RNA-seq Part II, an introduction to Single-cell RNA-seq analysis

Yunhua Zhu
computational genomics specialist, transcriptomics. BCBB

National Institute of Allergy and Infectious Diseases

Our helpdesk, Bioinformatics@niaid.nih.gov
Work email, zhuy16@nih.gov

Self introduction

- Bachelor @ National University of Singapore(NUS) in Biochemistry 2000-03
- Ph.D @ NUS in stem cell biology | 2006 -10
  - 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
  - Learning R, Bash script statistics through Google and Youtube
- Computational Genomics Specialist @ BCBB | 2019
  - Single-cell RNA-seq of bile duct tumors
  - Single-cell CITE-seq, HASH-tag of lung tumors
  - Single cell RNA-seq of T cells and intestinal epithelium
  - Single cell RNA-seq of B cells hyperplasia

  - SMARTseq2 pipeline with FASTQC, RSEM, SENIC
  - 10X genomics: Cellranger, Seurat, Monocle, RNA Velocity, CSI-microbes.
  - Deconvolution Bulk RNA-seq using ABIS and CybersortX.
  - Functional annotation with clusterProfiler.
  - Deep learning, SAUCIE, Solo, Autoencoders.

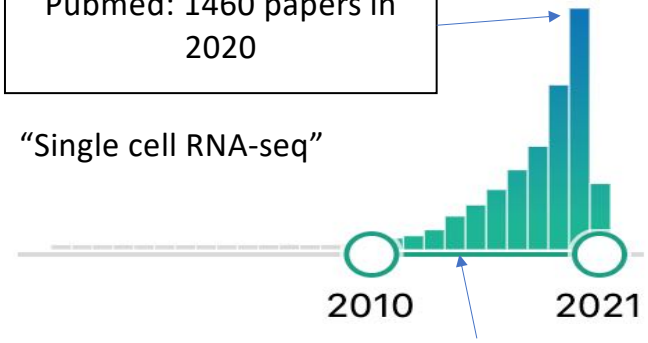
National Institute of Allergy and Infectious Diseases

An inter-disciplinary field

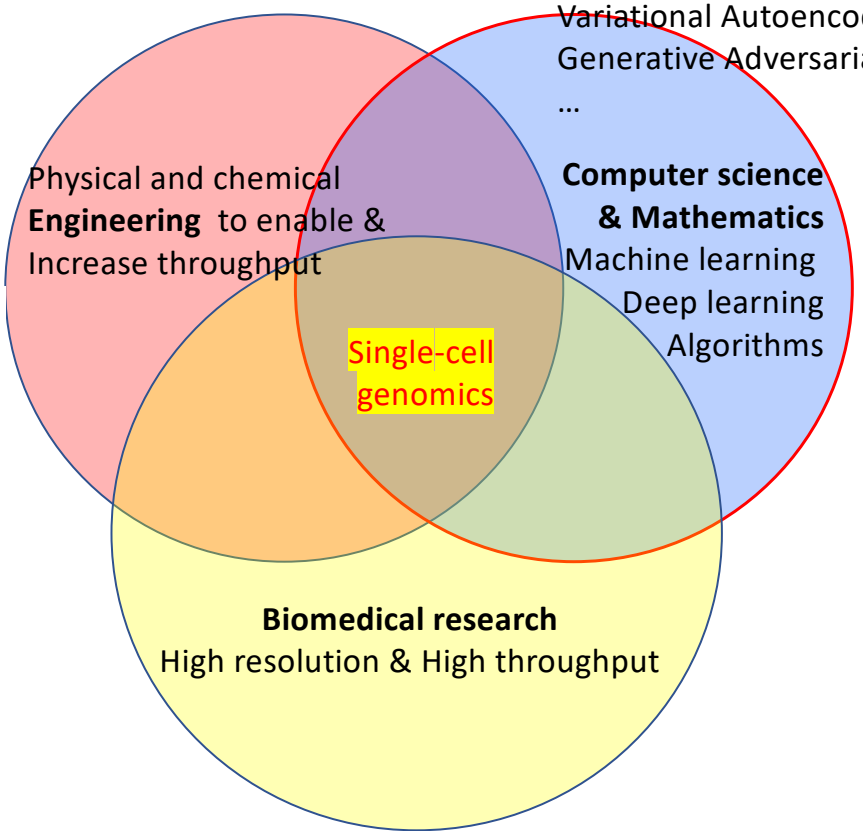
Single cell sequencing examines the **sequence information from individual cells** with NGS, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment. --wikipedia

Pubmed: 1460 papers in 2020

"Single cell RNA-seq"



Method of the Year 2013
Nature Methods

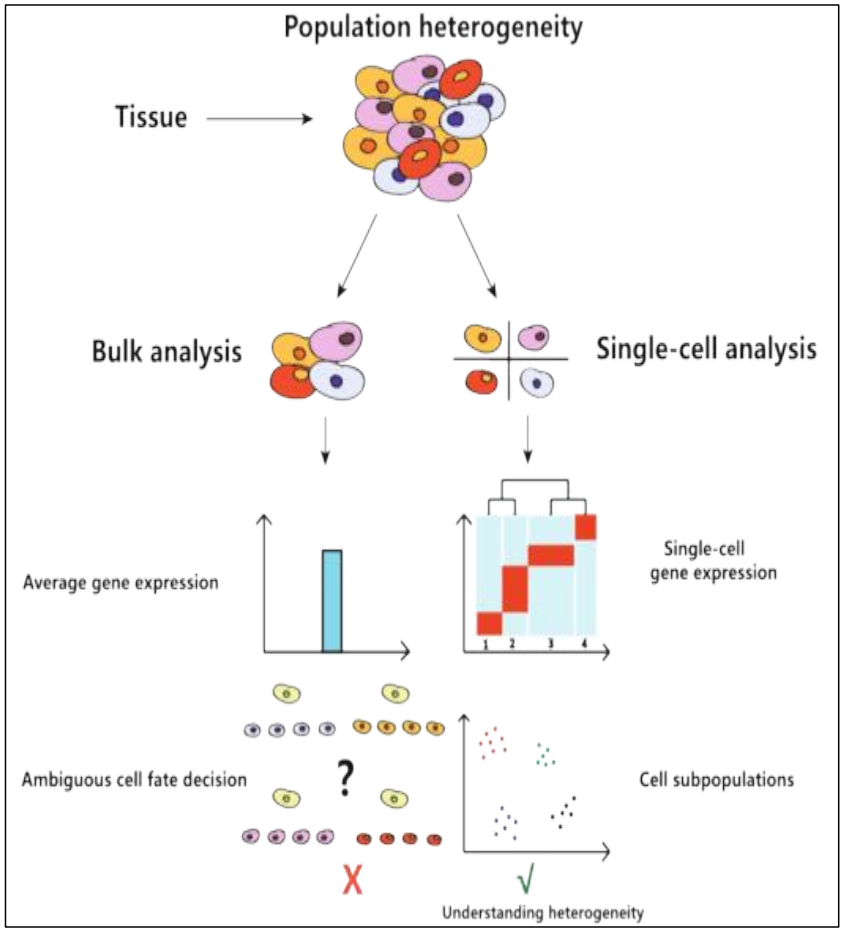


Information theory
K-Nearest Neighbors
Unsupervised clustering
Expectation maximization
Variational Autoencoder
Generative Adversarial Network
…

National Institute of Allergy and Infectious Diseases

Why single-cell RNA-seq?

- Advantages
  - High resolution for novel details
    - Cellular types and states
    - Reveal minor populations
    - Reveal gradual cell state 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).
- Challenges and opportunities
  - Low depth due to the highly multiplexed system
    - Genes with lower expression may not be reliably detected
  - High dropout rates, reads highly sparse
    - Not all genes can be picked up and amplified (5%-10%)
  - Huge amount of data requiring specific knowledge
    - Stay focused on your biology and extract valuable insight

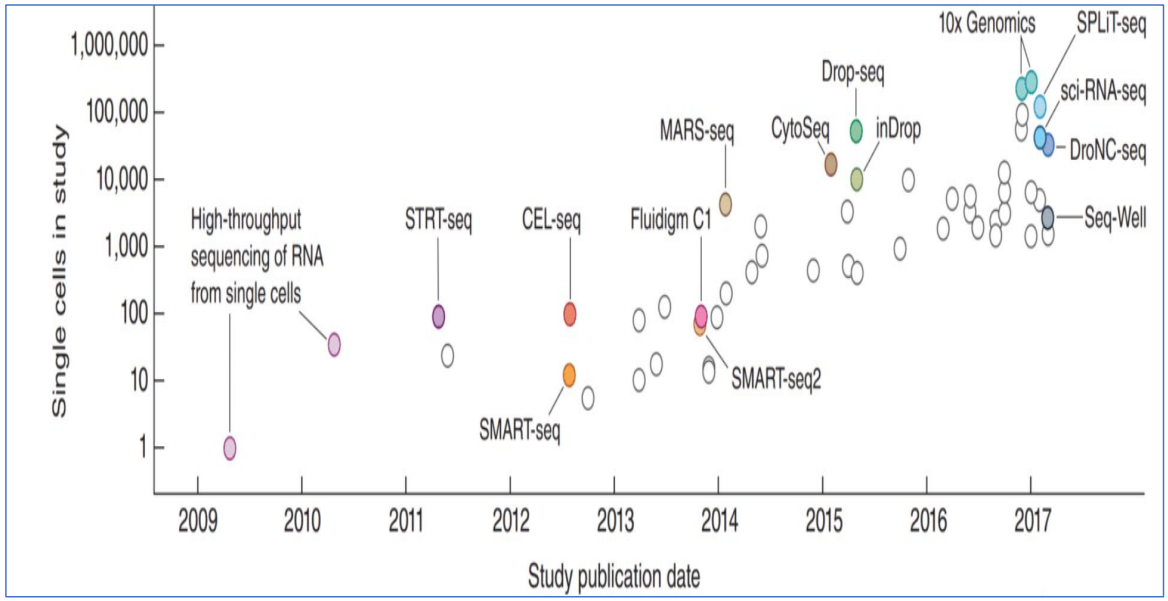


National Institute of Allergy and Infectious Diseases

Towards designing scRNA projects, What should we consider?

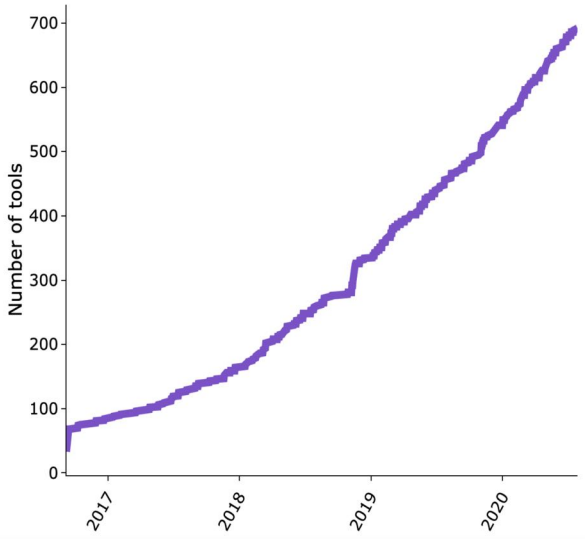
https://www.researchgate.net/figure/Single-cell-analysis-reveals-heterogeneity-Traditional-experiments-on-bulk-samples-mask_fig1_312664044

Evolution of Technologies



Exponential scaling of single-cell RNA-seq in the past decade. Nature protocols 13;4;599-604)

Number of analytic tools



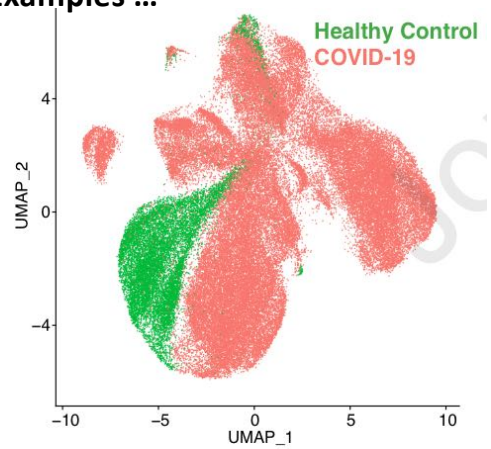
https://www.scrna-tools.org/analysis

THE HUMAN CELL ATLAS
White Paper
The HCA Consortium
October 18, 2017

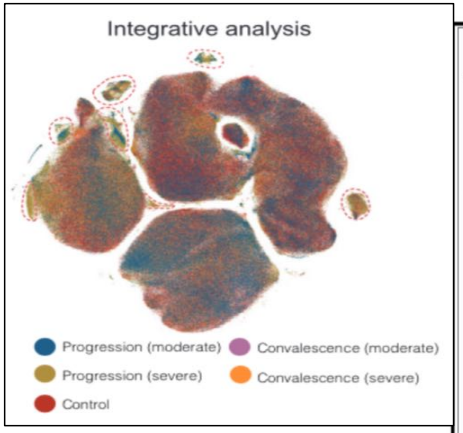
National Institute of Allergy and Infectious Diseases

10 billion cells, covering all tissues, organs, and systems

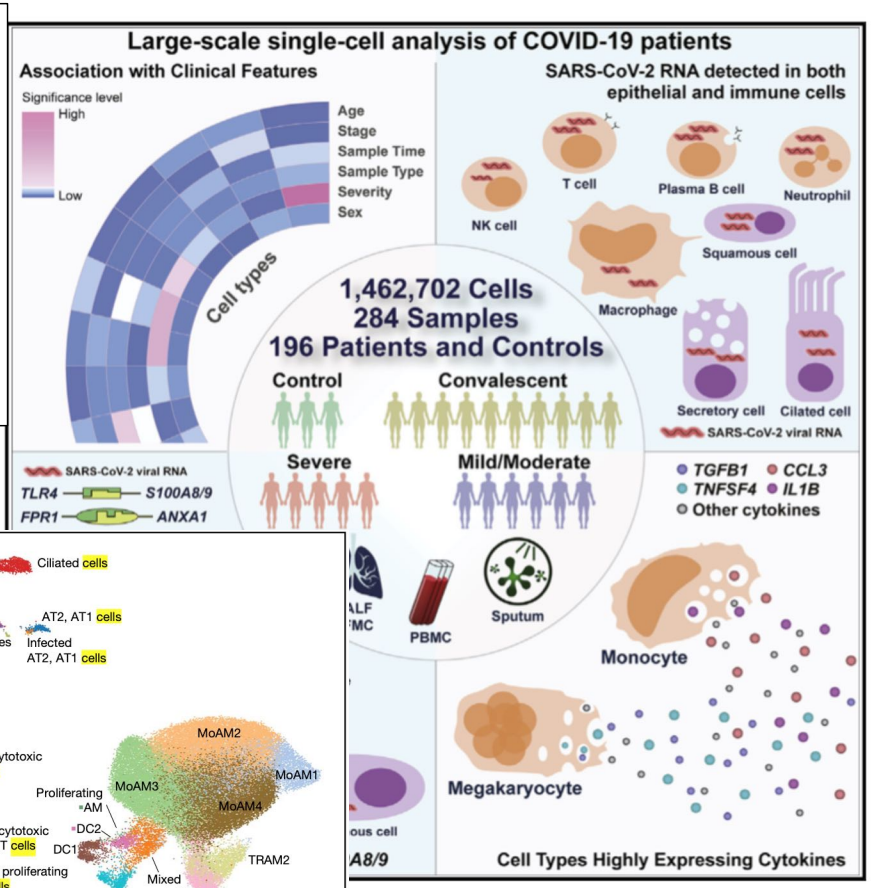
Examples …

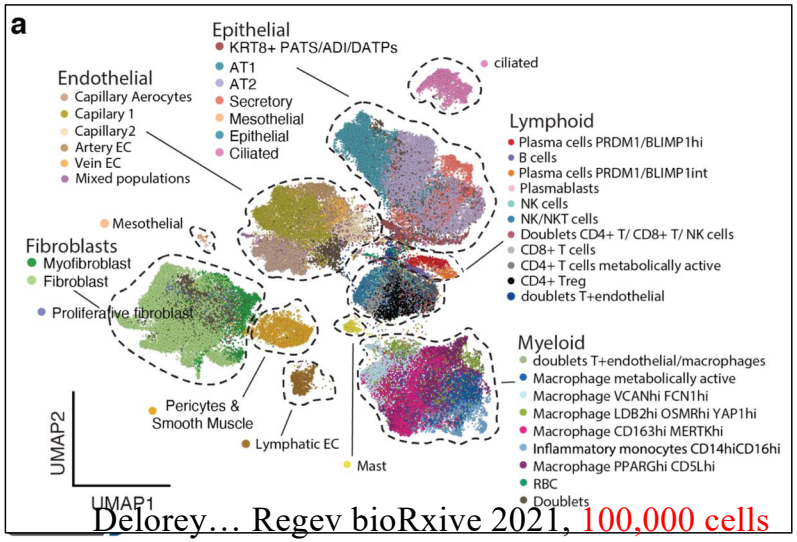


Liu … Tsang, Cell 0.4 million

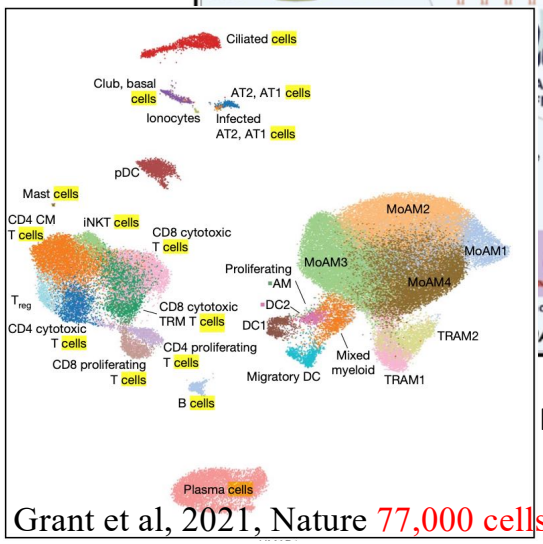


Zhang et al, Cell 1.4 million





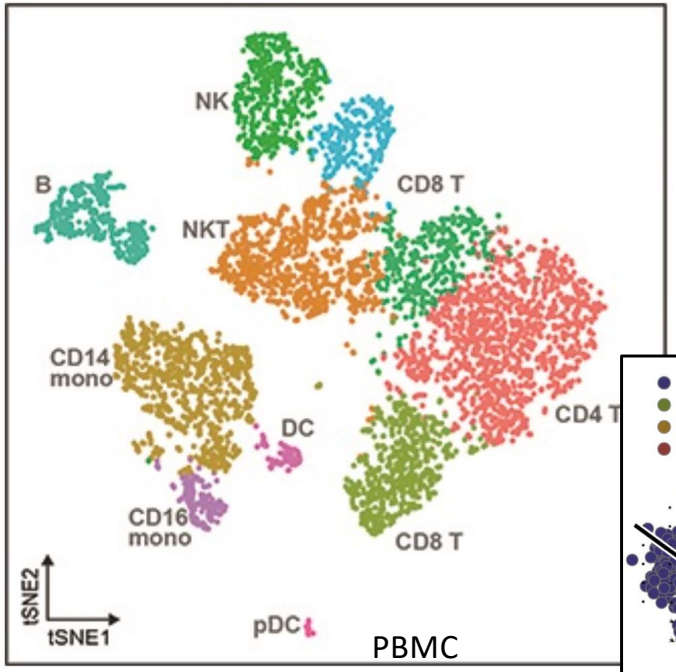
Delorey… Regev bioRxive 2021, 100,000 cells



