STAT718/BIOL703: Genomic Data Science

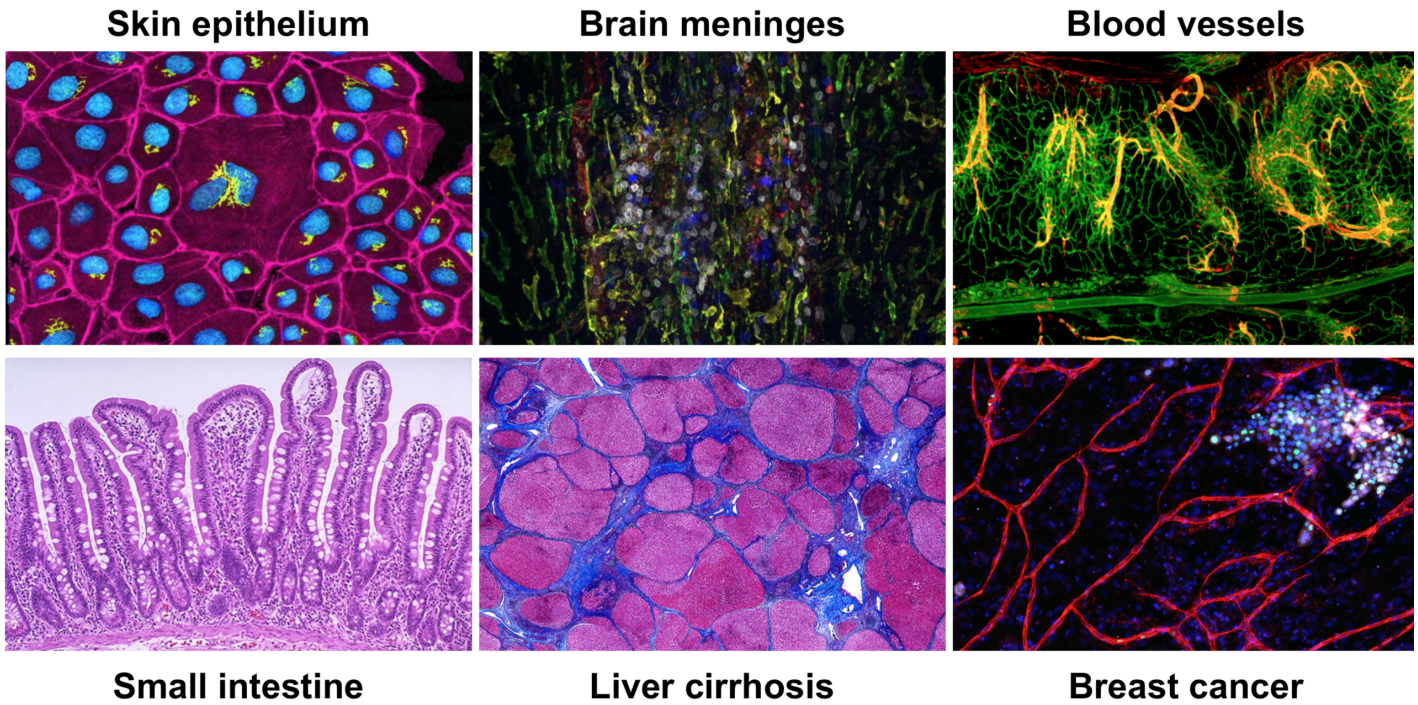
Introduction to single-cell RNA-seq

Yen-Yi Ho (hoyen@stat.sc.edu)

Introduction

Why single-cell RNA-seq

- Diversity of cell types, states, and interactions across tissues
- Single-cell RNA-seq (scRNA-seq) provides a high resolution view of transcriptomic program within a single individual cell.

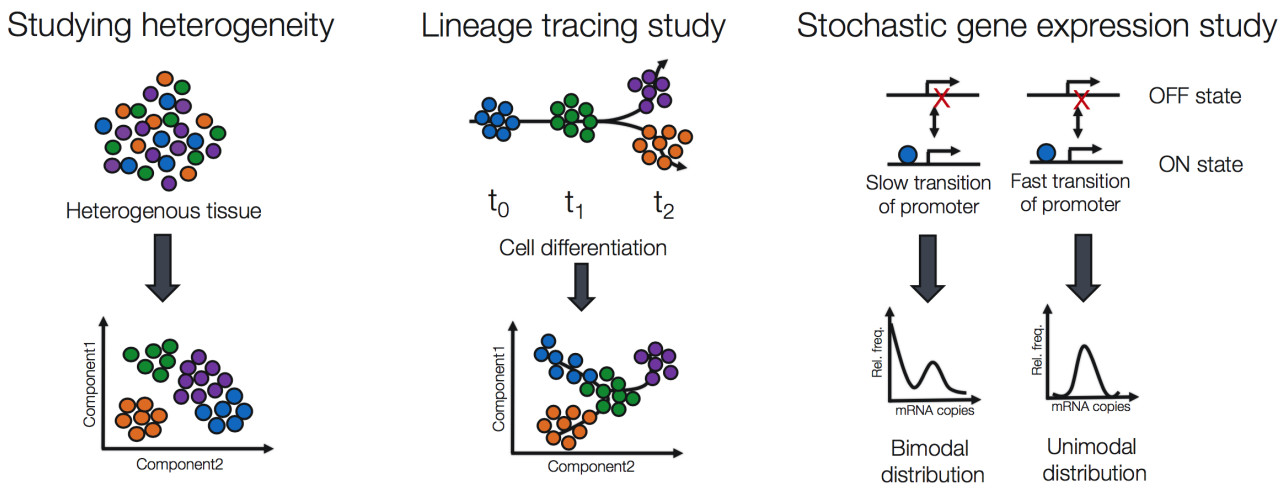


http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

Applications

This exciting and cutting-edge method can be used to: - explore which cell types are present in a tissue - identify unknown/rare cell types or states - elucidate the changes in trascriptomic program during differentiation processes or across time or states - identify genes that are differentially expressed in particular cell types between conditions (e.g. treatment or disease)

