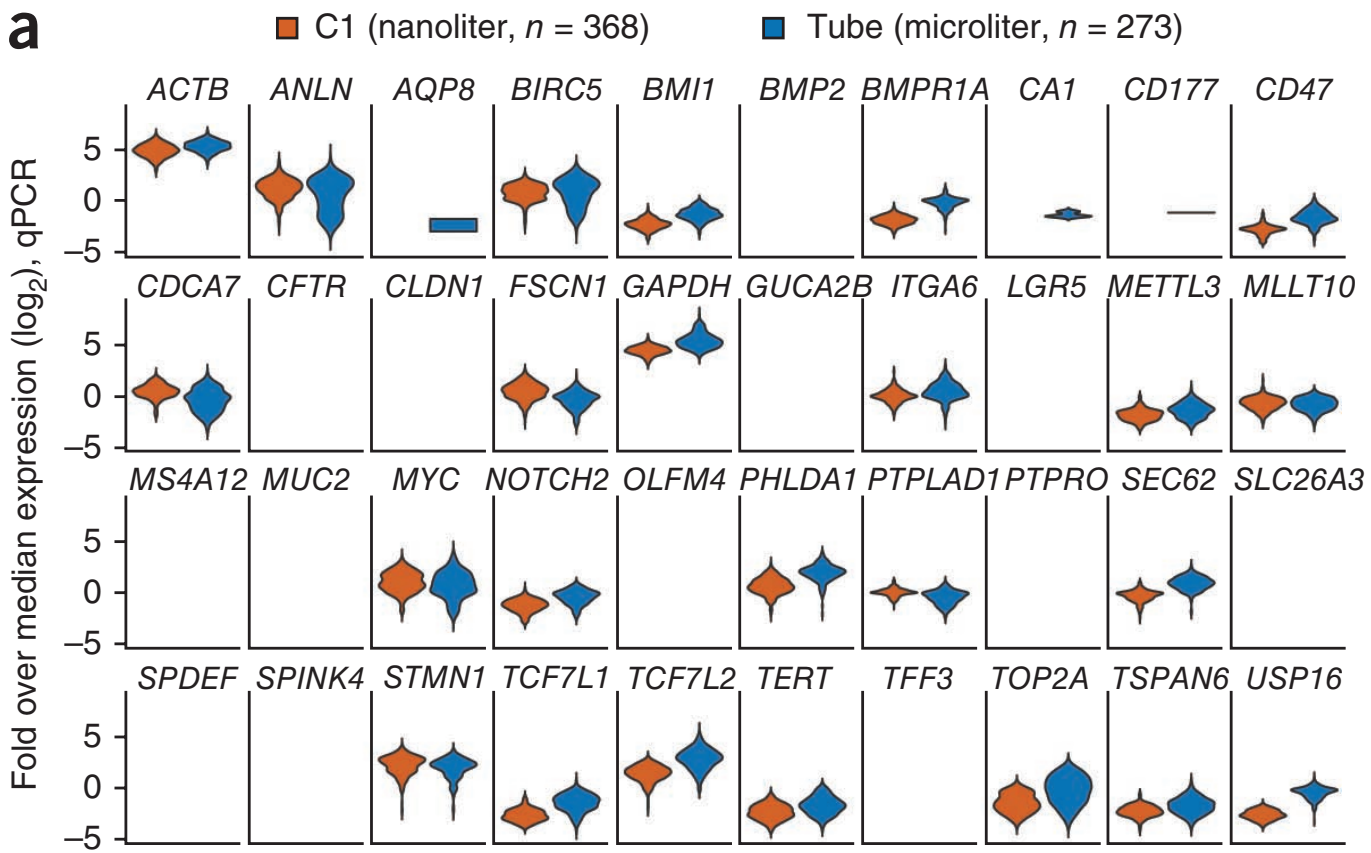
Wu et al. 2013 Nature Method

- Bulk and aggregated single cell expressions have good correlation.



Wu et al. 2013 Nature Method

- Expression levels for a gene cross cells sometimes show bimodal distribution.



