Single Cell Transcriptomics
ANGSD

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2CUdS9z[1]

March 19, 2019

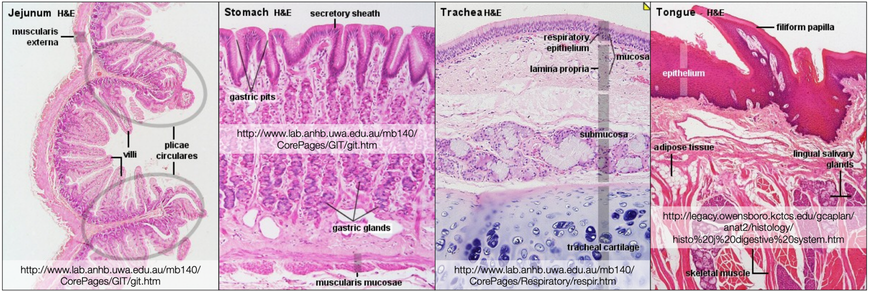
Weill Cornell Medicine

[1] http://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2018/

Why measure single cells?

Bulk RNA-seq returns the average expression of an entire cell population.

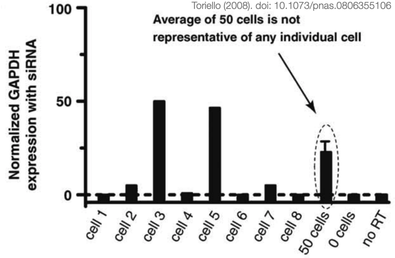
1. Tissues/organs[2] are usually made up of **very different** types of cells that are often **difficult to separate** prior to the experiment.
   - endothelial cells, osteocytes, myocytes, neurons, lymphocytes, macrophages, erythrocytes, oocytes, alveolar cells, chondrocytes, ...
   - stem cells, secreting cells, metabolizing cells, pacemaker cells, ...

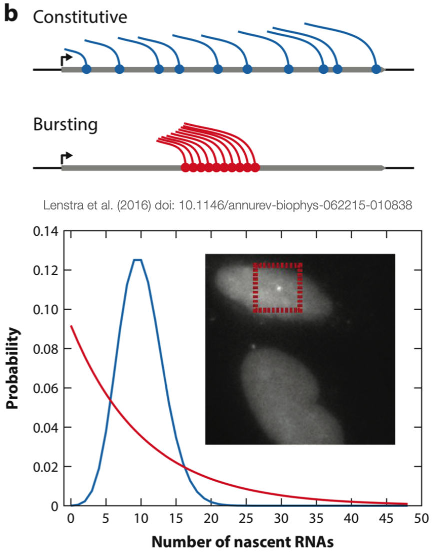


[2] Many solid tumors, too.

Why is bulk RNA-seq not enough?

2. Even very similar cells/clonal cell cultures display **heterogeneity at the molecular level** when interrogated at a defined time point.
   - cell cycle, age, exposure to environmental stimuli/stress, metabolic state





Why is bulk RNA-seq not enough?

The average behavior measured in millions of cells does not necessarily reflect the behavior in individual cells

In theory, we should therefore apply single-cell approaches to **all** studies of cells because **transcription** is, fundamentally, a **stochastic** process and mammalian cells are known to have non-continuous, **bursting** transcription, which inherently leads to variable cellular states.

Why is bulk RNA-seq not enough?

In practice, most scRNA-seq studies published to date deal with the higher-level complexities of organs and tissues:

- characterizing **developmental** processes
  - traditionally hampered by extremely low cell numbers
- cell type catalogues of **entire organs** or very **heterogeneous tissues**
  - pancreas, brain, liver, lung, retina
- **immune cell** studies
  - often coupled with single-cell clonotyping
  - helps distinguish numerous activation states of T/B cells
- **tumor** studies
  - so far, mostly distinguishing between malignant and physiological cells (e.g. infiltrating immune cells)

> "Single-cell analyses are needed to fully understand the cellular specificity and complexity of tissue microenvironments."

Why is bulk RNA-seq not enough?

In practice, most scRNA-seq studies published to date deal with the higher-level complexities of organs and tissues:

- characterizing **developmental** processes
  - traditionally hampered by extremely low cell numbers
- cell type catalogues of **entire organs** or very **heterogeneous tissues**
  - pancreas, brain, liver, lung, retina
- **immune cell** studies
  - often coupled with single-cell clonotyping
  - helps distinguish numerous activation states of T/B cells
- **tumor** studies
  - so far, mostly distinguishing between malignant and physiological cells (e.g. infiltrating immune cells)

"Single-cell analyses are needed to fully understand the cellular specificity and complexity of tissue microenvironments."

Why is bulk RNA-seq not enough?

In practice, most scRNA-seq studies published to date deal with the higher-level complexities of organs and tissues:

- characterizing **developmental** processes
  - traditionally hampered by extremely low cell numbers
- cell type catalogues of **entire organs** or very **heterogeneous tissues**
  - pancreas, brain, liver, lung, retina
- **immune cell** studies
  - often coupled with single-cell clonotyping
  - helps distinguish numerous activation states of T/B cells
- **tumor** studies
  - so far, mostly distinguishing between malignant and physiological cells (e.g. infiltrating immune cells)

"Single-cell analyses are needed to fully understand the cellular specificity and complexity of tissue microenvironments."

Why is bulk RNA-seq not enough?

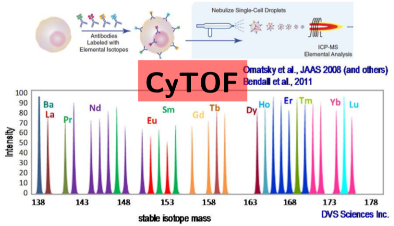
In practice, most scRNA-seq studies published to date deal with the higher-level complexities of organs and tissues:

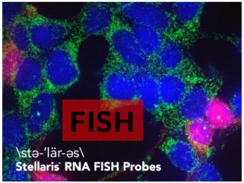
- characterizing **developmental** processes
  - traditionally hampered by extremely low cell numbers
- cell type catalogues of **entire organs** or very **heterogeneous tissues**
  - pancreas, brain, liver, lung, retina
- **immune cell** studies
  - often coupled with single-cell clonotyping
  - helps distinguish numerous activation states of T/B cells
- **tumor** studies
  - so far, mostly distinguishing between malignant and physiological cells (e.g. infiltrating immune cells)

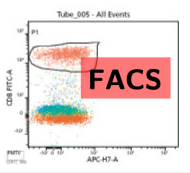
> "Single-cell analyses are needed to fully understand the cellular specificity and complexity of tissue microenvironments."

“Traditional” single-cell methods

Microscopy and **cytometry** have been used for decades to understand properties of single cells. The major limitations have been **throughput** and the number of **features** that could be assessed simultaneously.







	FACS	CyTOF	qPCR
Cell capture method	Laser	Mass cytometry	Micropipettes
Number of cells per experiment	Millions	Millions	300–1,000
Cost	\$0.05 per cell	\$35 per cell	\$1 per cell
Sensitivity	Up to 17 markers	Up to 40 markers	10–30 genes per cell

Papalexi & Satija (2018) doi: 10.1038/nri.2017.76

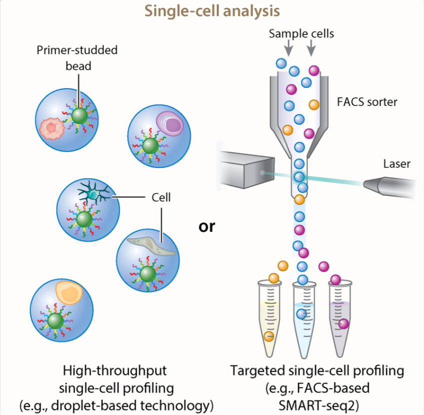
How to sequence the transcriptome of single cells?

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

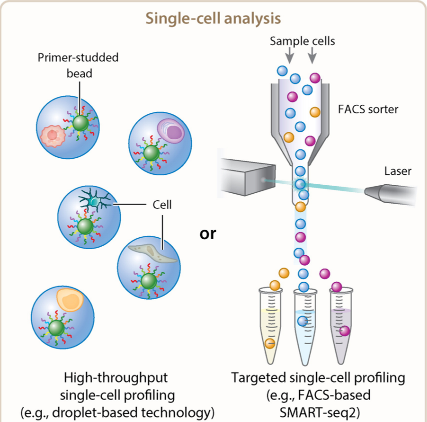
Major **innovations**:

- microfluidics
- random cell capture

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Major **innovations**:

- microfluidics
- random cell capture

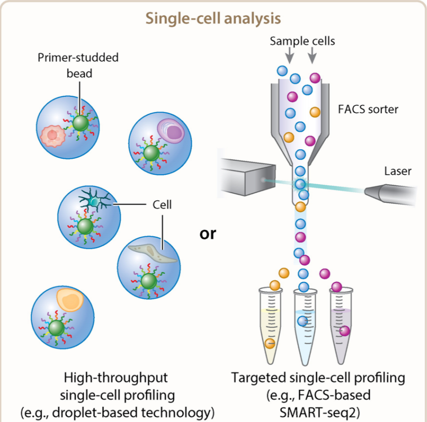
Details: Saliba et al. [2014] & Chen et al. [2018].

