HAU901I

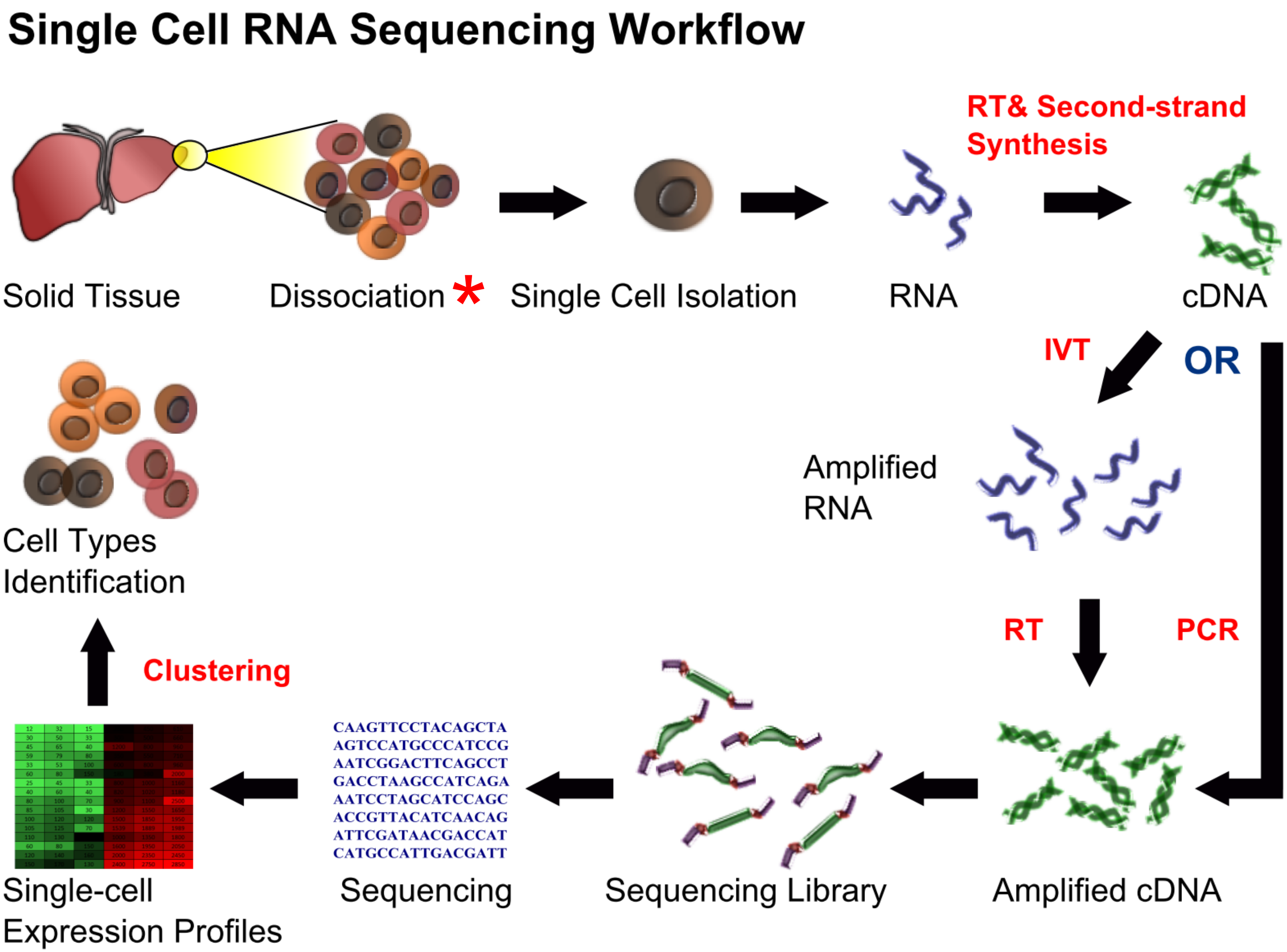
Single-cell transcriptomics

Introduction

Pr Jacques Colinge

UM/Inserm - IRCM - ICM

scRNA-seq general workflow



Source: Wikipedia

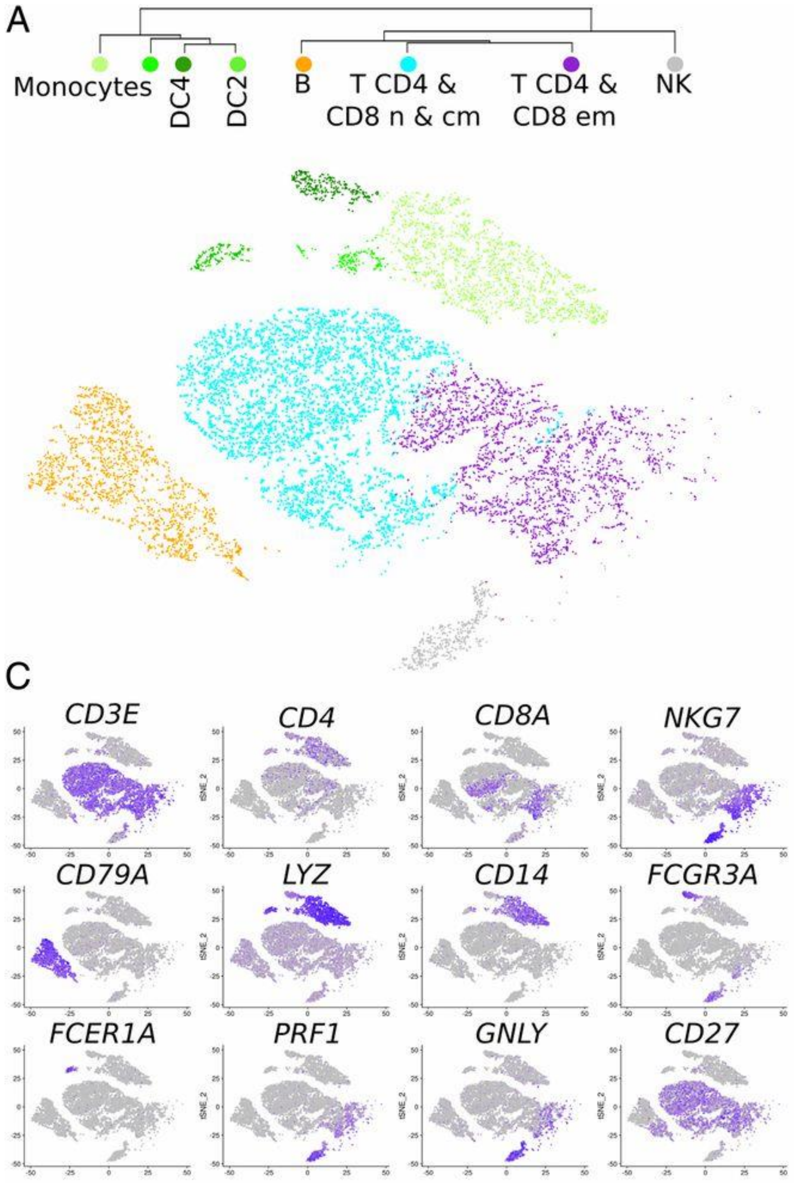
* Not required if cells already in suspension, *e.g.*, blood cells

