

STAT588/BIOL588 Genomic Data Science

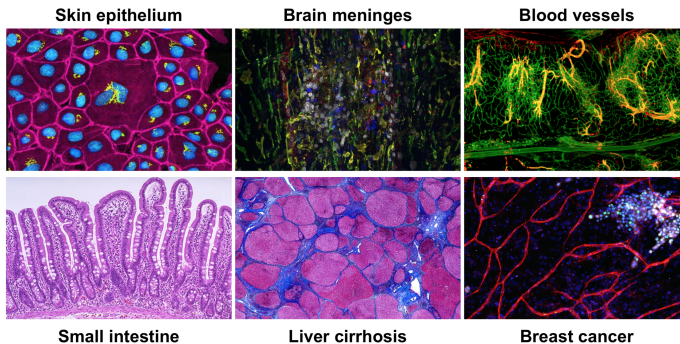
Lecture 20 Introduction to single-cell RNA-seq

Yen-Yi Ho

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.



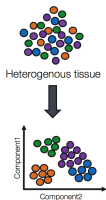
<http://www.cell.com/pictureshow/skin> | <https://library.med.utah.edu/WebPath/webpath.html>

Applications

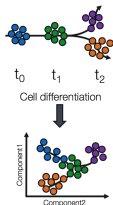
This exciting and cutting-edge method can be used to: - explore which cell types are present in a tissue - identify unknown/rare cell types or states - elucidate the changes in transcriptomic program during differentiation processes or across time or states - identify genes that are differentially expressed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

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

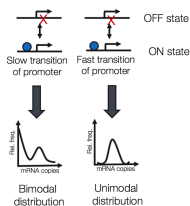
Studying heterogeneity



Lineage tracing study



Stochastic gene expression study



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 versus bulk RNA-seq

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

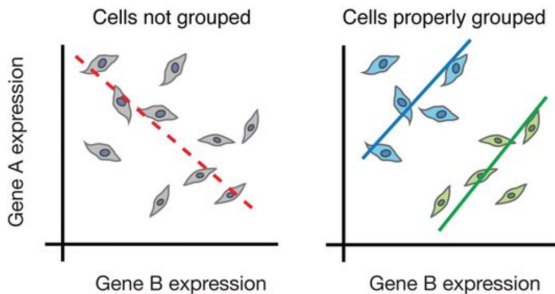


Image credit: Trapnell, C. Defining cell types and states with single-cell genomics, Genome Research 2015 (doi: <https://dx.doi.org/10.1101/gr.190595.115>)

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

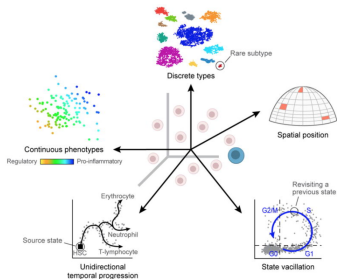


Image credit: Wagner, A, et al. Revealing the vectors of cellular identity with single-cell genomics, Nat Biotechnol. 2016 (doi:<https://dx.doi.org/10.1038%2Fnb.3711>)

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

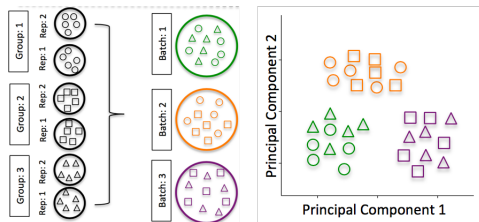


Image credit: Hicks SC, et al., bioRxiv (2015)](<https://www.biorxiv.org/content/early/2015/08/25/025528>)