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Major **innovations**:

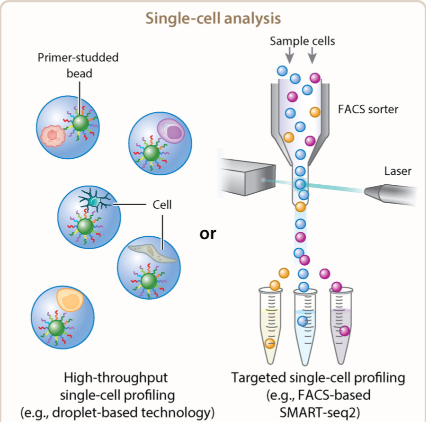
- microfluidics
- random cell capture

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Major **innovations**:

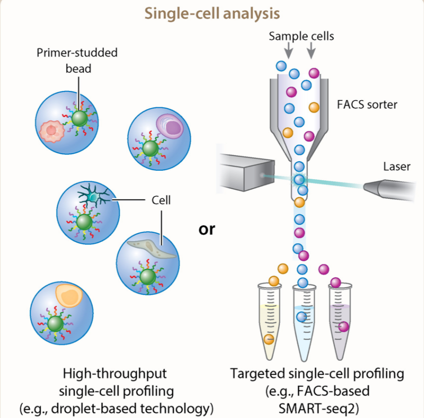
- microfluidics
- random cell capture

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Major **innovations**:

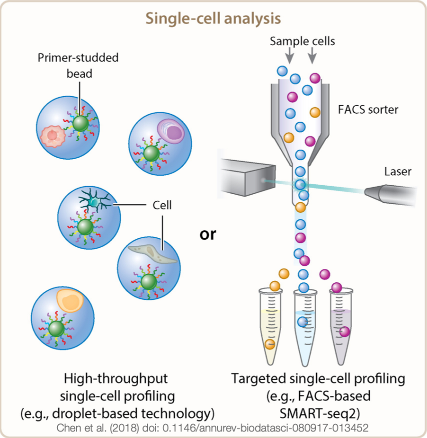
- microfluidics
- random cell capture

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].



Major **innovations**:

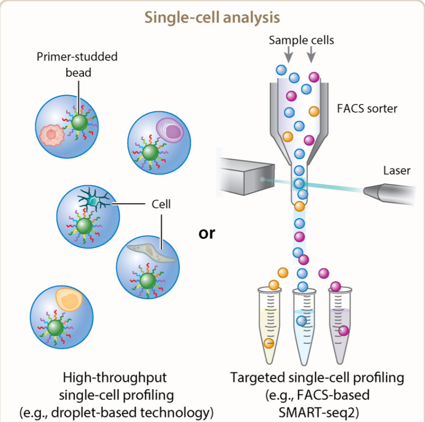
- microfluidics
- random cell capture

From bulk to single cell transcriptomes

The main **challenges**:

- automated **cell isolation**
  - FACS vs. microfluidics
- untargeted whole transcriptome **amplification**
  - required input: 0.1–1 µg total RNA
  - [RNA] per cell: 0.1–50 pg (!)
- **parallel processing**
  - individual cell lysis & RT carried out in wells (<100 cells), microchambers (Fluidigm chip), nanochambers, or droplets (>10,000s cells)

Details: Saliba et al. [2014] & Chen et al. [2018].

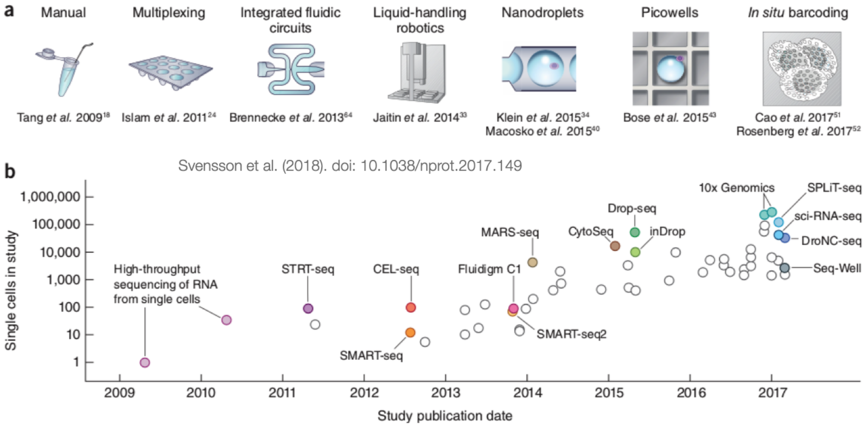


Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Major **innovations**:

- microfluidics
- random cell capture

Numerous solutions have been proposed in the past decade

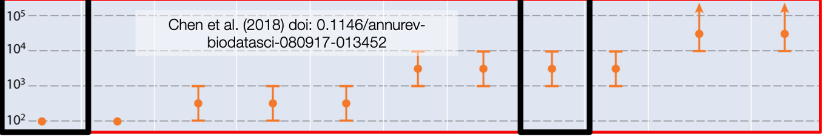


100s cells thanks to **multiplexing**, ca. 1,000 cells thanks to **fluidics**, 10,000s cells thaks to random cell captures techniques with **nanodroplets** and picowells, 100K cells thanks to ***in situ* barcoding**

The most popular scRNA-seq methods

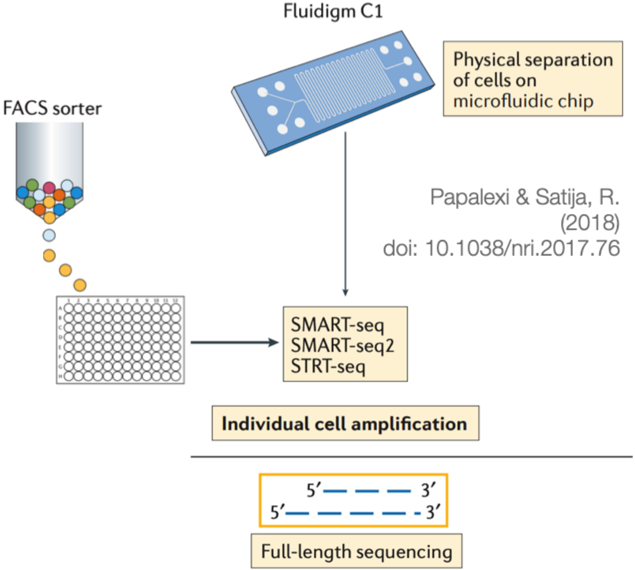
	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'

Number of cells per assay



Chen et al. (2018) doi: 0.1146/annurev-biodatasci-080917-013452

Smart-seq

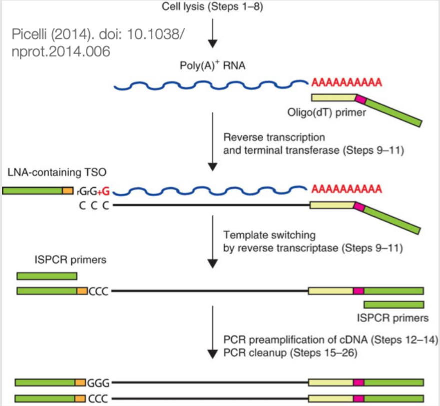


Papalexi & Satija, R. (2018)
doi: 10.1038/nri.2017.76

Smart-seq

RNA capture and cDNA synthesis

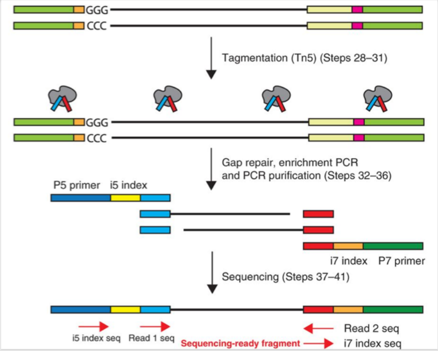
“SMART”: **S**witching **M**echanism **A**t the 5’ end of the **R**NA **T**ranscript



- RT reaction: poly(A) capture with oligo(dT) primer
- the MMLV-RT adds 3-4 C’s to the 3’ end of the **cDNA**
- these CCC hybridize with GGG-tail of template-switching-oligos (TSO)
- the TSO then serve as a template for the MMLV-RT to add the complementary sequence of the TSO plus universal PCR primers* to the cDNA (* same sequence as on the 5’ end of the cDNA)

Smart-seq

Library preparation

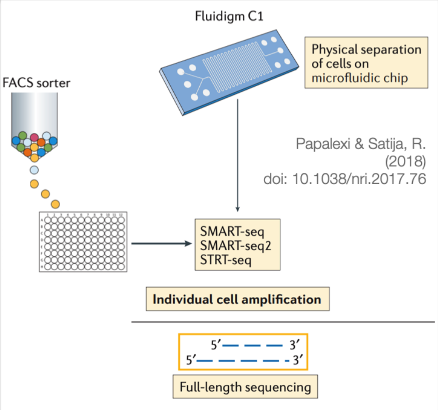


Picelli (2014). doi: 10.1038/nprot.2014.006

- amplification with few PCR cycles
- **tagmentation**: combining fragmentation and sequencing adapter integration
  - hyperactive derivative of the the Tn5 transposase **cuts** the cDNA and **ligates** sequencing adapters

Smart-seq

Summary



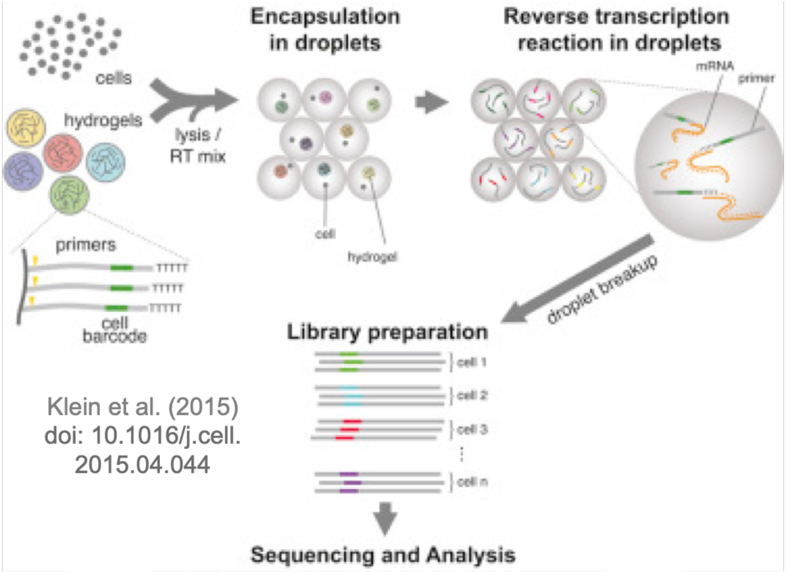
Advantages:

- high sensitivity
- full-length transcript sequencing
- usually better coverage per cell

Disadvantages:

- only poly(A)
- no strand-specificity
- somewhat "low-through-put": 100s of cells
- labor-intensive
- every cell gets its own library prep!