Popular methods to address some of the more common investigations include:



Liu S and Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges, F1000 Research 2016 (doi: 10.12688/f1000research.7223.1)
Junker and van Oudenaarden; Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology, Cell 2014 (doi: 10.1016/j.cell.2014.02.010)

scRNA-seq versu bulk RNA-seq

- Bulk RNA-seq measures average gene expression (homogeneous cell composition)
- heterogeneous cells

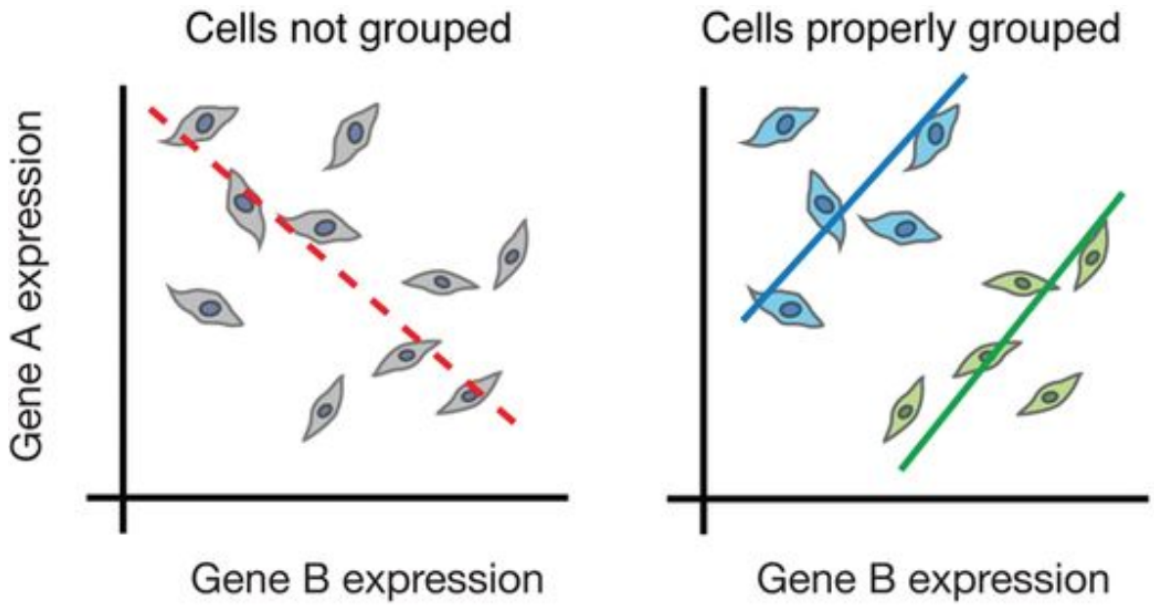


Image credit: Trapnell, C. Defining cell types and states with single-cell genomics, Genome Research 2015 (doi: https://dx.doi.org/10.1101/gr.190595.115)

Challenges of scRNA-seq analysis

- Sample generation and library preparation is more expensive
- The analysis is much more complicated and more difficult to interpret
- Large volume of data
- Low sequencing depth per cell (zero-inflation)
- Biological variability across cells/samples
- Technical variability across cells/samples

Potential sources of biological variability

- Transcriptional bursting
- Varying rates of RNA processing
- Continuous or discrete cell identities (e.g. the pro-inflammatory potential of each individual T cell)
- Environmental stimuli
- Temporal changes

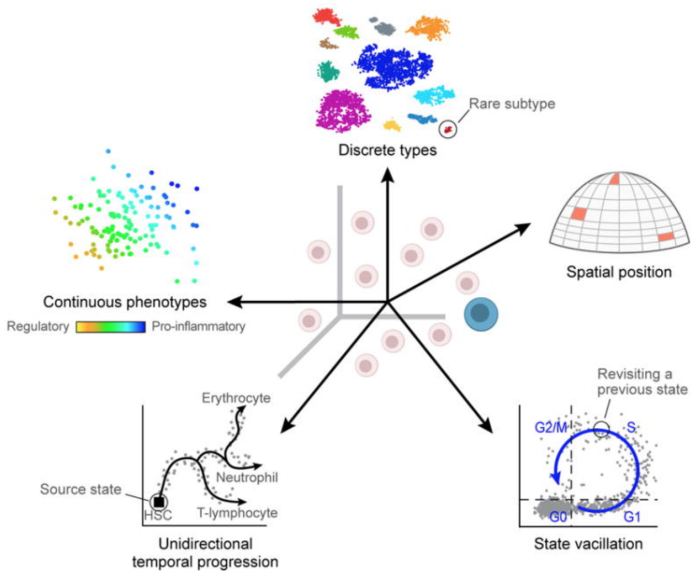


Image credit: Wagner, A, et al. Revealing the vectors of cellular identity with single-cell genomics, Nat Biotechnol. 2016 (doi:https://dx.doi.org/10.1038%2Fnbt.3711)

Technical variability

- Cell-specific capture efficiency
- Library quality: degraded RNA, low viability/dying cells, lots of free floating RNA, poorly dissociated cells, and inaccurate quantitation of cells can result in low quality metrics
- Amplification bias
- Batch effects