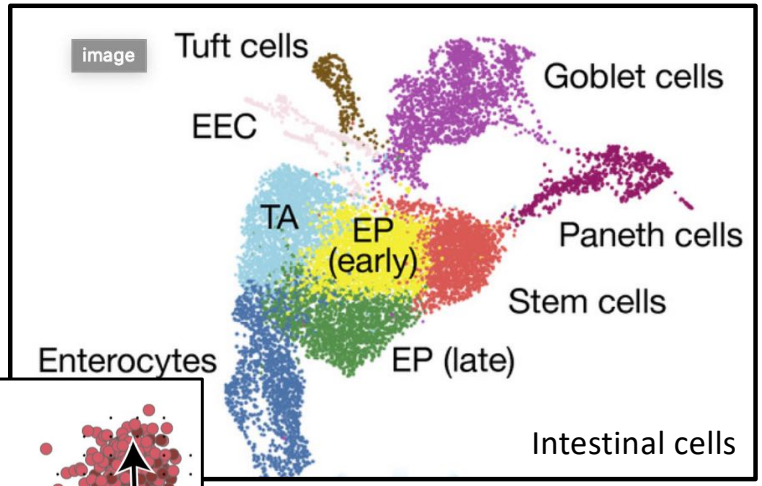
Grant et al, 2021, Nature 77,000 cells

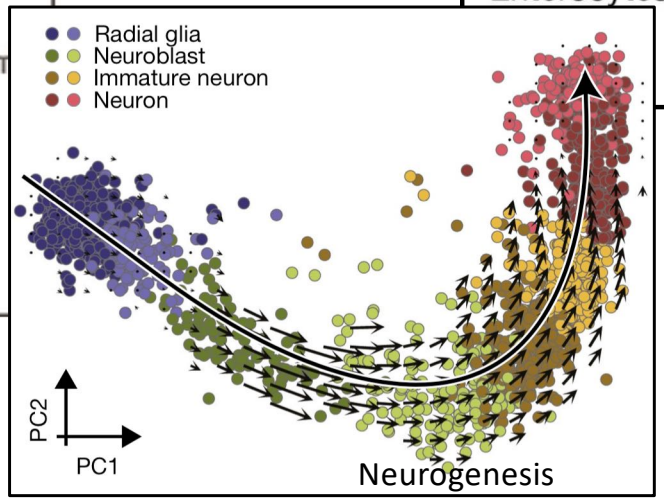
Recent SARS-Cov-2 single-cell RNA-seq Studies
Big Data

Examples of single cell visualization--



PBMC



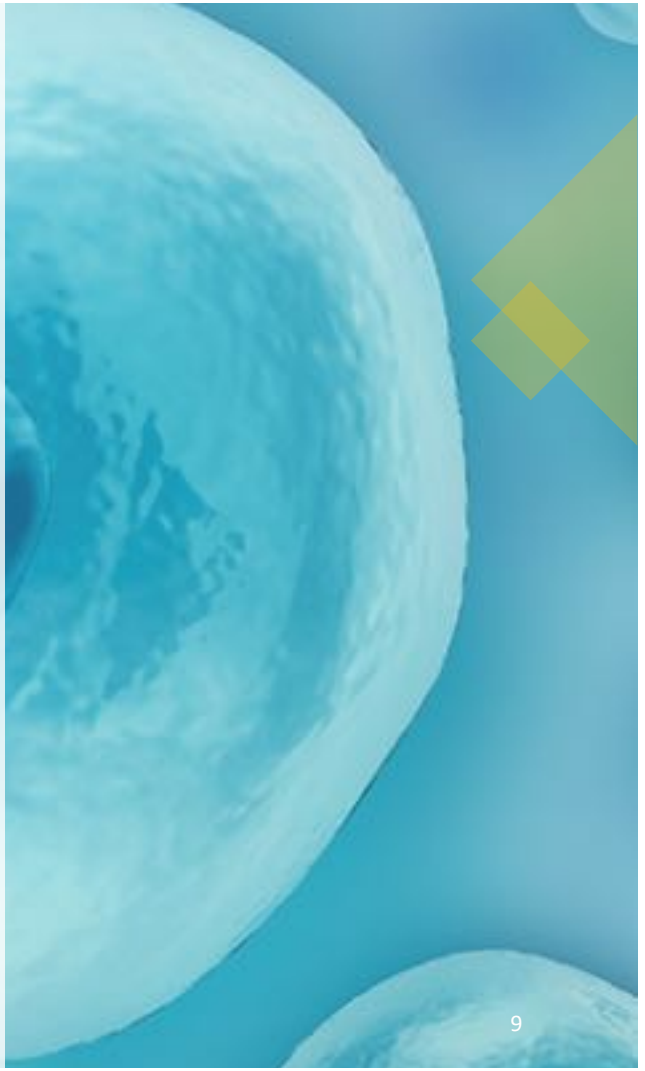


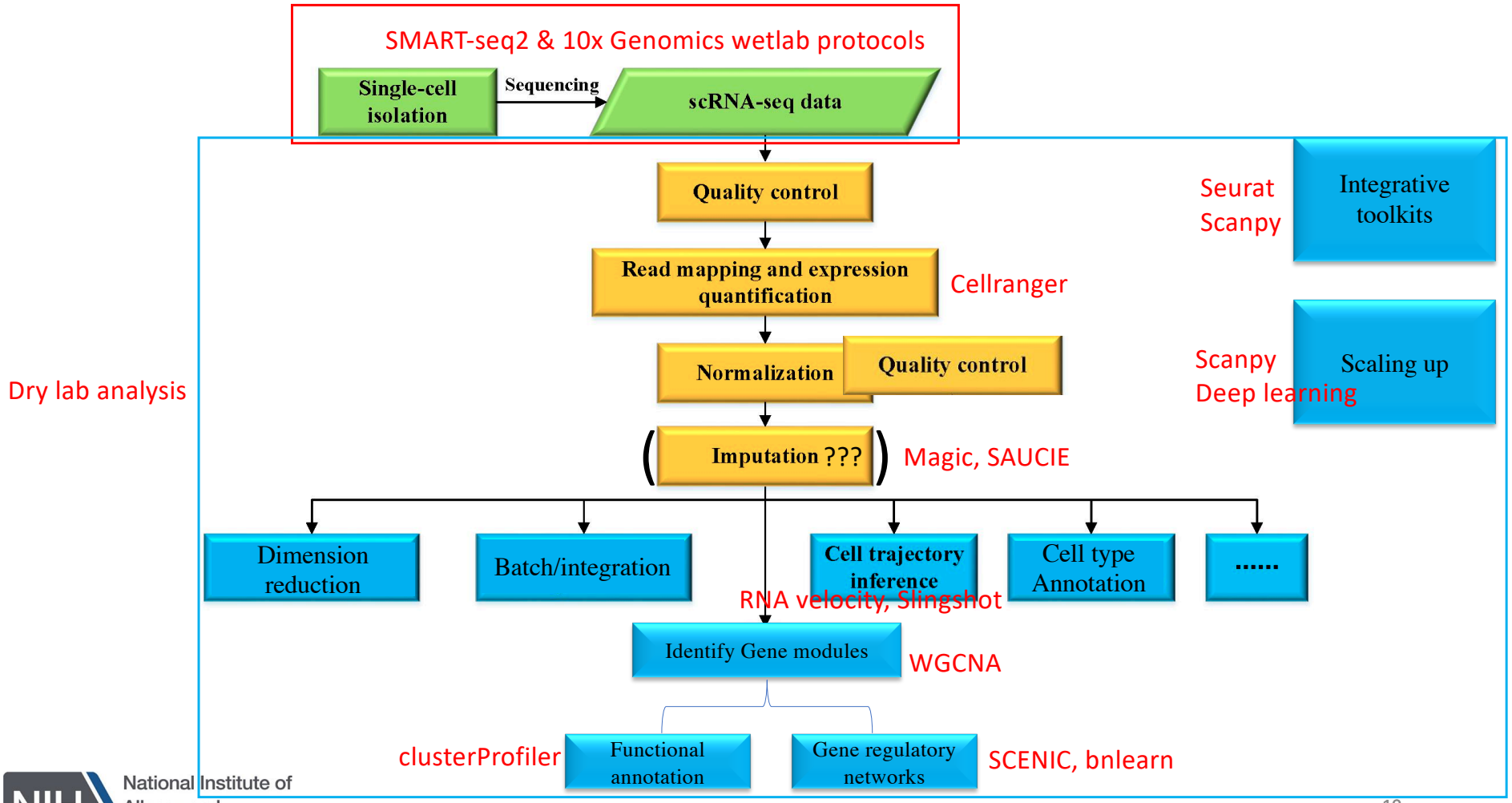
Neurogenesis

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
  - **Dry lab, overview of the workflow**
    - FASTQ files and FASTQC
    - Cellranger to get expression matrix
    - Dimension reduction
    - Trajectory analysis
    - Functional annotation
    - Gene regulatory network analysis
    - **Comprehensive tools**
  - **Summary**



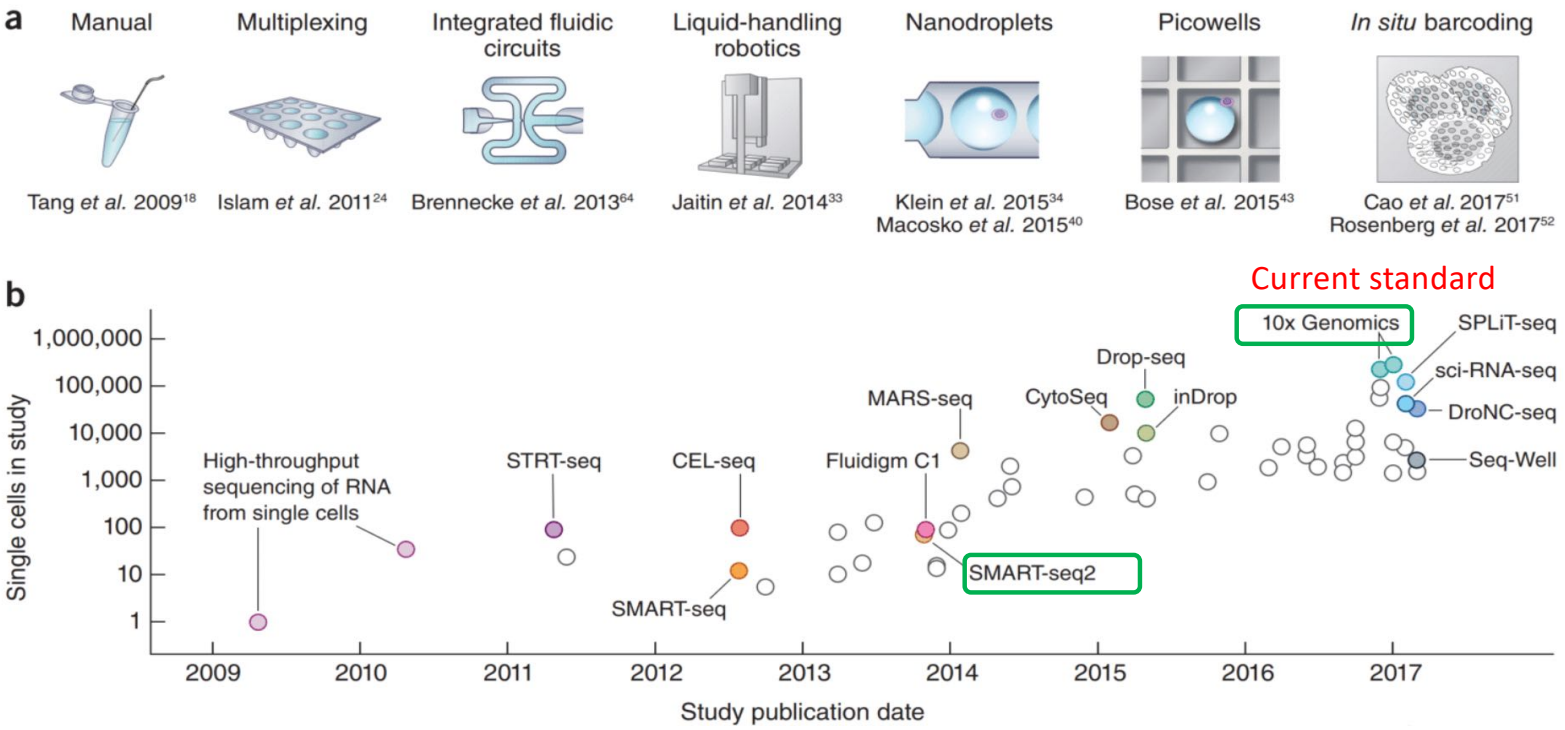


https://www.frontiersin.org/files/Articles/441123/fgene-10-00317-HTML/image_m/fgene-10-00317-g001.jpg

Wet lab-library construction

- The Smart-seq protocol to amplifying single cell (sc) mRNA (10pg)
- 10X genomics--the current standard

Development of platforms



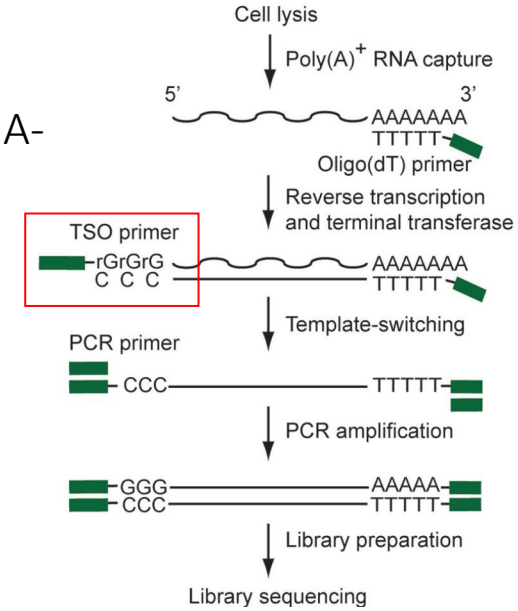
National Institute of Allergy and Infectious Diseases

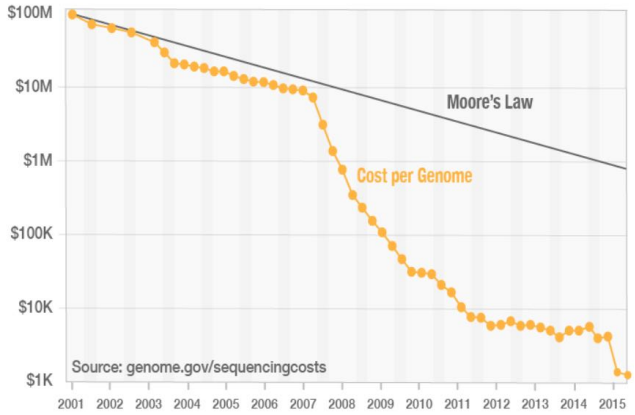
https://figshare.com/articles/Single_Cell_Present_and_near_Future/12121674

Where everything started—

SMART-seq protocol for efficient mRNA amplification made scRNA-seq affordable

- Core challenge:
  - How to get enough signal from 10pg RNA/cell?
- Advances in engineering
  - SMARTer reverse transcriptase that add 3x C at the end of cDNA.
  - Template switching oligo (TSO and LNA modification) enabled efficient amplification of mRNA
  - Barcode-mediated multiplexing enables combining many samples together and greatly reduced the cost for each cell
  - Drastic cost reduction of next generation sequencing (NGS)