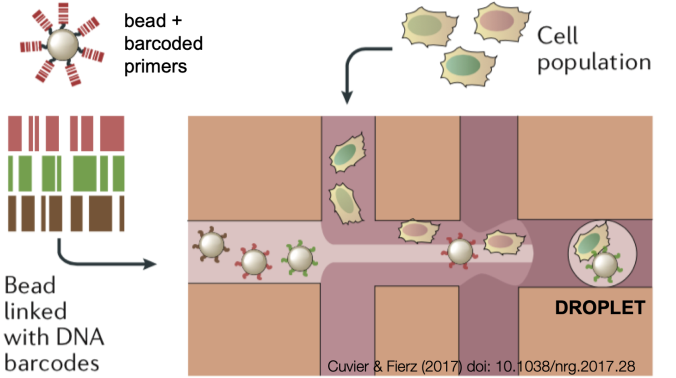
Droplet-based sequencing



Klein et al. (2015)
doi: 10.1016/j.cell.2015.04.044

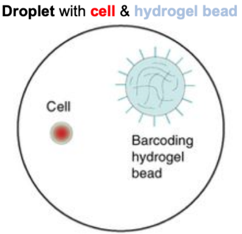
Droplet generation

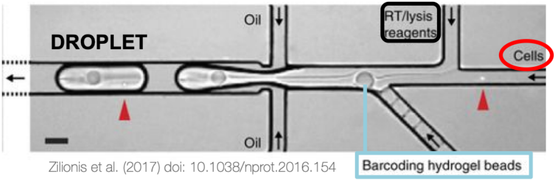
Using microfluidics, individual cells are captured together with a large set of (barcoded) poly(dT) primers (that are attached to hydrogel beads for the purpose of delivery).



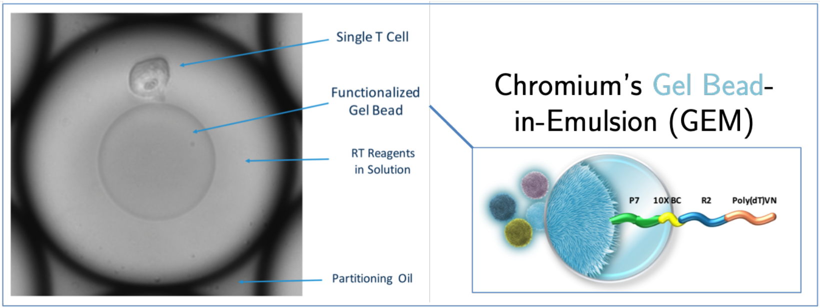
The final droplet contains **cell** + **primers** + **reagents** for cell lysis and RT.

Droplet content





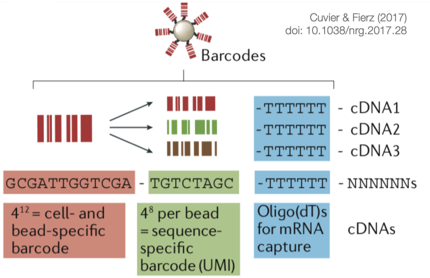
Zilionis et al. (2017) doi: 10.1038/nprot.2016.154



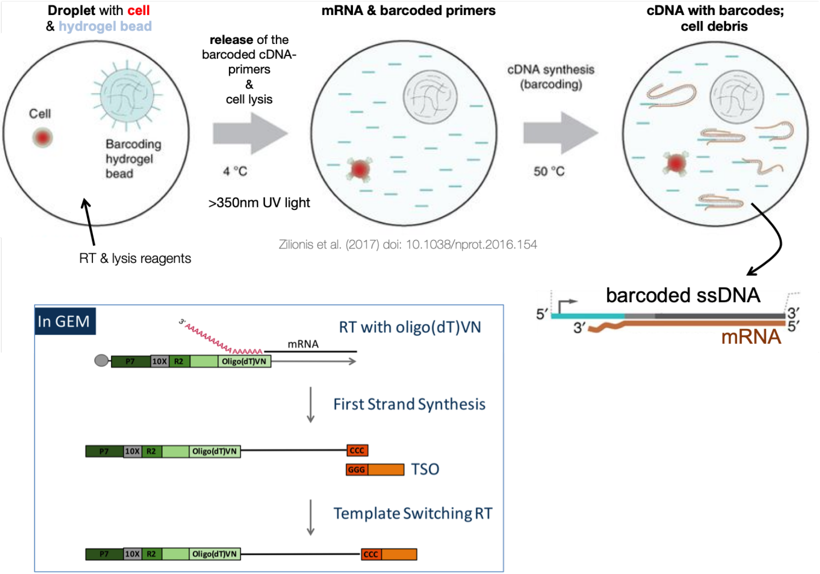
inDrop: Barcode details

Barcode diversity can be increased through multiple rounds of oligo-additions (see [Zilionis et al., 2017] for details).

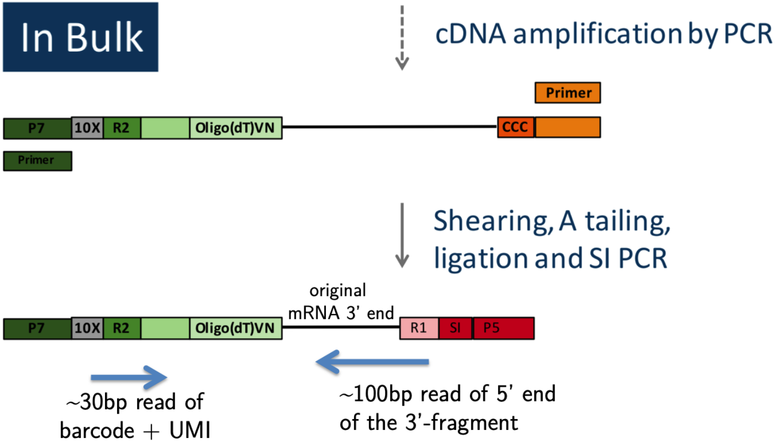
1. bead-specific **barcode** (–> cell)
2. primer-specific unique molecular identifier (**UMI**) (–> individual transcripts!)
3. (Illumina adapters)
4. **oligo(dT)** for poly(A)-mRNA capture



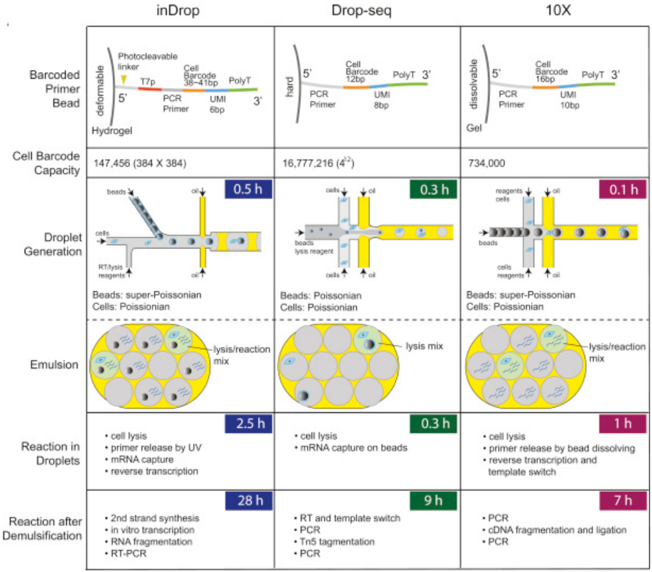
Droplets: Capturing and barcoding mRNA transcripts



Droplets: Library preparation



Comparing the most popular droplet-based technologies



Zhang (2019). doi: 10.1016/j.molcel.2018.10.020

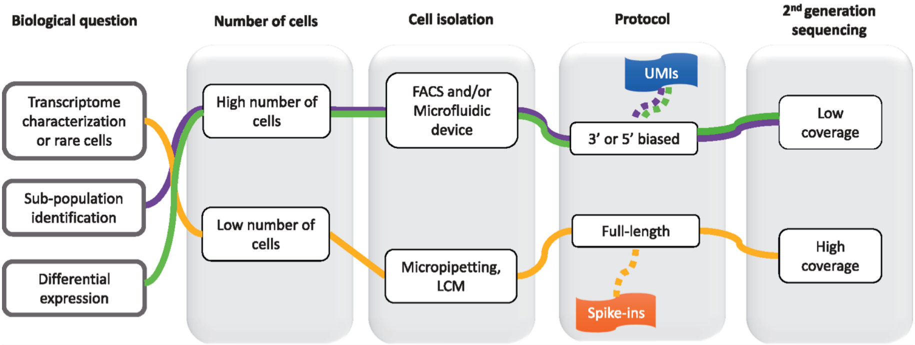
Pros:

- 1000s of cells
- fairly fast, highly automated
- UMI = no PCR bias!

Cons:

- 3' coverage only
- shallow seq. depth per cell

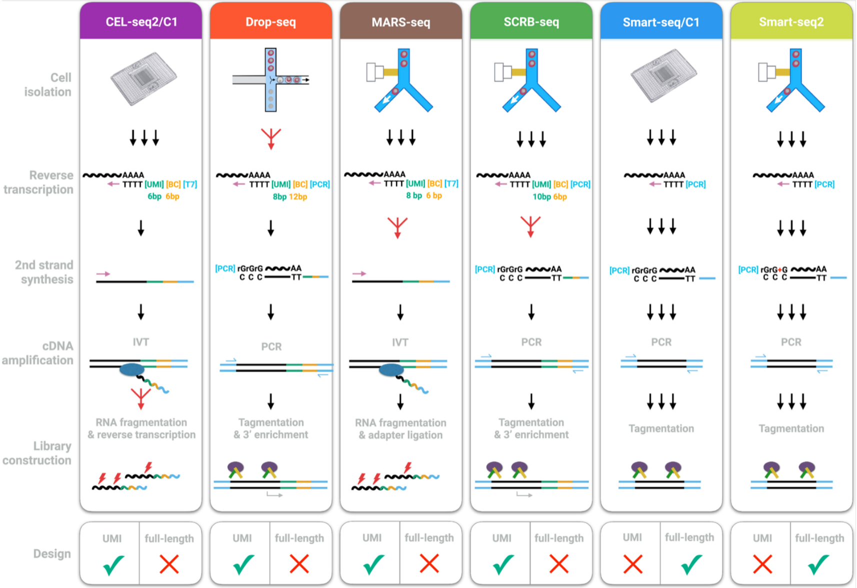
How to choose



Dal Molin (2018). doi: 10.1093/bib/bby007

See Chen et al. [2018], Svensson et al. [2018], Ziegenhain et al. [2017], Zhang et al. [2019] for good overviews and reviews of different platforms.