Best practices regarding batches

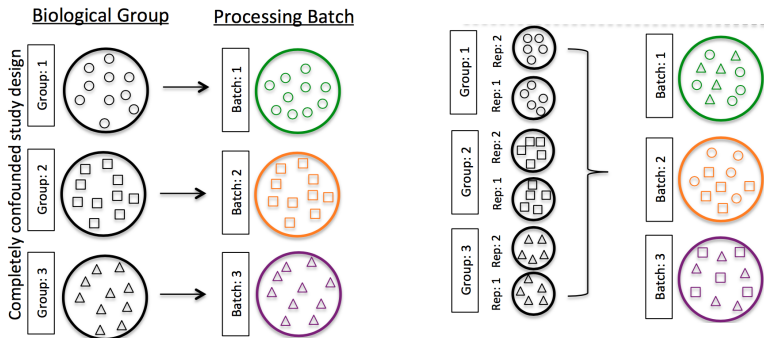


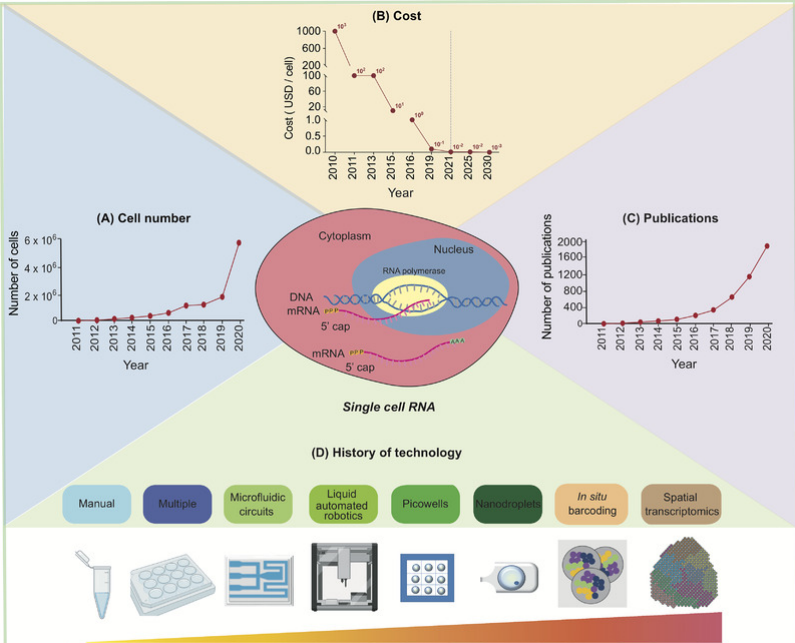
Image credit: Hicks SC, et al., bioRxiv (2015)(<https://www.biorxiv.org/content/early/2015/08/25/025528>)

Conclusions

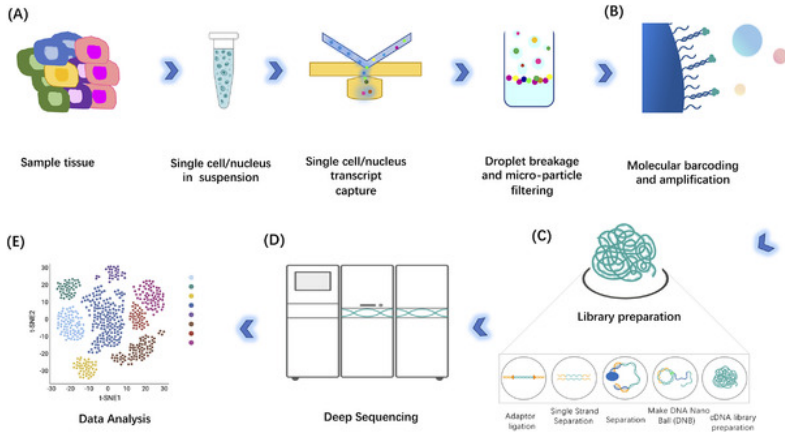
- ▶ scRNAseq is a powerful and insightful method
- ▶ Many challenges and sources of variation
- ▶ Avoid technical sources of variability, if possible:
 - ▶ Discuss experimental design with experts prior to experiment
 - ▶ Isolate RNA from samples at the same time
 - ▶ Prepare libraries at same time/alternative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