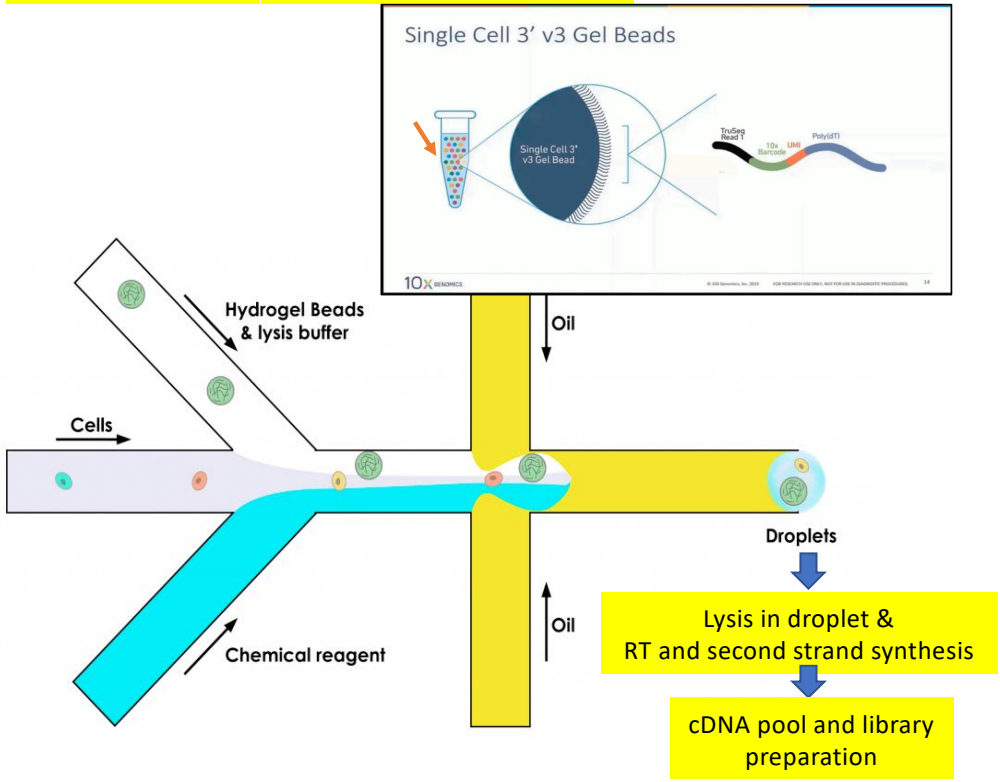




National Institute of Allergy and Infectious Diseases

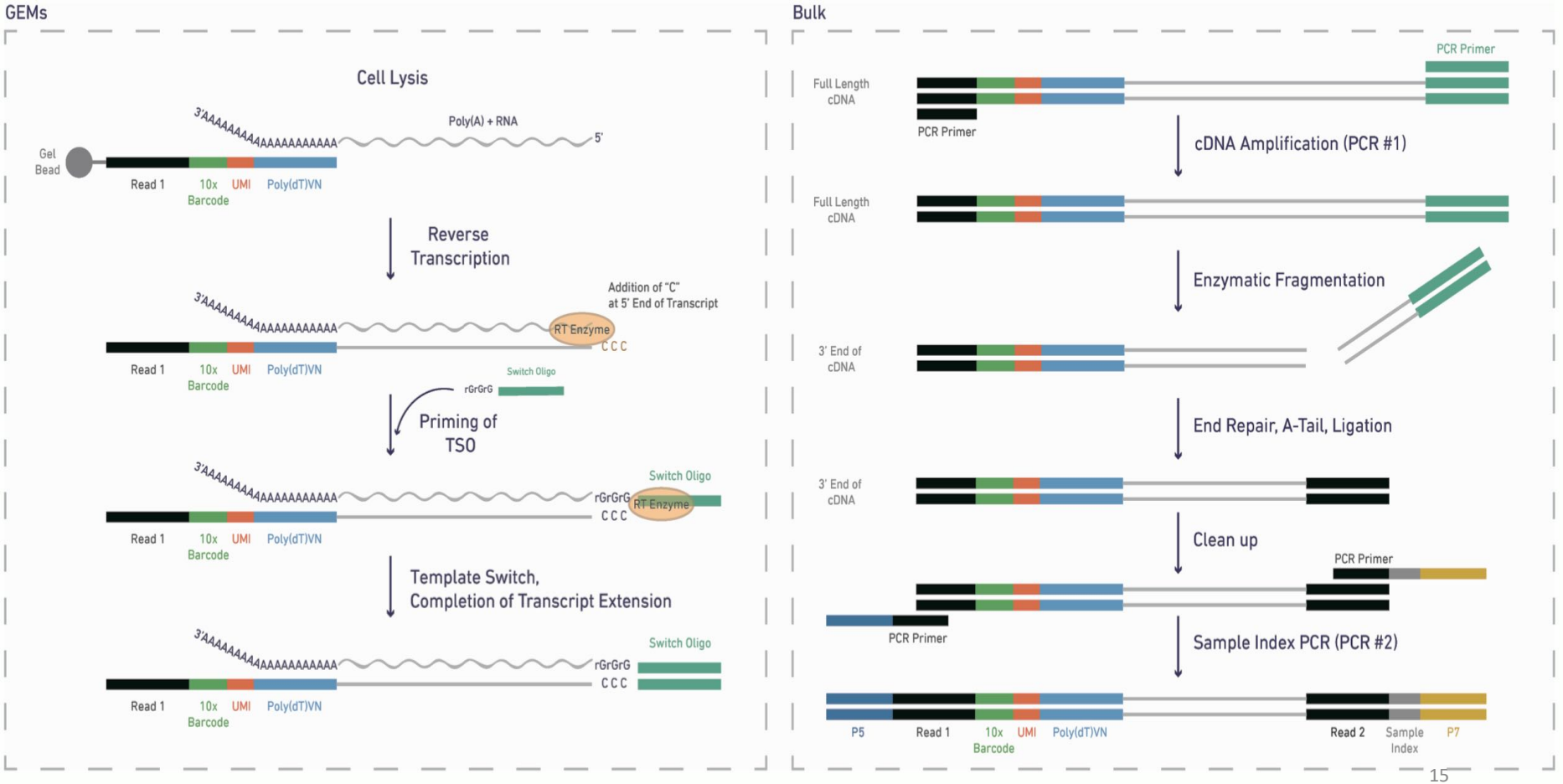
10X Genomics– commercial solution that facilitates automatic generation of Gel Bead-in-Emulsion (GEM)

- 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 (~1nL) 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



Inside individual GEMs (Gel Bead-in-Emulsion)

Pooled cDNA processed in bulk

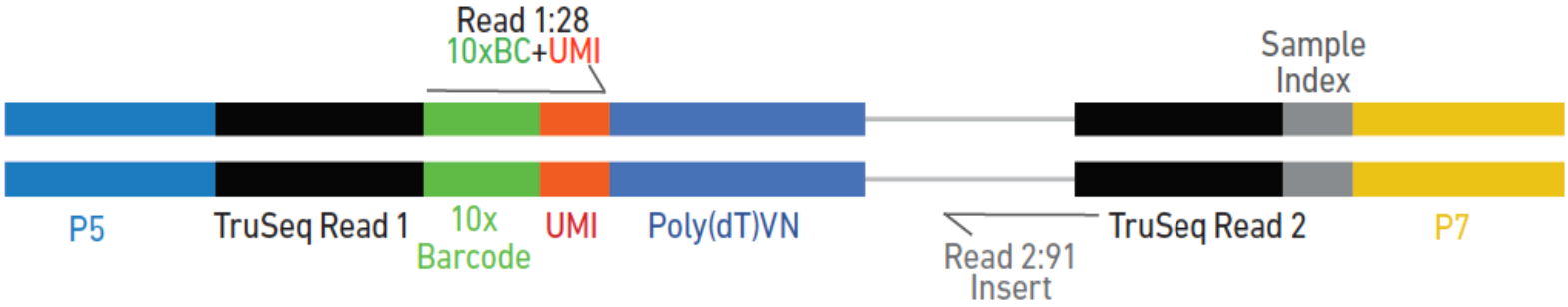


Infectious Diseases

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

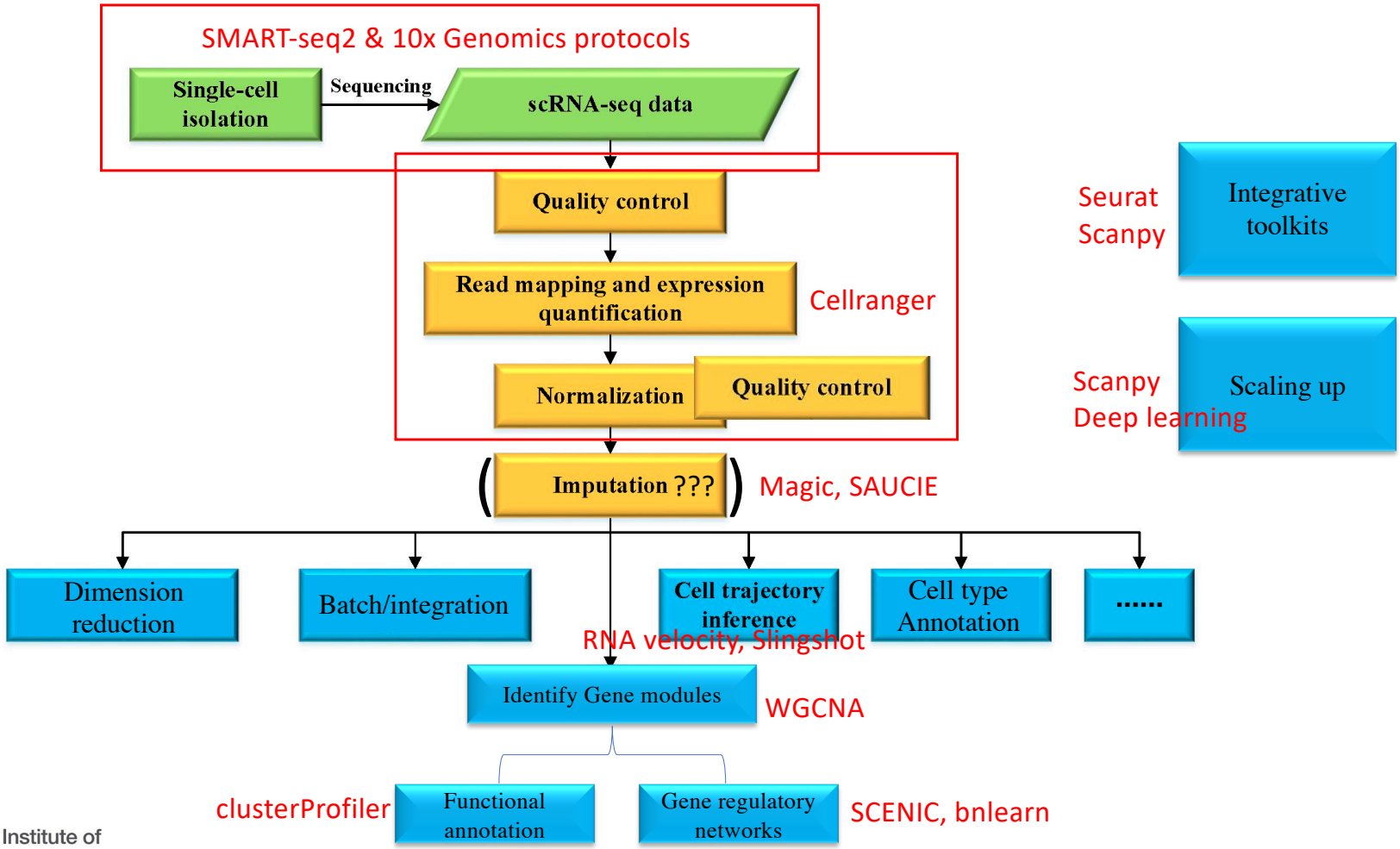
Next-seq reading the paired ends

- 10XBC: 16 bp barcodes 2^16=65536 possible unique cells
- UMI, 2^12=4096 unique copies of mRNA can be distinguished for each gene
- Read2 will read into cDNA to identify the identity of the gene
- Sample barcode, identify the batch of your library sample
- All information will be summarized by the Cellranger software



National Institute of Allergy and Infectious Diseases

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



National Institute of Allergy and Infectious Diseases

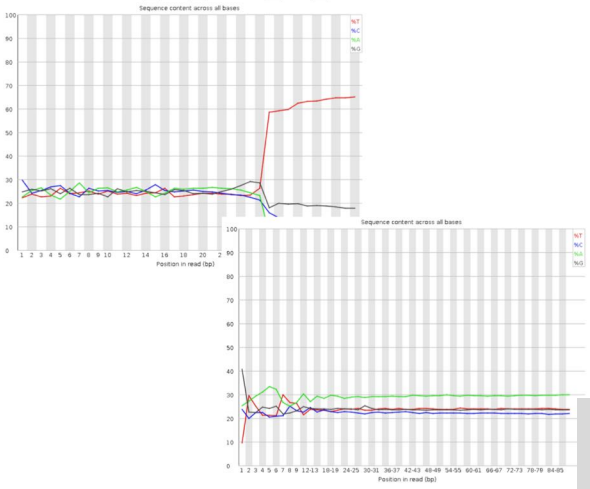
https://www.frontiersin.org/files/Articles/441123/fgene-10-00317-HTML/image_m/fgene-10-00317-g001.jpg

Alignment results and Quality Controls

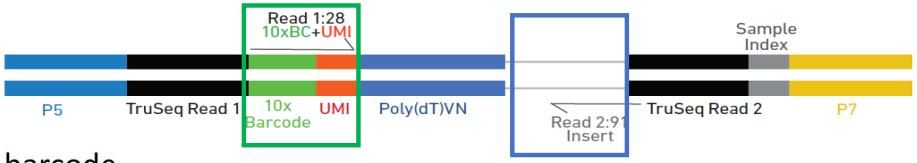
- Cellranger
  - cellranger mkfastq
  - cellranger count
  - cellranger aggr
- Quality controls
  - FASTQC on fastq files
  - Number of cells per experiment
  - Number of UMI per cell
  - Number of genes per cell
  - Percentage of mitochondrial reads
  - Removal of doublets/aggregates

QC on sequencing results

```
scrALI001_S1_L001_I1_001.fastq.gz
scrALI001_S1_L001_R1_001.fastq.gz
scrALI001_S1_L001_R2_001.fastq.gz
```



- I1
  - Index file. All identical (or one of 4) at Babraham
- R1
  - Barcode reads
    - 16bp cell level barcode
    - 10bp UMI
- R2
  - 3' RNA-seq read




```
$ cellranger count --id=sample1 \  # set name for your output folder
        --transcriptome=/opt/refdata-gex-GRCh38-2020-A \  # reference
        --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path \
        --sample=scrALI001 \ # prefix of the FASTQ file
        --expect-cells=5000 \ # optional
```

Cellranger count report

FATSTQ

I1.FASTQ
R1.FASTQ
R2.FASTQ
} Cellranger →

report summary …

Estimated Number of Cells

15,894

Mean Reads per Cell	Median Genes per Cell
11,380	2,174

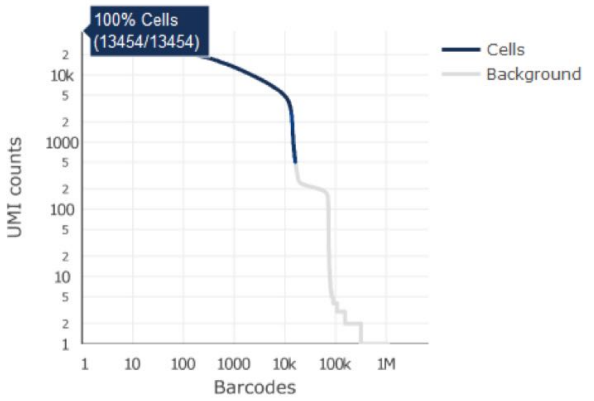
Sequencing

Number of Reads	180,878,636
Valid Barcodes	98.1%
Sequencing Saturation	10.3%
Q30 Bases in Barcode	98.4%
Q30 Bases in RNA Read	82.7%
Q30 Bases in UMI	98.7%

Mapping

Reads Mapped to Genome	95.4%
Reads Mapped Confidently to Genome	90.2%
Reads Mapped Confidently to Intergenic Regions	3.0%
Reads Mapped Confidently to Intronic Regions	12.8%
Reads Mapped Confidently to Exonic Regions	74.4%
Reads Mapped Confidently to Transcriptome	71.9%
Reads Mapped Antisense to Gene	0.9%

Cells



Estimated Number of Cells	15,894
Fraction Reads in Cells	88.1%
Mean Reads per Cell	11,380
Median Genes per Cell	2,174
Total Genes Detected	20,185
Median UMI Counts per Cell	5,742

Sample

Name	embryoid_d4
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.2

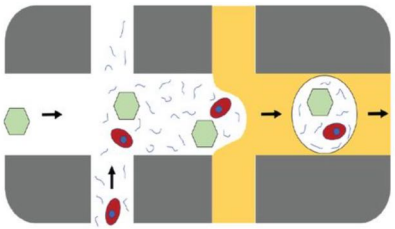
NIH National Institute of Allergy and Infectious Diseases

