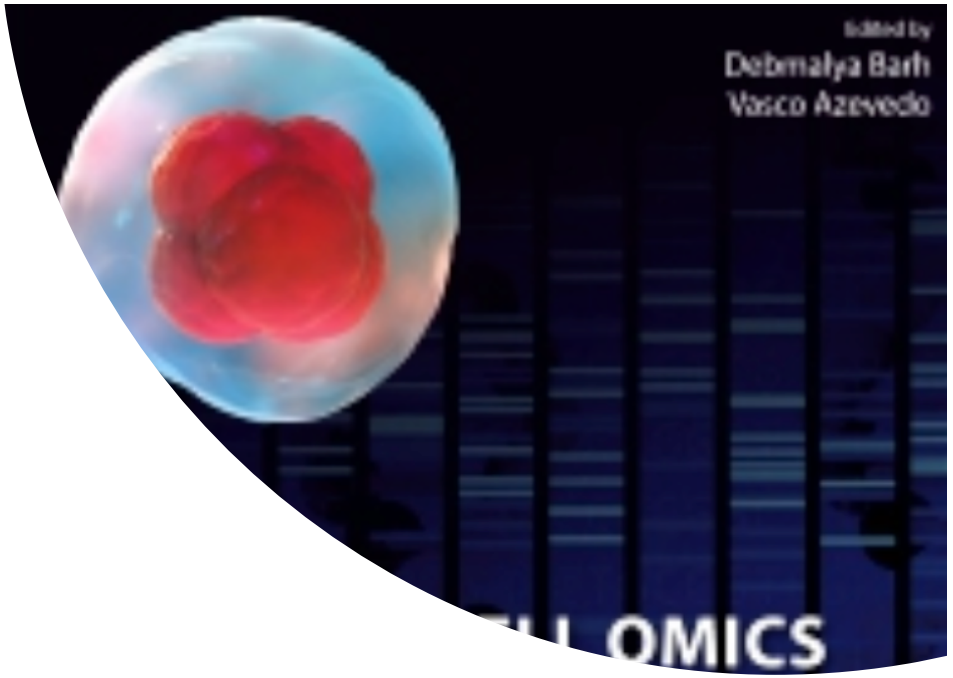


Single-cell RNA-seq --an introduction to workflow

Yunhua Zhu, Ph.D.

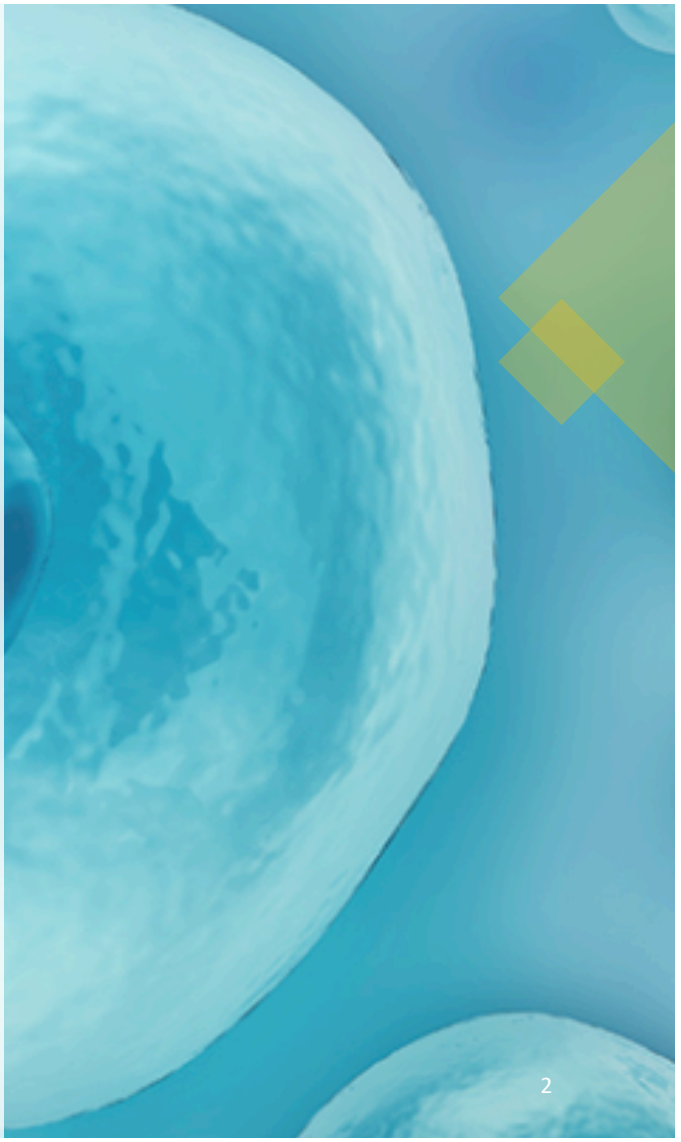
June 1, 2020.

BCBB, NIAID, NIH

National Institute of Allergy and Infectious Diseases

Outline

- Objective
  - An general introduction on single-cell RNA-seq
    - On the general workflow
    - To emphasize the differences compared with bulk RNA-seq
- Outlines
  - **Wet lab, technical advances**
    - What is single-cell RNA-seq– the start of this technology
    - Current platform – microfluidics and 10X Genomics
    - Advantages of the current technology and limitations
    - Considerations when designing a scRNA-seq project
  - **Dry lab, overview of the workflow**
    - Getting expression matrix
    - Dimension reduction
    - Trajectory analysis
    - Functional annotation
    - Gene regulatory network analysis
    - **Comprehensive tools**
  - **Summary and general discussion**



Self introduction

- Ph.D @ NUS in stem cell biology | 2006 -13
  - Aging of neural progenitors
  - Intestinal stem cells
- Postdoc @ Hopkins with wet lab & dry lab | 2014 - 2019
  - Neurogenesis w/t single-cell RNA-seq
  - Neurodegeneration w/t single-nucleus RNA-seq
- Computational Genomics Specialist | 2019
  - Single-cell CITE-seq
  - Single-cell RNA-seq on bile duct tumor
  - Bulk RNA-seq on IR response of monkey brain samples
  - Bulk RNA-seq data of HIV blood samples

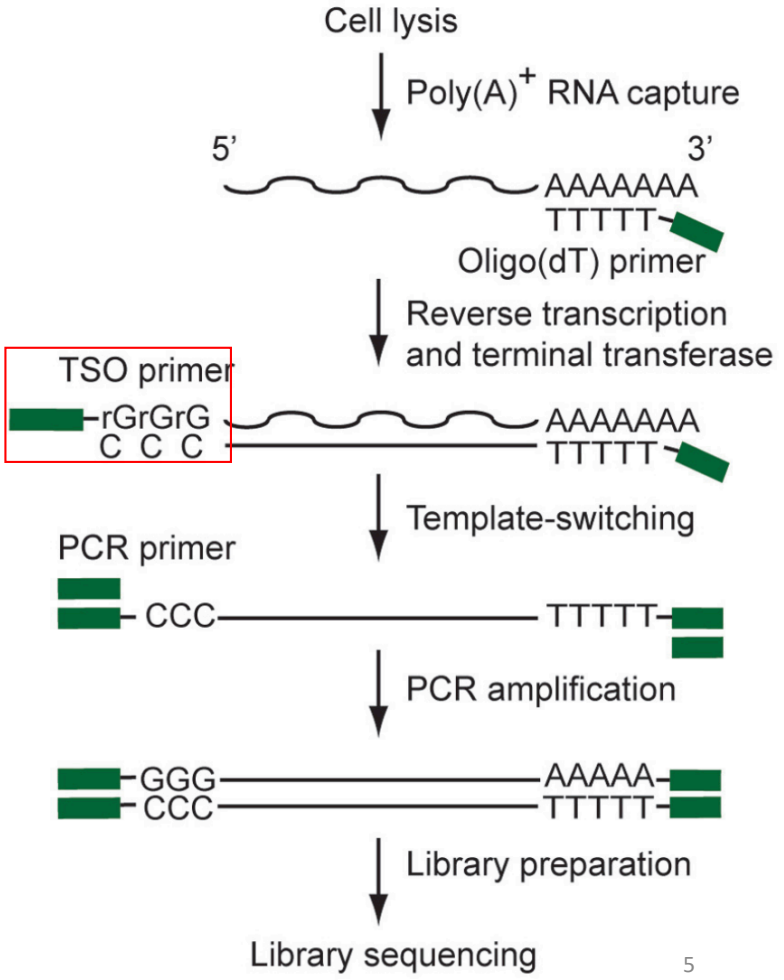
National Institute of Allergy and Infectious Diseases

Wet lab

- The Smart-seq2 protocol to amplifying single cell (sc) mRNA (10pg)
- 10X genomics--the current standard
- What biological problems can you solve with scRNA-seq
- Considerations when designing a project involving scRNA-seq

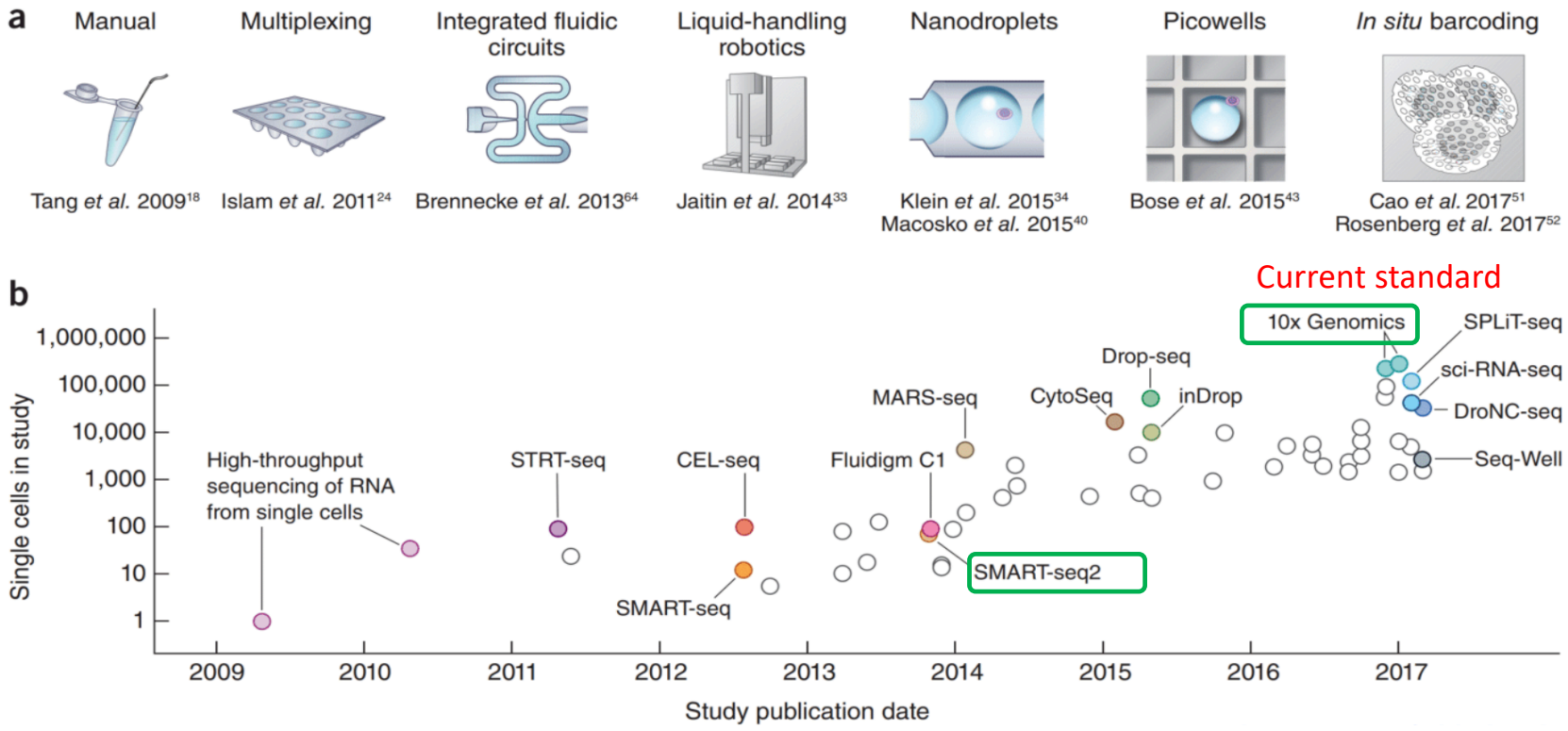
RT vs mRNA amplification

- Core challenge:
  - How to get enough signal from 10pg RNA/cell?
- Advances in engineering
  - Template switching oligo (TSO) enables efficient amplification of mRNA
  - Barcode-mediated multiplexing enables combining many samples together and greatly reduced the cost for each cell
  - Cost reduction of next generation sequencing (NGS)