Pre-processing

Development of scRNA-seq technology

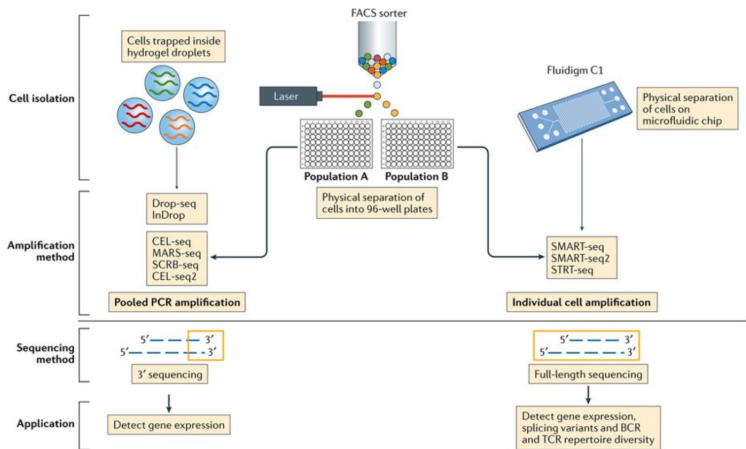


Experimental Procedure

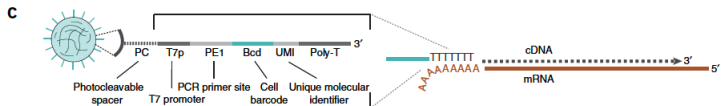
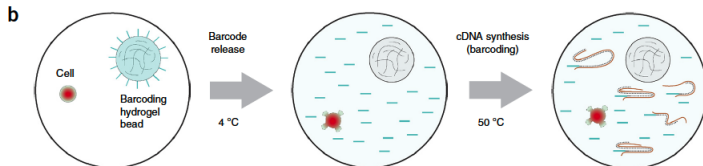
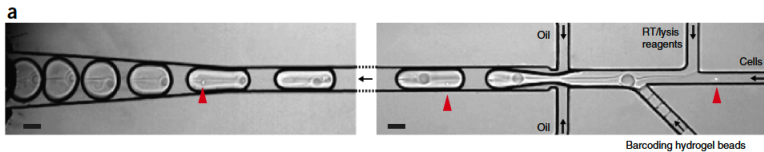


From raw data to count matrix

- ▶ Droplet: Drop-seq, 10X Genomics, CEL-seq2, Drop-seq, inDrops
- ▶ Plate-based: Fluidigm C1



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.

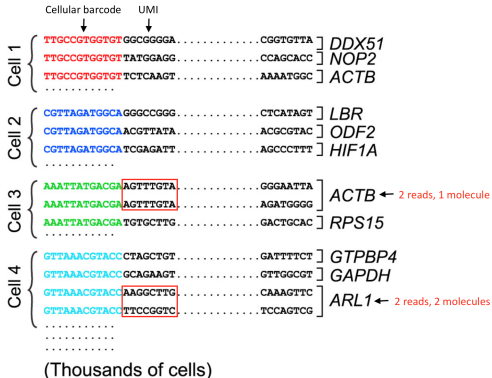
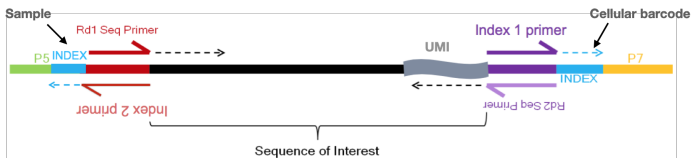


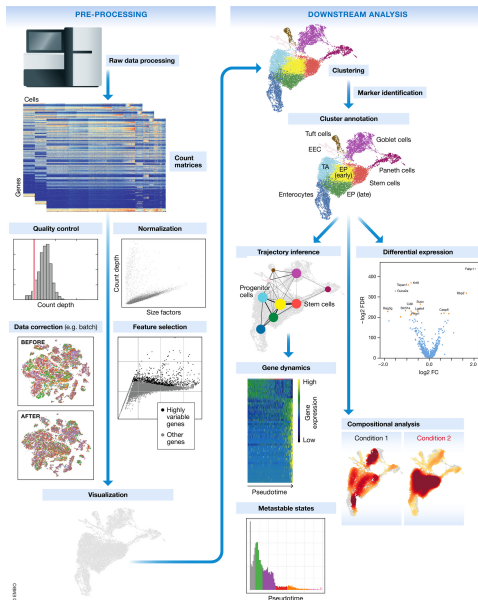
Image credit: modified from Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets, Cell 2015 (<https://doi.org/10.1016/j.cell.2015.05.002>)

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)

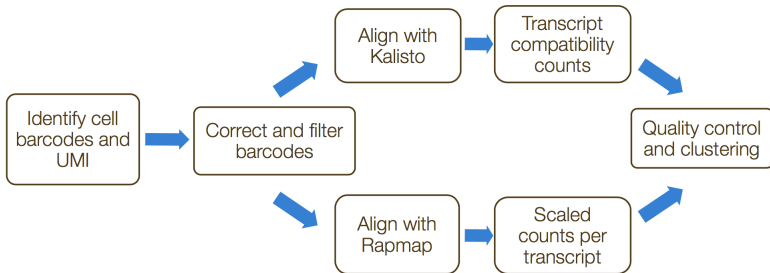
scRNAseq Analysis Workflow



- ▶ 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

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.



Cell Ranger Pipeline

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