Typical outcome

- Peripheral blood mononuclear cells (PBMC), Pizzolato et al., PNAS, 2019
- (No dissociation required here.)

- Each dot in (A) represents an individual cell's transcriptome
- Clustering of those transcriptomes has identified several well-defined clusters

- The specific expression of known cell population marker genes enables us to assign each cluster to a cell type as indicated in (C).



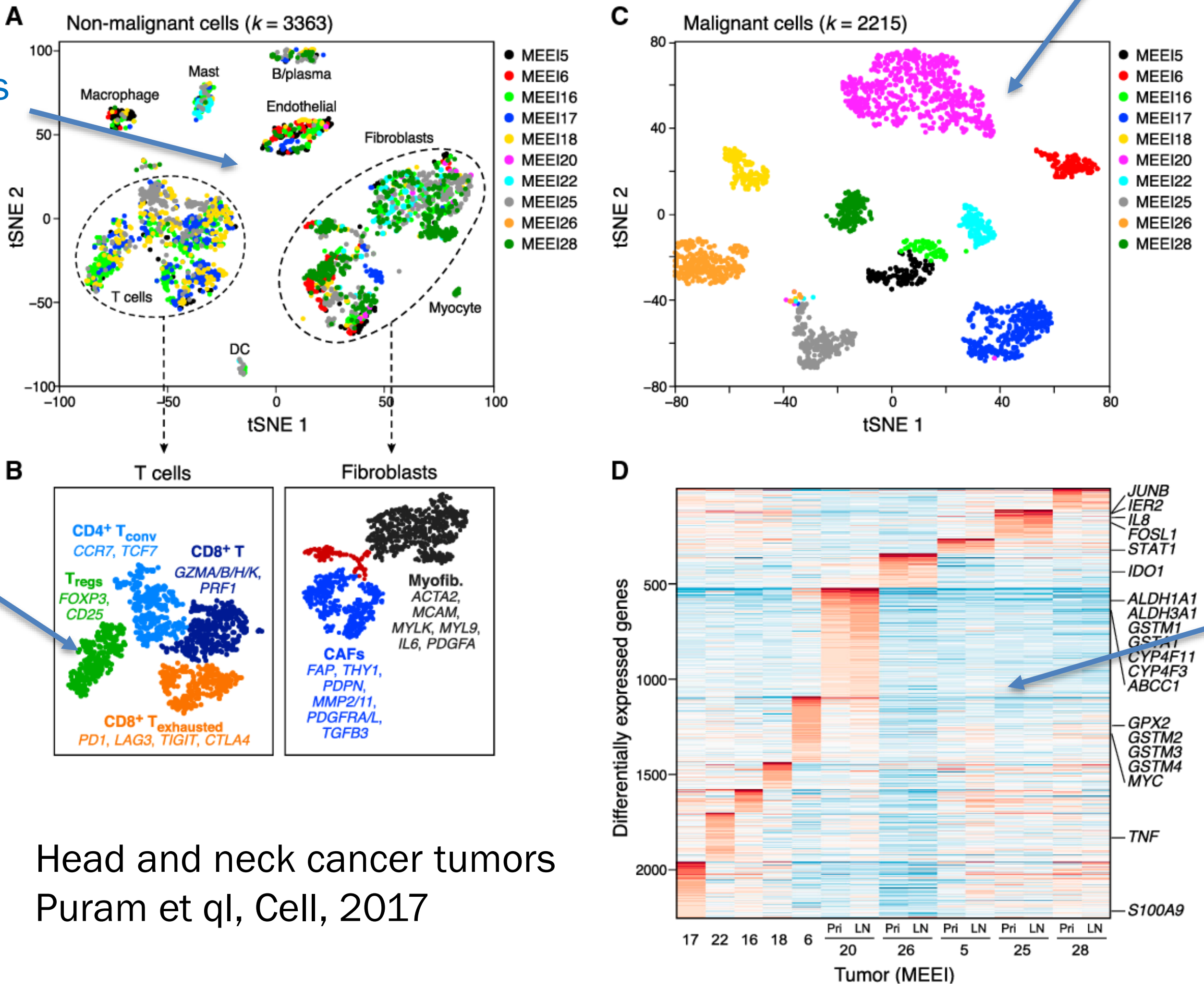
Typical outcome (2)

Comparable stromal cells across patients

Subclustering enables identifying phenotypically distinct subpopulations of cells