National Institute of Allergy and Infectious Diseases

Development of new platforms

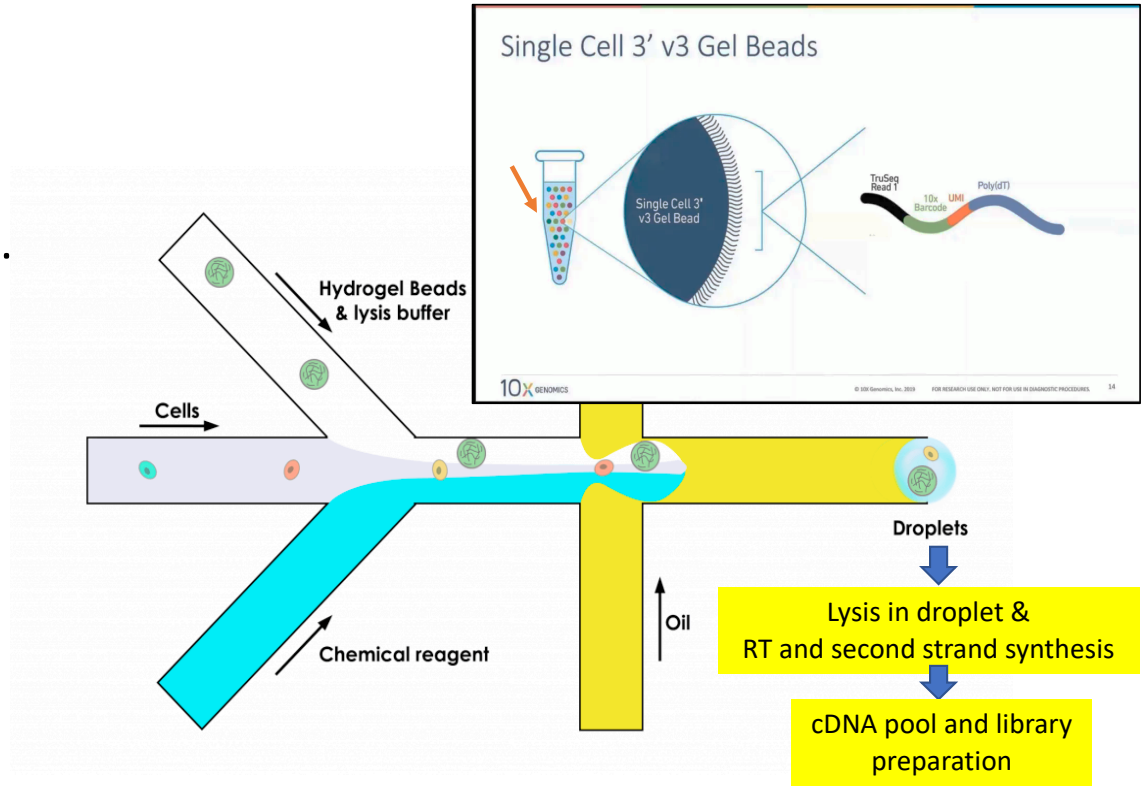


National Institute of Allergy and Infectious Diseases

https://figshare.com/articles/Single_Cell_Present_and_near_Future/12121674

10X Genomics– commercial company that facilitates automatic generation of **Gel Bead-in-Emulsion** (GEM)

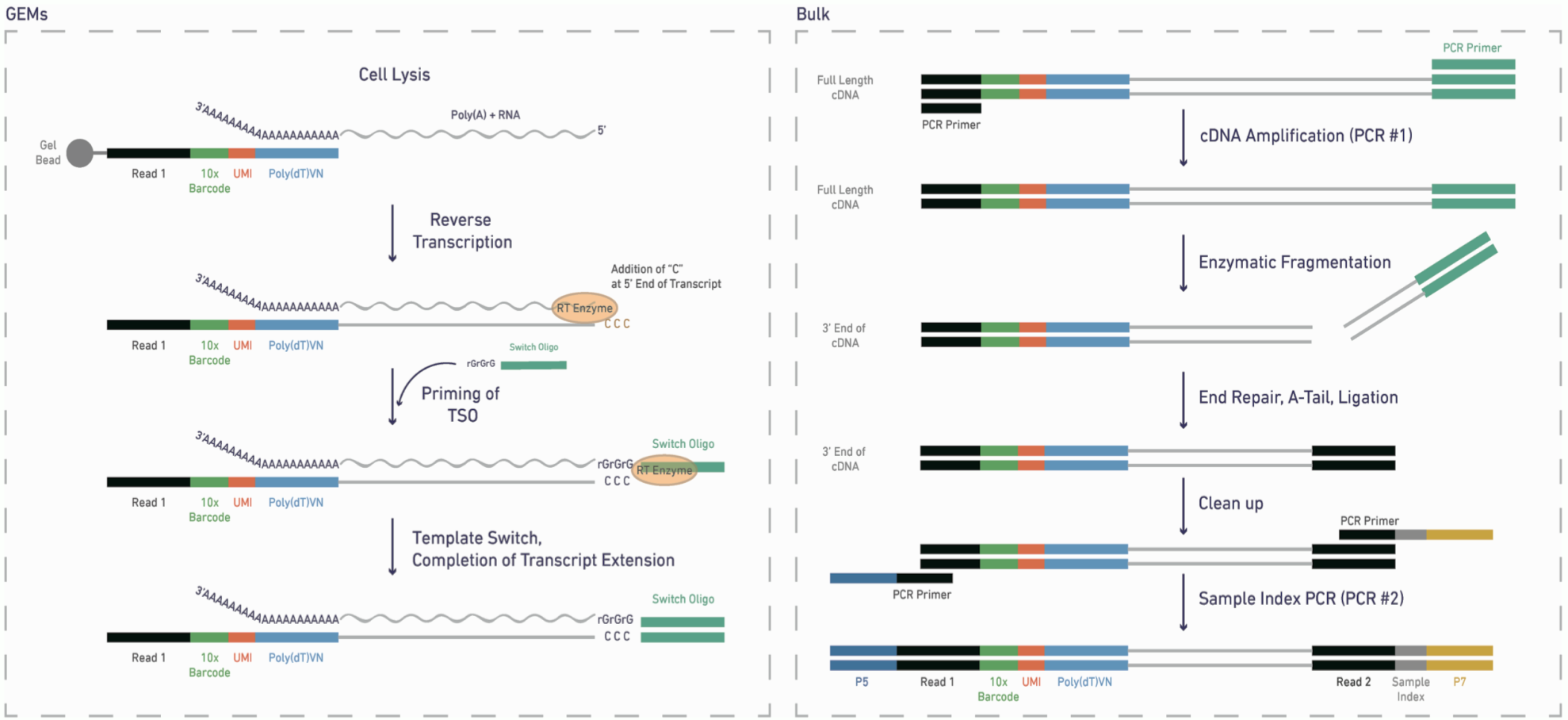
- In a GEM droplet, one hydrogel bead and one cell were captured
- One hydrogel bead is attached with millions of poly-T primers with an identical unique barcode.
- cDNA and the second-strand synthesis in the droplet
- Droplets are disrupted to collect all the barcoded samples for highly multiplexed library preparation and sequencing
- Standardized automation and reagent has made sequencing library preparation very efficient



National Institute of Allergy and Infectious Diseases

Inside individual GEMs (Gel Bead-in-Emulsion)

Pooled cDNA processed in bulk

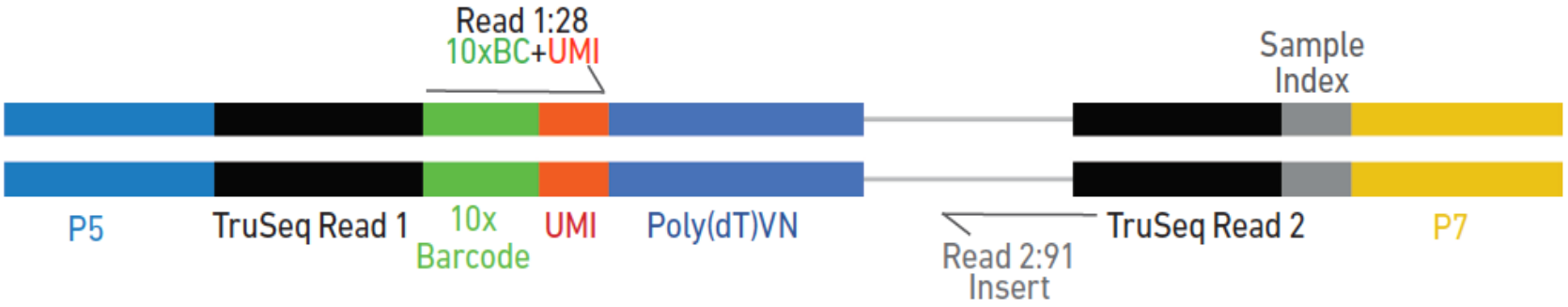


National Institute of Allergy and Infectious Diseases

https://assets.ctfassets.net/an68im79xiti/4fIy9tr6qQuCWamIii0iEa/40658acce7a6756e38537584897840e3/CG000108_AssayConfiguration_SC3v2.pd

Next-seq reading the paired ends

- 10XBC: 16 bp barcodes 2^16=65536 possible unique cells
- UMI, 2^10=2048 unique copy of mRNA for each each gene
- Read2 will read into cDNA
- Sample barcode, batch of your library
- All information will be summarized by the Cellranger software



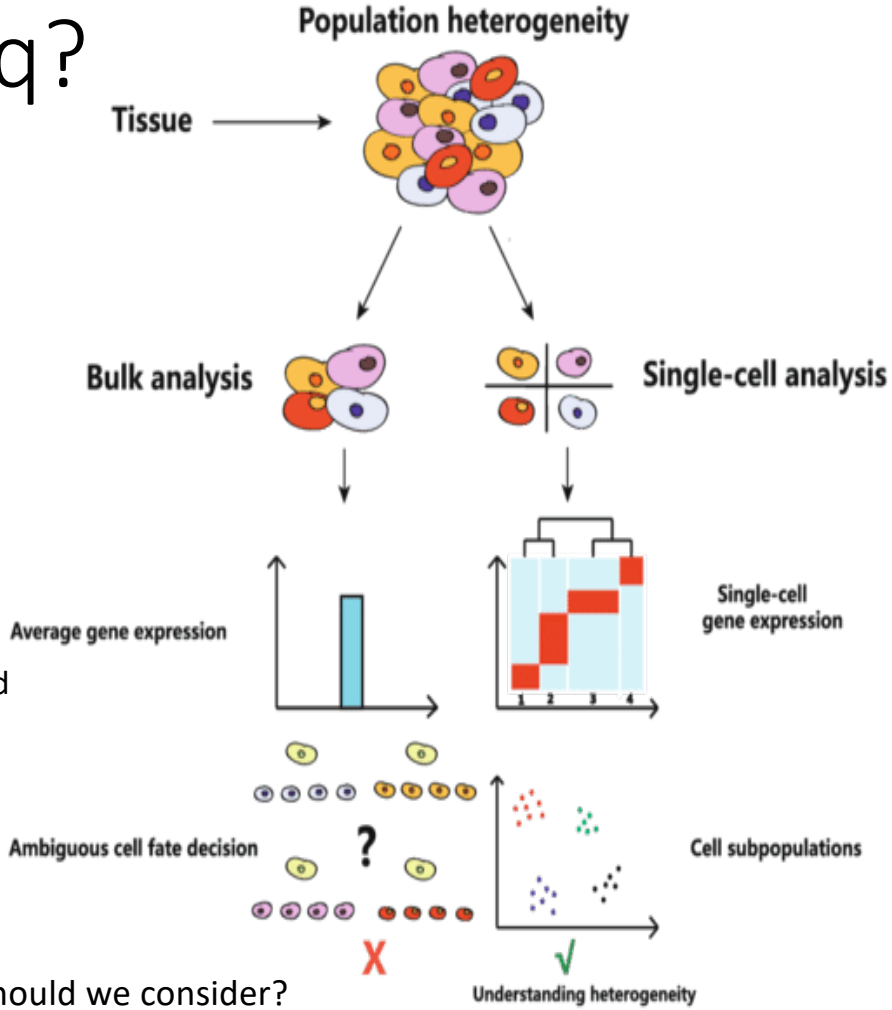
Complicated computation … why should we bother this complication??? …

National Institute of Allergy and Infectious Diseases

https://github.com/niaid/NGS_Intro/blob/master/notes/sequence_qc_class.md

Why single-cell RNA-seq?