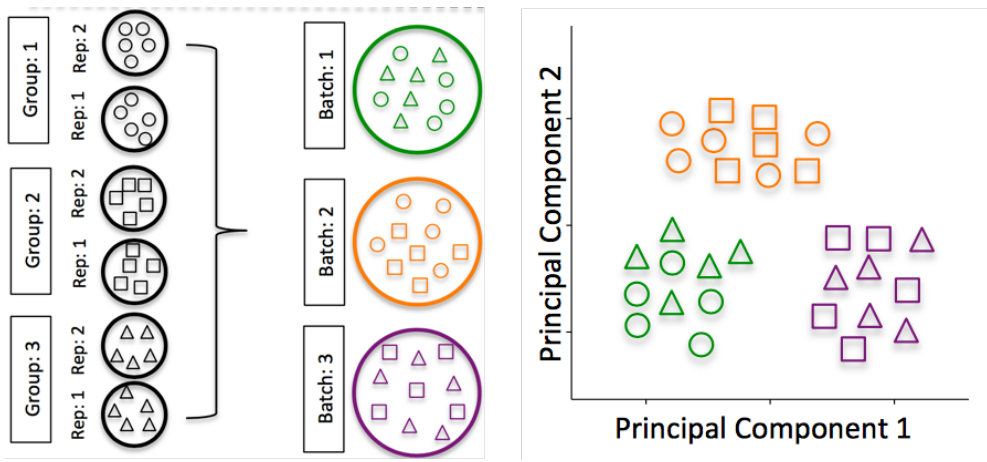


Image credit: Hicks SC, et al., bioRxiv (2015)](https://www.biorxiv.org/content/early/2015/08/25/025528)