Distinct cancer cells for each patient

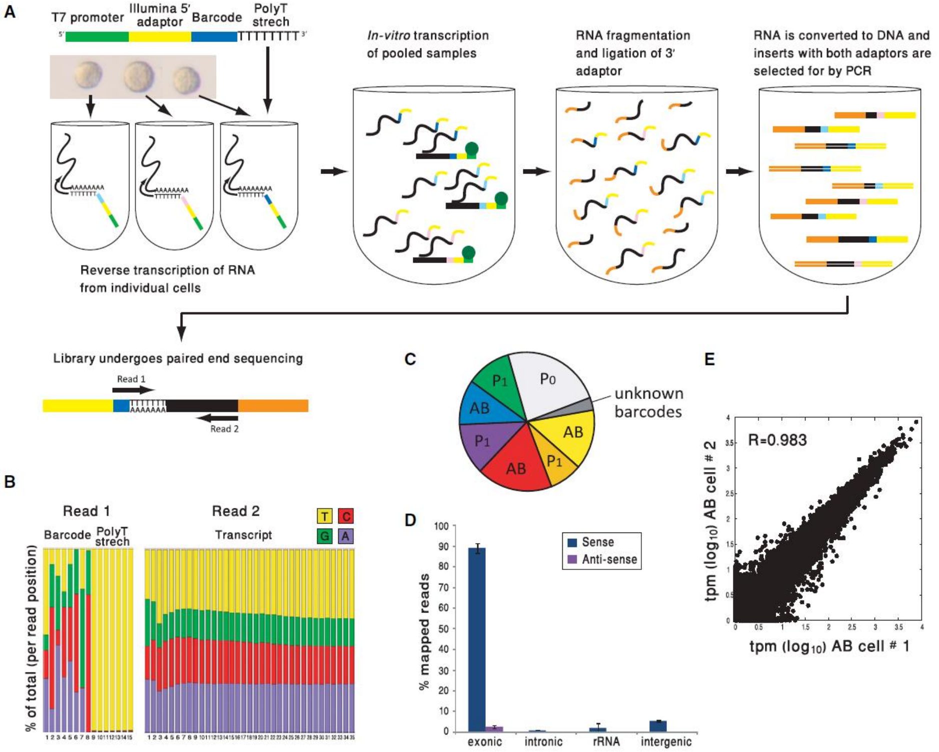
Assess the biological differences between cell clusters



Head and neck cancer tumors
Puram et ql, Cell, 2017

scRNA-seq technologies

- First released protocols aimed at sequencing one cell mRNA thanks to improved cDNA amplification and NGS (Lao, Nat Methods, 2009)
- Smart-seq (Ramsköld, Nat Biotechnol, 2021) enables the transcriptome profiling of roughly 100 cells
  - Full length cDNAs are captured, amplified, and sequenced by Illumina seq
  - One library per cell, typically >20 million uniquely mapped reads per cell
  - Roughly 8,000 genes per cell
- CEL-seq (Hashimshony, Cell Rep, 2012) introduces the idea of multiplexing the sequencing of several cells by appending a cell-specific barcode
  - Only the 3' end is sequenced
  - Multiplexing reduces the sequencing depth per cell, *i.e.*, the number of genes with reliable expression data

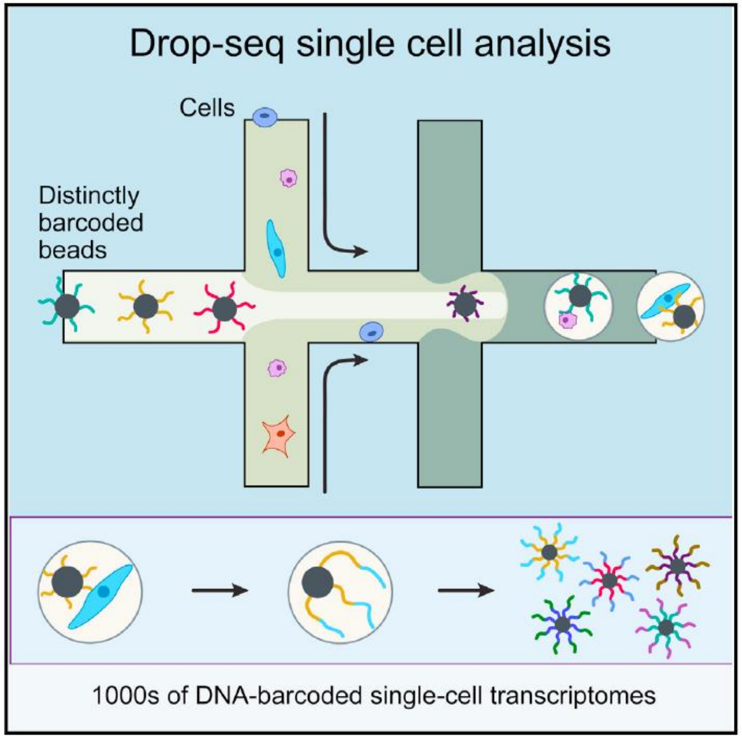

A
T7 promoter
Illumina 5' adaptor
Barcode
PolyT strech
5'
3'
Reverse transcription of RNA from individual cells
In-vitro transcription of pooled samples
RNA fragmentation and ligation of 3' adaptor
RNA is converted to DNA and inserts with both adaptors are selected for by PCR
Library undergoes paired end sequencing
Read 1
Read 2
AAAAAAA
B
Read 1
Barcode
PolyT strech
Read 2
Transcript
T
C
G
A
% of total (per read position)
C
P0
P1
AB
P1
AB
P1
AB
unknown barcodes
D
% mapped reads
Sense
Anti-sense
exonic
intronic
rRNA
intergenic
E
R=0.983
tpm (log10) AB cell # 2
tpm (log10) AB cell # 1