- Advantages
  - High resolution for novel details
    - Reveal minor populations
    - Reveal gradual transitions
  - High throughput → big data
    - Completeness for an atlas study of a target tissue or an entire organism -- ecosystem
    - High statistic power to infer relationships between genes
  - Connection to other research fields
    - Computing, mathematics, machine learning, (and visual arts).
- Disadvantages
  - Low depth in the highly multiplexed system
    - Genes with lower expression may not be reliably detected
  - High dropout rates
    - Not all genes can be picked up and amplified
  - Huge amount of data requiring substantial knowledge
    - Stay focused on your biology and extract valuable insight
    - Good collaboration and eagerness to learn
  - Timing

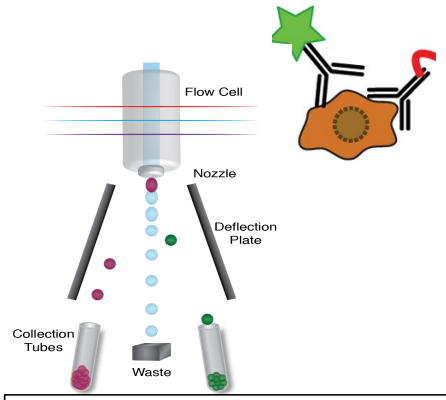


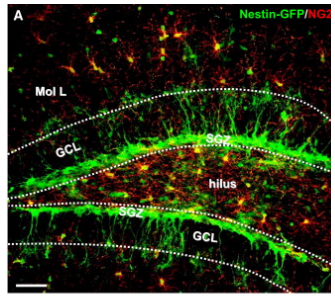
Towards designing scRNA projects, What should we consider?

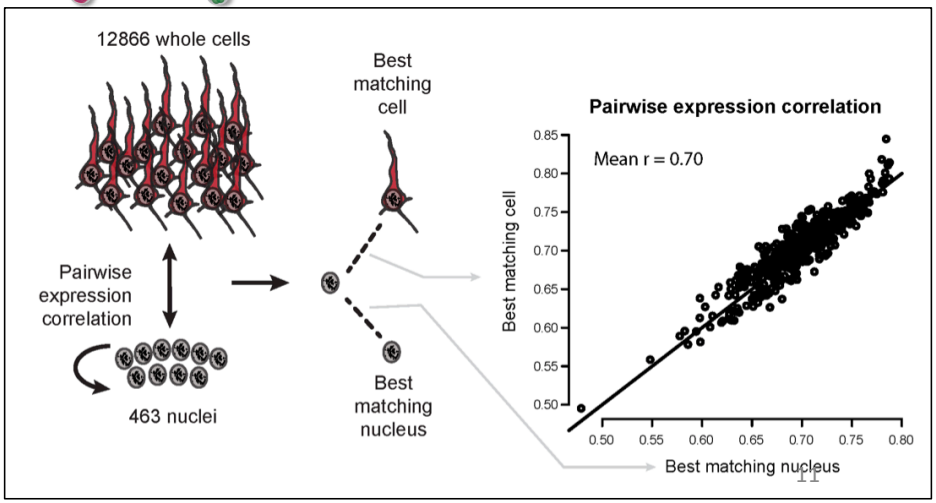
https://www.researchgate.net/figure/Single-cell-analysis-reveals-heterogeneity-Traditional-experiments-on-bulk-samples-mask_fig1_312664044

Considerations on wet-lab design

- Is your cell population under-represented in your tissue?
  - Unbiased assessment?
  - Genetic/antibody labeling combined with FACS sorting
- Integrity of your cells is critical
  - Intact and fully dissociated single cells
    - Short enzymatic dissociation procedure
      - Fully dissociated (minimal doublets/aggregates
      - No damage (no stain for DAPI)
    - Best for soft/liquid or developing tissue in which cells are not densely connected
    - Validate with microscopy
  - Single nucleus RNA-seq
    - For difficult tissues
      - Frozen human tissue
      - Mature neuron, in white matter
    - Simple mechanical dissociation
    - Clean up is crucial
    - 10% of total mRNA



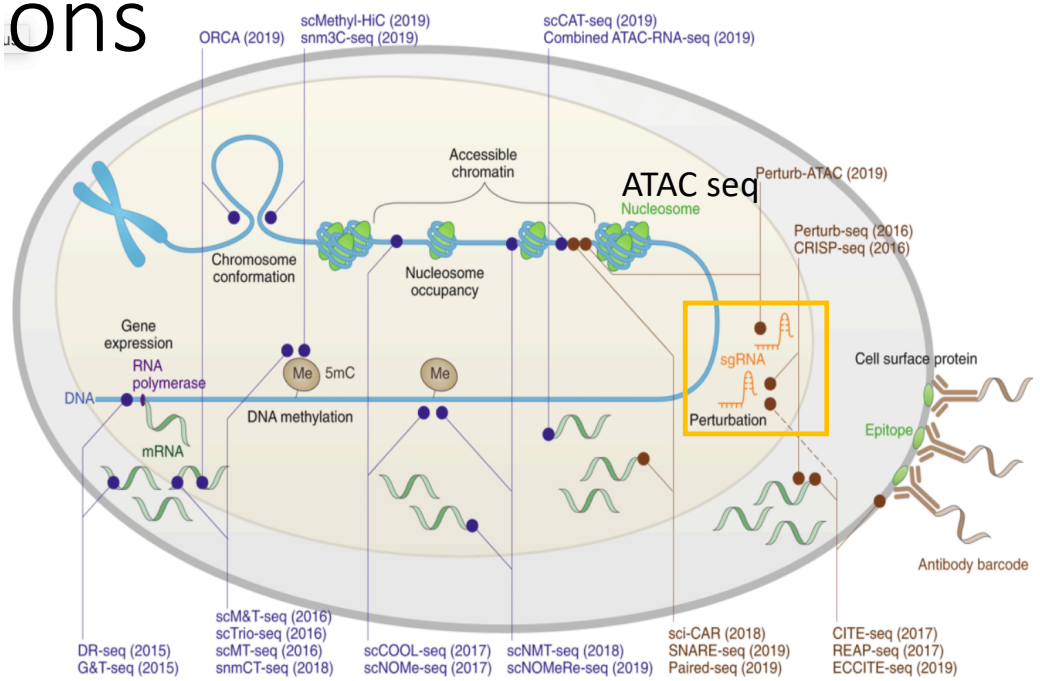




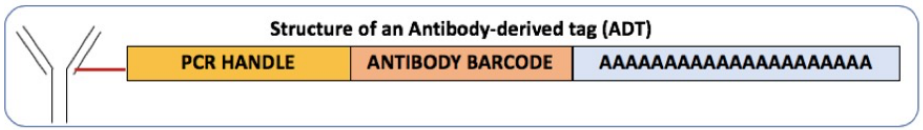
National Institute of Allergy and Infectious Diseases

Downstream validations

- Validate with another sc-omics?
  - 10X offers commercial reagents for
    - CITE-seq for surface protein antigens
    - ATAC-seq to access open chromatin
- Biological validations make an distinction
  - At expression level For key/novel molecular markers
    - Immunohistochemistry (IHC)
    - In Situ Hybridization (ISH)
  - In vitro functional validation?
    - Fast but in an artificial environment
  - In vivo functional validation?
    - Expensive and time consuming



https://doi.org/10.1038/s41592-019-0691-5



CITE-seq: Cellular Indexing of Transcriptomes and Epitopes by Sequencing

ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing

National Institute of Allergy and Infectious Diseases

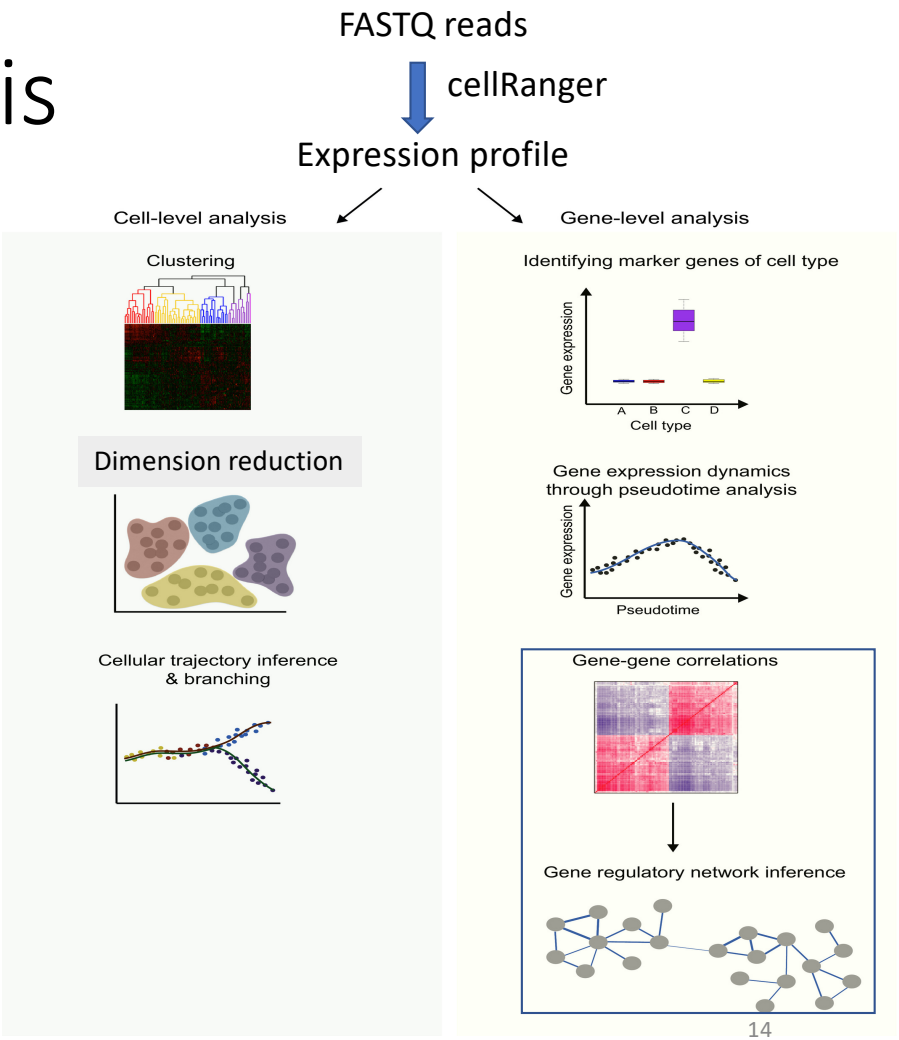
Dry-lab workflow

- General description of steps
  - Illustration of the overall workflow
  - Important quality controls! –commonly used ones
- To highlight a few important analysis
  - Dimension reduction → to find global clusters
  - Trajectory → to study gradual transition
  - Gene modules → to find meaningful gene modules
  - Meaning of gene modules → Functional annotation of gene modules
  - Gene Regulatory Networks → to find key nodes in your gene mocule
- Links for an extensive list of single-cell tools and further reading