Best practices regarding batches

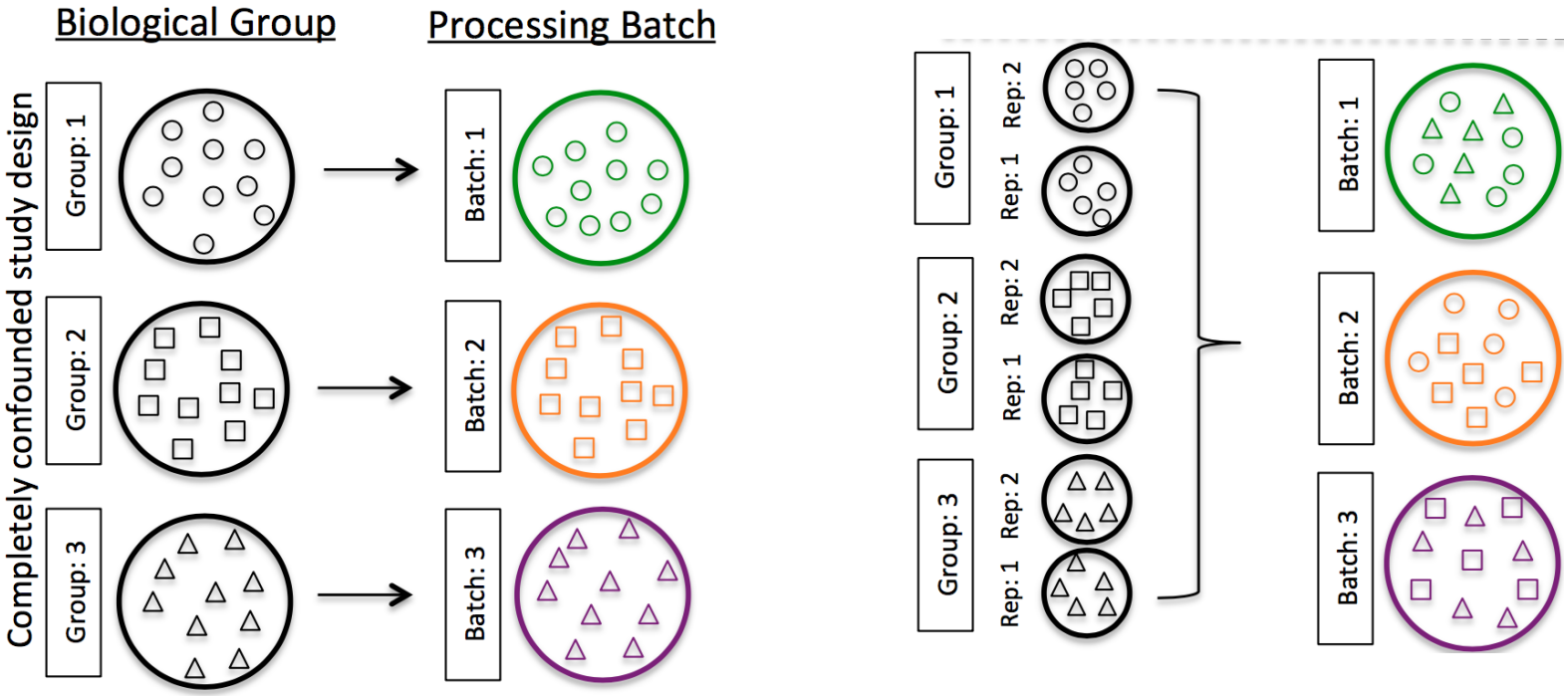


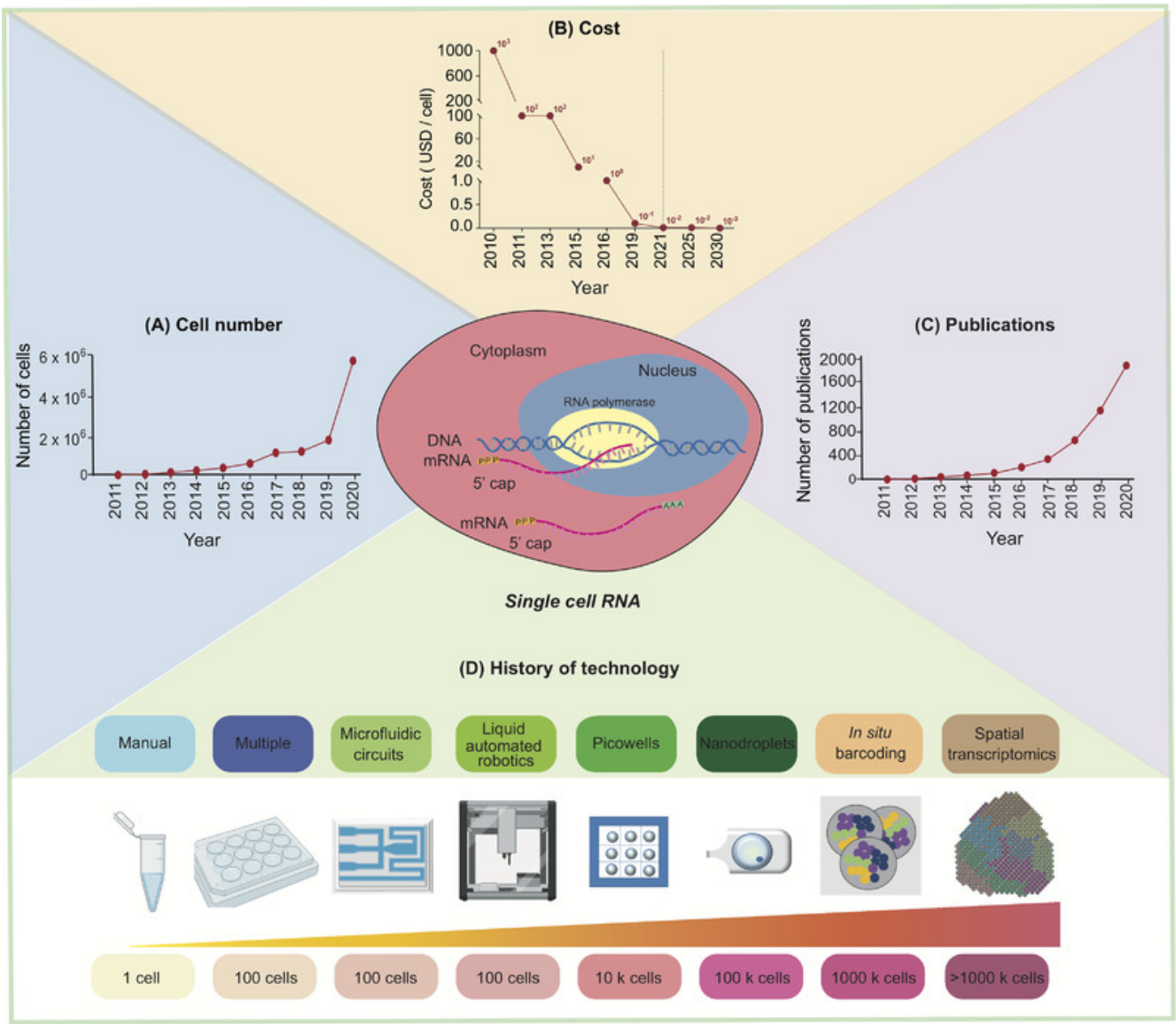
Image credit: Hicks SC, et al., bioRxiv (2015)](https://www.biorxiv.org/content/early/2015/08/25/025528)

Conclusions

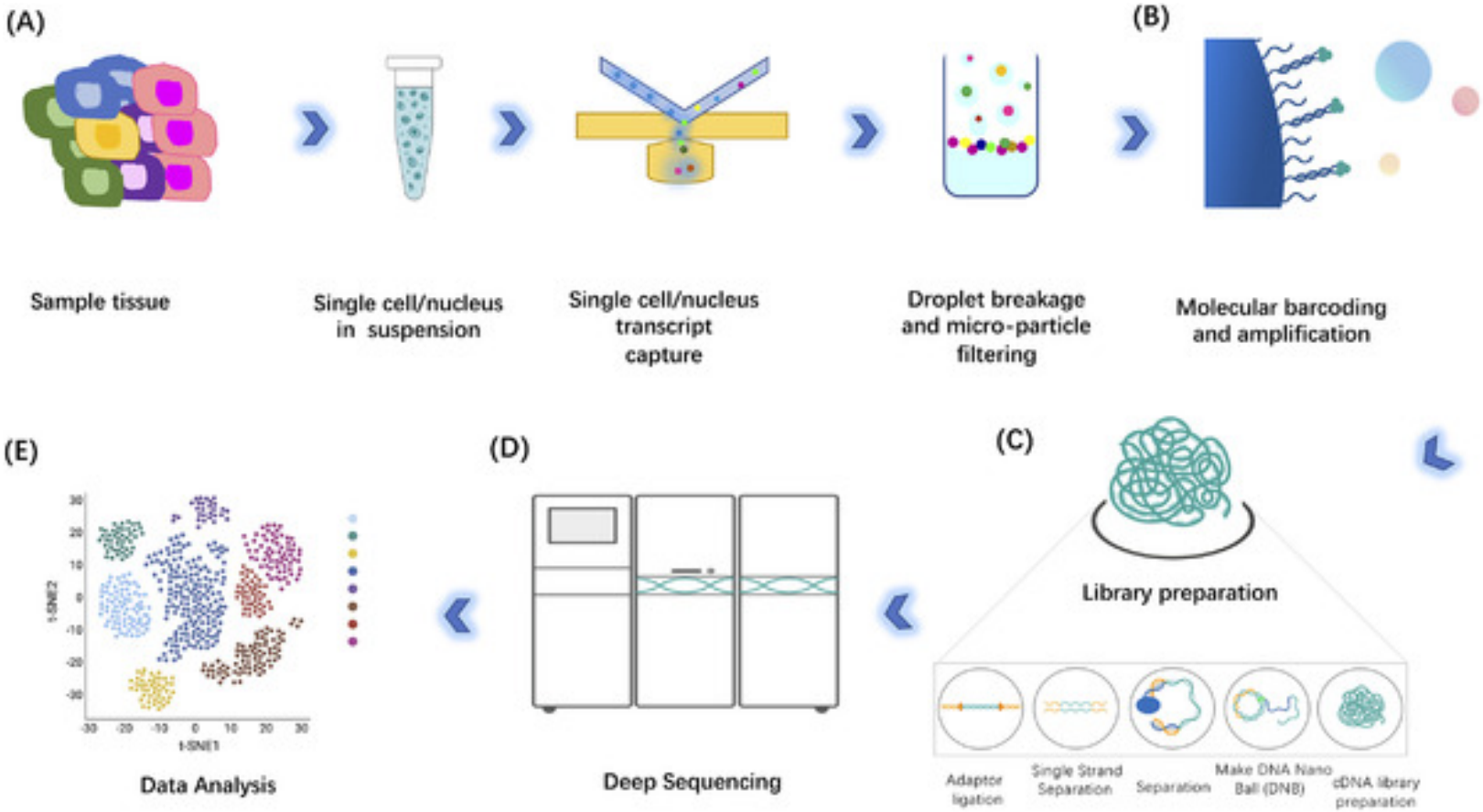
- scRNAseq is a powerful and insightful method
- Many challenges and sources of variation
- Avoid technical sources of variability, if possible:
  - Discuss experimental design with experts prior to experiment
  - Isolate RNA from samples at the same itme
  - Prepare libraries at same time/alternative sample groups
  - Do not confound sample groups by sex, age, or batch