scRNA-seq technologies

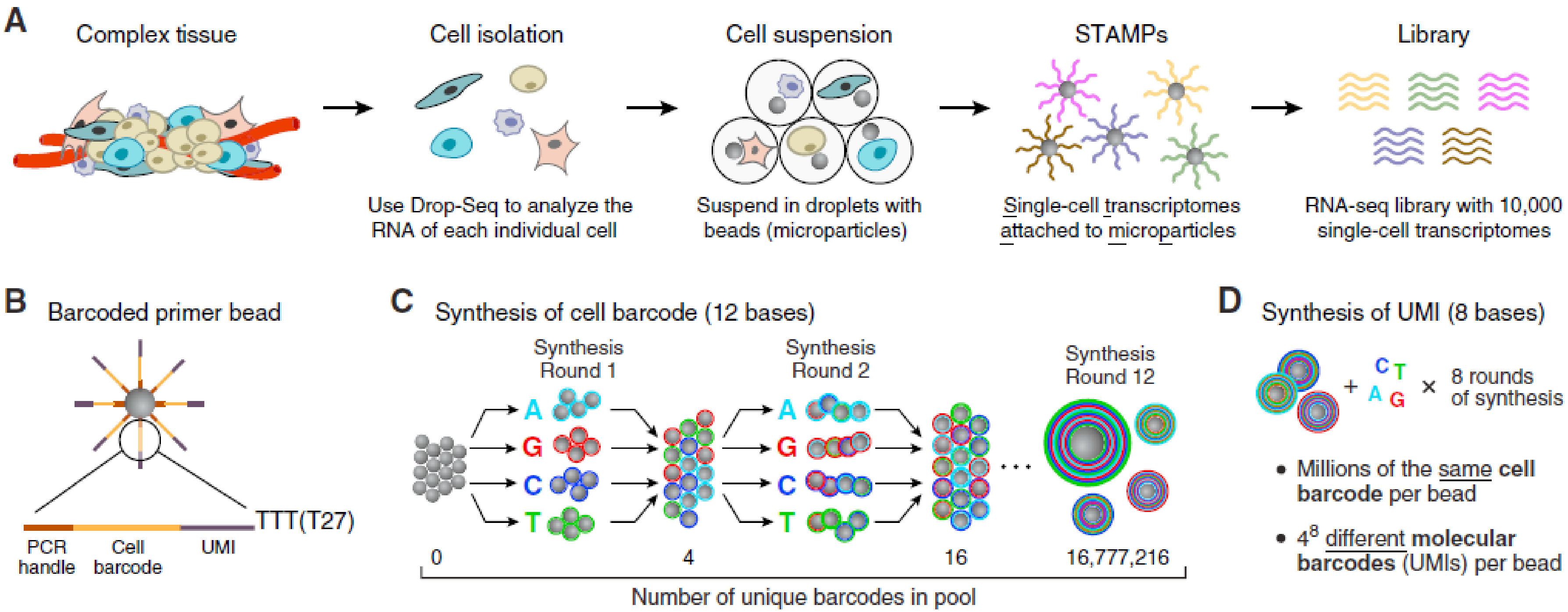
- DL-seq, QuartzSeq, MARS-seq, …
- Smart-seq-2 (Picelli, Nat Methods, 2013) improves sensitivity and full-length coverage of transcripts
- …
- Separating cells to prepare libraries requires microfluidics
- A commercial device is Fluidgm C1
  - Dissociated cells are loaded onto the device and flowed through microfluidic channels for size- and shape-dependent capture
  - In individual wells, cells are lysed, RT and cDNA synthesis are performed followed by PCR
  - Distinct cDNA libraries can be generated for each well
  - Introduction of a barcode in well and subsequent pooling is also possible

Droplet sequencing, highly multiplexed analyses

- A revolutionary idea (Macosko, Cell, 2015) is to accomplish all the cDNA and barcoding chemistry in a single droplet (containing one cell) instead of a physical well
- This allows to process thousands of cells instead of dozens when building a multiplexed library


Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

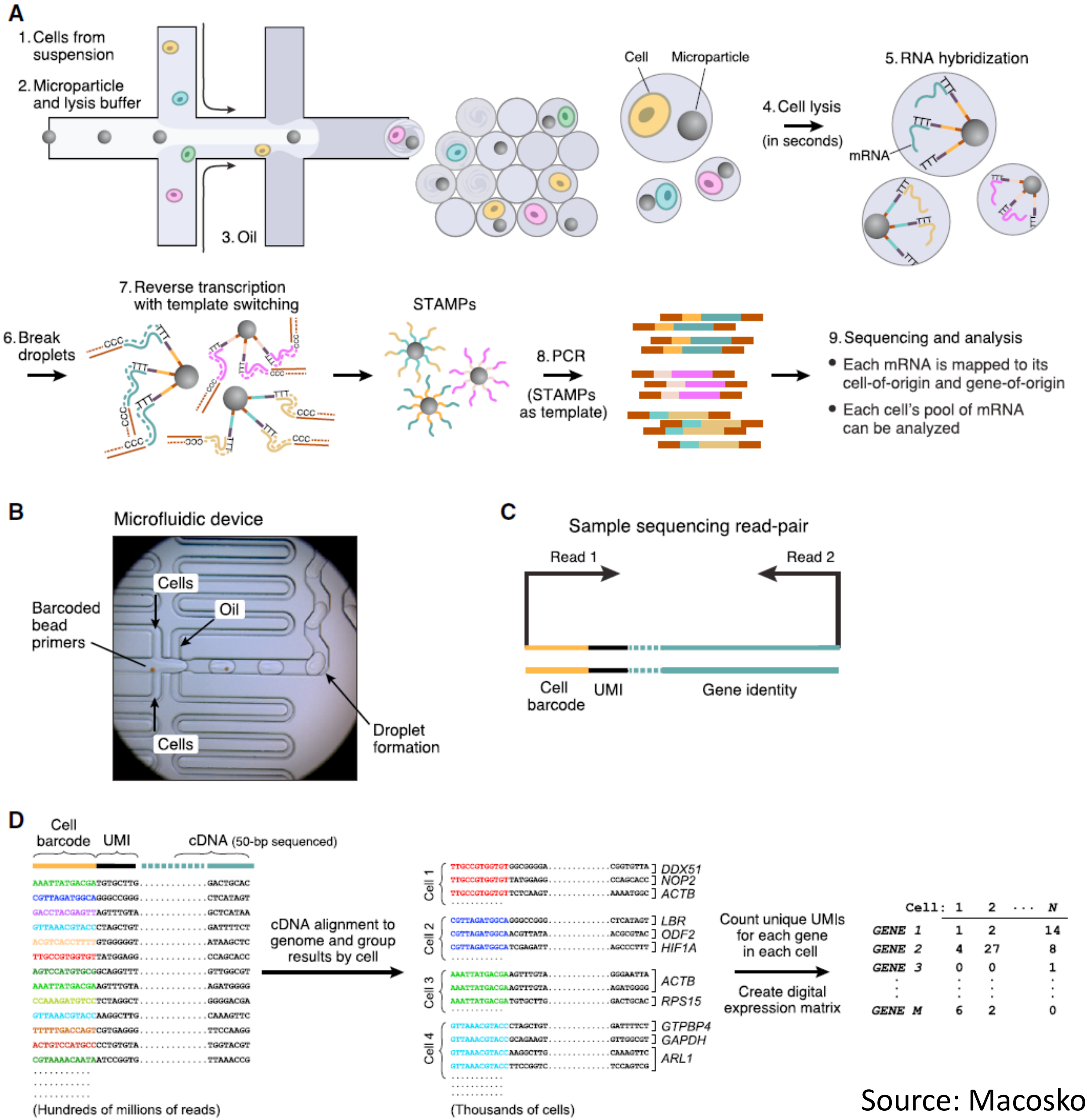
Drop-seq


A
Complex tissue
Cell isolation
Cell suspension
STAMPs
Library
Use Drop-Seq to analyze the RNA of each individual cell
Suspend in droplets with beads (microparticles)
Single-cell transcriptomes attached to microparticles
RNA-seq library with 10,000 single-cell transcriptomes
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool
D
Synthesis of UMI (8 bases)
+ C T A G × 8 rounds of synthesis
Millions of the same cell barcode per bead
4^8 different molecular barcodes (UMIs) per bead