  eg. https://github.com/sdparekh/awesome-single-cell

General steps of analysis

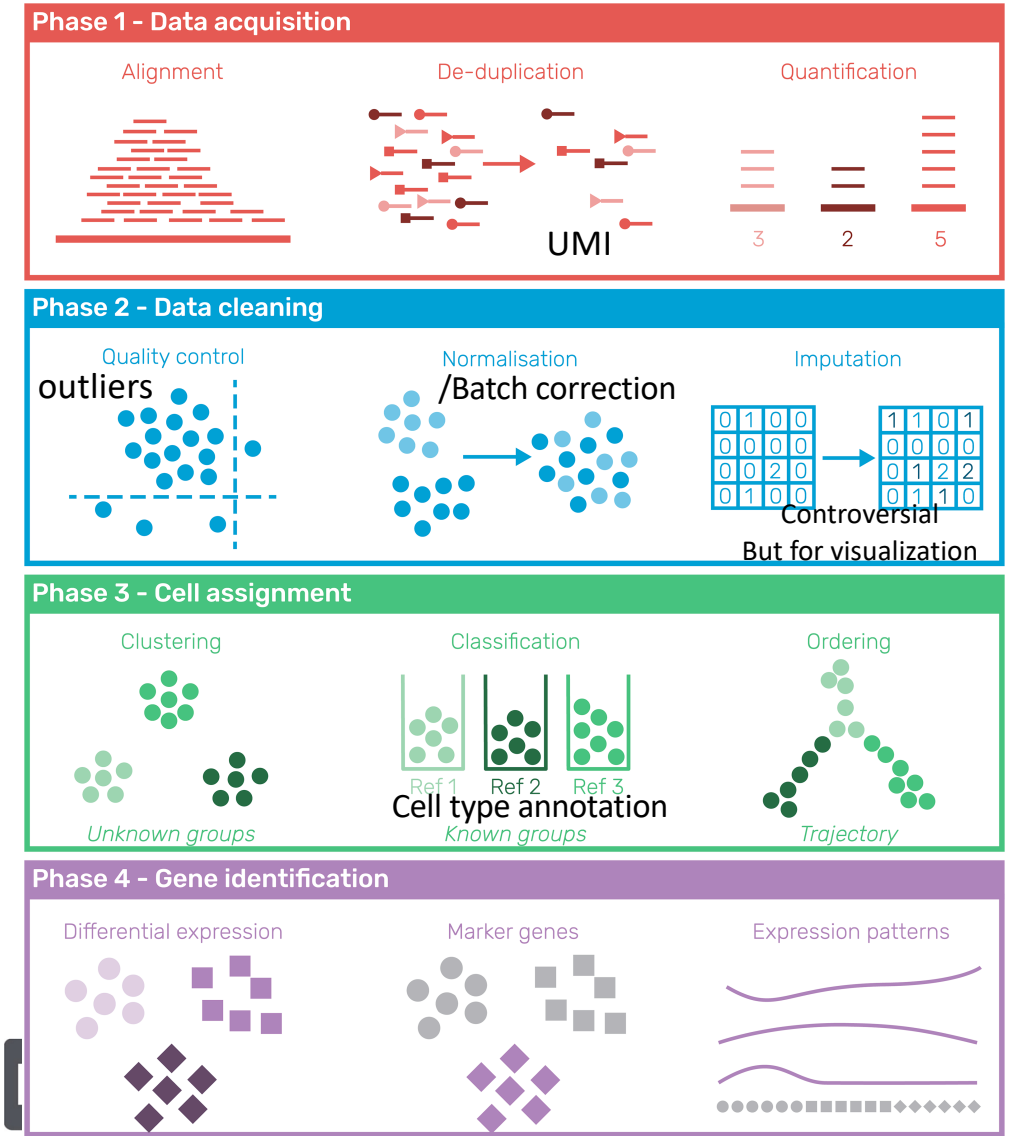
- Read-based
  - Fast QC
  - UMI, unique molecule identifier
  - spike-in, to estimate the absolute copy number of total RNA
- Cell-based
  - Contamination and denoising
    - Remove background noise and doublets.
  - Imputation (MAGIC) for visualization
  - Dimension reduction
    - PCA, TSNE, UMAP and Auto-encoder (NN).
  - Trajectory analysis
- Gene-based
  - Dynamic tree cutting to define clusters
  - Functional annotation of gene lists
  - Gene regulatory network analysis

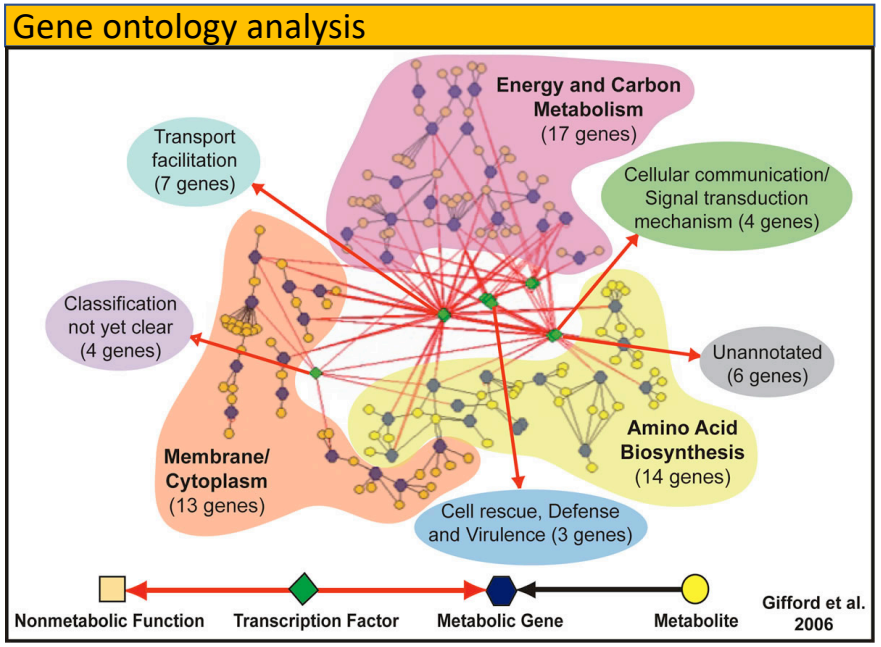


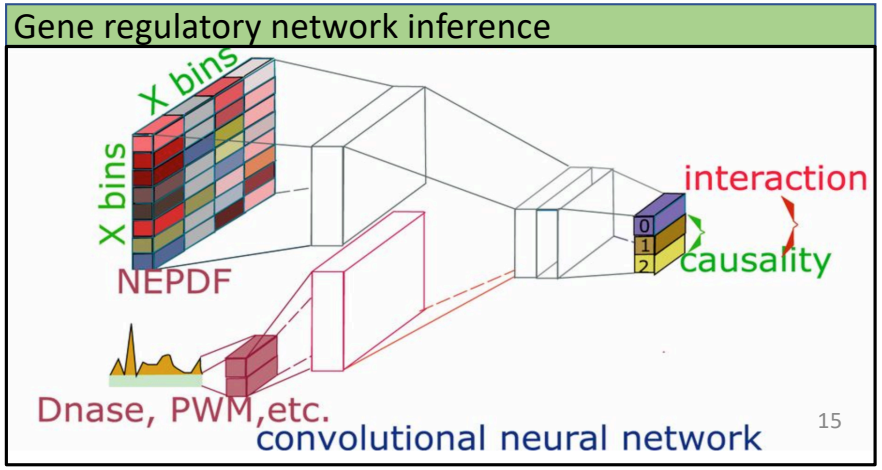
https://febs.onlinelibrary.wiley.com/doi/full/10.1002/1873-3468.12684

National Institute of Allergy and Infectious Diseases

More visual illustration…







https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1006245.g002

https://www.pnas.org/content/116/52/27151 https://public.ornl.gov/site/gallery/detail.cfm?id=323

QC are very important—common ones

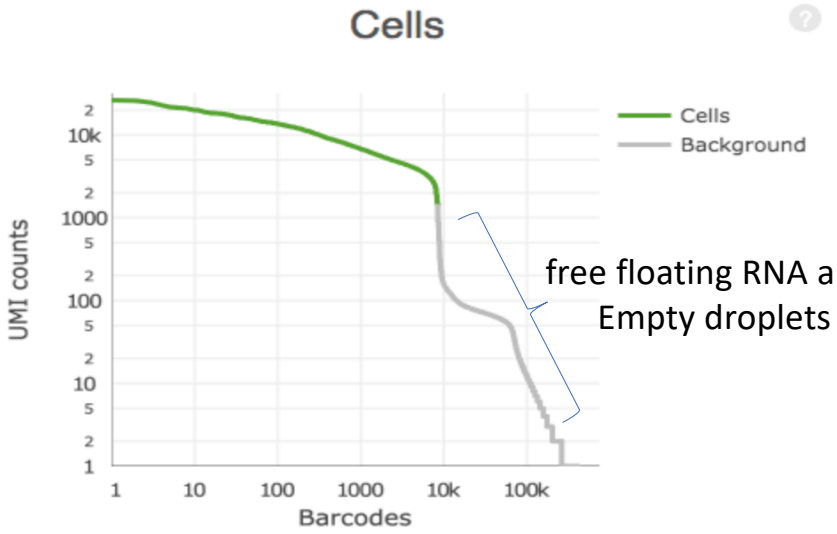
- Quality of reads – fastqc
- Mitochondrial reads?
  - Too much of mitochondria reads (mt reads >5-10%?) may indicate that cells are dying/dead/broken
- How many cells are you capturing?
  - Few thousands in each 10X run
- The sequencing depth
  - Are they acceptable in the field (minimal 1-2,000 reads /cell?) determined by the CellRanger
- Alignment to the genome and exons
  - Should be 90-100% to the genome
  - A reasonably narrow range 70-80% to the exons
  - Could be 30% to exons if you use nucleus, which contain lots of introns
- Expected markers expressed?
  - Highly expressed genes, cell type markers, automatic detection such as scMCA etc
  - Be prepared to see differences between RNA (because of the depth and dropouts) and proteins
- Batch effect? Confounding factors?
  - Can be evaluated by WGCNA and visualization in PCA or tSNE
  - Is your dimension reduction capturing biological or technical variations?