Pre-processing

Development of scRNA-seq technology



Experimental Procedure



From raw data to count matrix

- Droplet: Drop-seq,10X Genomics, CEL-seq2, Drop-seq, InDrop
- Plate-based: Fluidigm C1, SMART-seq (full-length sequencing)

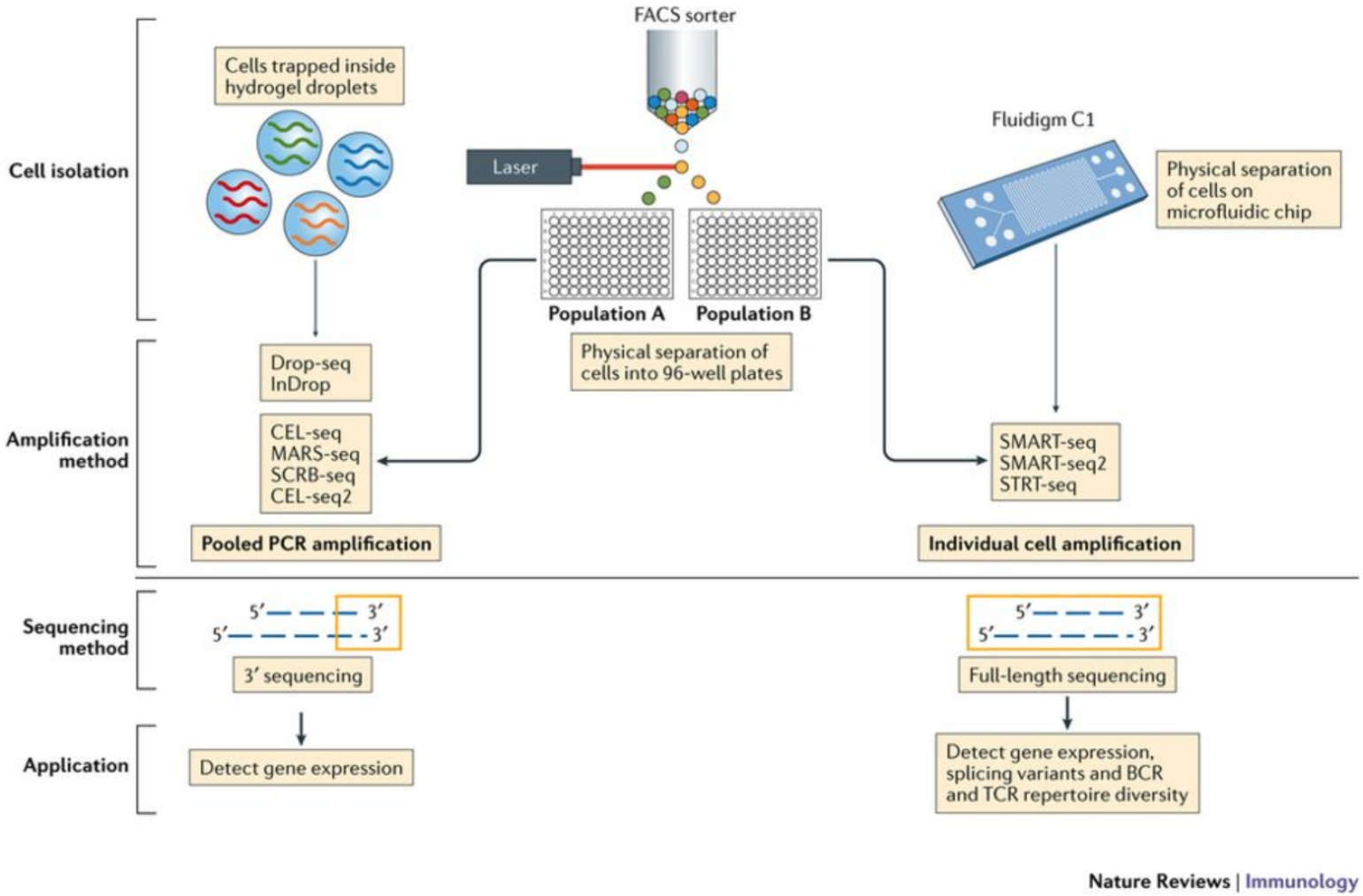
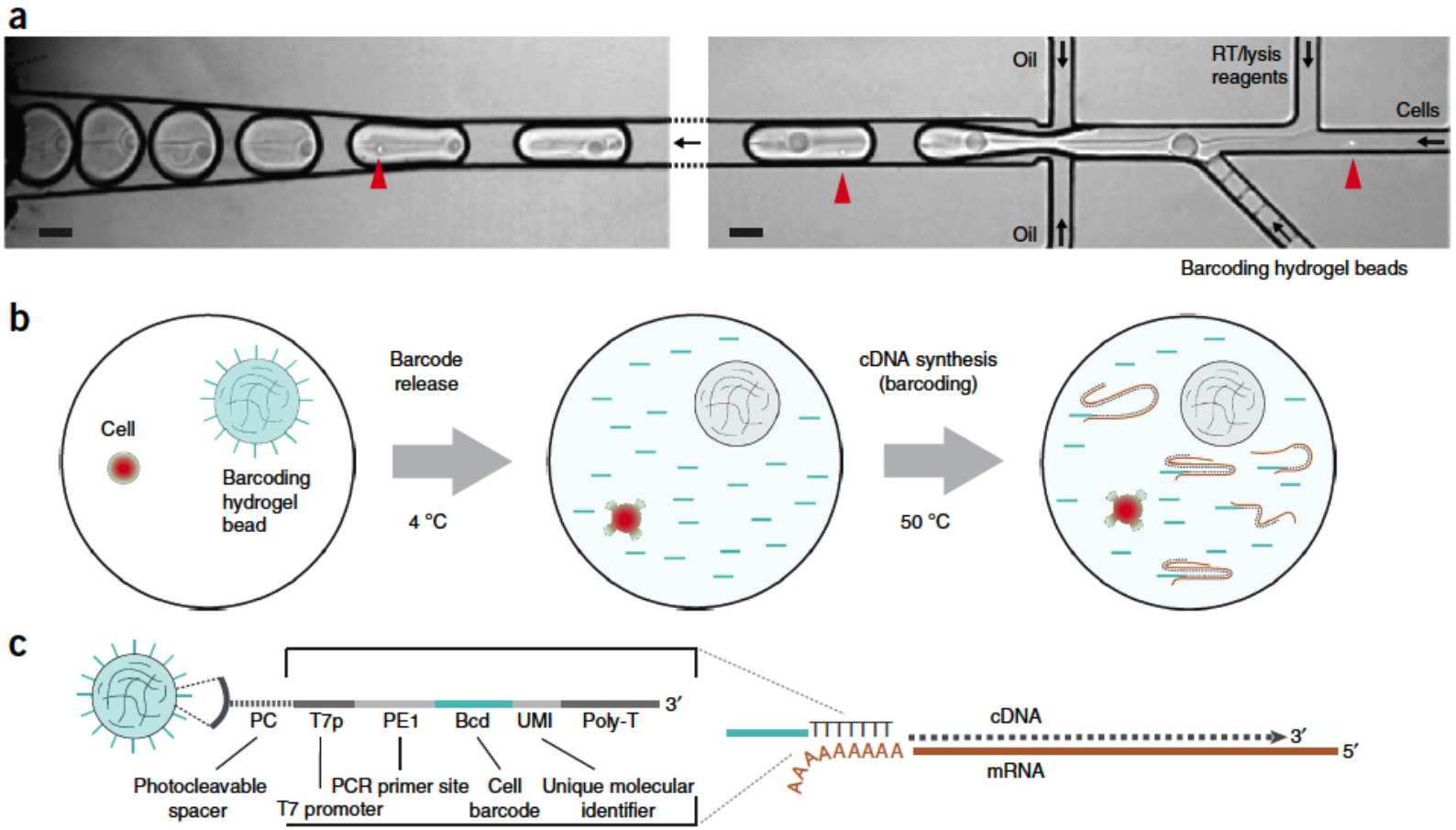