Source: Macosko, Cell, 2015

- Beads are coated with millions of primers, each bead has a distinct cell barcode
- Each cell is isolated in a droplet together with a single bead (microparticle)
- Due to extensive PCR, sequence bias is likely to happen that would distort the transcriptomes (this protocol only captures the 3’ end of each gene!)
- UMIs are added to avoid this bias: after sequencing, for a given cell transcriptome, each UMI only counts once

Drop-seq

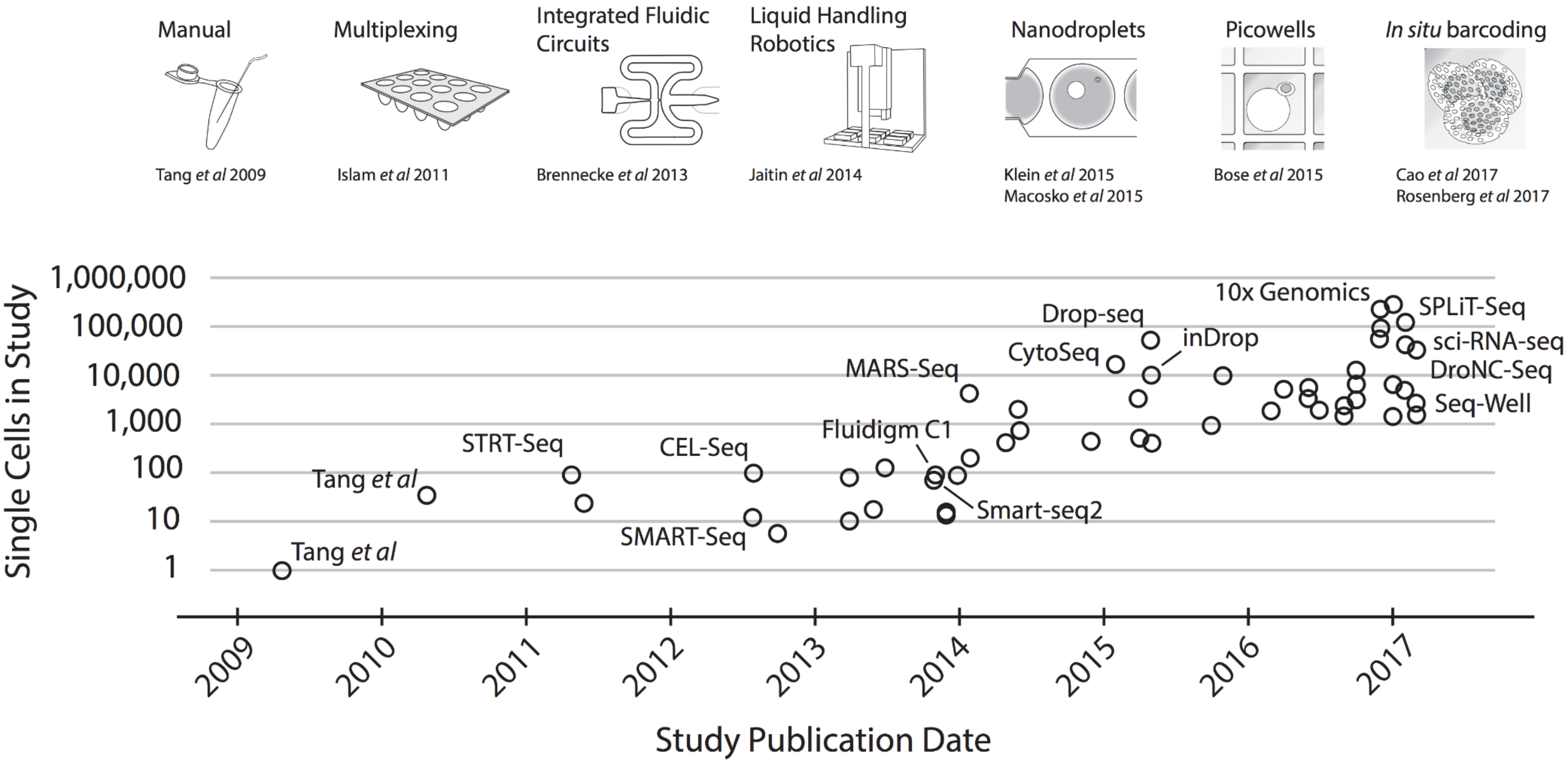

A
1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
6. Break droplets
7. Reverse transcription with template switching
CCC
STAMPs
8. PCR
(STAMPs as template)
9. Sequencing and analysis
• Each mRNA is mapped to its cell-of-origin and gene-of-origin
• Each cell's pool of mRNA can be analyzed
B
Microfluidic device
Cells
Oil
Barcoded bead primers
Cells
Droplet formation
C
Sample sequencing read-pair
Read 1
Read 2
Cell barcode
UMI
Gene identity
D
Cell barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG..............GACTGCAC
CGTTAGATGGCAGGGCCGGG..............CTCATAGT
GACCTACGAGTTAGTTTGTA..............GCTCATAA
GTTAAACGTACCCTAGCTGT..............GATTTTCT
ACGTCACCTTTTGTGGGGGT..............ATAAGCTC
TTGCCGTGGTGTTATGGAGG..............CCAGCACC
AGTCCATGTGCGGCAGGTTT..............GTTGGCGT
AAATTATGACGAAGTTTGTA..............AGATGGGG
CCAAAGATGTCCTCTAGGCT..............GGGGACGA
GTTAAACGTACCAAGGCTTG..............CAAAGTTC
TTTTTGACCAGTCGTGAGGG..............TTCCAAGG
ACTGTCCATGCCCCTGTGTA..............TGGTACGT
CGTAAAACAATAATCCGGTG..............TTAAACCG
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
TTGCCGTGGTGTGGCGGGGA..............CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG..............CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT..............AAAATGGC ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG..............CTCATAGT LBR
CGTTAGATGGCAACGTTATA..............ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT..............AGCCCTTT HIF1A
Cell 3
AAATTATGACGAAGTTTGTA..............GGGAATTA
AAATTATGACGAAGTTTGTA..............AGATGGGG ACTB
AAATTATGACGATGTGCTTG..............GACTGCAC RPS15
Cell 4
GTTAAACGTACCCTAGCTGT..............GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT..............GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG..............CAAAGTTC
GTTAAACGTACCTTCCGGTC..............TCCAGTCG ARL1
(Thousands of cells)
Count unique UMIs for each gene in each cell
Create digital expression matrix
Cell: 1 2 ··· N
GENE 1 1 2 14
GENE 2 4 27 8
GENE 3 0 0 1
GENE M 6 2 0