Read-based

- Reads level
  - CellRanger
    - cellranger mkfastq
      - Generate fastq files from image ".bcl" files
    - **cellranger count**
      - umi, unique molecule identifier
    - cellranger aggr
      - Combine count data from multiple batches
  - (For CITE-seq and HASH-tag)
    - Cite-seq-count

Output:

```
$ cd /home/jdoe/runs/sample345/outs
$ tree filtered_feature_bc_matrix
filtered_feature_bc_matrix
├── barcodes.tsv.gz     --cells
├── features.tsv.gz     --genes
└── matrix.mtx.gz       --sparse matrix
0 directories, 3 files
```

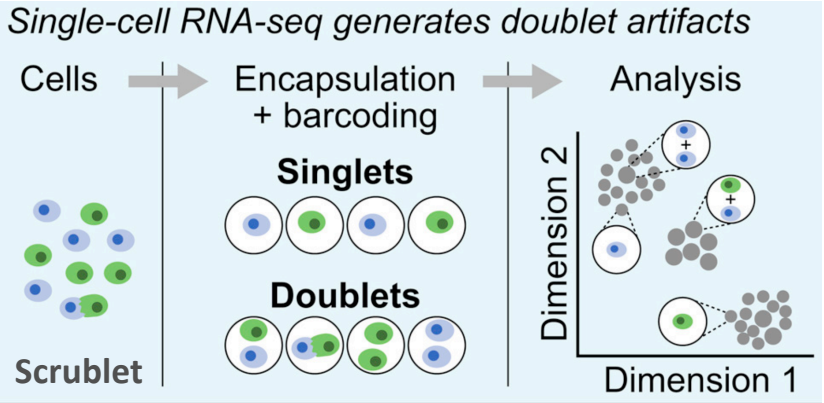


Ranked by number of associated UMIs

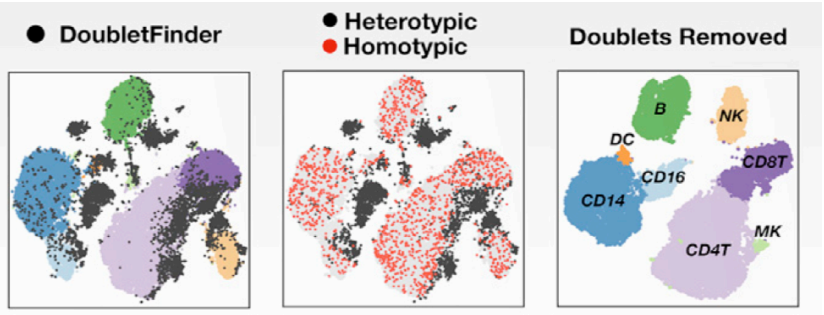
https://davetang.org/muse/2018/08/09/getting-started-with-cell-ranger/

Cell-based methods

- Cell type/state detection and annotation
  - PCA, tSNE, UMAP, autoencoder
  - Annotation of clusters
    - Known markers
    - Software such as scMCA
      - Based on correlation with known cell types
- Contamination
  - Removing doublets and aggregates
- Gradient
  - Monocle, slingshot, whishbone etc
  - Velocity
  - Customized codes



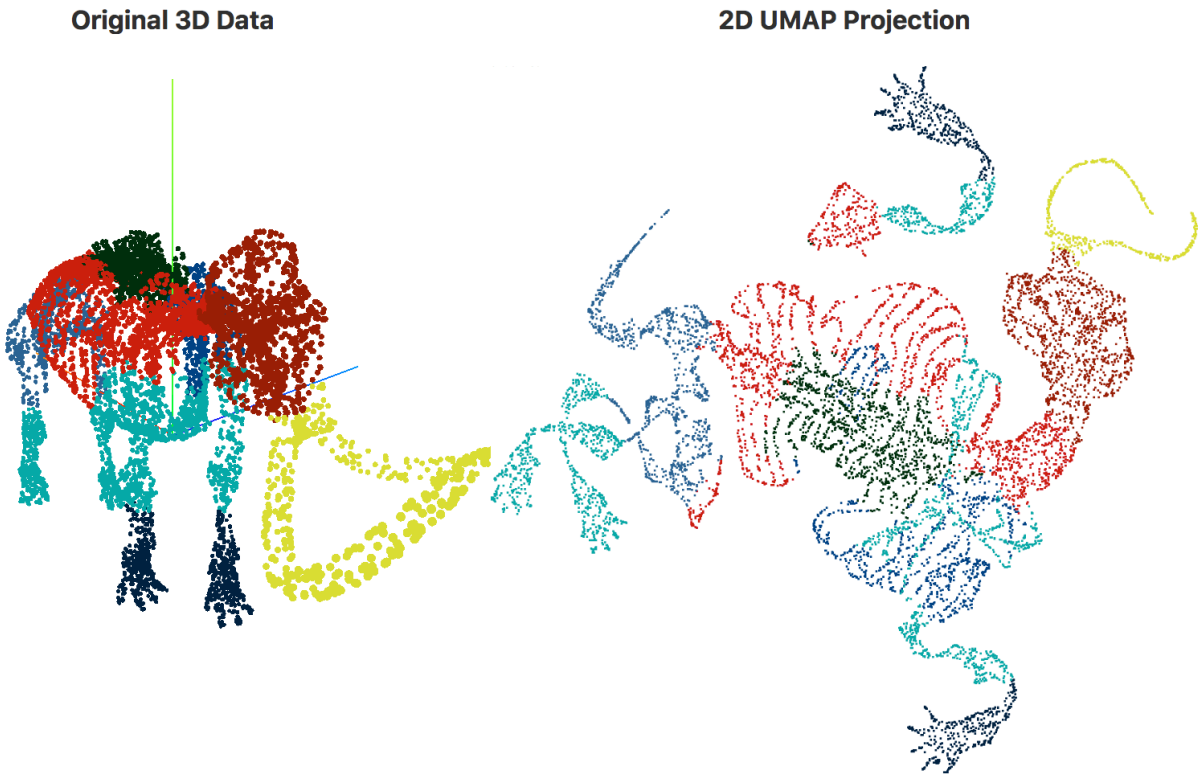
https://www.sciencedirect.com/science/article/pii/S2405471218304745



https://www.sciencedirect.com/science/article/pii/S2405471219300730?via%3Dihub

Careful when you want to find novel cell types

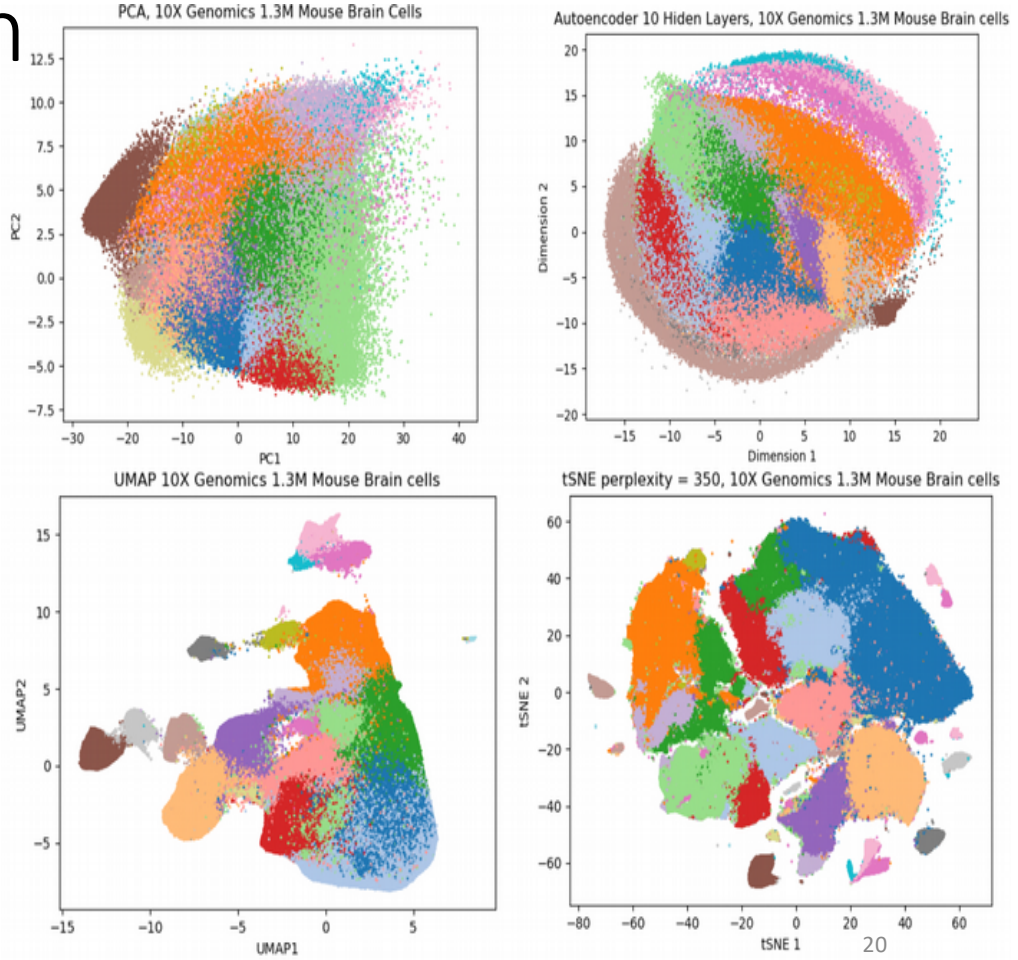
Why dimension reduction? An intuitive illustration



Screenshot by Mariam from
https://pair-code.github.io/understanding-umap/

Dimension reduction

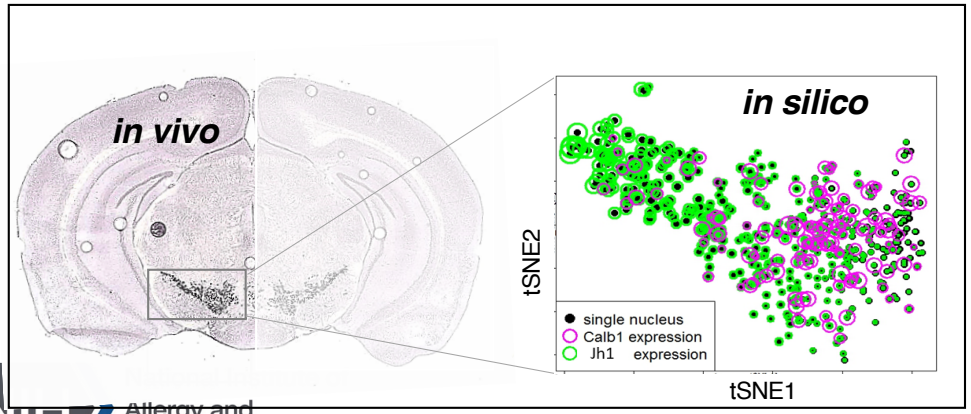
- Dimension reduction
  - PCA
    - Linear reduction
    - Based on distances
    - 2D structure in PCA depends on certain observed dominant variations
    - Often not great for big number of cells
  - tSNE
    - Non-linear reduction
    - Attention to local similarity
    - Global shape is less meaningful
    - Add new data changes the whole pattern
  - UMAP
    - Consider both global and local structure
    - Learnt embeddings can be saved for new batch of data
  - AutoEncoder
    - Fast algorithm for large number of cell

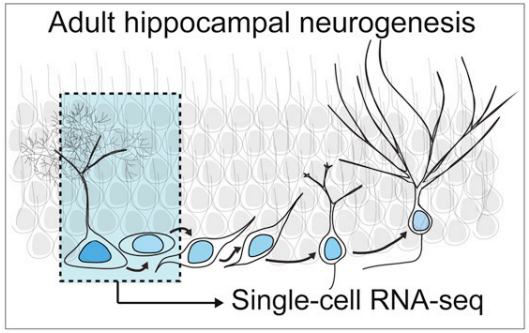


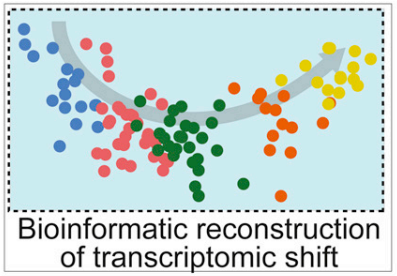
https://towardsdatascience.com/deep-learning-for-single-cell-biology-935d45064438