Image credit:Papalexi E and Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity, Nature Reviews Immunology 2018 (https://doi.org/10.1038/nri.2017.76)

Droplet-based sequencing



Sequencing Approaches

The following advantages are listed below for various methods:

- 3' (or 5')-end sequencing
  - More accurate quantification through use of unique molecular identifiers distinguishing biological duplicates from amplification (PCR) duplicates
  - Larger number of cells sequenced allows better identity of cell type populations
  - Cheaper per cell cost
  - Best results with $>$ 10,000 cells
- Full length sequencing
  - Detection of isoform-level differences in expression
  - Identification of allele-specific differences in expression
  - Deeper sequencing of a smaller number of cells
  - Best for samples with low number of cells

3'end reads: UMIs

To determine whether a read is a biological or technical duplicate, these methods use unique molecular identifiers, or UMIs.

- Reads with **different UMIs** mapping to the same transcript were derived from **different molecules** and are biological duplicates - each read should be counted.
- Reads with the **same UMI** originated from **the same** molecule and are technical duplicates - the UMIs should be collapsed to be counted as a single read.

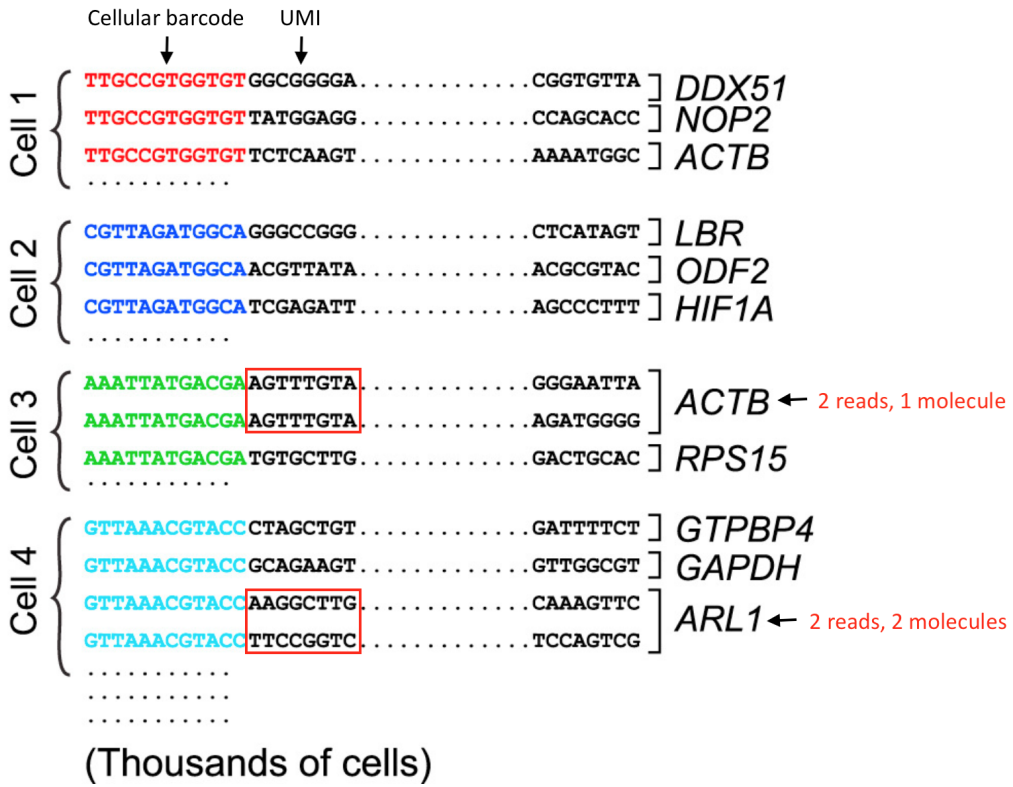
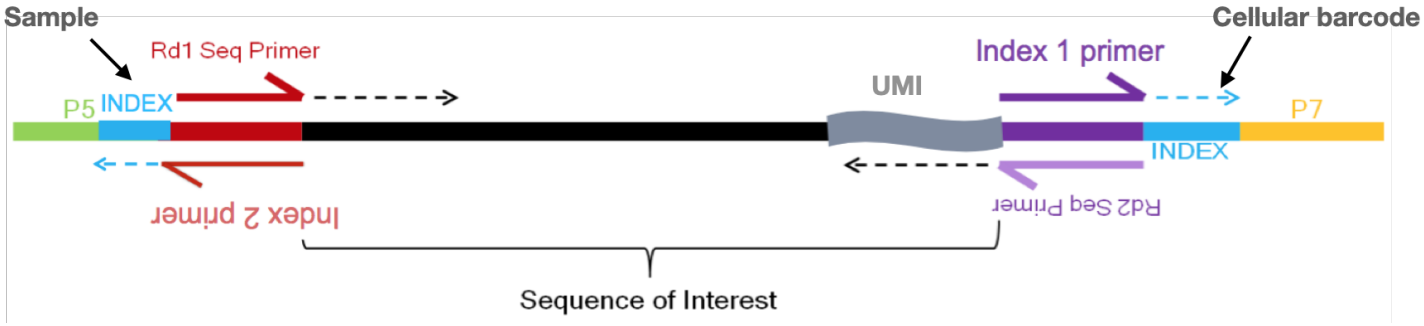


Image credit: modified from Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets, Cell 2015 (https://doi.org/10.1016/j.cell.2015.05.002)

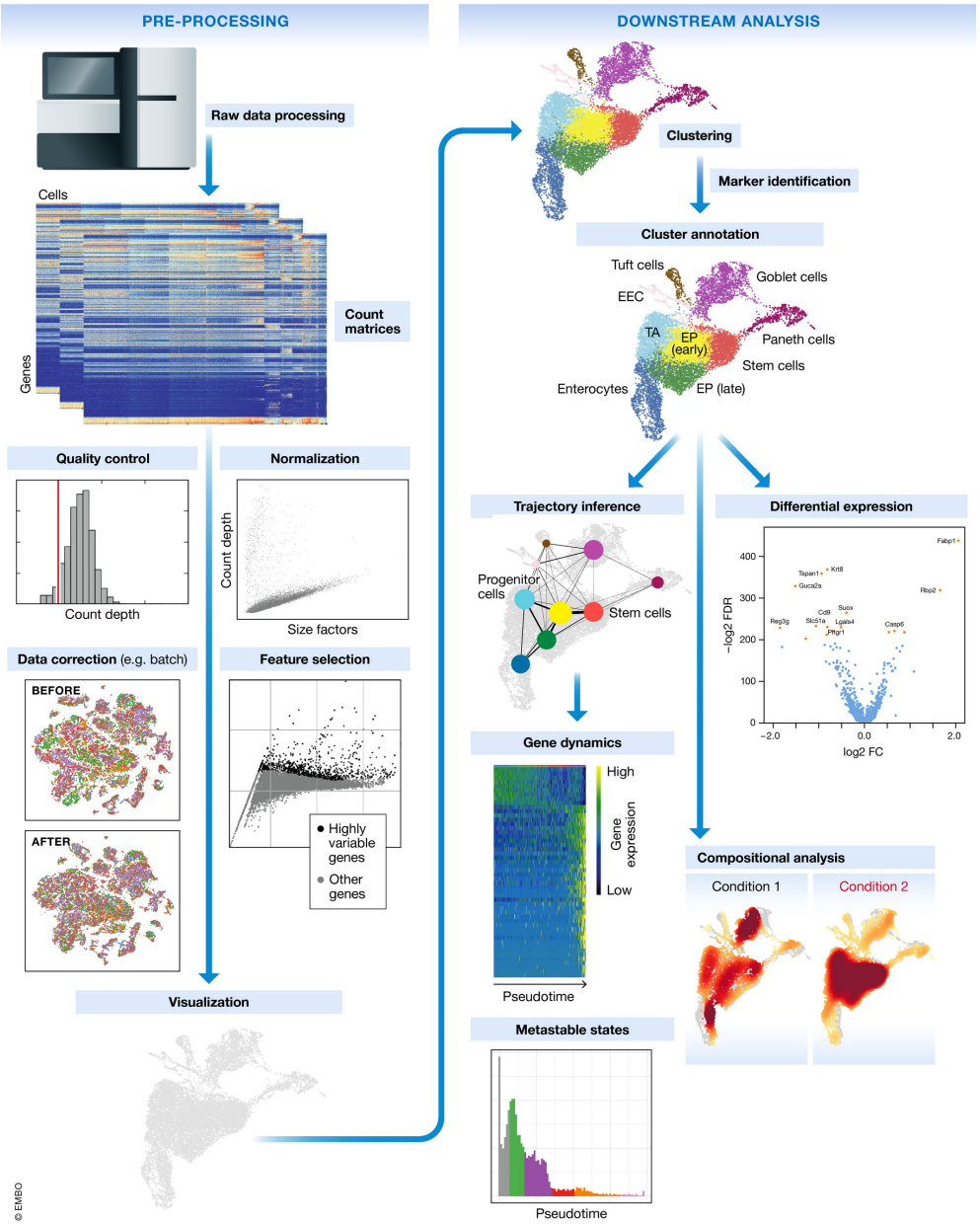
Barcodes



- Sample index: determines which sample the read originated from (red bottom arrow)
  - Added during library preparation (needs to be documented)
- Cellular barcode:determines which cell the read originated from (purple top arrow)
  - Each library preparation method has a stock of cellular barcodes used during the library preparation
- Unique molecular identifier (UMI): determines which transcript molecule the read originated from
  - The UMI will be used to collapse PCR duplicates (purple bottom arrow)
- Sequencing read1: the Read1 sequence (red top arrow)
- Sequencing read2: the Read2 sequence (purple bottom arrow)