Source: Macosko, Cell, 2015

Drop-seq

- A commercial platform has been developed by 10x Genomics: Chromium
- This company sells reagent kits and a device to bring drop-seq to the broadest public
- Big commercial success, one device @ MGX in Montpellier
- At the moment, this is the most commonly used platform
- Even with deep sequencing, the high multiplexing limits the depth of individual transcriptomes to a few 1000 genes

Technological developments



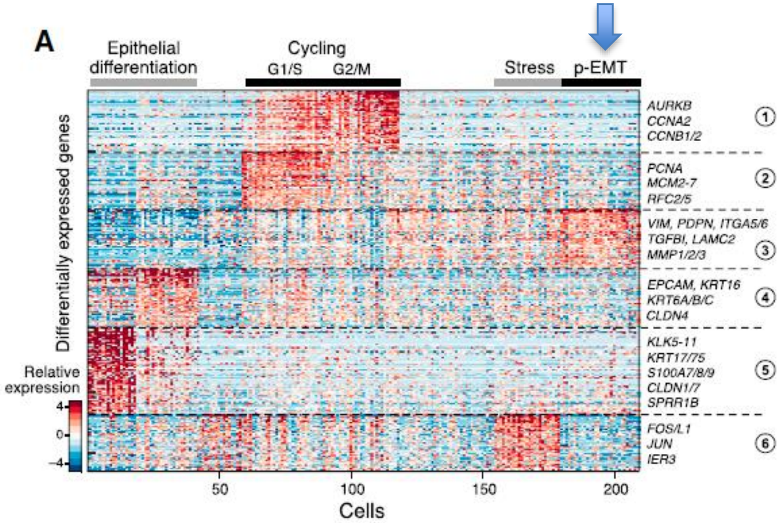
Source: Svensson, Nat Methods, 2017

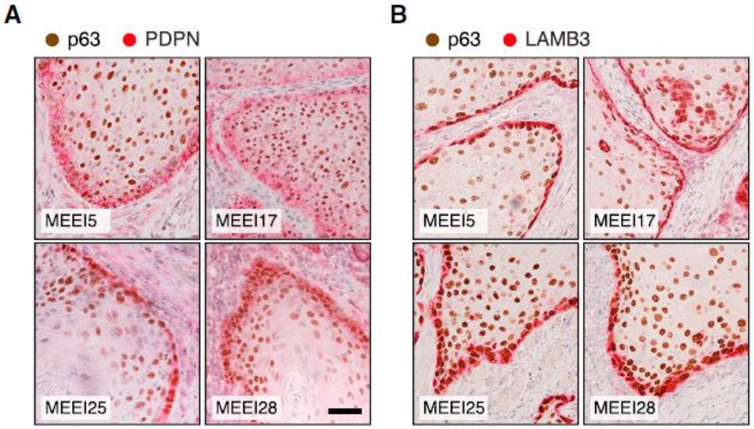
Why doing single-cell studies

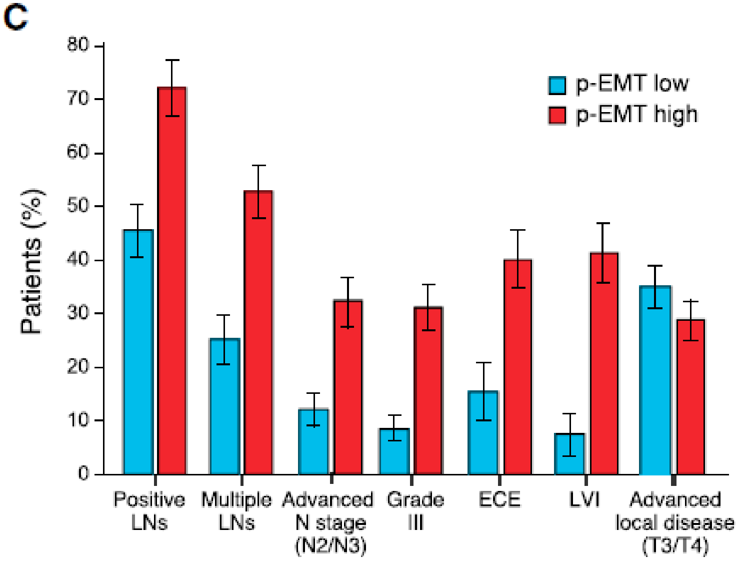
- To uncover the cellular composition of tissues
  - Bulk RNA-sequencing or proteomics only tell us about the average gene or protein expression in tissues (there are gene signatures but not accurate)
  - Techniques such as microdissections are tedious and not at single-cell resolution
  - Flow cytometry requires *a priori* known surface markers
  - Single-cell transcriptomics enable us to discover all (most) of the cell types present in a tissue *de novo*
  - This might reveal unexpected or new cell types that may have an important function in the tissue
  - The same cells in different states, *e.g.*, activated *versus* non-activated fibroblasts, are likely to display different transcriptomes → cellular state resolution
  - Variation in cellular composition might relate to important biological or clinical issues, *e.g.*, in tumors whose microenvironment composition may inform on treatment outcome (immunotherapy unlikely to work in immune poor tumors)
- To understand gene regulation at a single-cell level
  - Identical cells do not express all the same genes (stochastic control)
  - Differentiation processes, *e.g.*, hematopoiesis, gradients in tissues, cell localized at damages areas of an organ
  - Tissue and organism development

One application example

- Analyzing head and neck tumors, Puram et al, Cell, 2015, discovered a cluster of cells harboring a partial EMT program
- They found that those cells undergoing p-EMT were located at the leading edge of the tumors and they were engaging in cross-talk with CAFs