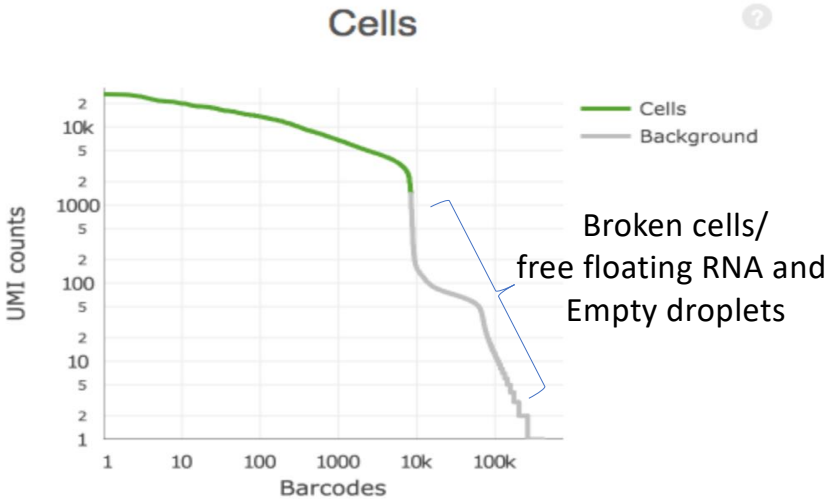
Gene counts

- Reads level
  - CellRanger
    - cellranger mkfastq
      - Generate fastq files from image ".bcl" files
    - **cellranger count → sparse matrix**
      - 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
```





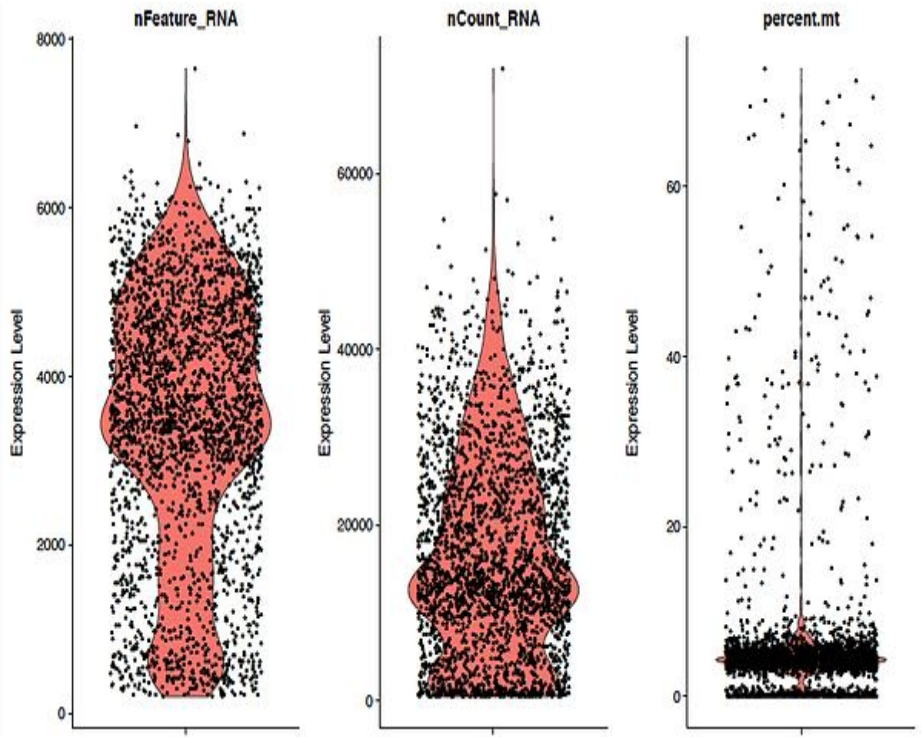
Broken cells/
free floating RNA and
Empty droplets

Ranked by number of associated UMIs

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

Quality Controls

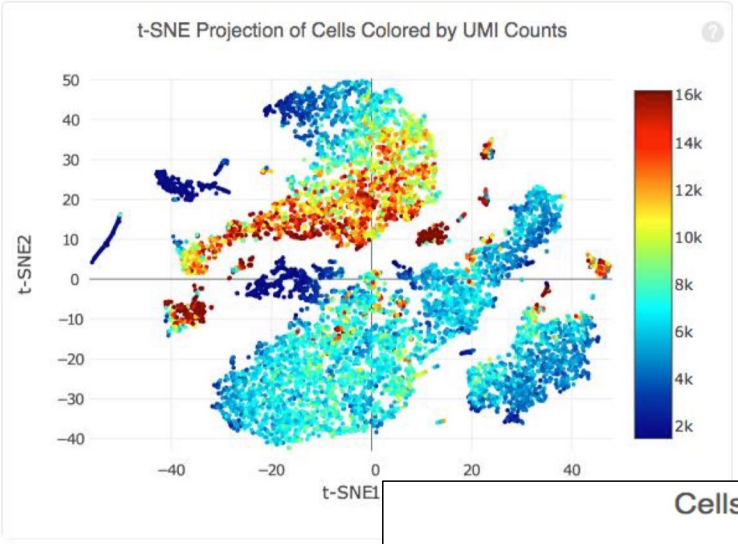
- Quality of reads – FASTQC
- Percentage of mitochondrial reads
  - Too many mitochondria reads may indicate that cells are dying/dead/broken
  - Case-by-case, the range may vary with method of library prep methods/cell type
- How many cells are you capturing?
  - Typically few thousands in each 10X run
- The sequencing depth
  - Are they acceptable in the field (minimal 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.
- Confounding factors?
  - Batch effect?
  - Is your dimension reduction capturing biological or technical variations?
  - Can be evaluated by WGCNA and visualization in PCA or tSNE

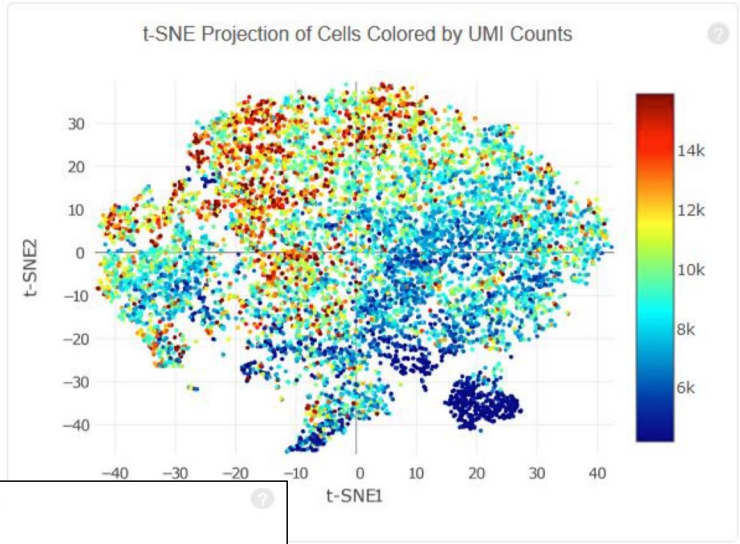


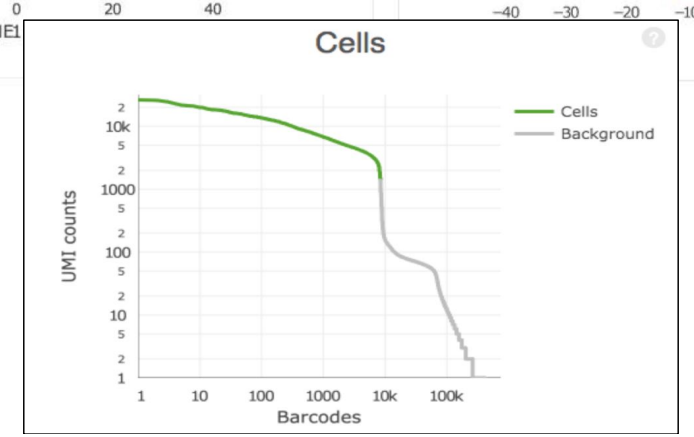
https://www.biostars.org/p/377422/

National Institute of Allergy and Infectious Diseases

Is coverage variation affecting your data?





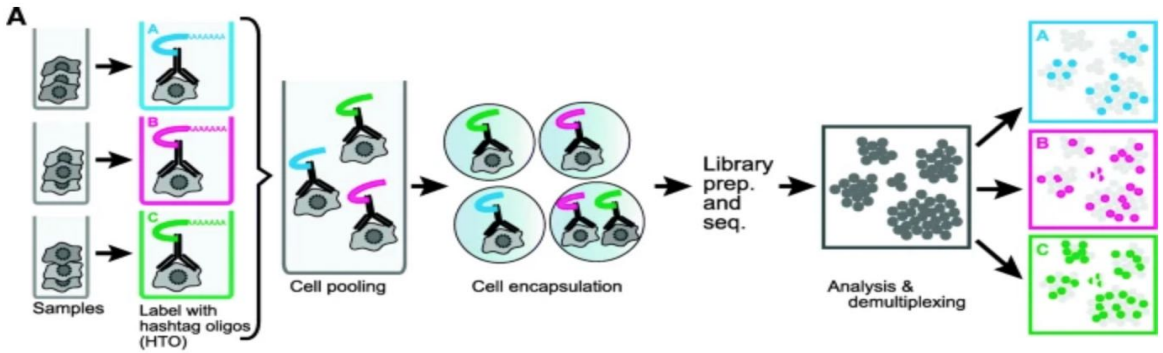


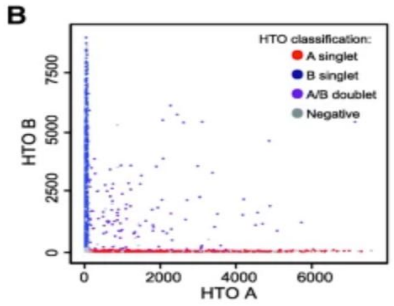
UMI dominate the variance?
Need to do batch correction?
Is the level of separation enough/expecte
(to much background RNA?

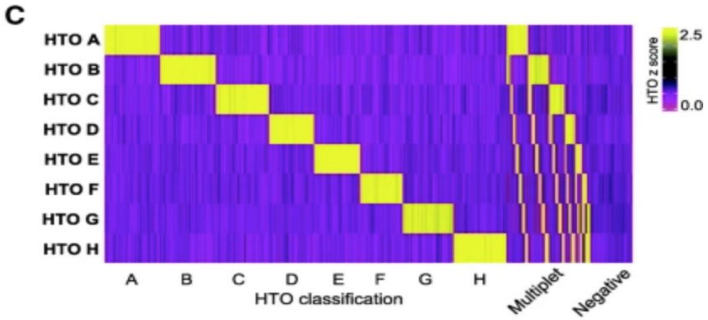
National Institute of Allergy and Infectious Diseases

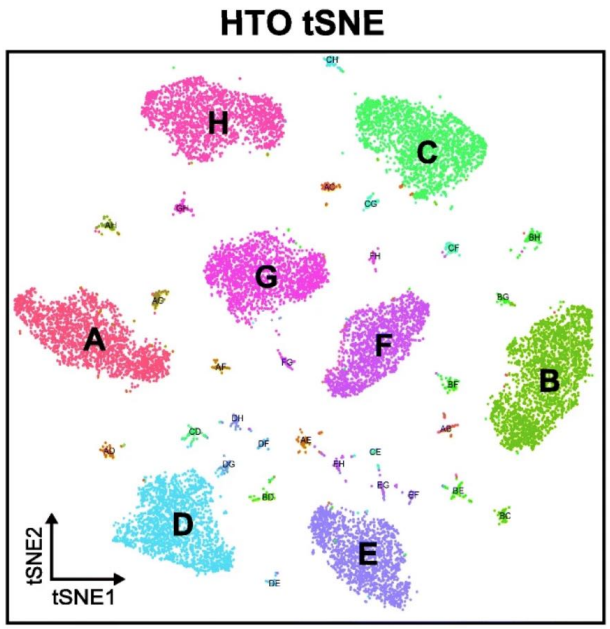
Doublet removal by HASH-tagging and computational tools

Using antibody-attached hash-tag to label cells at each individual batch, then combine batched for sequencing



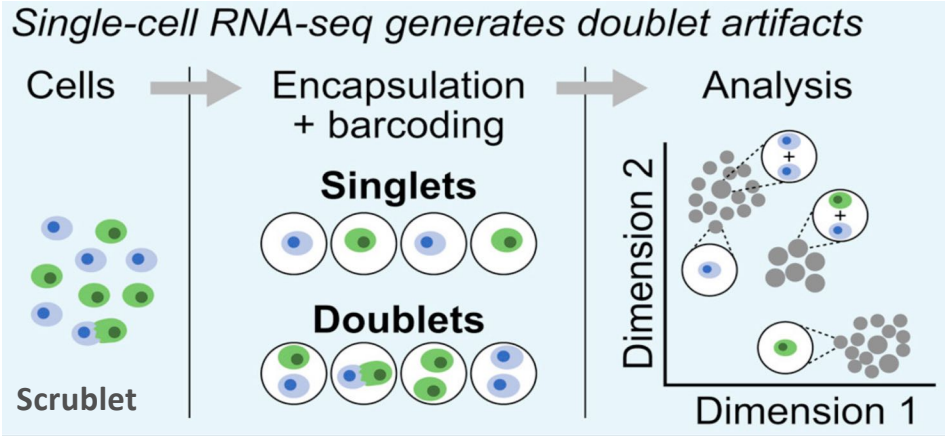




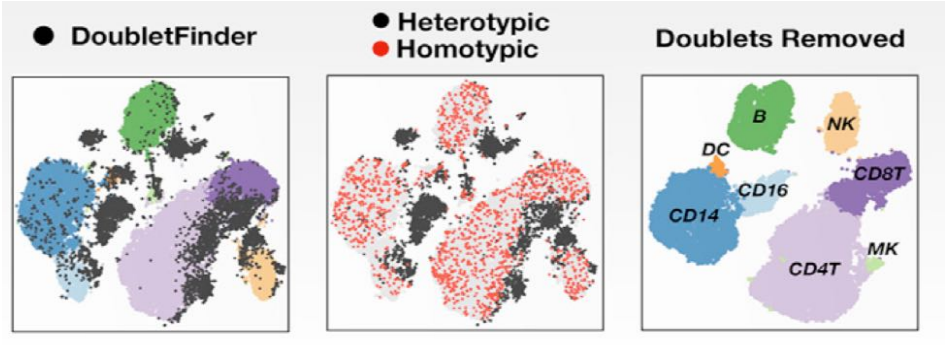


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1603-1

Doublet removal by computational methods



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



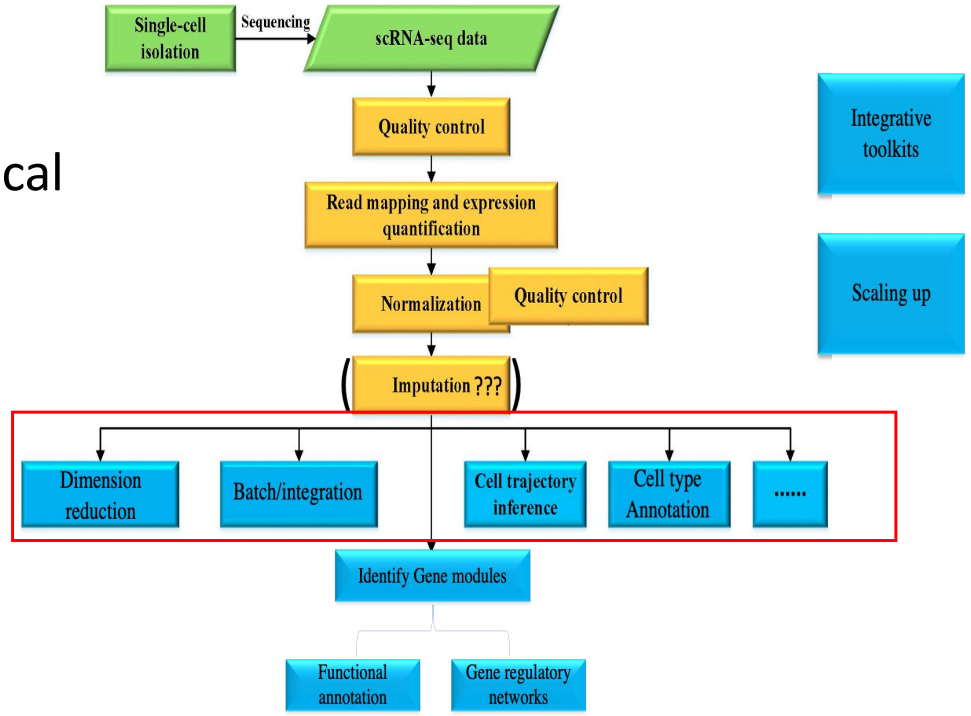
https://www.sciencedirect.com/science/article/pii/S2405471219300730?via%3Dihub

Careful when you want to find novel cell types

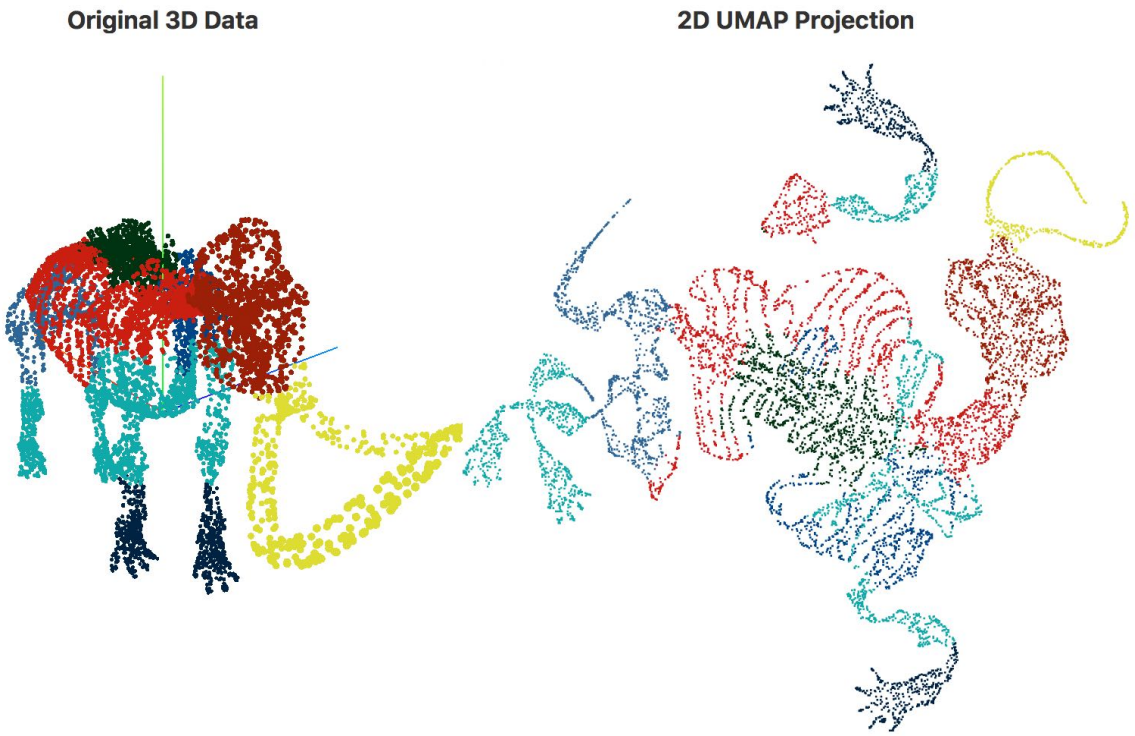
National Institute of Allergy and Infectious Diseases

Cell-based analysis

- Clustering and annotate the biological identities of clusters
- Inference of trajectory
- Batch correction/data integration.
- comprehensive workflow/toolkits



Dimension reduction--an intuitive illustration

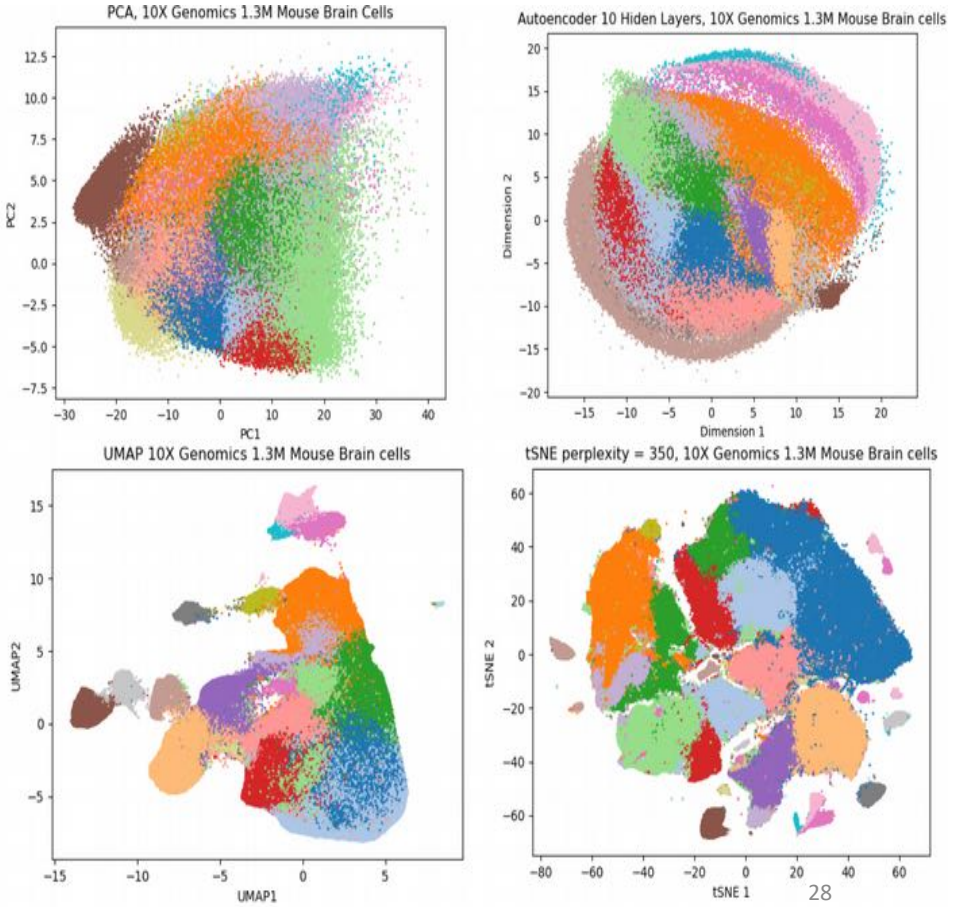


National Institute of Allergy and Infectious Diseases

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 sufficient for large 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 to handle up to millions of cells

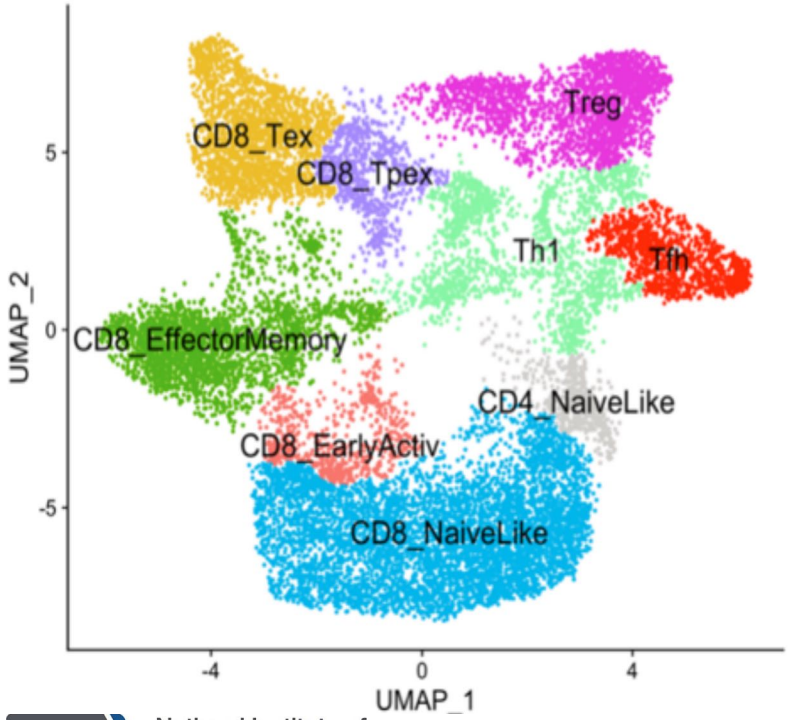


https://towardsdatascience.com/deep-learning-for-single-cell-biology-935d45064438

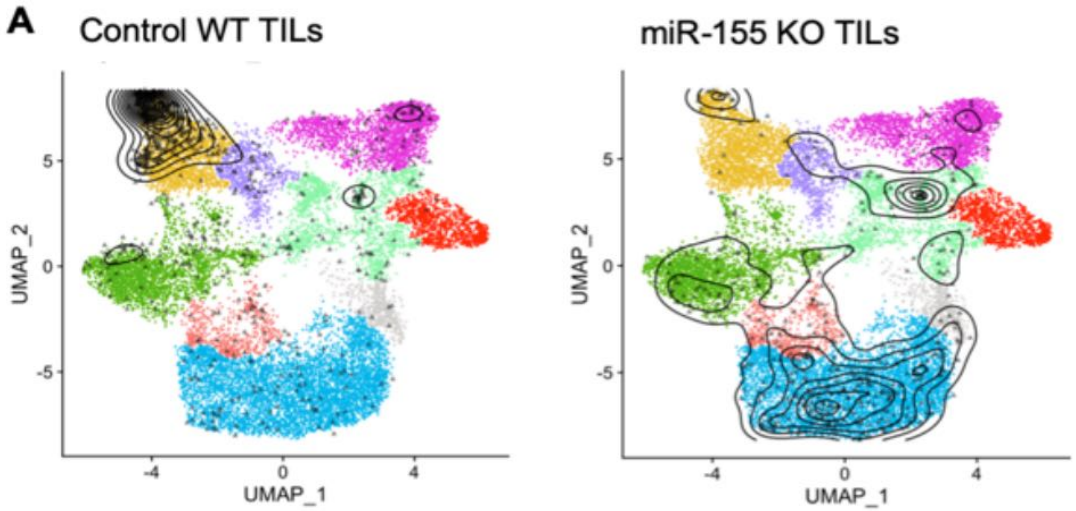
National Institute of Allergy and Infectious Diseases

New approach: projection of cells to a reference map

-- the map is determined by a set of marker genes