The most popular scRNA-seq methods



Ziegenhain (2017). doi: 10.1016/j.molcel.2017.01.023

The ideal single-cell transcriptomics method

From Beltrame et al. [2019]:

Feature	Smart-Seq2	10X Chromium
Universal in terms of cell size, type and state.	not yet	not yet
In situ measurements.	not yet	not yet
No minimum input of number of cells to be assayed.	🙂	🙁
Every cell is assayed, i.e. 100% capture rate.	🙂	😐
Every transcript in every cell is detected, i.e. 100% sensitivity.	🙁	🙁
Every transcript is identified by its full-length sequence.	🙂	🙁
Transcripts are assigned correctly to cells, e.g. no doublets.	🙂	😐
Additional multimodal measurements.	not yet	not yet
Cost effective per cell.	🙁	🙂
Easy to use.	😐	😐
Open source.	🙂	🙁

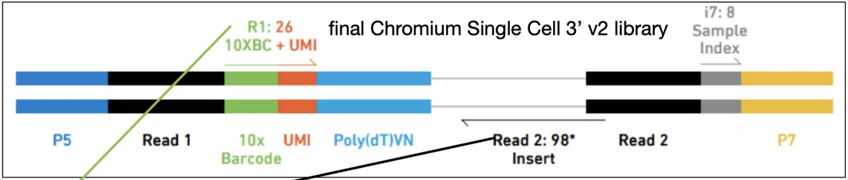
Obviously, the optimal solution does not exist. Pick the one that matches your needs most closely.

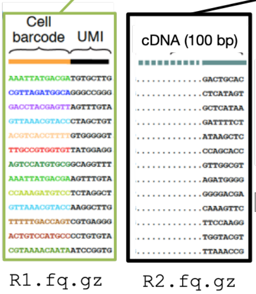
Future improvements

- **in situ barcoding**, which will make the cell isolation step redundant
- **molecular "crowding"*** within the RT-reaction chambers (e.g. droplets) to increase the capture efficiency of the transcripts
- **nuclear isolation** to reduce noise and increase the range of sample types that can be processed

What to do with scRNA-seq raw data?

Processing raw reads: “tagging” & aligning





- **cDNA alignment**, e.g. with STAR
- **tagging reads** with cell barcode & UMI info

10X Chromium provides CellRanger software for this

BAM file with

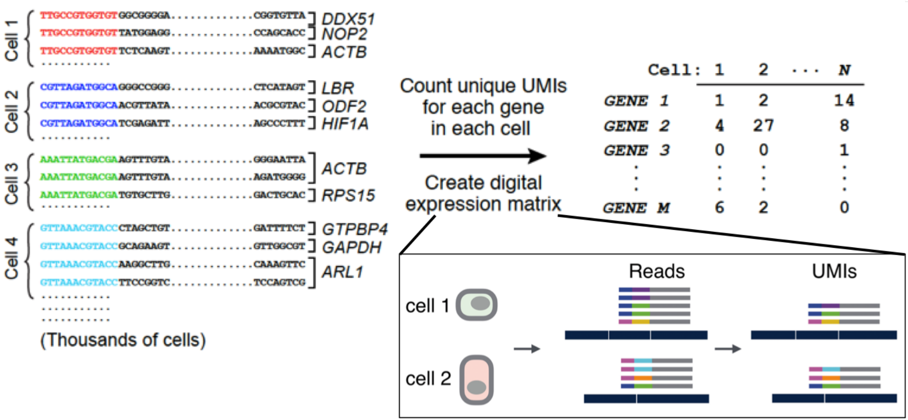
- **genome locations** of the reads
- every read has its corresponding information, e.g. original and error-corrected **barcode** and **UMI** sequences etc.

See https://github.com/mccrowjp/Dropseq, https://github.com/beiseq/baseqDrops, https://hemberg-lab.github.io/scRNA.seq.course/processing-raw-scrna-seq-data.html and Tian et al. [2018] for **pipelines** other than CellRanger.

Processed raw reads: how CellRanger stores the BC/UMI info

Tag	Type	**Description** cellular and molecular barcode information for each read is stored as `TAG` fields
CB	Z	Chromium cellular barcode sequence that is error-corrected and confirmed against a list of known-good barcode sequences.
CR	Z	Chromium cellular barcode sequence as reported by the sequencer.
CY	Z	Chromium cellular barcode read quality. Phred scores as reported by sequencer.
UB	Z	Chromium molecular barcode sequence that is error-corrected among other molecular barcodes with the same cellular barcode and gene alignment.
UR	Z	Chromium molecular barcode sequence as reported by the sequencer.
UY	Z	Chromium molecular barcode read quality. Phred scores as reported by sequencer.
BC	Z	Sample index read.
QT	Z	Sample index read quality. Phred scores as reported by sequencer.
TR	Z	Trimmed sequence. For the Single Cell 3' v1 chemistry, this is trailing sequence following the UMI on Read 2. For the Single Cell 3' v2 chemistry, this is trailing sequence following the cell and molecular barcodes on Read 1.

Generating the count matrix



How to QC and process the count matrices of scRNA-seq?

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

Main issues regarding CELLS and barcodes

- **barcode collisions** = 2 droplets with the same barcode
  - increase barcode diversity!
- barcode **sequencing errors**
- **empty droplets** = no cell was captured
  - in theory, these should yield 0 UMI
  - in practice, there's often plenty of **ambient** RNA[3] that will be amplified and sequenced
- **doublets** = 2 cells captured in the same well/droplet
  - will get the same barcode
  - resulting transcriptome for this barcode will be a random sample of both cells
  - influenced by the flow-rate during droplet generation
  - usually around 5%
  - often impossible to detect
- **overrepresentation** of mitochondrial transcripts = dying cells/cells with lots of background noise

[3] Mostly released from dying cells [Lun et al., 2018].

CellRanger's basic QC

Estimated Number of Cells
11,769

Mean Reads per Cell	Median Genes per Cell
54,286	1,906

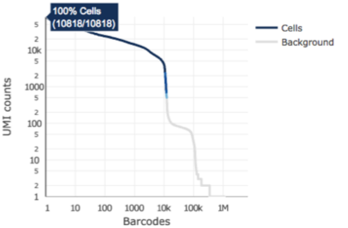
Sequencing

Number of Reads	638,901,019
Valid Barcodes	97.4%
Sequencing Saturation	68.2%
Q30 Bases in Barcode	93.7%
Q30 Bases in RNA Read	90.1%
Q30 Bases in Sample Index	90.1%
Q30 Bases in UMI	92.4%

Mapping

Reads Mapped to Genome	95.5%
Reads Mapped Confidently to Genome	92.5%
Reads Mapped Confidently to Intergenic Regions	5.0%
Reads Mapped Confidently to Intronic Regions	34.7%
Reads Mapped Confidently to Exonic Regions	52.7%
Reads Mapped Confidently to Transcriptome	49.7%
Reads Mapped Antisense to Gene	1.3%

Cells



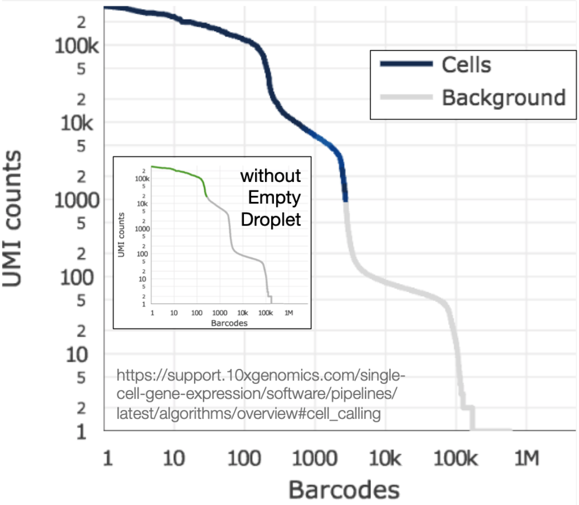
Estimated Number of Cells	11,769
Fraction Reads in Cells	95.1%
Mean Reads per Cell	54,286
Median Genes per Cell	1,906
Total Genes Detected	23,036
Median UMI Counts per Cell	6,521

Sample

Name	pbmc_10k_v3
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Transcriptome	GRCh38
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.0

https://support.10xgenomics.com/img/single-cell-gex/web-summary-gex-3.0a.png

Separating empty droplets from truly amplified cell content



300 high RNA content 293T cells were mixed with 2,000 low RNA content PBMC cells

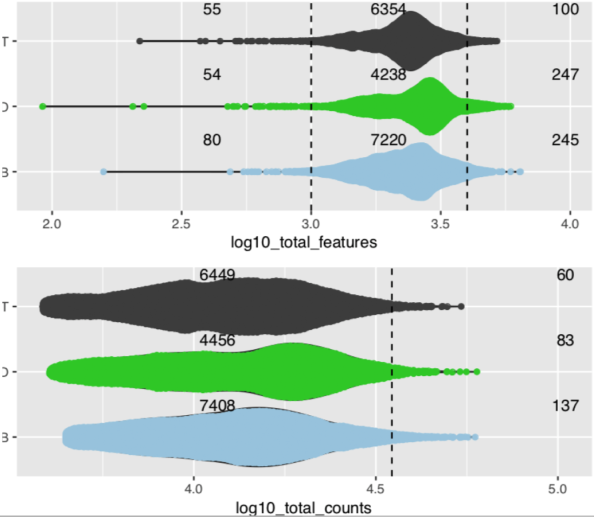
How to separate background & signal:

1. nUMI cut-off
2. determining difference from ambient droplet RNA content

Based on insights from Lun et al. [2018].