Wu et al. 2013 Nature Method

scRNA-seq data after processing

- A matrix of read counts: rows are genes and columns are cells

	AACGGTACCTTCGC_1	AGAGAAACGCCCTT_1	AGGCAGGACGAATC_1
ENSG00000228463	0	0	0
ENSG00000230021	0	0	0
ENSG00000237491	0	0	0
ENSG00000177757	0	0	0
ENSG00000225880	0	0	0
	ATACCTTGCCGATA_1	ATAGGCTGGCTTCC_1	
ENSG00000228463	0	0	
ENSG00000230021	0	0	
ENSG00000237491	0	0	
ENSG00000177757	0	0	
ENSG00000225880	0	0	

A few useful R packages

- SingleCellExperiment:
  - Bioconductor package. Defines "SingleCellExperiment" class for storing single cell data: expression matrix, gene and cell information, etc.
- Visualization tools:
  - tSNE
  - UMAP

SingleCellExperiment

- Installation:

```
BiocManager::install("SingleCellExperiment")
```

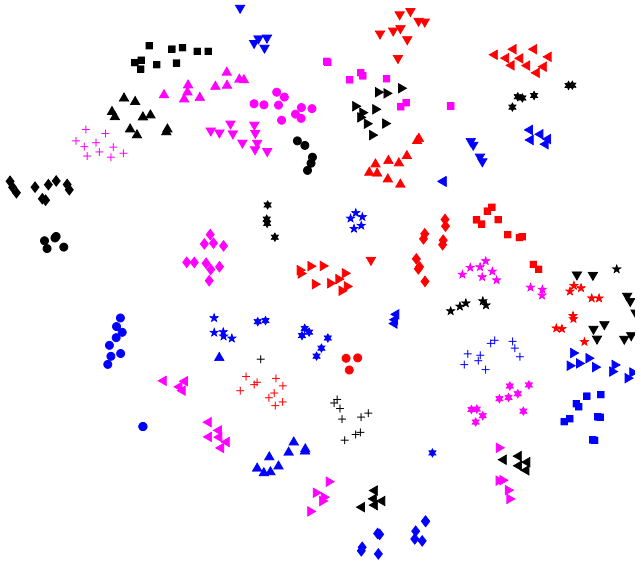
- Create SingleCellExperiment object:

```
sce <- SingleCellExperiment(list(counts=counts),
    colData=DataFrame(label=celllabels),
    rowData=DataFrame(genenames=genenames),
    metadata=list(study="GSE111111")
)
```

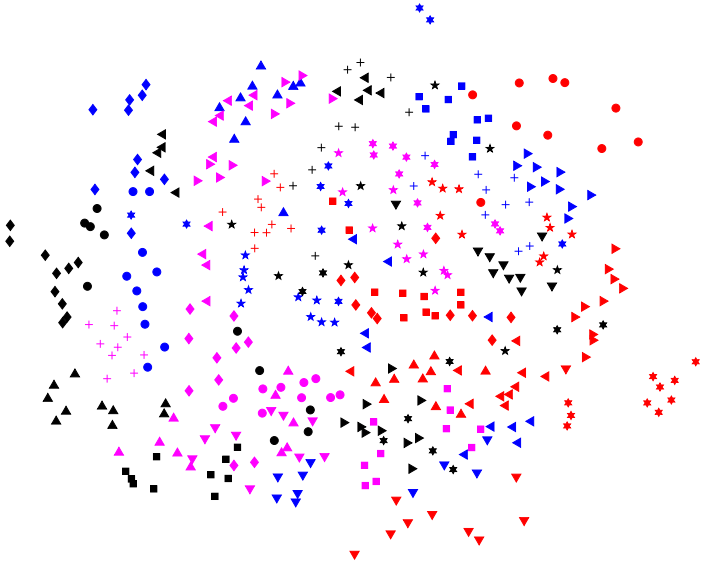
- Functions to access contents of the object: `counts`, `rowData`, `colData`, etc.