National Institute of Allergy and Infectious Diseases

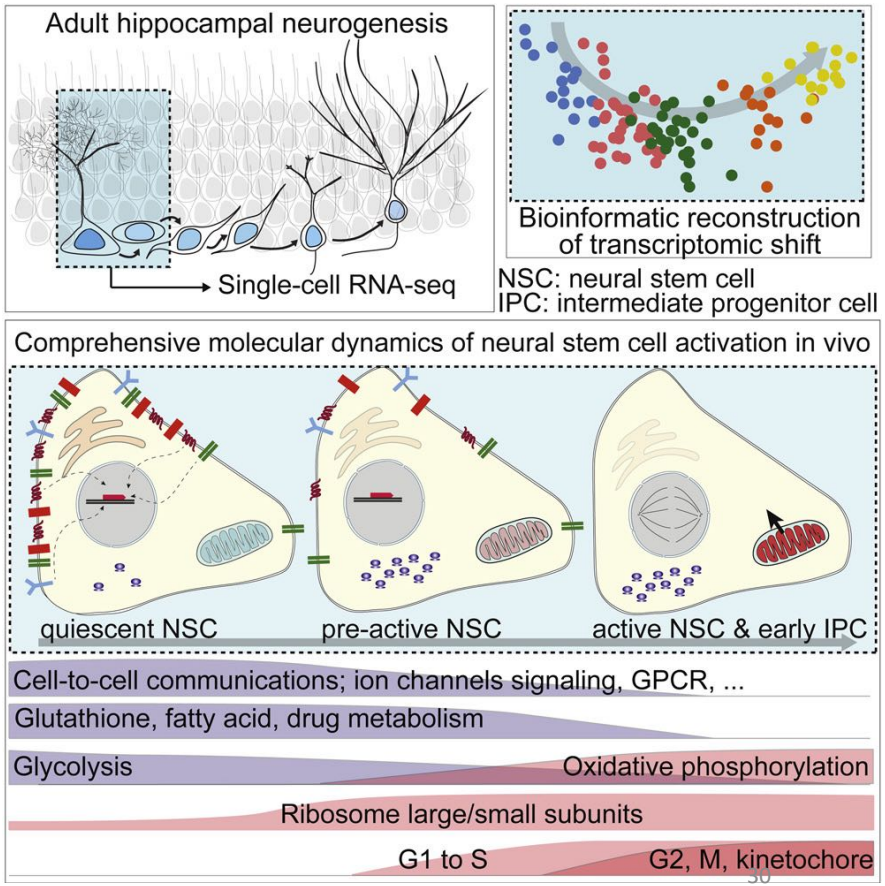


https://www.biorxiv.org/content/10.1101/2020.06.23.166546v1.full.pdf

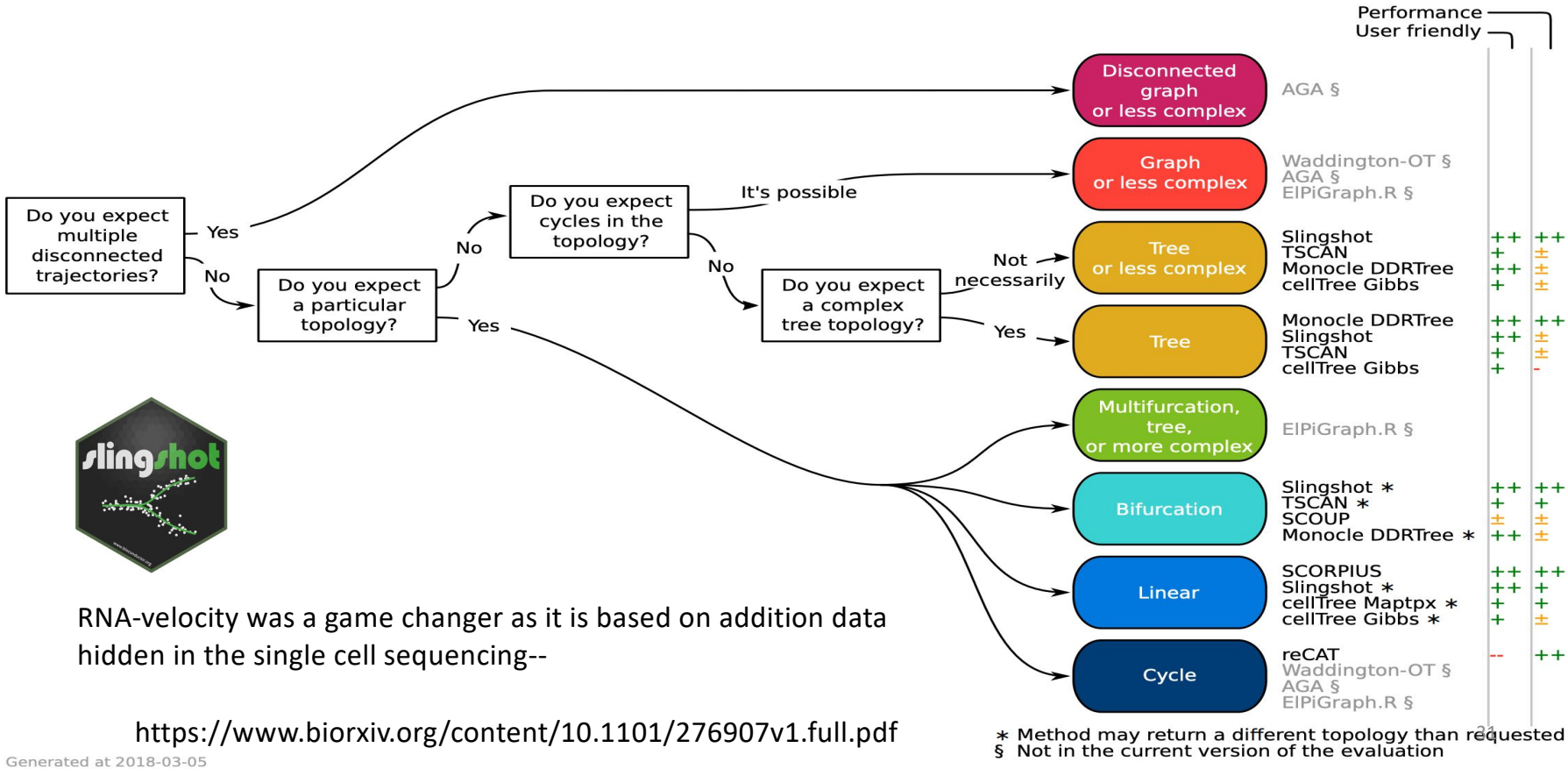
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02280-8 iMAP, 2021 used autoencoder and GAN framework to scale up

Trajectory analysis

- Temporal and spatial gradient
  - Observed after dimension reduction
  - Use known markers to annotate the pattern interested, and assign directions
- Aim
  - Find relationship between cells
  - Further delineate the cells and genes with temporal/spatial information
- Packages
  - Monocle
  - Slingshot
  - RNA Velocity – based on intron/exon reads in the data



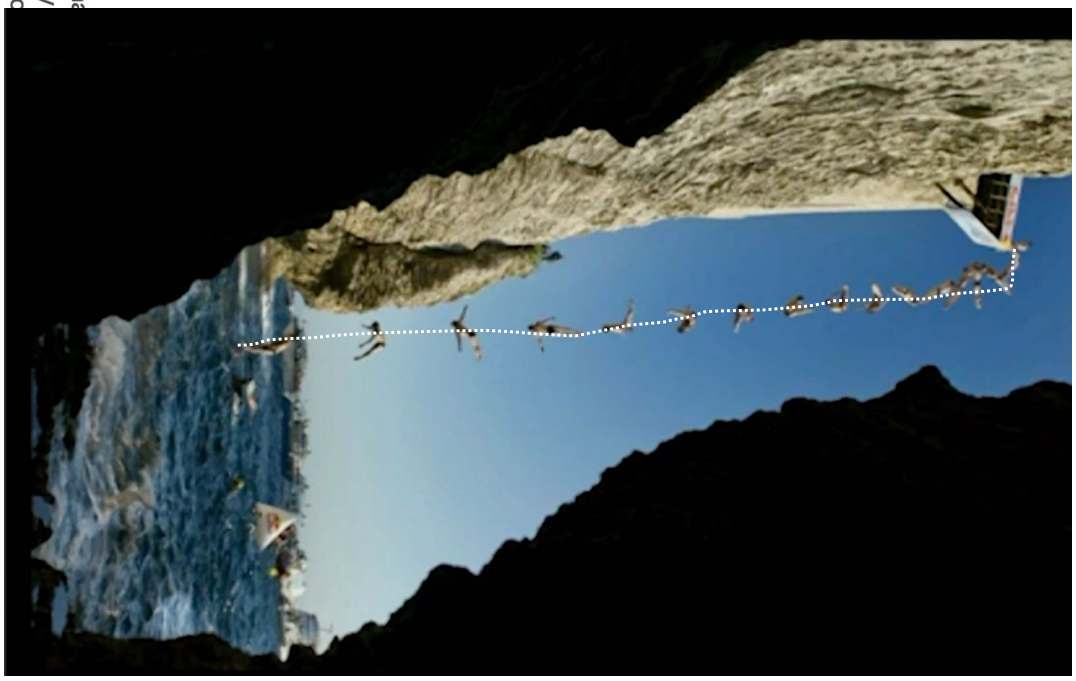
Many packages have been developed to extract trajectory purely based on computation, slingshot seems standing out



RNA-velocity was a game changer as it is based on addition data hidden in the single cell sequencing--

https://www.biorxiv.org/content/10.1101/276907v1.full.pdf

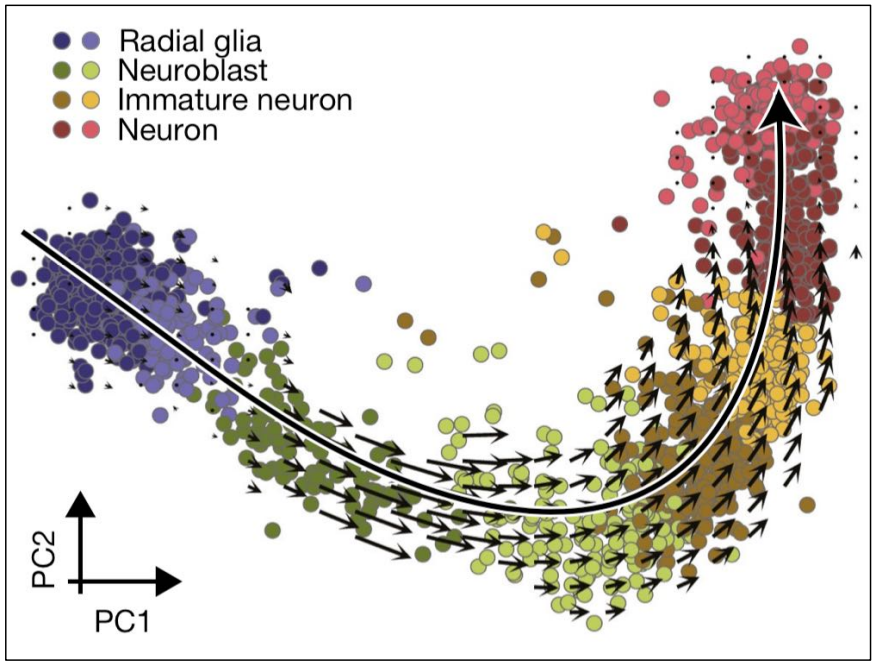
Traditionally – connecting the data points



RNA Velocity – infer the directionness

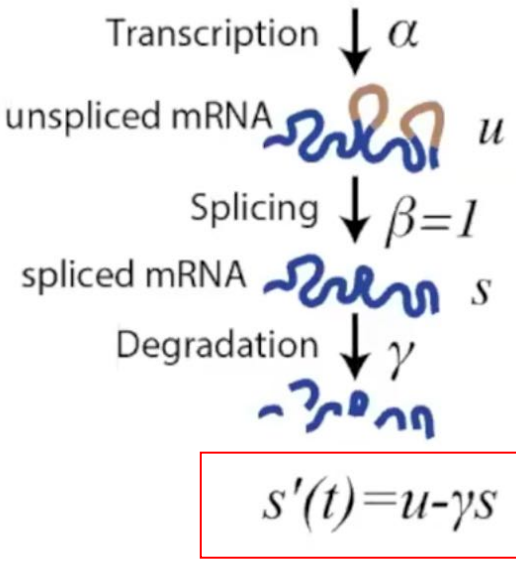
Trajectory analysis by RNA velocity

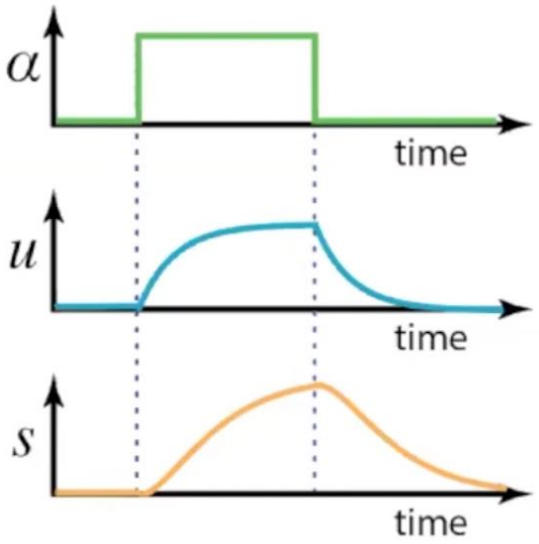
- When cells differentiate, **new genes** will start to be expressed
- Transcripts have introns and will be spliced off given time
- Through assessing the present percentage of reads in introns, increase or decrease of expression can be modeled

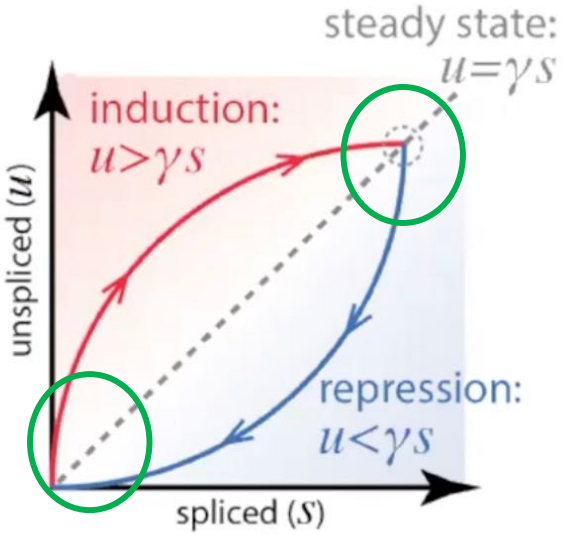


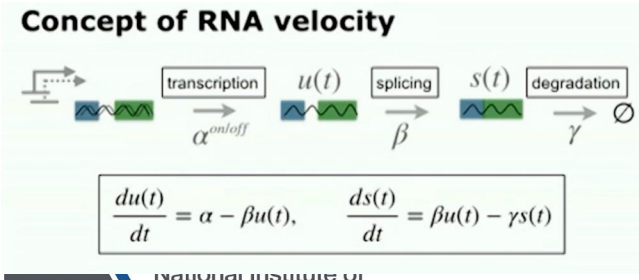
https://liorpachter.wordpress.com/tag/velocyto/
http://pklab.med.harvard.edu/software.html

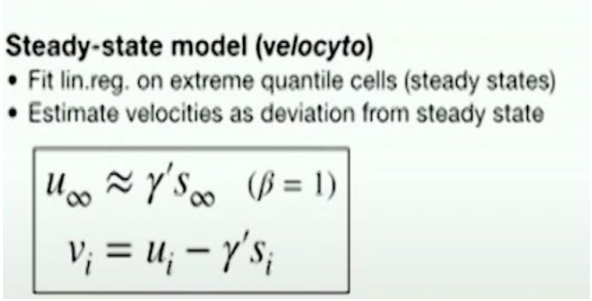
Modeling RNA dynamics









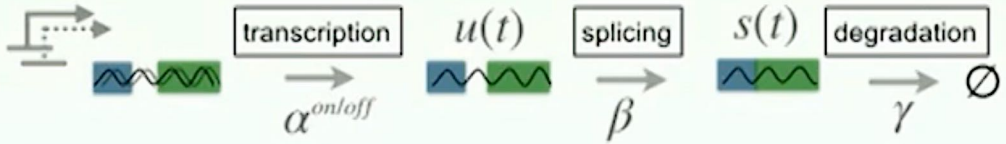


NIH National Institute of Allergy and Infectious Diseases

2 assumptions:
steady states has been observed
a constant splicing rate β across all RNA

scvelo.org

Generalizing RNA velocity to dynamical populations

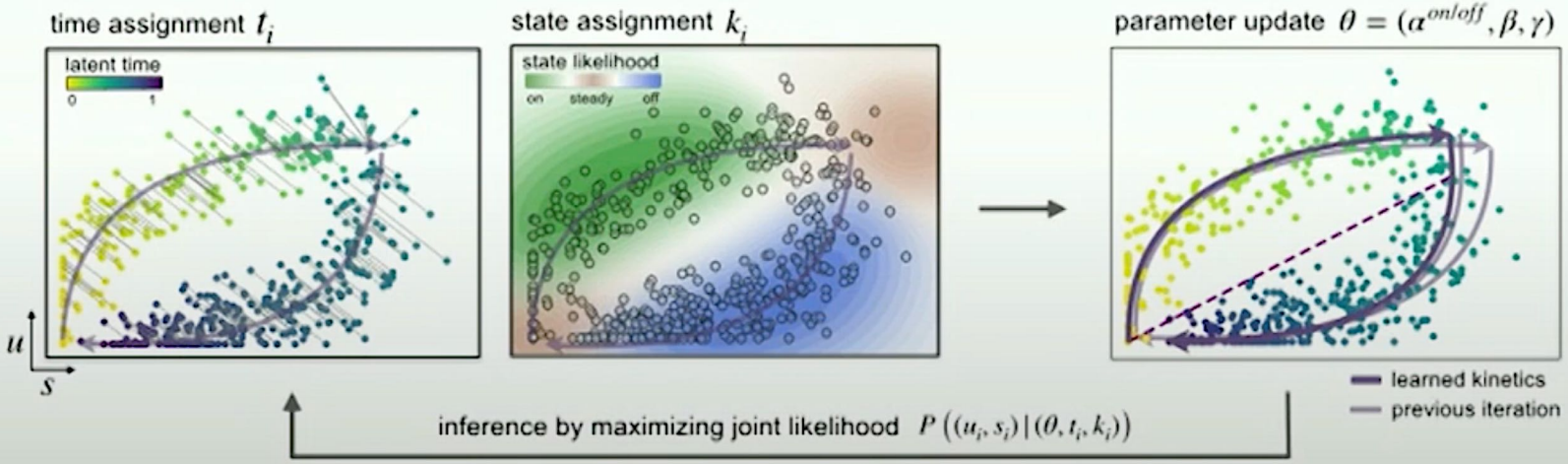


$$u(t) = u_0 e^{-\beta\tau} + \frac{\alpha}{\beta}(1 - e^{-\beta\tau})$$

$$s(t) = s_0 e^{-\gamma\tau} + \frac{\alpha}{\gamma}(1 - e^{-\gamma\tau}) + \frac{\alpha - \beta u_0}{\gamma - \beta}(e^{-\gamma\tau} - e^{-\beta\tau}) \qquad \tau = t - t_0$$

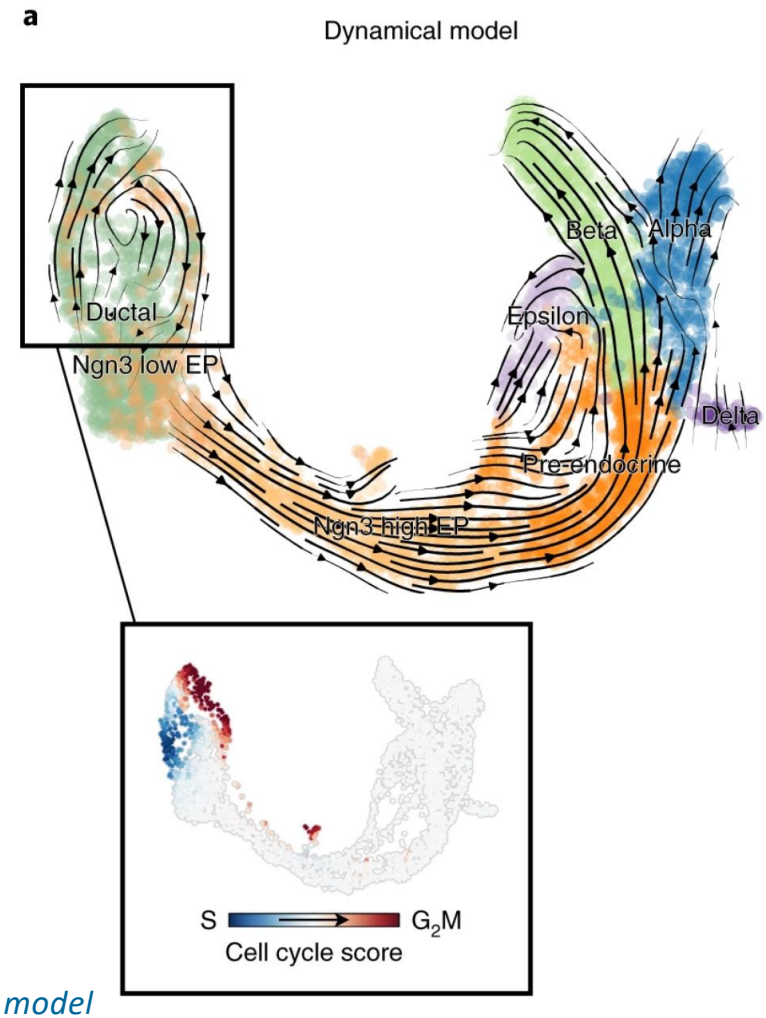
parameters of **reaction rates** $\theta = (\alpha^{off}, \alpha^{on}, \beta, \gamma)$

cell-specific **latent variables** (switch, time, state) $\eta_i = (t_0^{(i)}, t_i, k_i)$



Steps of RNA-velocity

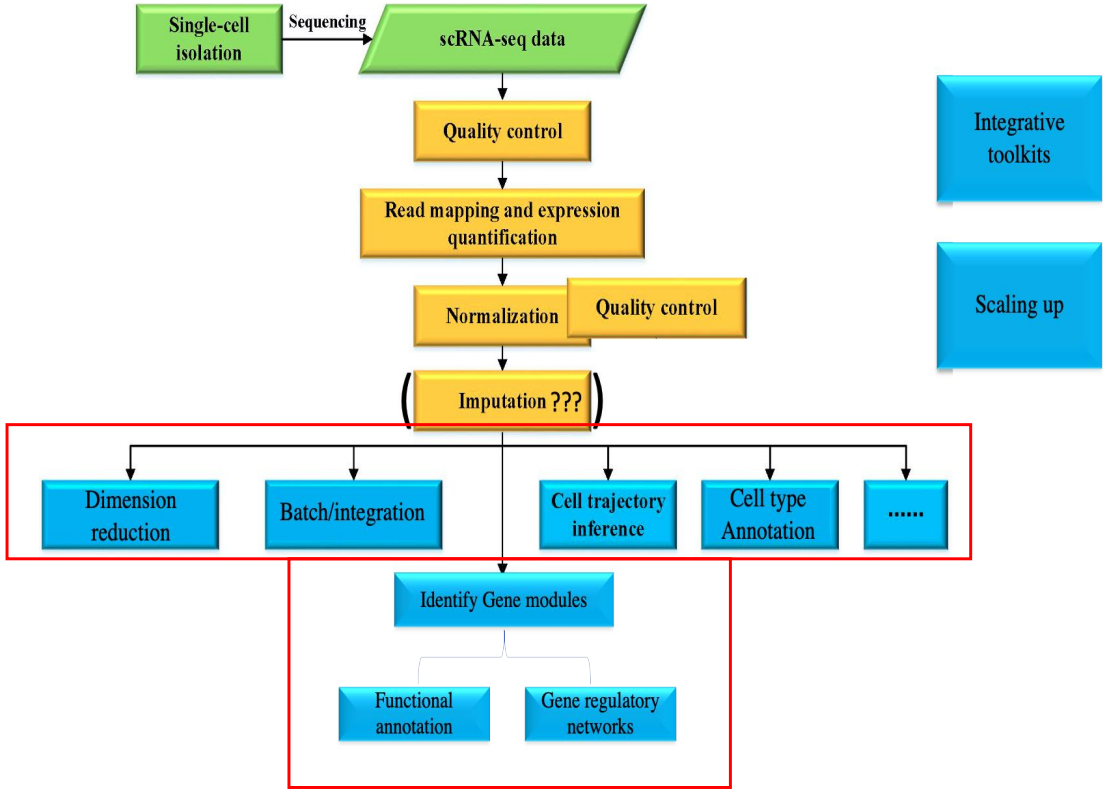
- From cellranger produced bam files
  - Sort bam files
  - Using Velocyto CLI to identify exonal/intronal reads as loom files.
  - Use scVelo to model the data and recover RNA dynamics
  - Through Markov process to predict which neighbor is the most probabal destiny for he cell.
  - Backtracking and forward tracking to get the destiny and ancestor of the cell.



National Institute of Allergy and Infectious Diseases

Nature, 2018. invented the concept
Nature Biotechnology, 2020. a generalized model

Comprehensive toolkits



National Institute of Allergy and Infectious Diseases

Gene-oriented approach

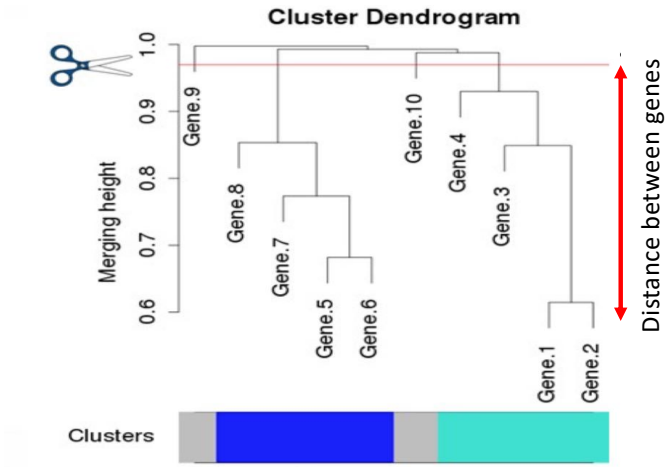