- Indeed, CAFs co-localized with cells expressing the p-EMT program and specific ligand-receptor interactions were found (TGFB3-TGFBR2, FGF7-FGFR2, CXCL12-CXCR7) that may promote EMT (ligands secreted by the CAFs)
- Finally, p-EMT signature was searched in large cancer transcriptomic data sets and found associated with nodal metastasis and adverse pathologic features

Data sources

- Transcriptomic, RNA-seq data repositories such as GEO include scRNA-seq ; many datasets can be downloaded from there

- Some large, community efforts have their own repositories, *e.g.*, the Human Cell Atlas

HUMAN
CELL
ATLAS

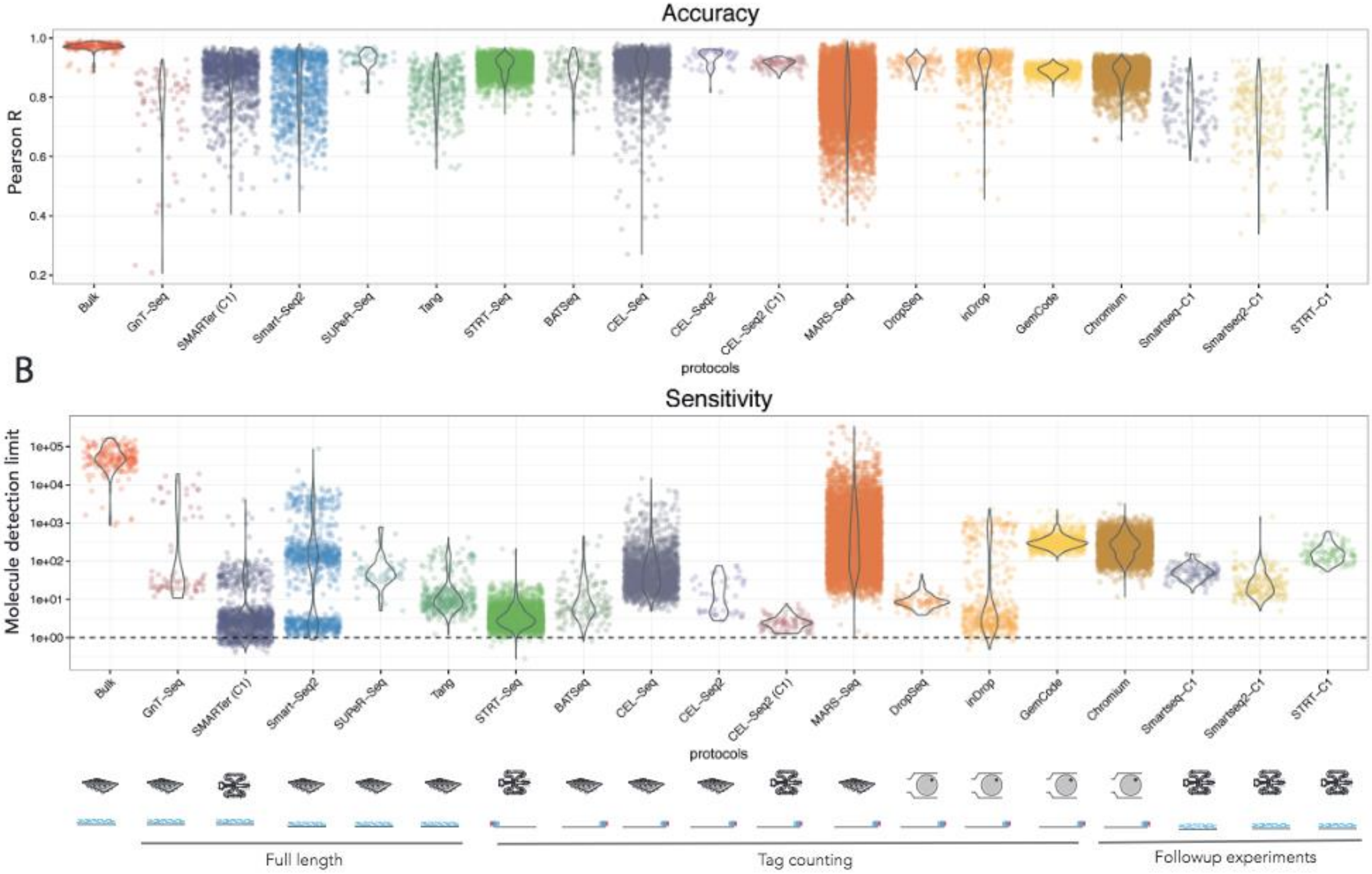
- Some authors or companies have started to integrate scRNA-seq data to offer query tools, and sometimes download functionalities

scRNA-seq data properties

- Due to limited depth (close to 10,000 genes for Smart-seq-2, but 800 to 4,000 genes for drop-seq), scRNA-seq are 0-inflated
- Due to a limit of detection, many genes remain undetected or data are too sparse to be used. This phenomenon is called dropout. One can also says that data are censored.
- This limitation must be taken into consideration in the data modeling (censored statistics) and biological interpretation
- Dropouts are also present in bulk RNA-seq, but much less prominent

- After read alignment and proper processing of individual cell barcodes as well as UMIs if present, scRNA-seq data take the form of a classical read count matrix, each column being a single cell and each row a gene
- Different data processing packages might export their output in different formats
- Data downloaded from GEO are usually in a tabular format. We will look at 10x Genomics CellRanger aligner output format in the exercises

Sequencing depth comparison



Accuracy = Pearson correlation coefficient with bulk RNA-seq ; sensitivity = number of detected genes
Source: Svensson, Nat Methods, 2017

Basic processing steps

- Alignment of the reads against the proper reference genome, management of cell barcodes and UMIs (if applicable)
- Filtering of dead cell (typically by eliminating cells with excessive mitochondrial gene contents)
- Filtering of doublets: two cells in a same well (seldom) or to cells in a same droplet (more frequent)
- Normalization (sequencing depth can be very different between cells, especially in drop-seq)
- Clustering and 2D-projection
- Cell type calling

- And project dependent analyses!