t-SNE: a useful visualization tool

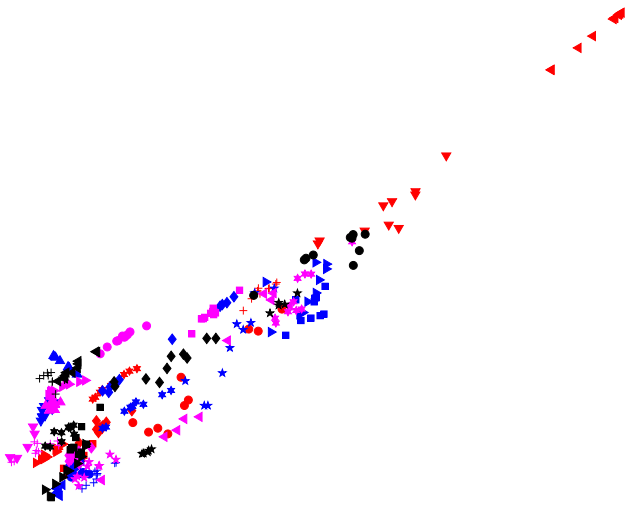
- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map.
- When project high-dimensional data into lower dimensional space, preserve the distances among data points.
  - Try to make the pairwise distances of points similar in high and low dimension.
- Has “tsne” and “Rtsne” package on CRAN.



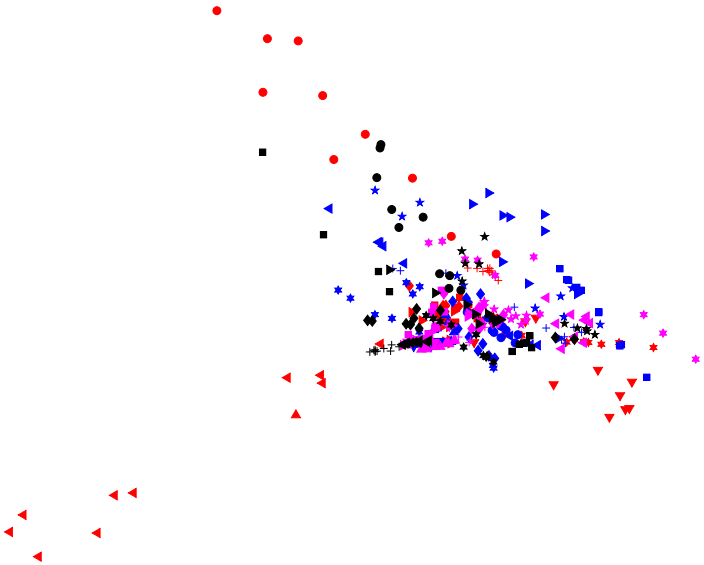
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



(c) Visualization by Isomap.



(d) Visualization by LLE.