- Most often biologist researchers want to
  - isolated groups of associated genes
  - Find out the biological implication of the genes.
  - Find the master regulators within a module, gene at the top of the regulatory hierarchy
- To find gene modules using
  - Conventional cluster identification using tree cutting is of little use.
  - Weighted correlation network analysis (WGCNA)
- Identify gene modules using WGCNA
- Functional annotation using clusterProfiler
- Find internal relationship between genes using gene regulatory network analysis
  - Bnlearn
  - SCENIC
  - PIDC

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

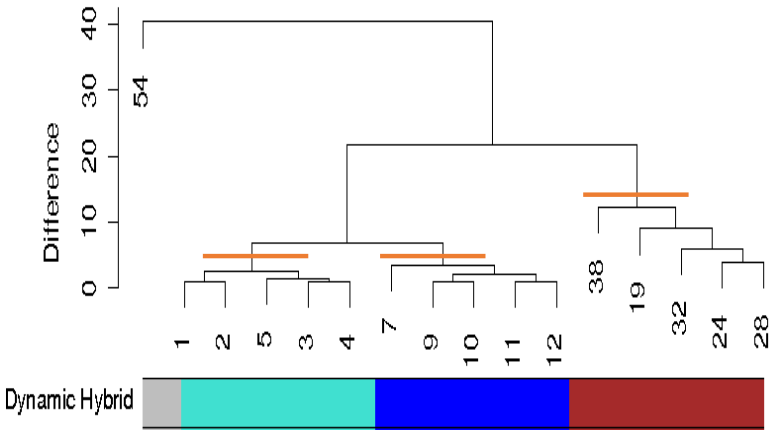
Finding gene clusters using weighted correlation network analysis (WGCNA)

Conventional tree cutting

Just a few genes to determine manually



Dynamic tree cutting: clusters determined by adaptive tree cutting

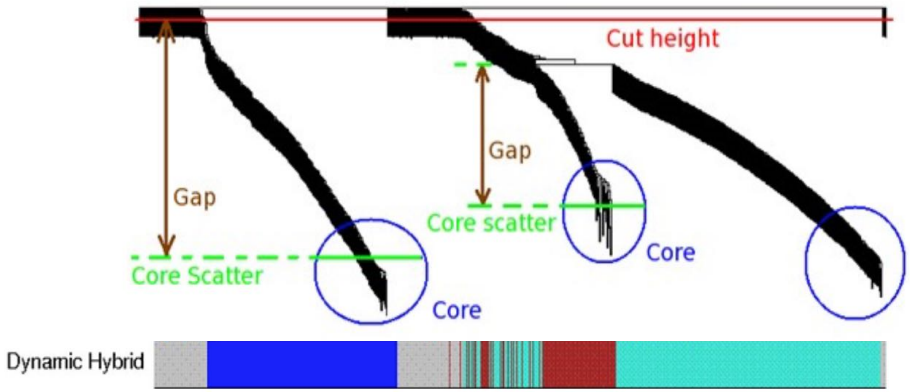


National Institute of Allergy and Infectious Diseases

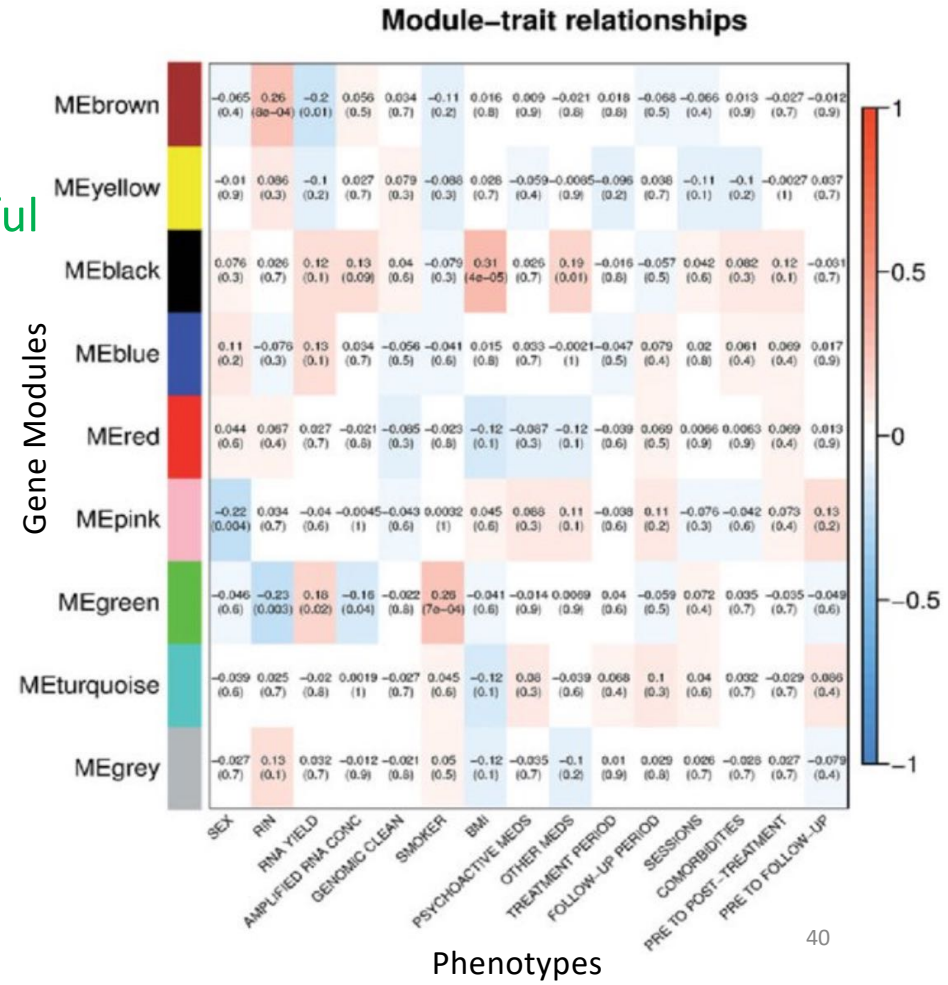
https://slideplayer.com/slide/7402978/

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



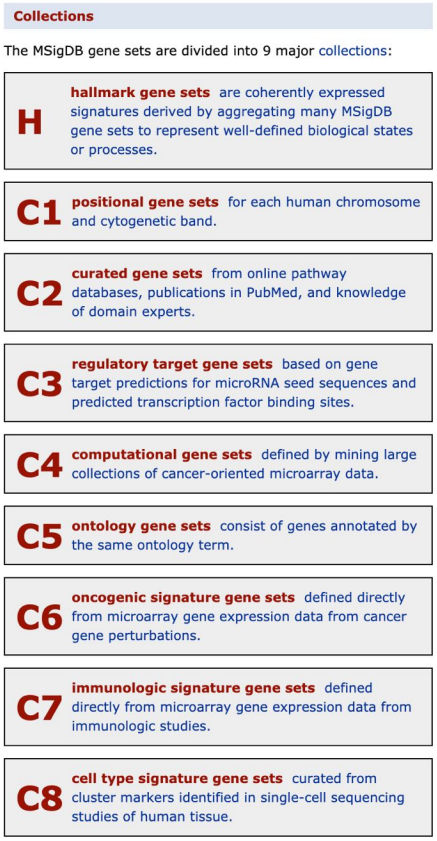
A nice feature of WGCNA



National Institute of Allergy and Infectious Diseases

Functional annotation with ‘clusterProfiler’

- Biological implication of differential expressed genes
  - p-value dependent method – enricher
  - Ranking of fold changes – GSEA analysis
- Gene-sets corresponding to biological processes
  - Kegg, canonical pathway, molecular hallmark, GO_biological processes et al. 13 of them.
- clusterProfiler allows functional profiling with R
  - https://yulab-smu.github.io/clusterProfiler-book/index.html
  - Customized codes to do automatic profiling against 13 databases.



National Institute of Allergy and Infectious Diseases

MsigDB: https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

Functional annotation of genes
-- Gene Set Enrichment Analysis

- Rank-based enrichment analysis
  - Detect consistent global changes based on ranking of fold changes, independent of P-value in differential expression analysis.
  - Genes in the leading edge are considered top enriched genes for that term
  - clusterProfiler will find all processes defined in multiple annotation database
    - Databases include: Hall mark, transcriptional factor, canonical pathways etc.
    - Found term can be positive or negatively correlates with the ranking.
- Based on several databases, used the clusterProfiler package
  - Wikipath: https://www.wikipathways.org/index.php/WikiPathways
  - GO, KEGG etc: https://yulab-smu.github.io/clusterProfiler-book/chapter1.html

Example:

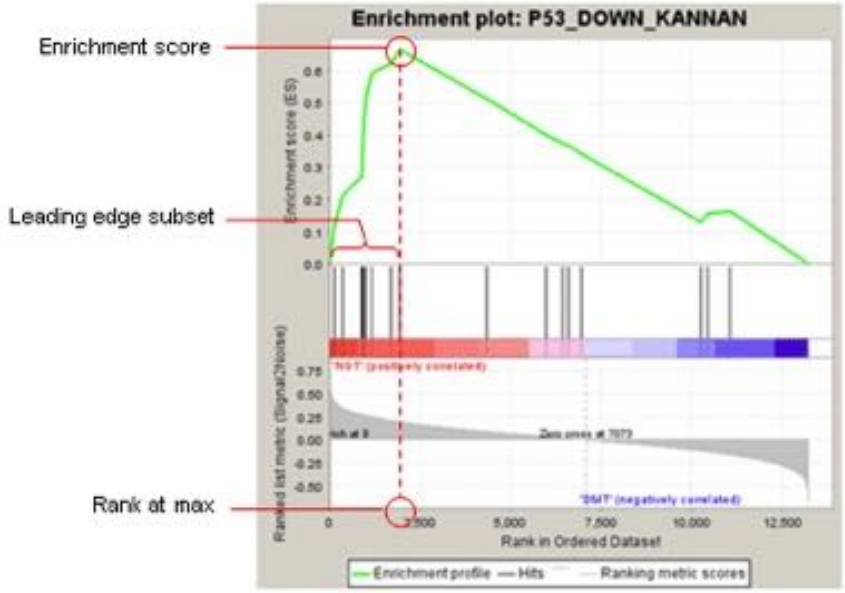


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

www.pnas.org › content

Gene set enrichment analysis: A knowledge-based approach ...

Sep 30, 2005 - **Gene set enrichment analysis**: A **knowledge-based approach** for **interpreting genome-wide expression profiles**. Aravind Subramanian, Pablo ...
by A Subramanian - 2005 - Cited by 21658 - Related articles

National Institute of Allergy and Infectious Diseases

clusterProfiler gives a summary of overall enrichment and nice visualization tools

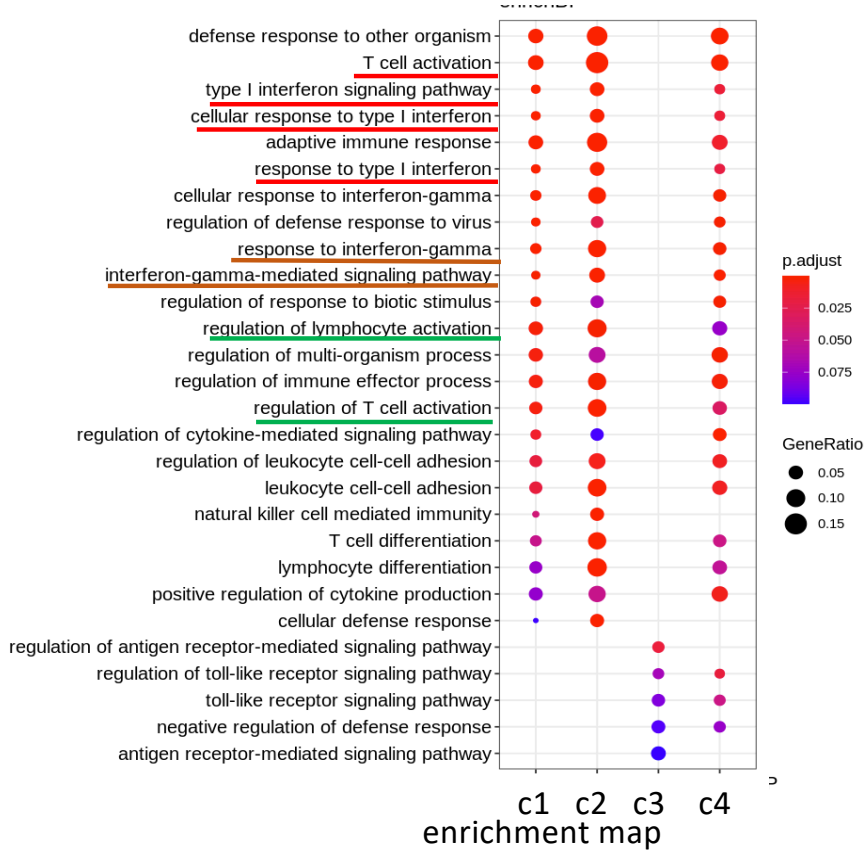