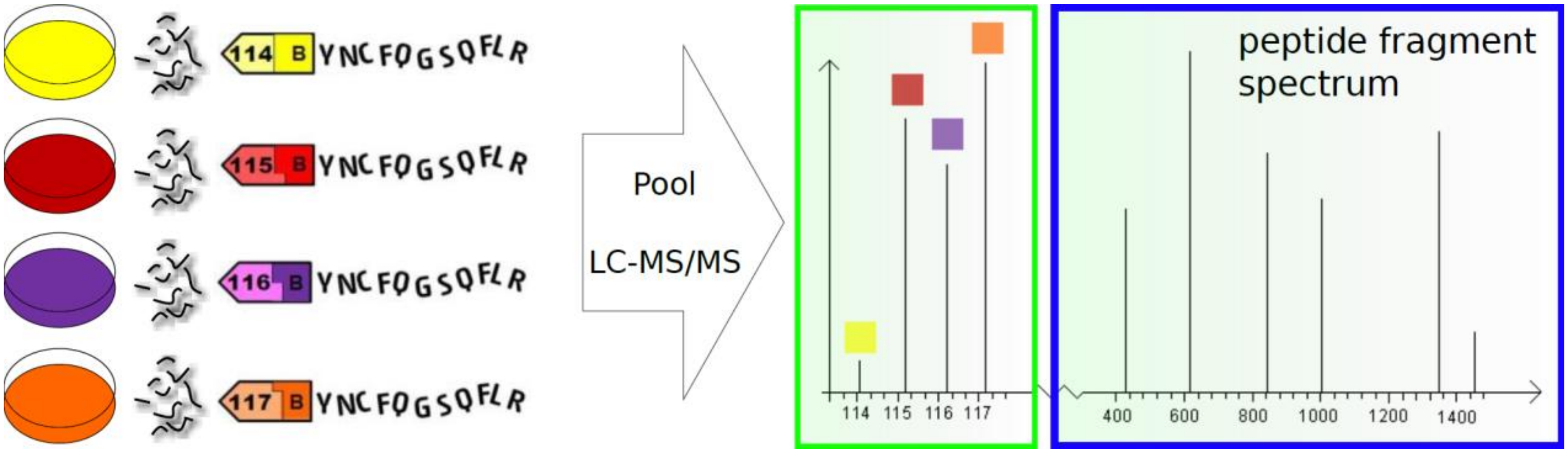
Tool sources

- Different complete or generic packages have been proposed to analyze scRNA-seq data
  - Seurat, open-access, very popular, the most widely used generic tool
  - Monocle, Scanpy, RaceID, ASAP, …
  - CellRanger & Loupe, solutions from 10x Genomics
  - There are also web servers: alona, Granatum, …

- Plenty of specific tools addressing one specific need
  - Some are interoperable or they can import data from Seurat or other generic tools (nice)
  - Can be much better than some of the integrated algorithms proposed by a generic package (often the case)
  - Check https://www.scrna-tools.org/

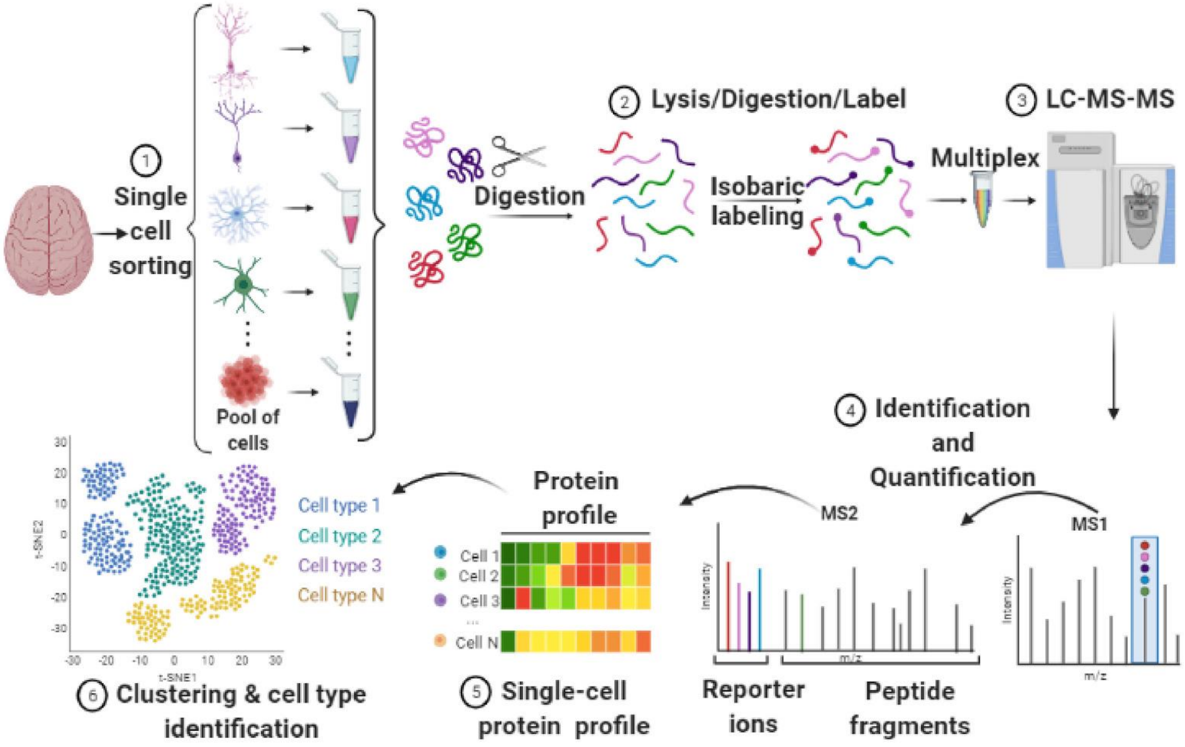
Can we do the same with proteins?

Multiplexed proteomics with isobaric labeling



Principle of single-cell proteomics

+ There is no PCR and we cannot introduce nucleotidic barcodes!
+ Use isobaric labels as barcodes



Source: Slavov Lab

+ We can only process 10-12 cells at a time!

Normalization issue

- Since each group of 10-12 samples is analyzed separately, advanced batch effect correction must be applied to pool multiple analyses in100- to 1000-cell data sets
- To analyze multiple samples in a pool increases detection sensitivity
- Also works with phospho-proteins
- ~1000 proteins detected/sample
- Now up to 3000-5000 proteins with the latest orbitrap astral and Bruker TIMS-TOF 2! This is comparable to transcriptomics

- Lastly, label-free attempts are developing based on data independent acquisition protocols applied to one cell at a time, less missing data and normalization issues, but more MS time!
- Same protein numbers