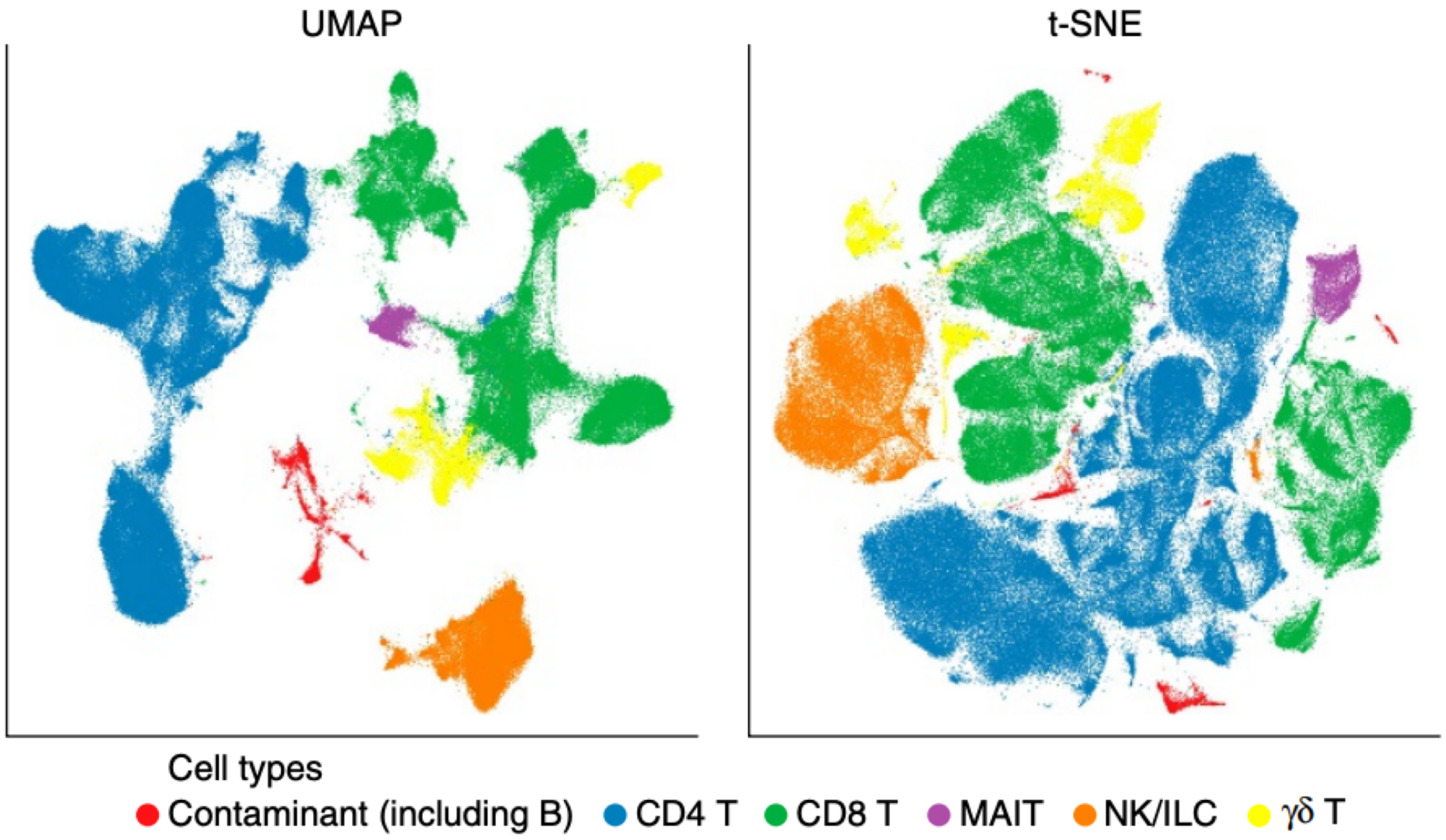
Example code for t-SNE

```
library(Rtsne)
tsne_model_1 = Rtsne(datamatrix,
                     check_duplicates=FALSE, pca=TRUE,
                     perplexity=30, theta=0.5, dims=3)
tsne_out = as.data.frame(tsne_model_1$Y)

plot(tsne_out$V1, tsne_out$V2,
    pch = 19, cex = 0.4, col = mycolor)
legend("bottomleft", col = mycolor,
       legend = uniqCT, pch = 19,
       cex = 0.5, bty = "n")
```

UMAP: a newer (and better?) visualization tool

- UMAP (uniform manifold approximation and projection): a recently developed dimension reduction tool
- "*Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters.* " ---- Betcht et al. 2018 Nat Biotech
- "*UMAP, which is based on theories in Riemannian geometry and algebraic topology, has been developed, and soon demonstrated arguably better performance than t-SNE due to its higher efficiency and better preservation of continuum.*" ---- Mu et al. 2018 GBP
- Has "umap" package on CRAN.



Betcht et al. 2018 Nat Biotech

Example code for UMAP

```
library(umap)
sim_umap <- umap(datamatrix)
sim_umap2 <- sim_umap$layout
colnames(sim_umap2) <- c("UMAP1", "UMAP2")

plot(sim_umap2[,1], sim_umap2[,2],
    pch = 19, cex = 0.4, col = mycolor)
legend("bottomleft", col = mycolor,
       legend = uniqCT, pch = 19,
       cex = 0.5, bty = "n")
```