list of DEGs from multiple comparison groups

	c1	c2	c3	c4
gse_KEGG	19	16	12	30
gse_CP	91	69	41	127
gse_Wiki	37	27	18	68
gse_GO_BP	0	0	0	0
gse_GO_CC	45	60	0	73
gse_GO_MF	70	93	62	85
gse_H	18	16	12	21
gse_MESH	290	362	0	401
gse_Trans	31	10	8	29
gse_DO	104	110	0	110
gse_DGN	218	225	0	297
enrich_KEGG	43	15	1	4
enrich_CP	128	32	23	29
enrich_Wiki	30	8	1	3
enrich_GO_BP	143	153	23	127
enrich_GO_CC	15	24	7	16
enrich_GO_MF	3	9	2	4
enrich_H	4	5	2	5
enrich_MESH	133	72	16	110
enrich_Trans	468	7	42	121
enrich_DO	0	13	0	7
enrich_DGN	50	72	2	83
gse_Marker	38	49	36	38
gse_Imm	2888	2784	2391	2974
enrich_Marker	14	28	14	13
enrich_Imm	3321	1273	909	1794

Rank-dependent

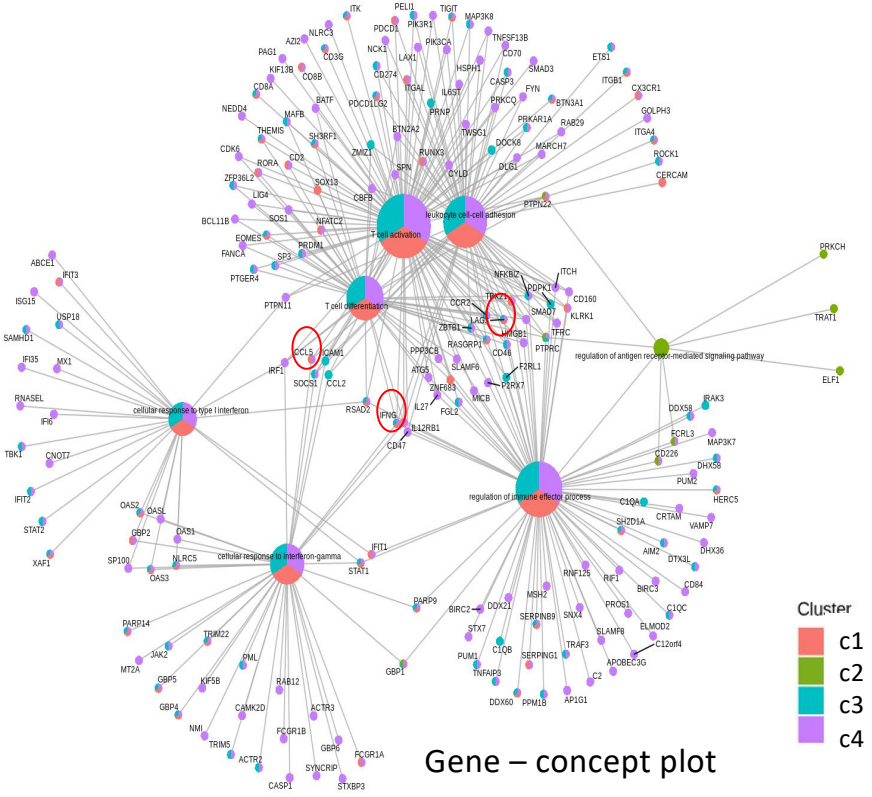
P-dependent

N
A
Infectious Diseases

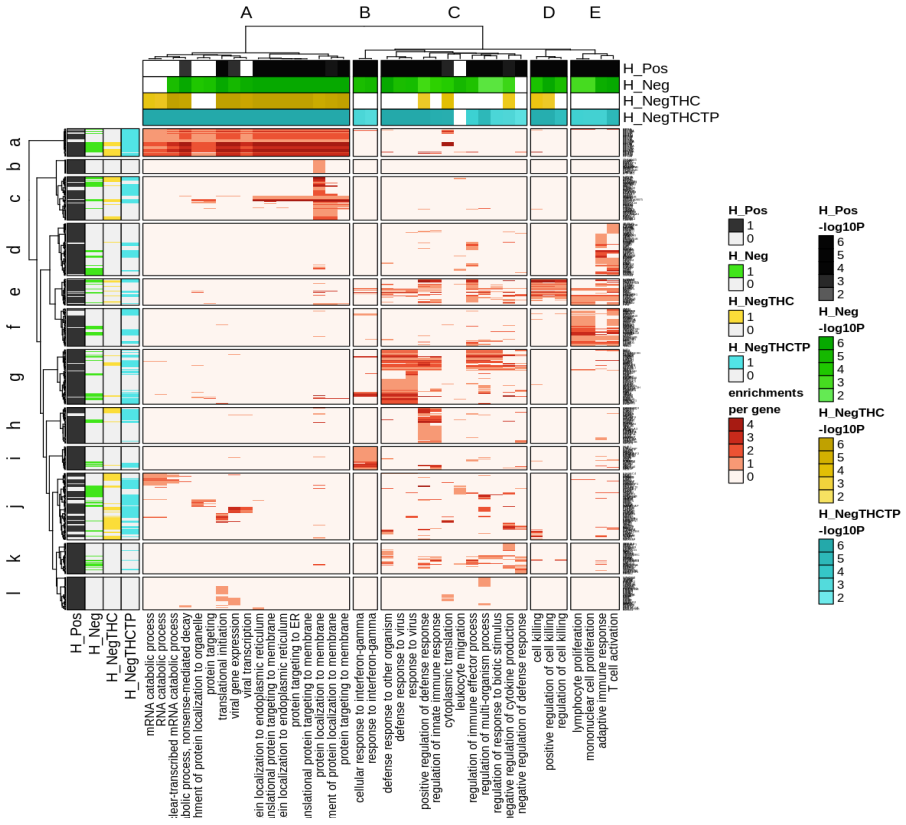


enrichment map

Overlapping genes in different GO/Pathway (redundancy) can be visualized in gene concept plots and the gene-pathway incidence heatmap



Gene – concept plot



'multienrichjam', r package

Gene – pathway incidence heatmap

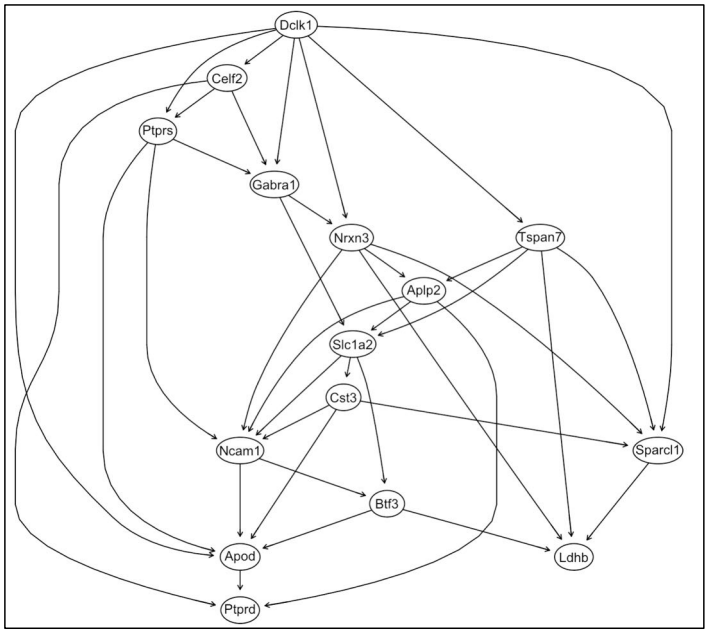
Useful packages for functional annotation of multiple gene sets

- clusterProfiler
  - High-throughput functional annotation of differential gene expression (enricher)
    - Visualize with enrichment map
    - Biological theme visualization with dotplots
    - Gene-concept network visualization
  - Based on ranking of fold changes (GSEA)
    - Visualize with enrichment map
    - Gene-concept network
- Multienrichjam
  - Gene-pathway incidence heatmap is a useful way to deal with redundancy issue of pathway analysis.
  - Importing IPA enrichment results to R and visualize it with
    - Gene-concept network analysis (cnetplot)
    - Gene-pathway incidence heatmap

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 can be assess with multiple ways
  - 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 Causal Inference
From R package “bnlearn”



National Institute of Allergy and Infectious Diseases

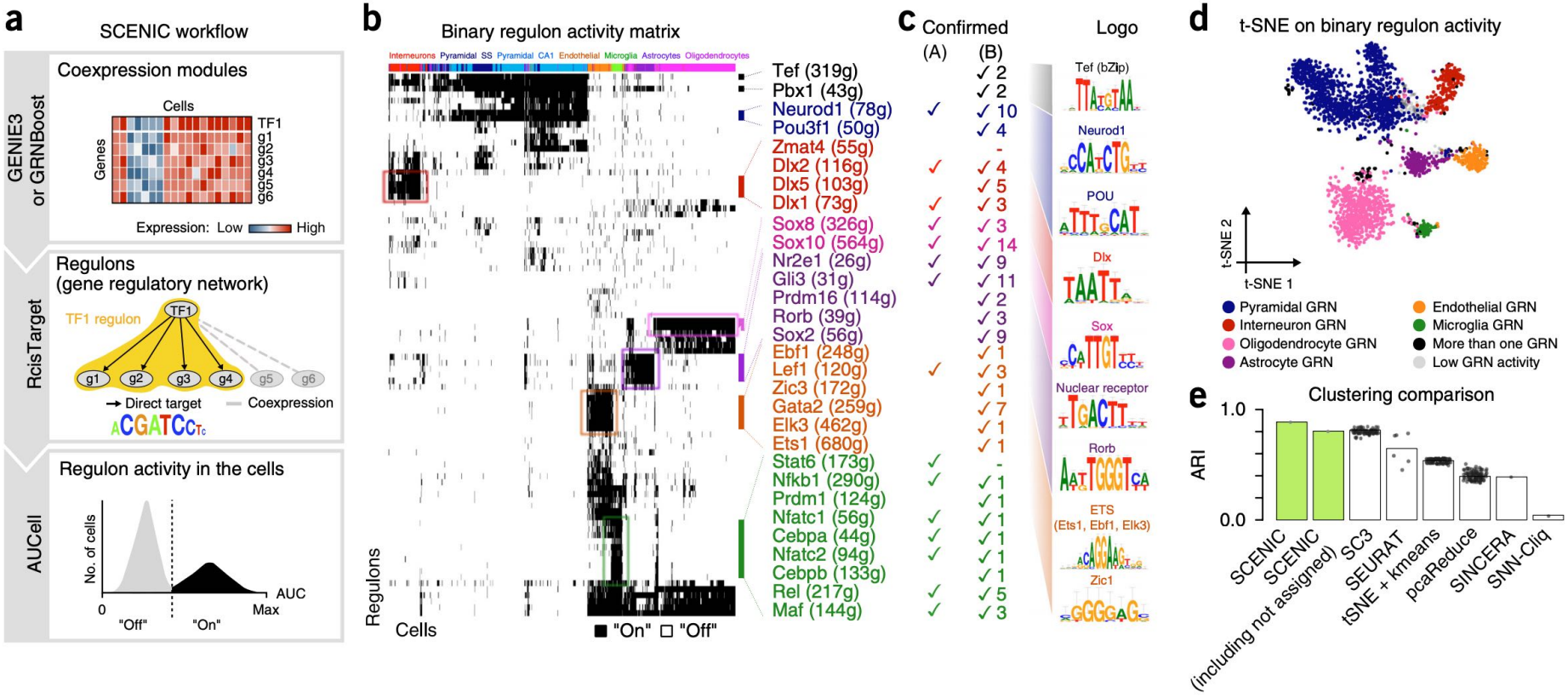
Packages for GRN inference

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

National Institute of Allergy and Infectious Diseases

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

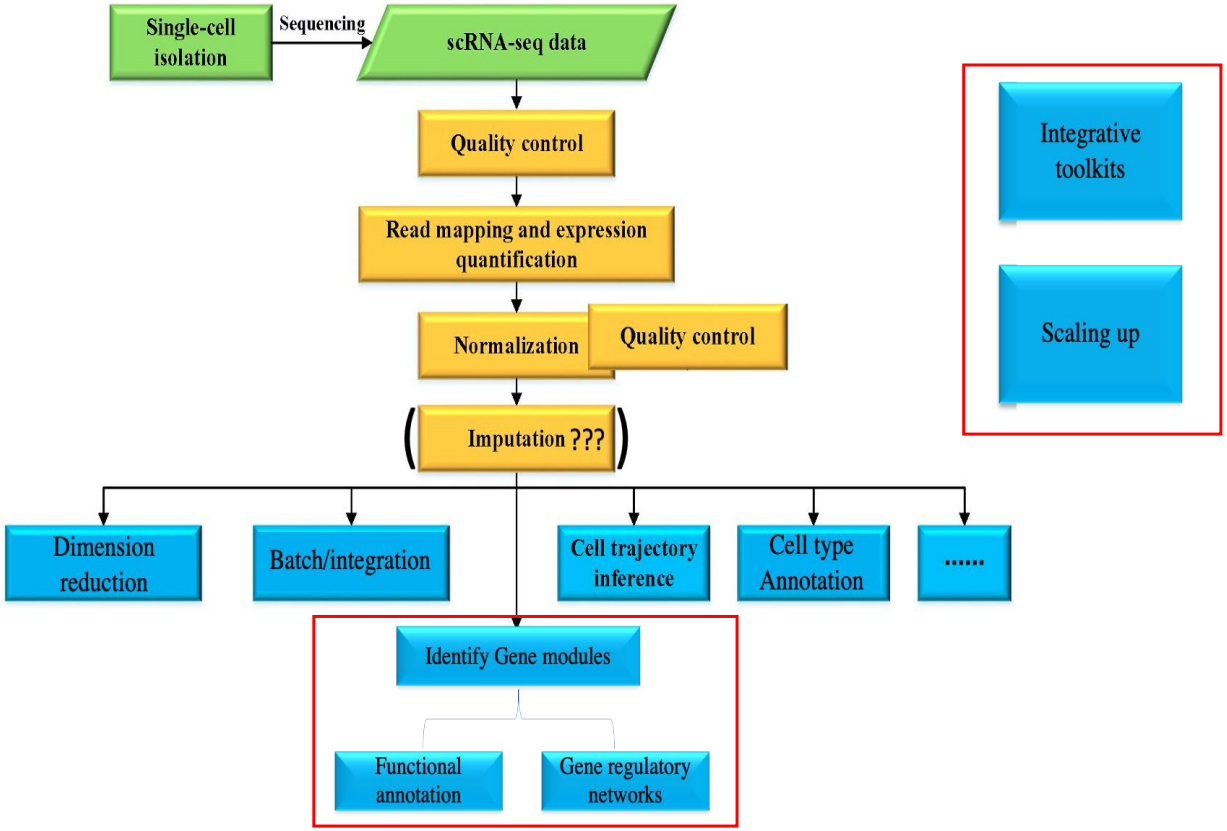
SCENIC identifies major transcriptional factors in clusters



National Institute of Allergy and Infectious Diseases

Nature method, 2017; Nature method, 2020

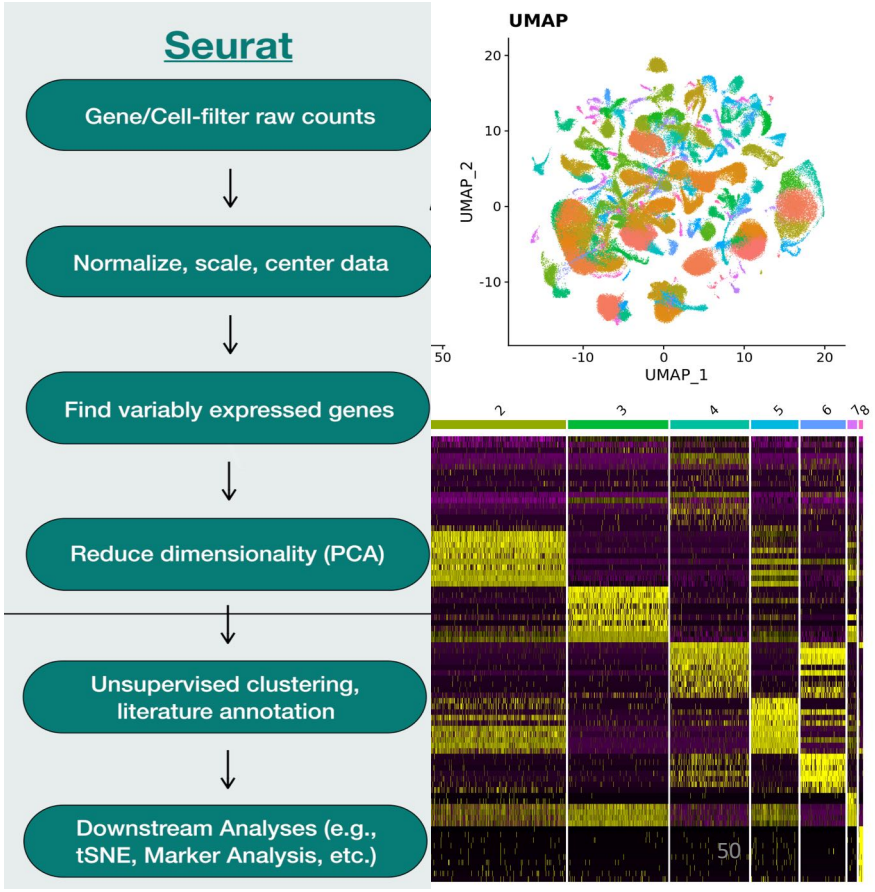
Comprehensive toolkits



National Institute of Allergy and Infectious Diseases

Comprehensive pipeline tools for explorative analysis -- Seurat

- Seurat pipeline
  - General QC assessment
  - Cell type annotation
  - Batch correction and meta analysis
  - Multimodal analysis (for CITE-seq, Hash-tagging, ATAC-seq)
  - Comparative analysis across different conditions



National Institute of Allergy and Infectious Diseases

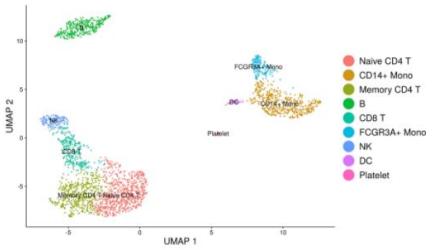
https://satijalab.org/seurat/vignettes.html

Multiple vignettes for different tasks

Basic pipeline: QC,
Dimension reduction
Clustering
Marker identifications

Integrating with CITE-seq,
HASH-seq

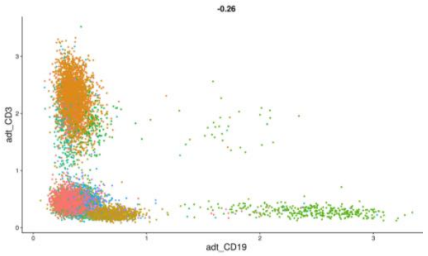
Guided tutorial — 2,700 PBMCs