Image credit: Sarah Boswell](https://scholar.harvard.edu/saboswell), Director of the Single Cell Sequencing Core at HMS

scRNAseq Analysis Workflow

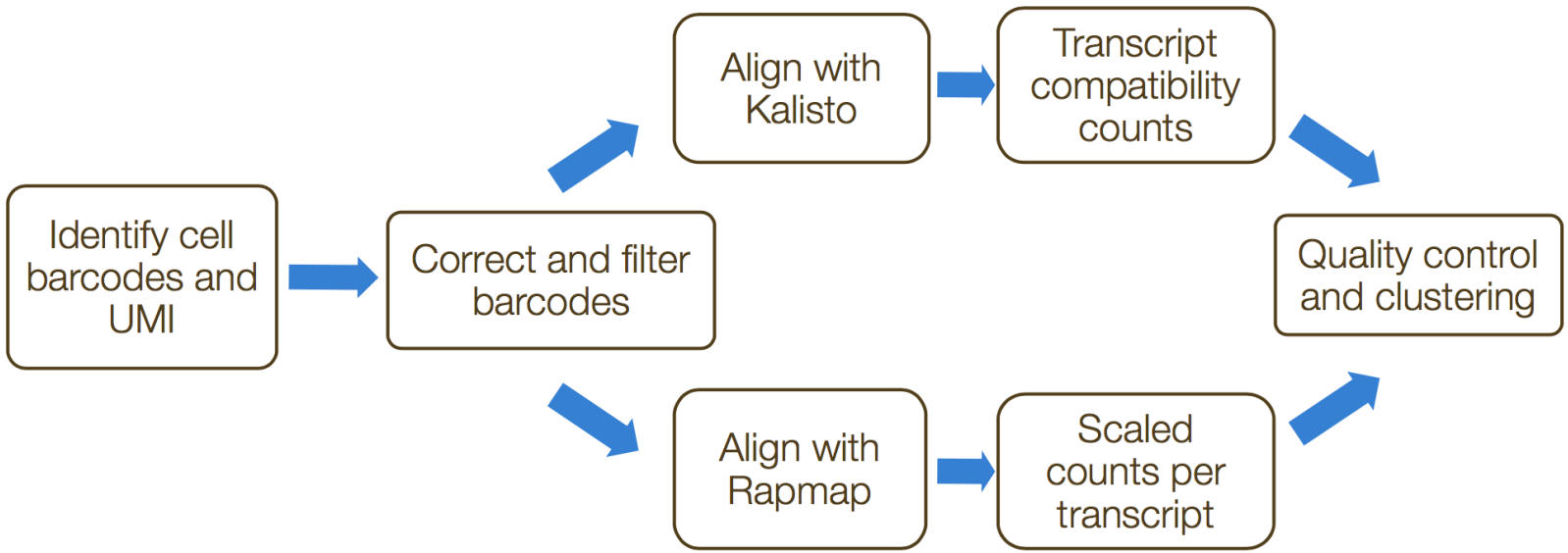


- Generation of the count matrix
- Quality control of the raw counts filtering of poor quality cells
- Clustering of filtered counts: clustering cells based on similarities in transcriptional activity
- Marker identification and cluster annotation:
- Optional downstream steps

Image credit: Luecken, MD and Theis, FJ. Current best practices in single-cell RNA-seq analysis: a tutorial, Mol Syst Biol 2019 (doi:

Generation of count matrix

The first part of the workflow is generation the count matrix from the raw sequencing data. We will focus on the 3' end sequencing used by droplet-based methods, such as inDrops, 10X Genomics, and Drop-seq.



Many barcodes . . .

- Formatting reads and filtering noisy cellular barcodes
- Demultiplexing the samples
- Mapping to transcriptome
- Collapsing UMIs and quantification of read

Formatting reads and filtering noisy cellular barcodes

For droplet-based methods, many of the cellular barcodes will match a low number of reads ($<$ 1000 reads) due to:

- Encapsulation of free floating RNA from dying cells
- Simple cells (RBCs, etc.) expressing few genes
- Cells that failed for some reason

```
@HWI-ST808:130:H0B8YADXX:1:1101:2088:2222:CELL_GGTCCA:UMI_CCCT
AGGAAGATGGAGGAGAGAAGGCGGTGAAAGAGACCTGTAAAAAGCCACCGN
+ @@@DDBD>=AFCF+<CAFHDECII:DGGGHGIGGIIIEHGIIIGIIDHII#
```

Demultiplexing sample reads

The next step of the process is to demultiplex the samples, if sequencing more than a single sample. This is the one step of this process not handled by the 'umis' tools, but is accomplished by 'zUMIs'. We would need to parse the reads to determine the sample barcode associated with each cell.

Mapping to cDNAs

To determine which gene the read originated from, the reads are aligned using traditional (STAR) or Kallisto/RapMap.

Collapsing UMIs and quantification of read

The duplicate UMIs are collapsed, and only the unique UMIs are quantified using a tool like Kallisto or featureCounts. The resulting output is a cell by gene matrix of counts:

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Image credit: extracted from Lafzi et al. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies, Nature Protocols 2018 (https://doi.org/10.1038/s41596-018-0073-y)