Trajectory analysis

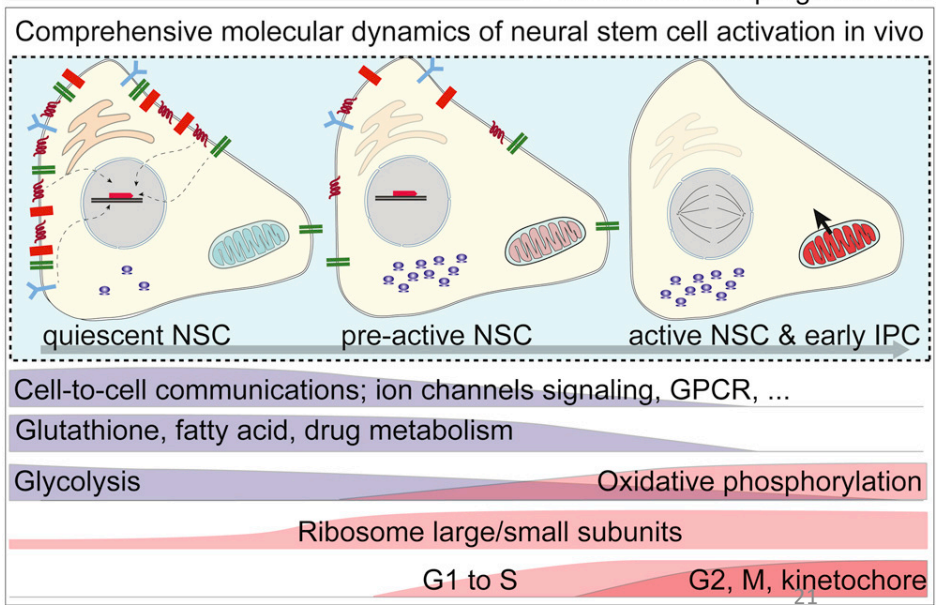
- Temporal and spatial gradient
  - Often seen when you are focusing on one cell type
  - Observed using an unbiased algorithm
  - Use known markers to annotate the pattern interested, and assign direction
- Aim
  - Further delineate the cells and genes with temporal/spatial information





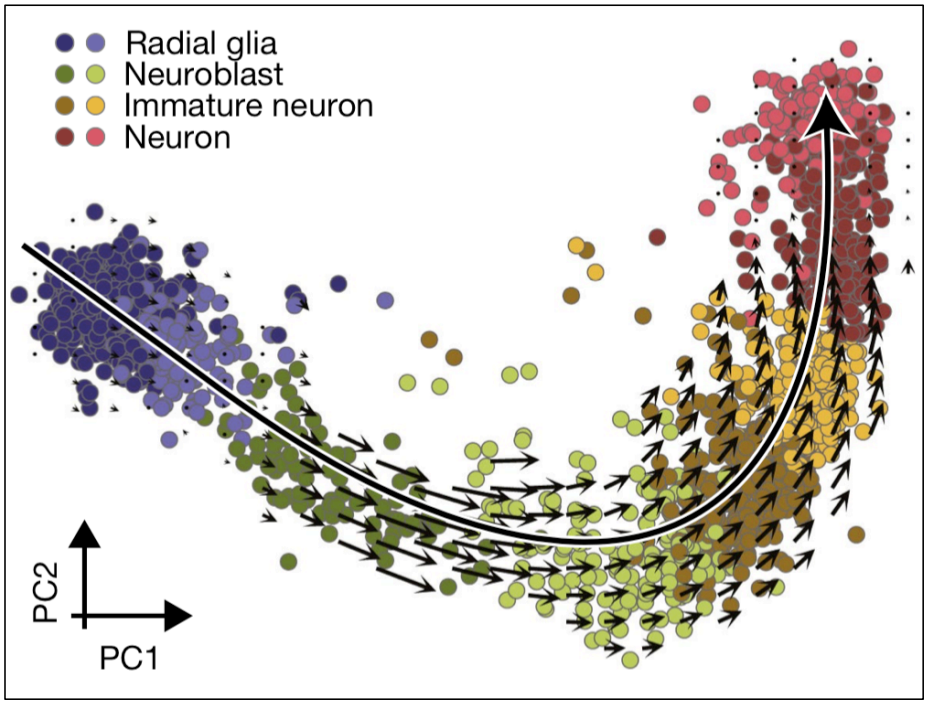


NSC: neural stem cell
IPC: intermediate progenitor cell



Direction prediction by intron retention

- When cells differentiate, **new genes** will start to be expressed
- Transcripts have introns and will be spliced off given time
- Through assessing intron retention, a direction of development can be assigned
- May require higher **depth** than usual sequencing



https://liorpachter.wordpress.com/tag/velocyto/
http://pklab.med.harvard.edu/software.html

Gene-based approach

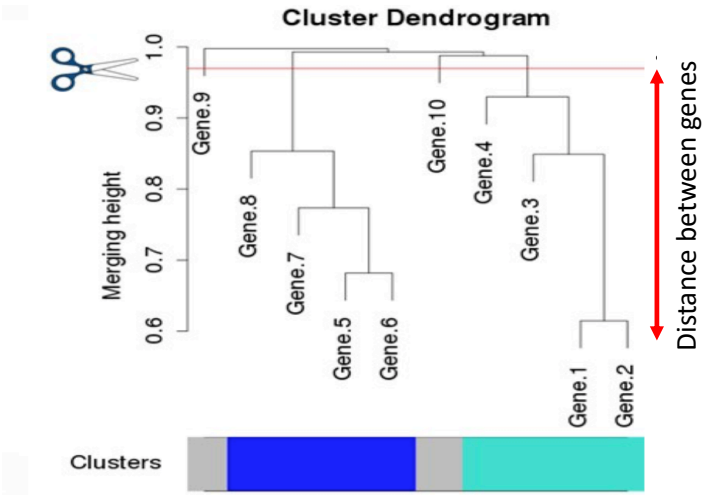
- Most often we want to know the key genes
- To find gene modules using
  - Conventional cluster identification using tree cutting is of little use.
  - Weighted correlation network analysis (WGCNA)
- To find biological meaning of gene modules using
  - clusterProfiler
- Find internal relationship between genes using gene regulatory network analysis
  - SCENIC
  - bnlearn
  - PIDC

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559

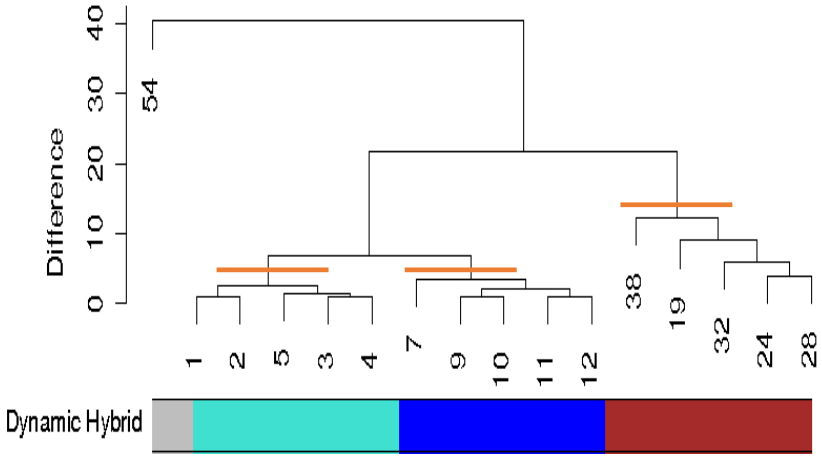
Finding gene clusters using WGCNA

Conventional treecutting

Just a few genes to determine manually



Dynamic tree cutting: clusters determined by adaptive tree cutting

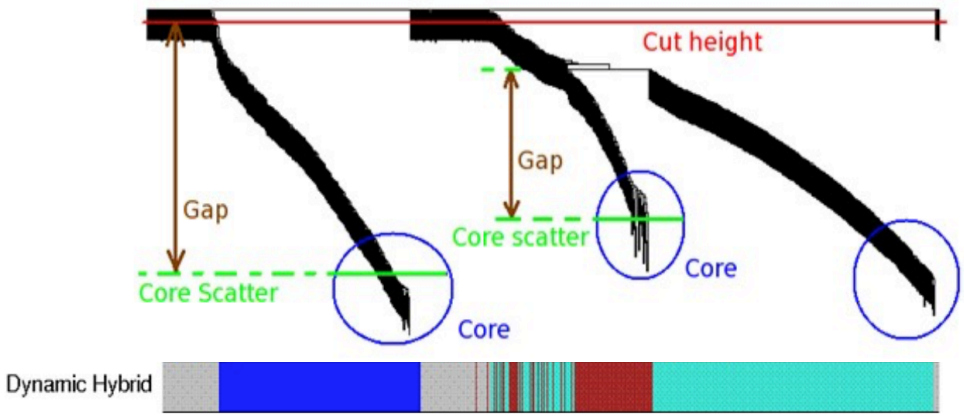


National Institute of Allergy and Infectious Diseases

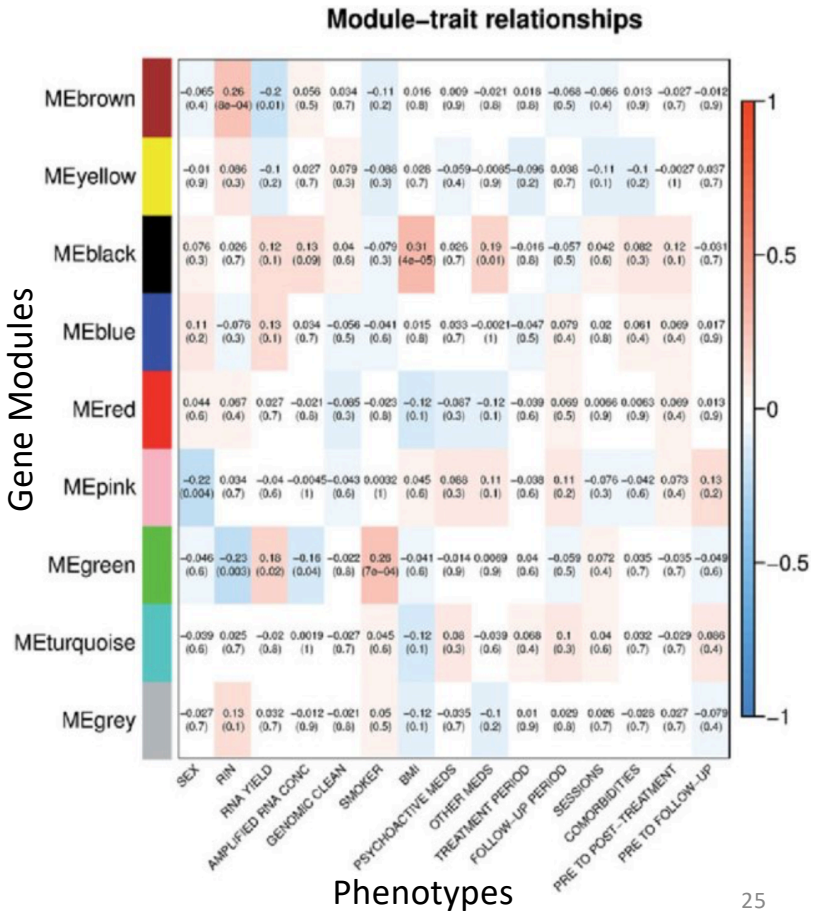
https://slideplayer.com/slide/7402978/

Gene modules <-> traits

- Systematic identification of meaningful gene modules in a complex network
- Systematic way to find the biological/technological correlation with specific gene modules