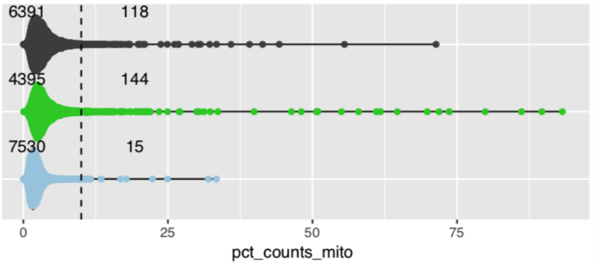
General QC: cells

All of this is optional, mostly based on "folklore" and not thoroughly benchmarked!



- remove cells with very **low UMI** counts – these days (as of Jan 2019), 10X is actually doing a pretty good job here
- remove cells with very **few genes**
- remove cells with very **high UMI** counts and genes – suspected doublets

QC: cells



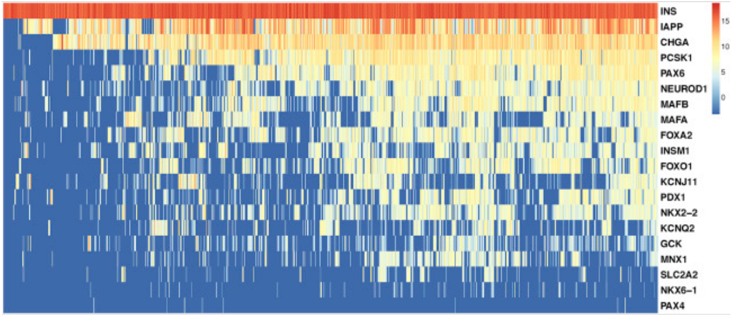
- remove cells with very **high mitochrondrial** content

The vignette of the bioconductor package scater offers a good QC workflow [Lun et al., 2016].

Main issues regarding GENES

- UMI **sequencing errors**
- **dropouts** = undetected transcripts
  - false negatives
  - nearly impossible to distinguish from true negatives
  - very common and not restricted to lowly expressed genes

Heatmap of β-Cell Markers Genes in β-Cells (Fluidigm 800HT)

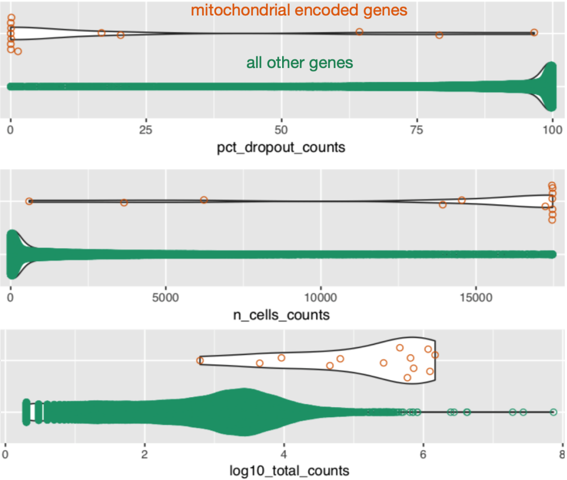


Wang & Kaestner (2018) doi: 10.1016/j.cmet.2018.11.016

All genes shown here are *known* to be expressed in pancreatic β cells.

QC: genes

Gene dropouts are VERY COMMON and NOT restricted to lowly expressed genes!

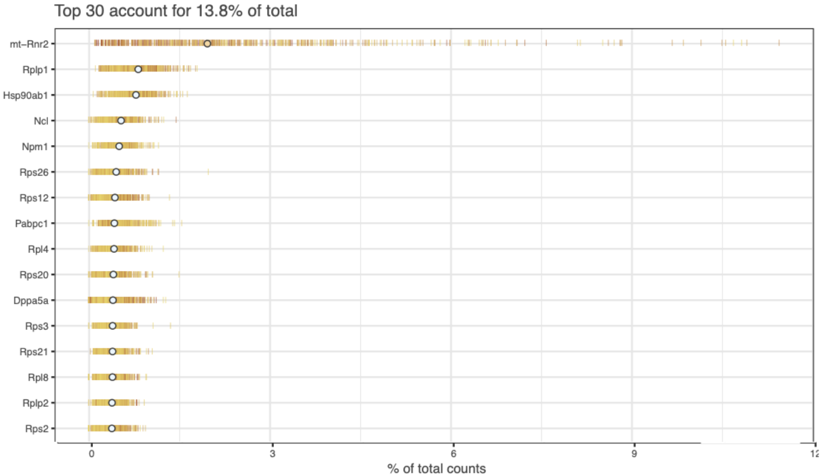


Currently, scRNA-seq is **not a transcriptome-wide** method; it is a technique that will return a **sample** of a cell's transcriptome! [Andrews and Hemberg, 2018]

Remove genes with extremely low capture rates because they can distort downstream analyses & identify possibly contaminating transcripts.

QC: genes (sanity check)

The most strongly expressed genes should encompass ribosomal proteins (!) and mitochondrial (housekeeping) genes and ideally some of the typical marker genes known for your sample type.



Normalization

Typical factors that influence downstream analyses are:

- **number of UMI/genes** within a cell – not just for technical reasons, this also correlates with cell size and general RNA content of a cell!
- biological factors: **cell cycle** status, cell size
- technical **batch effects** such as time of preparation, experimenter, sequencing lane/machine/day

Technical noise affecting the cell-wide profiles is difficult to estimate because every single cell (of every experiment) is considered a biological replicate.

For **biological confounders**, it's almost impossible to find a consensus of whether to ignore them or not.

Normalization: accounting for differences in seq. depth

Seurat's[4] workflow:

- `NormalizeData(object = mySeuratObjectWithCountMatrix)`
  - divide by sequencing depth (`colSums`) and
  - log-transform

```
## Normalization
log1p(double(it.value()) / colSums[k] * scale_factor)
```

- `ScaleData(object = mySeuratObjectWithCountMatrix)`
  - optional: regressing out possible confounders/non-interesting variables, e.g. cell cycle status, nUMI
  - z-score transformation

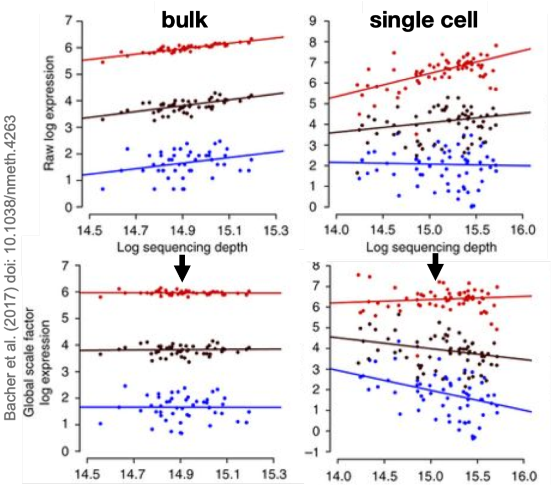
```
## Scaling -------------------------------
if(scale == true){
      if(center == true){
        rowSdev = sqrt((r - rowMean).square().sum() / (mat.cols() - 1));
      }
scaled_mat.row(i) = (r - rowMean) / rowSdev;
```

[4] https://cran.r-project.org/web/packages/Seurat/index.html

Regressing: accounting for systematic confounders

```
### extract the factor to regress out (could also be present in rownames)
latent.data <- latent.data[colnames(x = object), , drop = FALSE]
### create formula for regression
vars.to.regress <- colnames(x = latent.data)
fmla <- paste('GENE ~', paste(vars.to.regress, collapse = '+')) %>% as.formula

regression.mat <- cbind(latent.data, data.expr[1,])
colnames(regression.mat) <- c(colnames(x = latent.data), "GENE ~")

### fit a linear model and extract only the QR decomposition
qr <- lm(fmla, data = regression.mat, qr = TRUE)$qr
rm(regression.mat)

### Make results matrix
data.resid <- matrix(nrow = nrow(x = data.expr), ncol = ncol(x = data.expr))

### extract residuals via the function qr.resid --> QR decomposition of a matrix
regression.mat <- qr.resid(qr = qr, y = data.expr[x,])
data.resid[i, ] <- regression.mat
dimnames(x = data.resid) <- dimnames(x = data.expr)
```

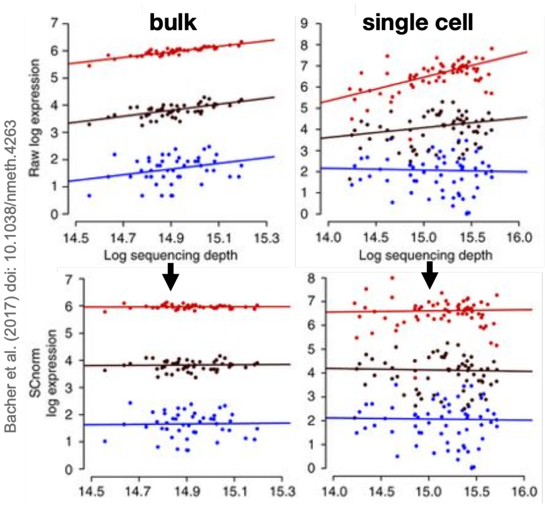
Normalization: effect of global scale factor



scRNA-seq shows systematic variation between transcript-specific expression & sequencing depth ("**count-depth relationship**")