A nice feature of WGCNA

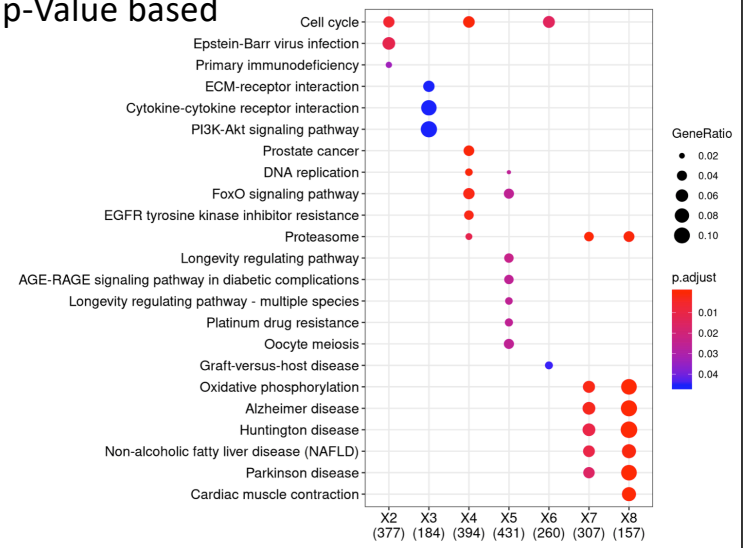


National Institute of Allergy and Infectious Diseases

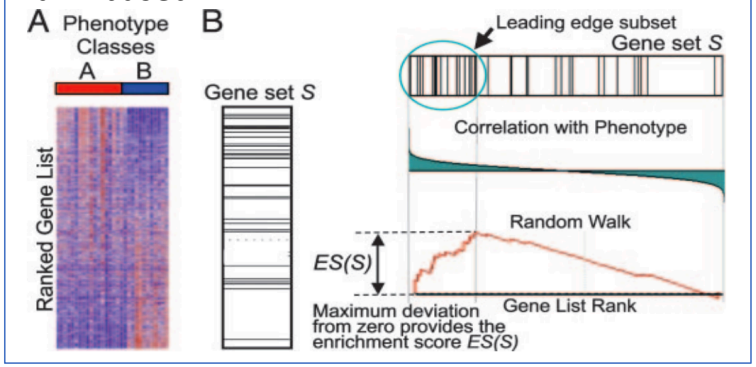
Functional annotation

- Differential expression, GO and KEGG enrichment, wikiPath, disease ontology etc
  - p-value based method to detect significantly changed genes
  - Good to detect large changes and the significance
  - Sometime difficult with single cell data as the number of cells (n) in different clusters constantly differ, which will affect P values
- Gene Set Enrichment Analysis (GSEA)
  - Rank-based, not p-value based
  - Good to detect small but consistent change in gene expression

p-Value based



Rank-based



National Institute of Allergy and Infectious Diseases

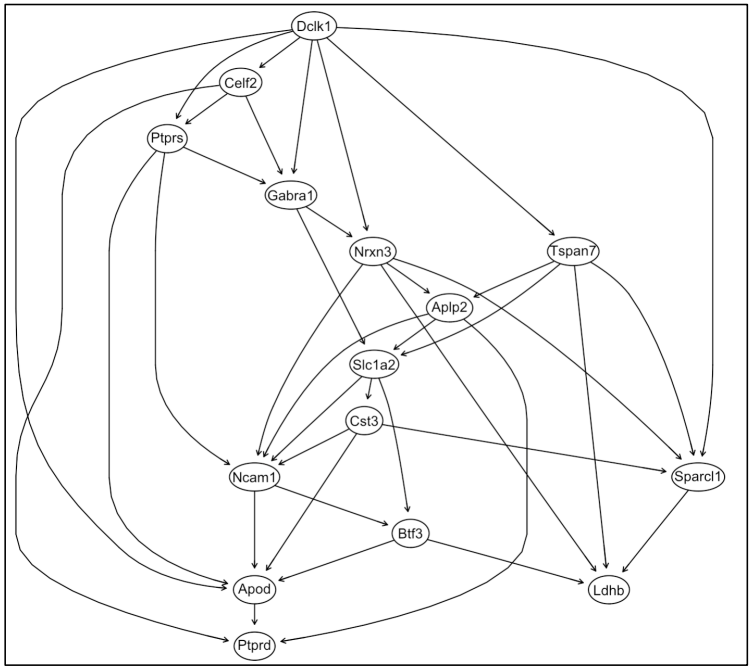
For detailed tutorial: https://yulab-smu.github.io/clusterProfiler-book/index.html

A jupyter notebook with codes and results: https://github.com/zhuy16/clusterProfiler_notebooks

Gene Regulatory Network Analysis

- Single cell data are inheritably suitable for assessing statistical relationships
  - High number of data points for statistical inference
- Statistical relationship between genes
  - Mutual information
  - Bayes theorem
  - Regression forest
  - Auto encoder
- Base on prior knowledge, binding motifs of transcriptional factors on promoter of a list of genes
  - cisTarget
  - SCENIC

Statistical relationship inferred by Bay's Theorem



National Institute of Allergy and Infectious Diseases

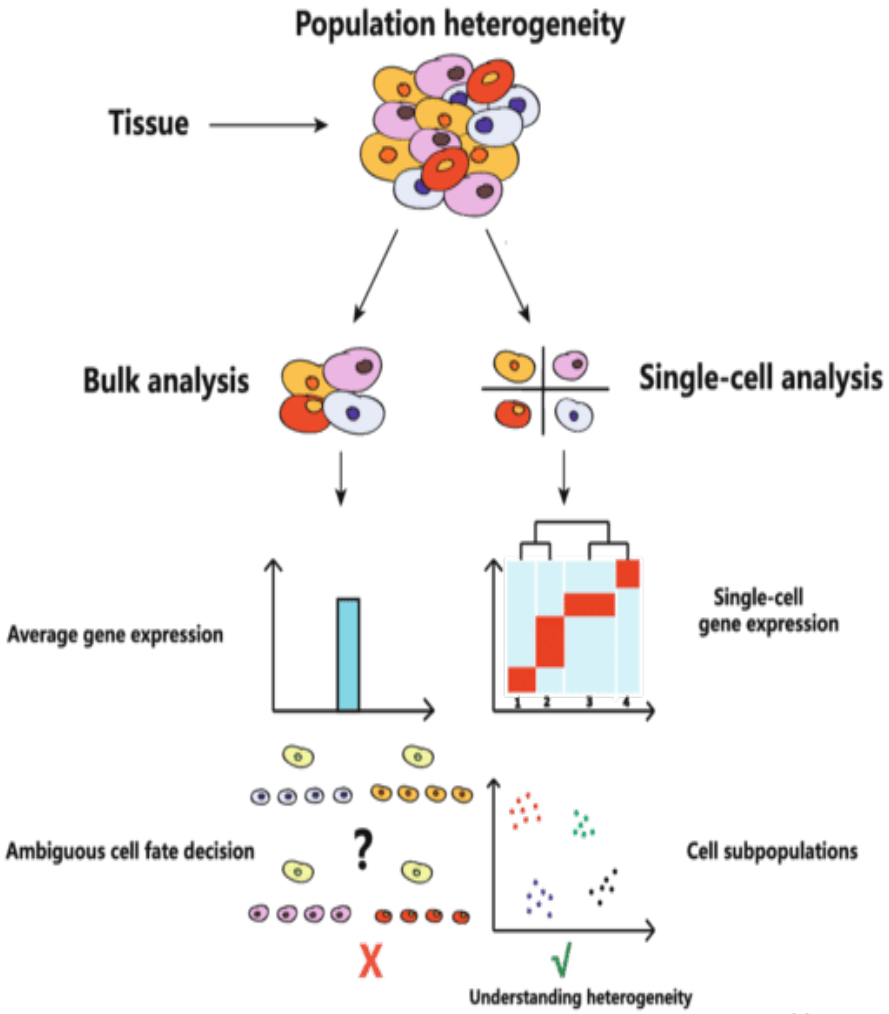
Gene-regulatory networks

Software	ARACNE	NetworkInference/ PIDC	bnlearn	GENIE3	iRegulon	SCENIC
semantics	Mutual information	Partial Information Decomposition	Bayes theory	Random Forest, Regression tree	Promoter and TF binding sequence, database	Combination of regression and promoter sequence
years published	2006	2017	2009	2010	2014	2017
No. of cited	2179	82	894	658	337	265
FullName/ explanation	Algorithm for the Reconstruction of Accurate Cellular Networks	Using proportional unique contribution (PUC) to a target gene	Bayes net structure and parameter learning, causality	GEne Network Inference with Ensemble of trees	reverse-engineer the transcriptional regulatory network with regulatory sequence analysis	single-cell regulatory network inference and clustering
Implementation	GUI (geWorkbench)	Julia	R	R	GUI (Cytoscape)	R, Python
type of experiment	Microarray, bulk RNA-seq	Single cell data	General	single cell data	a list of gene names	single cell data
input format	csv	csv	csv	csv	a list	csv/loom file
output	network file	network file	directed network file	network file	network file/binding sequences	network file/heatmap

For detailed tutorial:
https://github.com/niaid/Gene_Regulatory_Networks

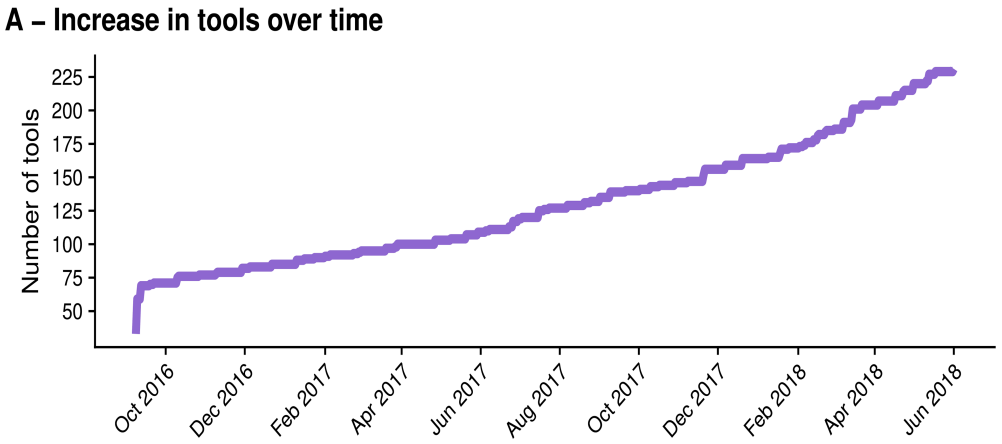
Summary

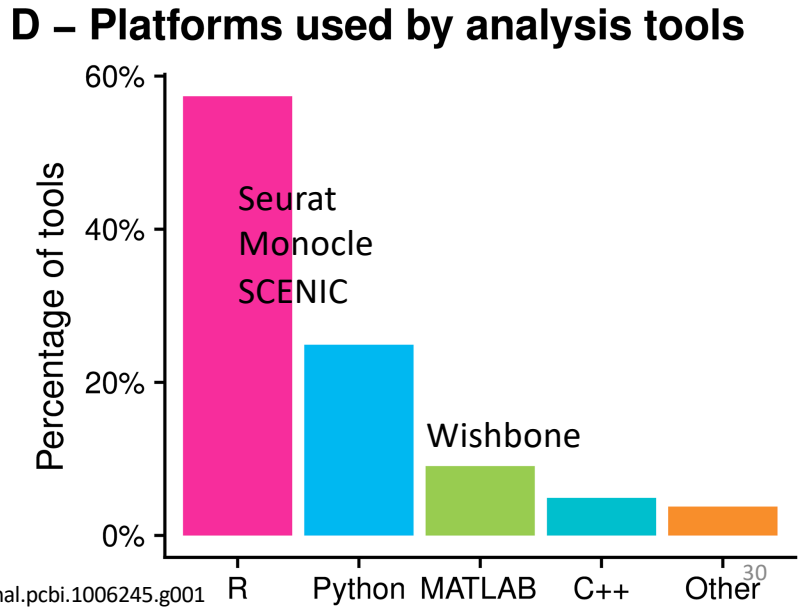
- Wet lab
  - **What**, made sequencing at single cell level possible
  - **Why**, single cell RNA-seq
  - **How**, to design single cell RNA-seq
- Dry lab
  - Read-based methods
    - QC
    - Deduplication
    - Determining barcodes/cells
  - Cell-based methods
    - Remove doublets
    - Dimension reduction
    - Trajectory analysis
  - Gene-based methods
    - WGCNA
    - Functional annotation – clusterProfiler
    - Gene regulatory network analysis
  - **Comprehensive software**
    - **GUI- Partek**
    - **R-Seurat**
  - **General considerations**



Implementation

- R vs Python
  - R is more for scientist
    - Intuitive and small
      - Visualization, data frames.
      - Big community
      - Good documentation
      - But comparatively slow
        - Use parallel computing
    - Good visualization
  - Python is more for big data and deep learning
    - Data is getting bigger and bigger – millions of cells, and >10,000 genes
    - More deep learning methods involved

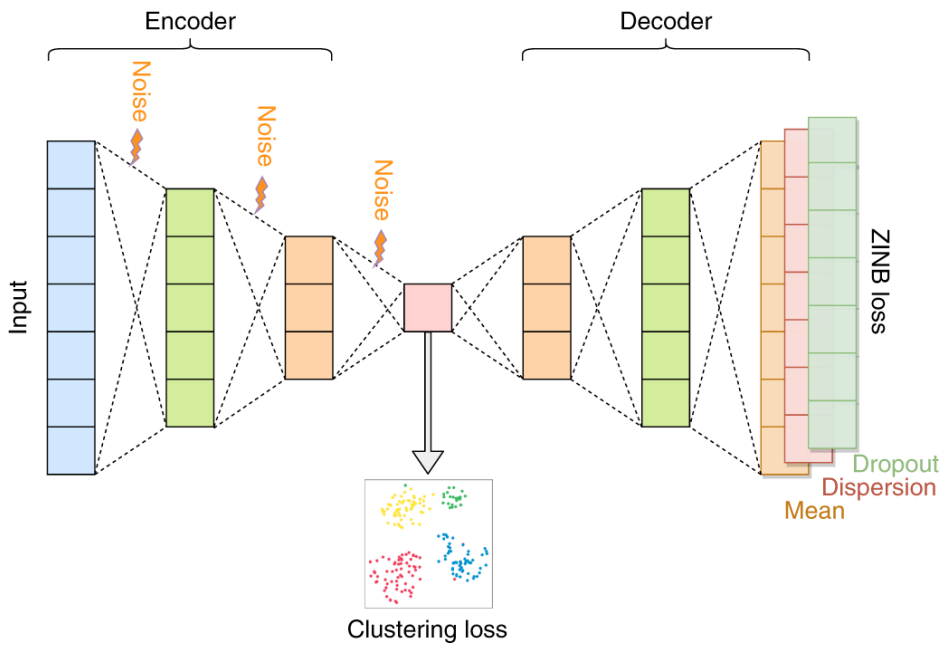




https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1006245.g001

Deep learning in single cell biology

- Why deep learning
  - Large sample size for statistical inference
  - High dimensionality
    - needs representation in lowdimensional space
  - High noise
- Application
  - Dimension reduction
  - Imputation
  - Gene regulatory network



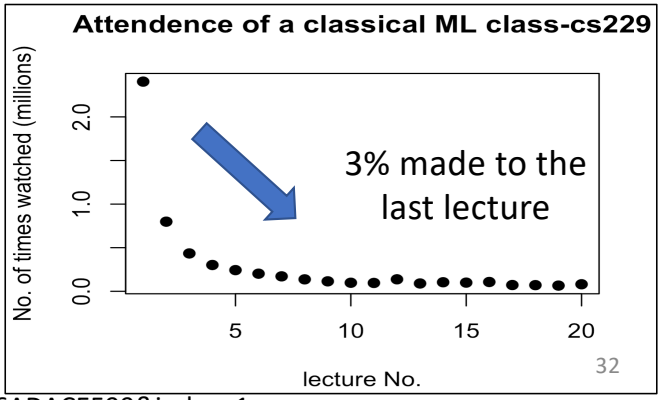
National Institute of Allergy and Infectious Diseases

Theoretical considerations

- Evaluating software on theoretical bases can be hard for biologist…
  - Linear algebra
  - Calculus
  - Statistics and probability
  - Regression
  - Information theory
  - Network theory
  - Bayes theory
  - Topology and manifold
  - Deep Neural Net
- Assessing meaningfulness through output
  - Ultimate judge

Equations

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{i \neq j} (p_{ij} - q_{ij}) \frac{\mathbf{y}_i - \mathbf{y}_j}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}$$

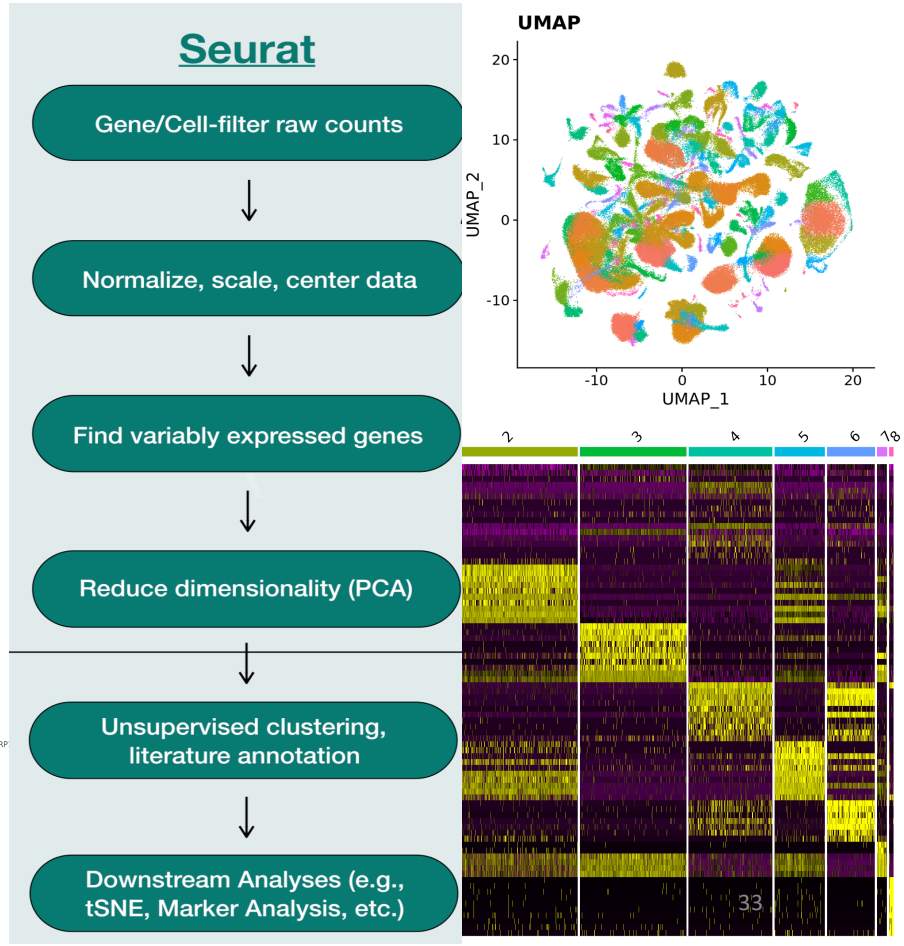


National Institute of Allergy and Infectious Diseases

https://www.youtube.com/watch?v=UzxYlbK2c7E&list=PLA89DCFA6ADACE599&index=1

Seurat – comprehensive pipelines for single cell data -- Verena

- Seurat pipeline
  - General assessment
  - Cell type annotation
  - Batch correction and meta analysis
  - Multimodal analysis (for CITE-seq, Hash-tagging, ATAC-seq)
  - Comparative analysis across different conditions
- Translating to discovery
  - Spatial temporal trajectory
  - Functional annotation of gene sets – GO, KEGG, GSEA
  - Gene regulatory network analysis

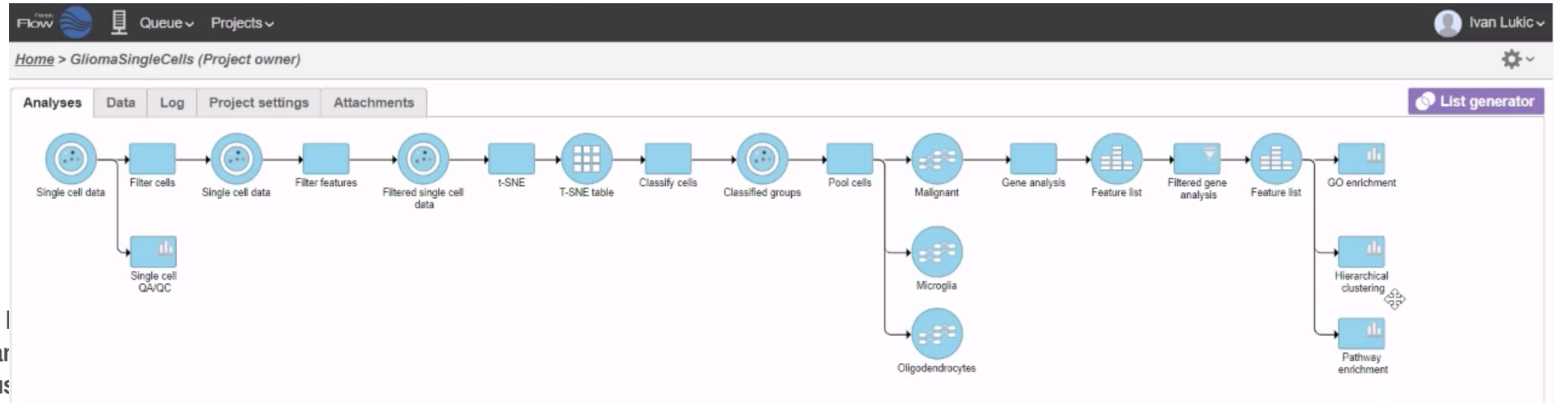


National Institute of Allergy and Infectious Diseases

https://satijalab.org/seurat/vignettes.html

Graphic User Interface

- Partek analysis
  - Access to biowulf, NIH library to get an account
    - https://www.youtube.com/watch?v=cj9M--9zzgI
  - Besides, NIH Library has a license for Partek, and people who need it can get an account.
    - https://www.nihlibrary.nih.gov/resources/tools/partek-flow
  - Biowulf has an instruction page how to deploy it.
    - https://partekflow.cit.nih.gov/



Practical considerations

- What language used in analytic tools?
- What is the strategy/mechanism behind?
- Is there a Jupyter notebook/ or R markdown in a Github to use?
- What are the input format/requirement?
- What is the format of the output? Where is information of my interest?
- How to visualize the result?
- Biological interpretation?
  - What QC should I check to help interpretation?
  - How to avoid false interpretation?

Bottlenecks overall…

- Tissue
  - To get the cells that are best interest to answer your question
  - A appropriate model system for your question
- Analysis
  - scRNA-seq generate huge amount of data
  - Substantial expertise, with good understanding in both computation and biology
    - Readiness to learn and spirit to collaborate
- Timing
  - With a fast evolvement, publication is also time sensitive

Further readings

- Single cell softwares
  - https://github.com/seandavi/awesome-single-cell
- Some personal tips for learning bioinformatics
  - https://github.com/zhuy16/learning_notes
- Gene Regulatory network analysis
  - https://github.com/niaid/Gene_Regulatory_Networks
- clusterProfiler & Functional annotation of gene lists
  - https://yulab-smu.github.io/clusterProfiler-book/index.html
  - Converted to notebook: https://github.com/zhuy16/FunctionalAnnotation_notebooks/tree/master/notebooks
- Sean Davis's overview on scRNA-seq
  - https://figshare.com/articles/Single_Cell_Present_and_near_Future/12121674