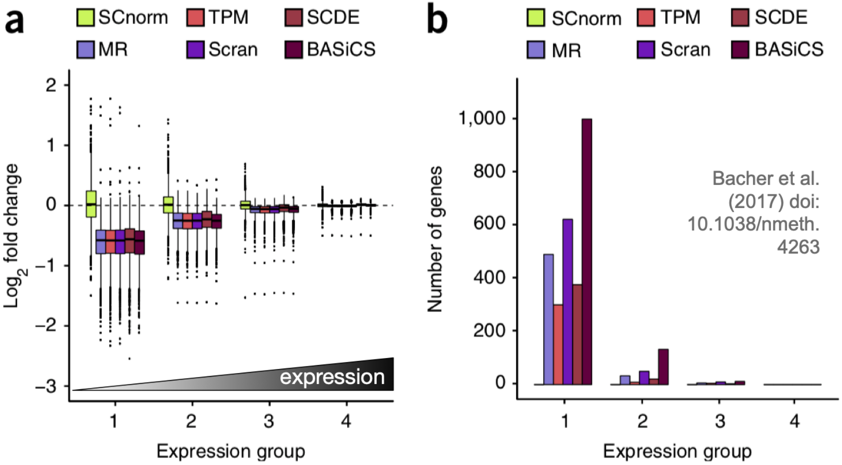
Global scale factor works well for bulk RNA-seq, but less so for scRNA-seq

Normalization: applying different scale factors for different groups of genes

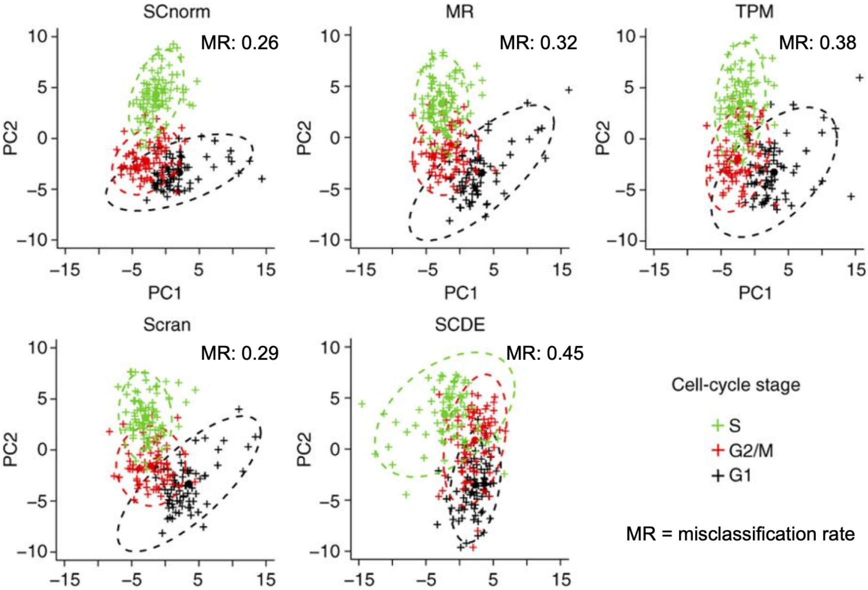


scNorm caclulates different scale factors for different groups of genes (grouping based on count-depth-relationships)

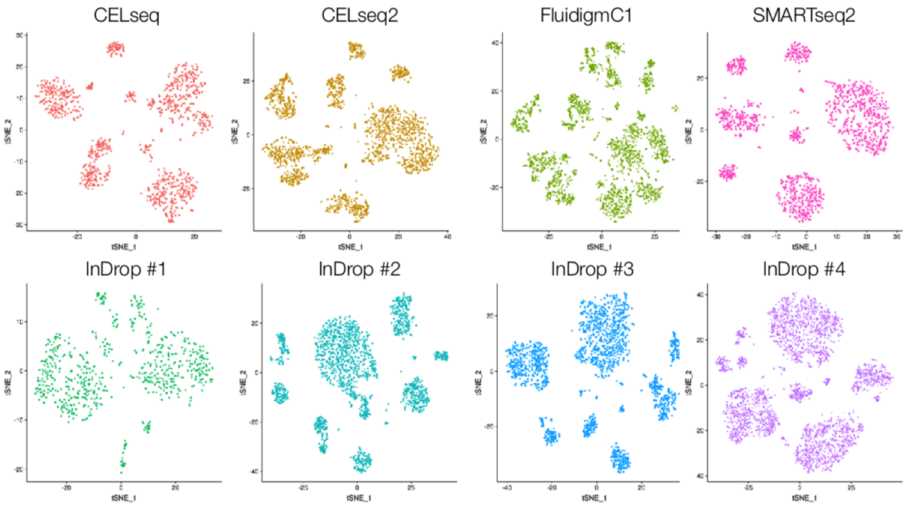
Normalization: effect on logFC and marker gene detection



Normalization: effect on PCA & clustering



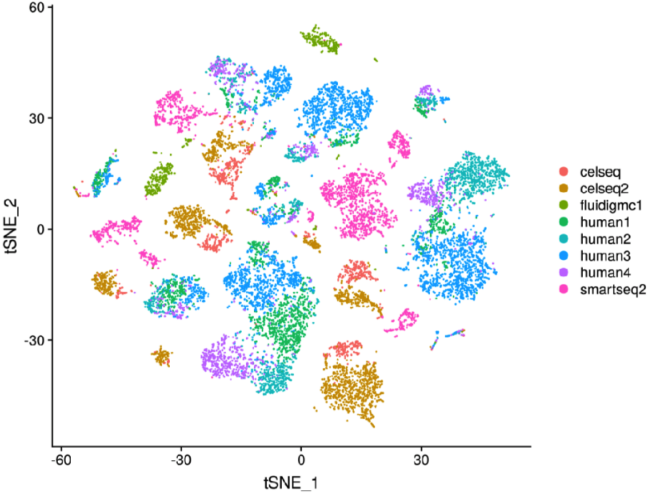
Batch correction



Data from Baron et al. 2016, Cell Syst.; Lawlor et al. 2017, Genome Res.; Grün et al 2016, Cell Stem Cell; Muraro et al. 2016 Cell Syst.

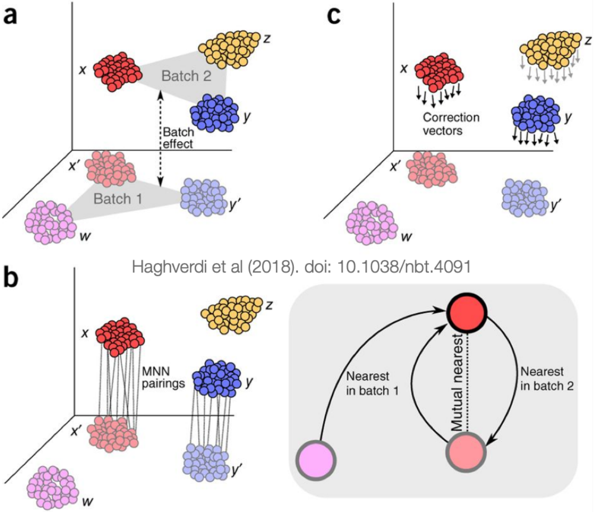
images courtesy Tim Stuart

Batch correction for integrative analyses



All samples were derived from pancreas. Merging all samples into one matrix without additional batch correction will lead to artificial clusters.

Batch correction for integrative analyses



Haghverdi et al (2018). doi: 10.1038/nbt.4091

1. Mutual Nearest Neighbors = most similar cells across batches

2. mean difference between cells in an MNN pair $\sim$ batch effect

3. correction vector applied to the expression values = batch correction

Summary of basic count matrix processing steps

Filtering cells

- require certain # UMI and genes per cell
- remove cells with high mitochondrial content

Filtering genes

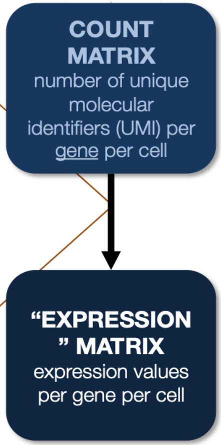
- require minimal detection threshold for individual genes

Adjusting for different library sizes (N) per cell

- e.g. scNorm (Bacher 2017), scater (Lun 2016)

Possibly batch effect removal/sample alignment

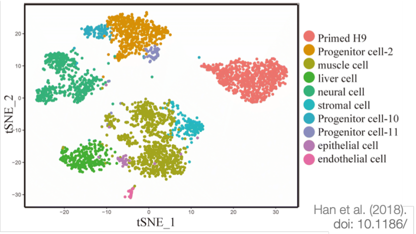
- e.g. MNNcorrect (Haghverdi 2018), Seurat v3 (Stuart 2018)



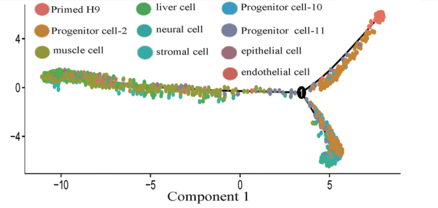
How to draw biologically meaningful insights from scRNA-seq?

Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt), tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

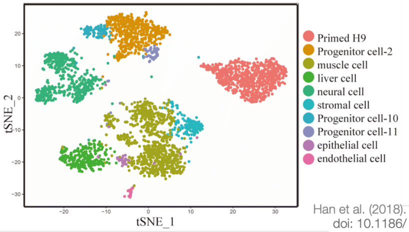


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

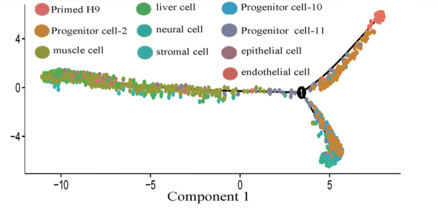


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

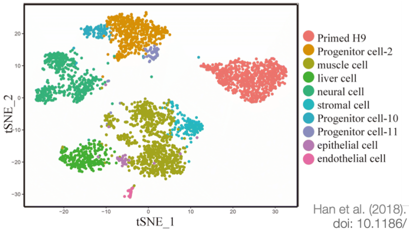


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

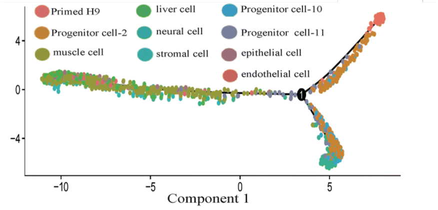


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

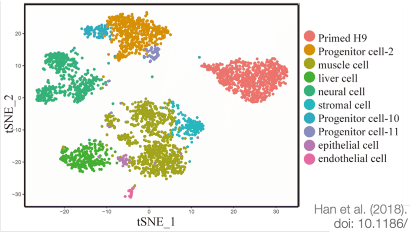


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

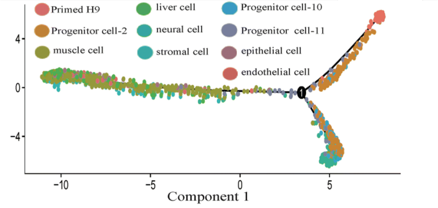


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

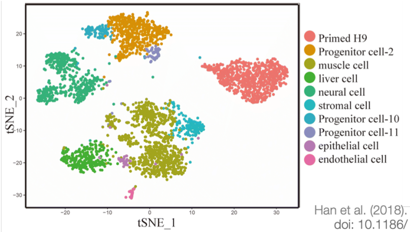


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

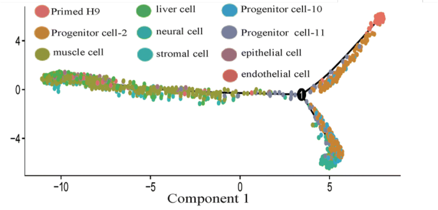


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

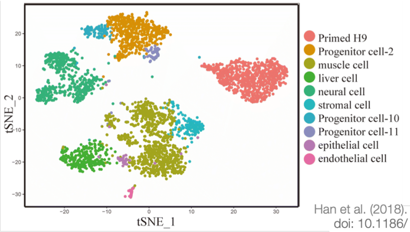


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

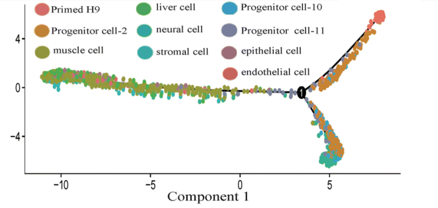


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

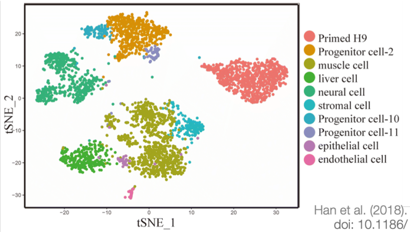


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

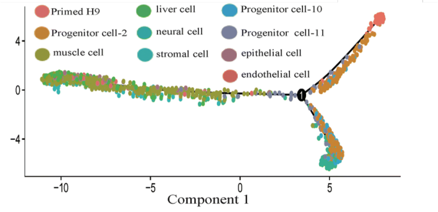


Identifying cell types and/or cell states of interest

- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature

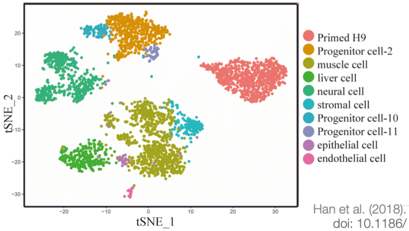


Han et al. (2018). doi: 10.1186/s13059-018-1426-0

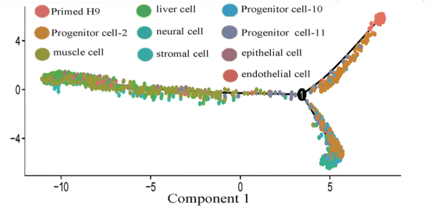


Identifying cell types and/or cell states of interest

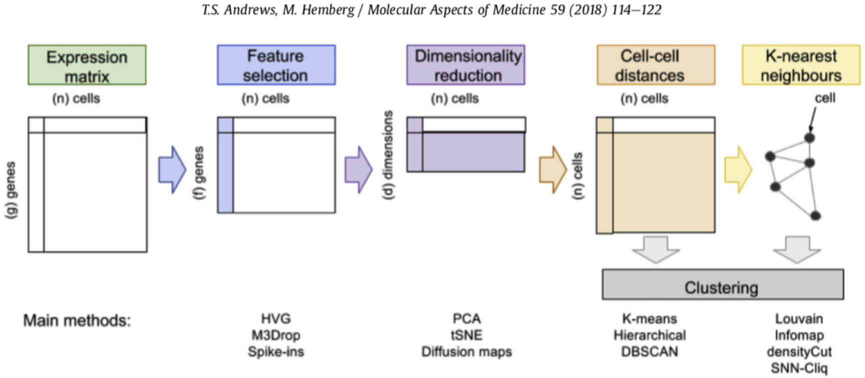
- insights are usually based on:
  - visualizations of **dimensionality reduction**
    - PCA (bit.ly/2W7XHwt) , tSNE (bit.ly/2Hm2ZRF, https://distill.pub/2016/misread-tsne/), UMAP (bit.ly/2qGhBkk), Diffusion Maps
  - **clustering**
    - k-means, hierarchical clustering, graph-based community detection
  - **marker gene identification**
    - DGE detection between clusters of interest
    - GO term & pathway enrichment analyses, comparison to literature



Han et al. (2018). doi: 10.1186/s13059-018-1426-0



Common workflow for identifying clusters



See Shekhar and Menon [2019] for a detailed Seurat-based workflow; https://hemberg-lab.github.io/scRNA.seq.course/biological-analysis.html for an even more detailed protocol using both bioconductor packages as well as Seurat.

Clustering methods implemented for scRNA-seq

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3[22]	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2[115], RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

DBSCAN, density-based spatial clustering of applications with noise; PCA, principal component analysis; scRNA-seq, single-cell RNA sequencing.

Kiselev (2019). doi: 10.1038/s41576-018-0088-9

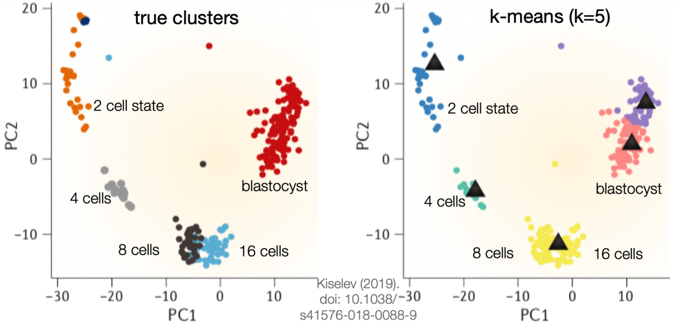
For assessments of the different clustering techniques for scRNA-seq data, see Freytag et al. [2018], Duò et al. [2018], Menon [2018].

No size fits all, but Seurat works reasonably well for high-throughput, droplet-based approaches.

Clustering

k-means

- pairwise similarity measures (e.g. 1−Pearson corr)
- cells are iteratively assigned to the nearest cluster center, followed by recomputation of the new cluster center (centroid)
- very fast
- number of clusters must be pre-determined

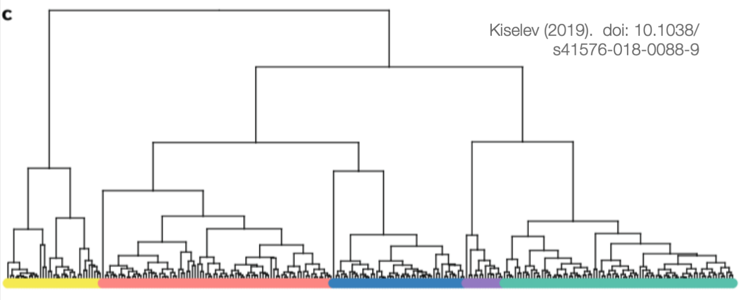


Kiselev (2019). doi: 10.1038/s41576-018-0088-9

Clustering

Hierarchical clustering

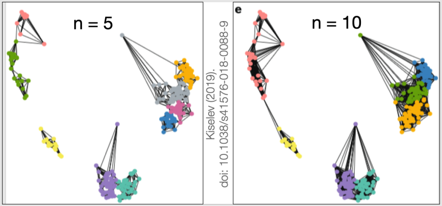
- can determine relationships between clusters of different granularities
- difficult to determine at which level to cut the tree, may suggest artificial clusters within cells of the same cell population
- prohibitively computationally expensive to use hierarchical clustering for large data sets



Clustering

Graph clustering/community detection

- clusters = groups of **nodes** that are densely connected
- density is a user-specified parameter
- works well on many (>1000) cells



1. select the top x PCs that capture the majority of the *gene* signatures
2. construct a graph where nodes = *cells*, edges = similarity measures (based on PCs)
3. for every cell, identify its k-nearest-neighbours (**SNN** graph), i.e. every cell::neighbor pair gets a weight that captures the similarity of the two cells' neighborhoods (that consist of k NN each!)
4. use the iterative Louvain community detection method to identify groups of nodes that are densely connected

See Andrews and Hemberg [2018] and Kiselev et al. [2019] for details for the clustering techniques.

Clustering

Graph clustering/community detection

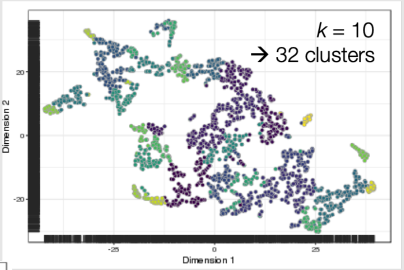
Dimensionality reduction with **PCA**

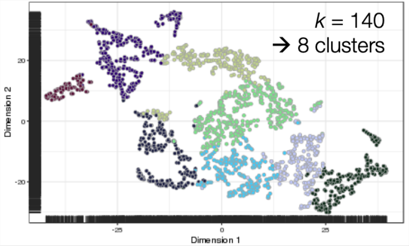
↓

pairwise cell-cell distances based on shared **nearest neighbors**

↓

identifying **communities** of densely connected cells (Louvain)

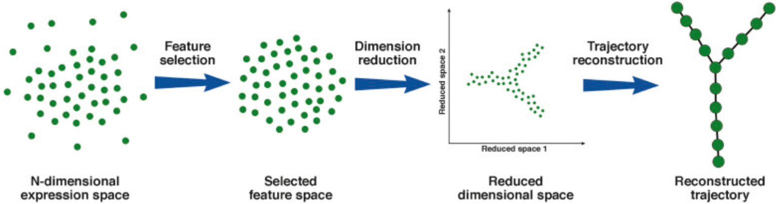




- How many PCs for the graph construction?
- **How many (k) neighbors for the initial graph?**
- How many iterations of the Louvain algorithm?

Pseudotime

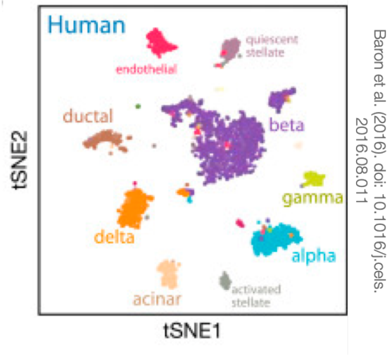
- based on concept of **diffusion maps**
  - data points are represented (embedded) into Euclidean space where the distance between two points will reflect their connectivity/similarity [Haghverdi et al., 2015, Angerer et al., 2016, Herring et al., 2018]
- can handle non-linear processes; more appropriate than clustering for **continuous** data along a **trajectory**
- pseudotime != real time [5]
- absolutely depends on **cells representing the transitional states** to be present in the data!



[5] A longer branch can simply reflect a lineage with more cells.

Getting some feeling for replicability and biological significance of CELL TYPES/POPULATIONS

- repeated runs (incl. different tools) of clusterings etc. will only give you an idea of the **technical** robustness of your parameter choices
- cell types may be compared across different species
- *known* marker genes may give some insights into significance of individual clusters



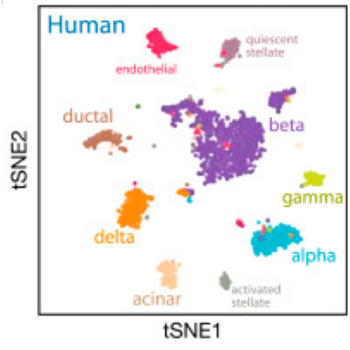
If cell types differ by few genes, we will not pick them up!

Getting some feeling for replicability and biological significance of MARKER GENES

Typical cell identity signals are robust & low-dimensional! [Crow and Gillis, 2018, Heimberg et al., 2016]

- ca. 100 genes: distinguish glia vs. neurons (1st PC)
- ca. 1,000 genes: distinguish neuron subtypes (PC1-3)

The genes you identify as "markers" may just have highly correlated expression patterns with the true drivers of the cell identity.

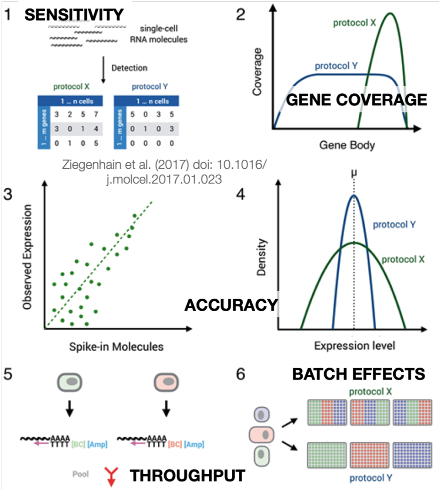


Baron et al. (2016). doi: 10.1016/j.cels.2016.08.011

Novel marker gene identifications must be followed up by **additional experiments**.

Conclusions

Every scRNA-seq technique has unique pros & cons



Ziegenhain et al. (2017) doi: 10.1016/j.molcel.2017.01.023

Decision will depend on:

- sample availability
- experimental question
- access to the method
- possibly previously published studies

Limits of scRNA-seq

- Technical challenges
  - **sensitivity** is still low
  - **costs** are still somewhat prohibitive
- Numerous sources of **cell-to-cell variability**
  - cell cycle
  - cell size
  - transcription bursts
  - stress during isolation
- **Analysis methods** are in their infancy!

Have a rationale!

What is your **hypothesis**? How are you going to distinguish transient from permanent effects? Do you have a way of obtaining some idea of the "ground truth"?

When NOT to use scRNA-seq (yet?)

- fairly **homogeneous populations**, true interest is in identifying the main effect of a treatment/condition/genotype...
- **complex experimental designs** (e.g., many experimental variables)
- genes of interest are known to be **lowly expressed**/subtly changing

Beware!

If you are interested in **individual *genes***, scRNA-seq should **not** be your first choice.

See Lafzi et al. [2018] for lots of practical advice before planning your own scRNA-seq experiment!

Examples of publicly available scRNA-seq data collections

Consortia-style efforts:

- Tabula muris
- Human Cell Atlas
- Single Cell Expression Atlas
- Allen Brain Map

Repositories for **published data sets** (providing processed data):

- Single Cell Portal (Broad Institute) – processed by the individual groups themselves
- Conquer – uniformly processed samples, includes QC reports! [Soneson and Robinson, 2018]

References

[Andrews and Hemberg, 2018, Bacher et al., 2017, Chen et al., 2018, Coulon et al., 2013, Haghverdi et al., 2018, L. Lun et al., 2016, Kiselev et al., 2019, Lenstra et al., 2016, Papalexi and Satija, 2018, Picelli et al., 2014, Stuart et al., 2018, Wang and Kaestner, 2018, Zhang et al., 2019, Zilionis et al., 2017]

References

What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Systems*, 2017. ISSN 24054712. doi: 10.1016/j.cels.2017.03.006.

Tallulah S. Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 2018. doi: 10.1016/j.mam.2017.07.002.

Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. Destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 2016. doi: 10.1093/bioinformatics/btv715.

Rhonda Bacher, Li Fang Chu, Ning Leng, Audrey P. Gasch, James A. Thomson, Ron M. Stewart, Michael Newton, and Christina Kendziorski. SCnorm: Robust normalization of single-cell RNA-seq data. *Nature Methods*, 2017. doi: 10.1038/nmeth.4263.

Eduardo Beltrame, Jase Gehring, Valentine Svensson, Dongyi Lu, Jialong Jiang, Matt Thomson, and Lior Pachter. Introduction to single-cell rna-seq technologies. 2 2019. doi: 10.6084/m9.figshare.7704659.v1. URL https://figshare.com/articles/Introduction_to_single-cell_RNA-seq_technologies/7704659.

Xi Chen, Sarah A. Teichmann, and Kerstin B. Meyer. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. *Annual Review of Biomedical Data Science*, 2018. doi: 10.1146/annurev-biodatasci-080917-013452.

Antoine Coulon, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. Eukaryotic transcriptional dynamics: From single molecules to cell populations. *Nature Reviews Genetics*, 2013. doi: 10.1038/nrg3484.

Megan Crow and Jesse Gillis. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? *Trends in Genetics*, 2018. doi: 10.1016/j.tig.2018.07.007.

Angelo Duò, Mark D. Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 2018. doi: 10.12688/f1000research.15666.1.

Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 2018. doi: 10.12688/f1000research.15809.1.

Laleh Haghverdi, Florian Buettner, and Fabian J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv325.

Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 2018. doi: 10.1038/nbt.4091.

Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems*, 2016. doi: 10.1016/j.cels.2016.04.001.

Charles A. Herring, Bob Chen, Eliot T. McKinley, and Ken S. Lau. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *CMGH*, 2018. doi: 10.1016/j.jcmgh.2018.01.023.

Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 2019. doi: 10.1038/s41576-018-0088-9.

Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, 2016. doi: 10.1186/s13059-016-0947-7.

Atefeh Lafzi, Catia Moutinho, Simone Picelli, and Holger Heyn. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols*, 13(December), 2018. doi: 10.1038/s41596-018-0073-y.

Tineke L. Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R. Larson. Transcription Dynamics in Living Cells. *Annual Review of Biophysics*, 2016. doi: 10.1146/annurev-biophys-062215-010838.

Aaron Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John Marioni. Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *bioRxiv*, 2018. doi: 10.1101/234872.

Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 2016. doi: 10.12688/f1000research.9501.2.

Vilas Menon. Clustering single cells: A review of approaches on high-and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics*, 2018. doi: 10.1093/bfgp/elx044.

Efthymia Papalexi and Rahul Satija. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 2018. doi: 10.1038/nri.2017.76.

Simone Picelli, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 2014. doi: 10.1038/nprot.2014.006.

Antoine Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Research*, 2014. doi: 10.1093/nar/gku555.

Karthik Shekhar and Vilas Menon. Identification of Cell Types from Single-Cell Transcriptomic Data. In Guo-Cheng Yuan, editor, *Computational Methods for Single-Cell Data Analysis*, pages 45–78. Springer Nature, methods in edition, 2019. doi: 10.1007/978-1-4939-9057-3{_}4.

Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 2018. doi: 10.1038/nmeth.4612.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, William M Mauck Iii, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single cell data. *bioRxiv*, 2018. doi: 10.1101/460147.

Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 2018. doi: 10.1038/nprot.2017.149.

Luyi Tian, Shian Su, Xueyi Dong, Daniela Amann-Zalcenstein, Christine Biben, Azadeh Seidi, Douglas J. Hilton, Shalin H. Naik, and Matthew E. Ritchie. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Computational Biology*, 2018. doi: 10.1371/journal.pcbi.1006361.

Yue J. Wang and Klaus H. Kaestner. Single-Cell RNA-Seq of the Pancreatic Islets—a Promise Not yet Fulfilled? *Cell Metabolism*, 2018. ISSN 15504131. doi: 10.1016/j.cmet.2018.11.016.

Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, and Jianbin Wang. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, 2019. doi: 10.1016/j.molcel.2018.10.020.

Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 2017. doi: 10.1016/j.molcel.2017.01.023.

Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 2017. doi: 10.1038/nprot.2016.154.