A basic overview of Seurat that includes an introduction to common analytical workflows. Start from here

GO

Multimodal analysis

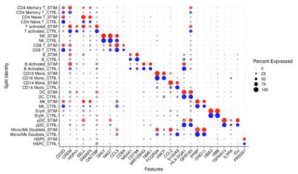


An introduction to working with multi-modal datasets in Seurat.

GO

Multiple pipelines for integrating data

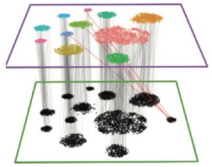
Introduction to scRNA-seq integration



An introduction to integrating scRNA-seq datasets in order to identify and compare shared cell types across experiments

GO

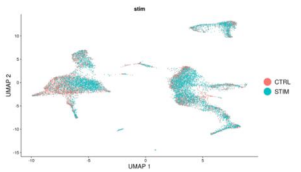
Mapping and annotating query datasets



Learn how to map a query scRNA-seq dataset onto a reference in order to automate the annotation and visualization of query cells

GO

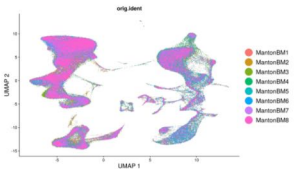
Fast integration using reciprocal PCA (RPCA)



Identify anchors using the reciprocal PCA (rPCA) workflow, which performs a faster and more conservative integration

GO

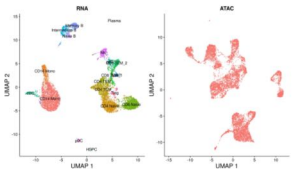
Tips for integrating large datasets



Tips and examples for integrating very large scRNA-seq datasets (including >200,000 cells)

GO

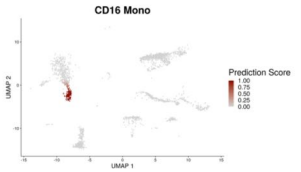
Integrating scRNA-seq and scATAC-seq data



Annotate, visualize, and interpret an scATAC-seq experiment using scRNA-seq data from the same biological system

GO

Multimodal Reference Mapping



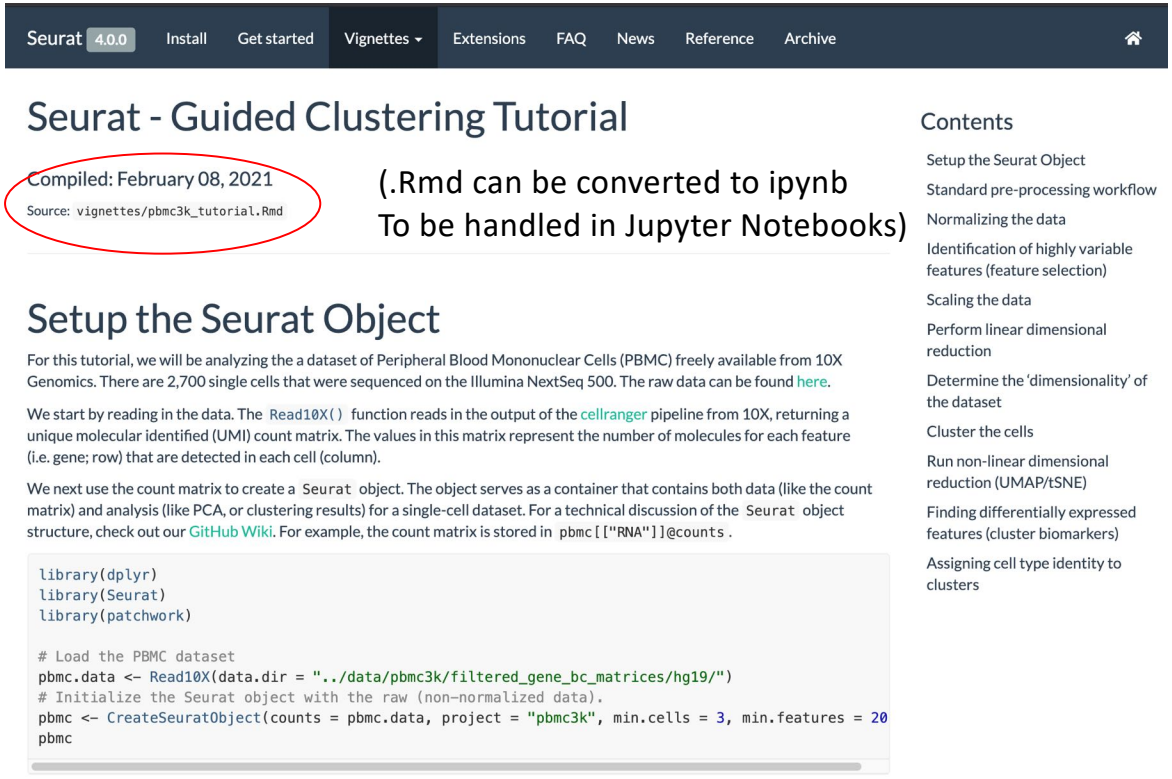
Analyze query data in the context of multimodal reference atlases.

GO

NIH National Institute of Allergy and Infectious Diseases

Download vignette/tutorial to use on your data

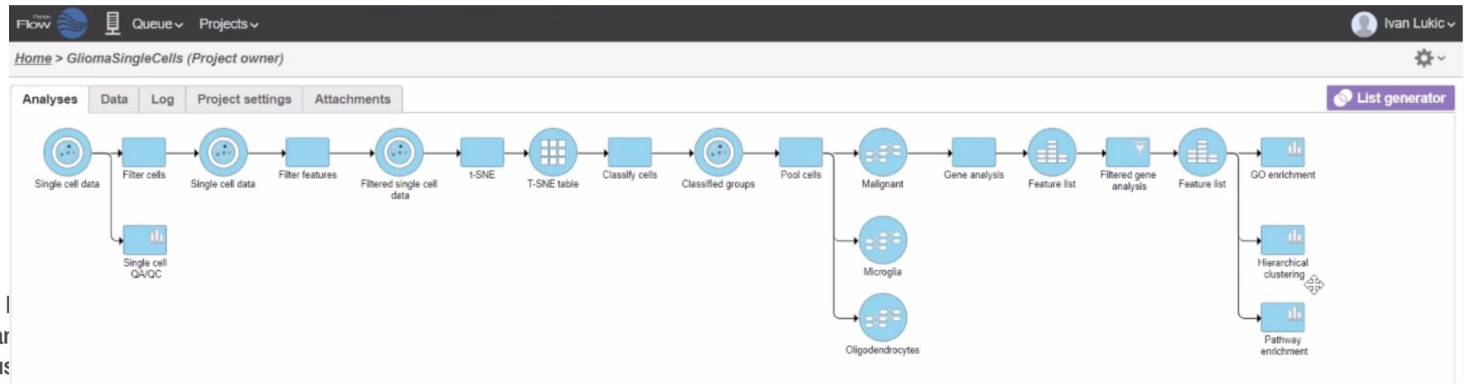
- Linked to github for you to download the code.
- Follow through the tutorial using sample data included in the package.
- Change the input file to your our data to use the analytic pipelines.



National Institute of Allergy and Infectious Diseases

Graphic User Interface– for researchers

- 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/



Scale up with python implementations

- Python packages/toolkits are increasingly popular
  - scanpy pipeline
  - scVelo pipeline
- Some has a R rapper.

scanpy in R

- Use python in R through Reticulate
- Use R in python through rpy2

tSNE plot of
1300000 neurons by scanpy

Scanpy vs. Seurat

Satija et al., Nat. Biotechn. (2015)

Scanpy is benchmarked with Seurat.

- preprocessing: <1 s vs. 14 s
- regressing out unwanted sources of variation: 6 s vs. 129 s
- PCA: <1 s vs. 45 s
- clustering: 1.3 s vs. 65 s
- tSNE: 6 s vs. 96 s
- marker genes (approximation): 0.8 s vs. 96 s

https://scanpy.readthedocs.io/en/stable/tutorials.html

scanpy

stable

Search docs

- Tutorials
  - Clustering
  - Visualization
  - Trajectory inference
  - Integrating datasets
  - Spatial data
  - Further Tutorials
- Usage Principles
- Installation
- API
- External API
- Ecosystem
- Release notes
- News
- Contributing
- Contributors
- References

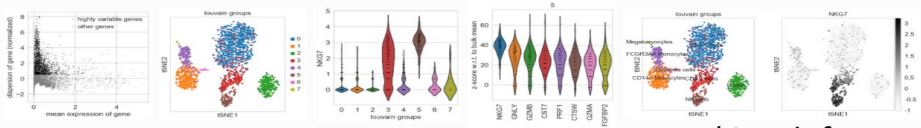
Read the Docs v: stable

» Tutorials

Tutorials

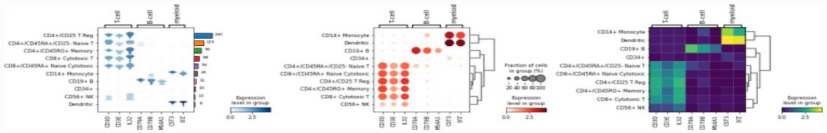
Clustering

For getting started, we recommend Scanpy's reimplementation → tutorial: pbmc3k of Seurat's [Satija15] clustering tutorial for 3k PBMCs containing preprocessing, clustering and the identification of cell types via known marker genes.



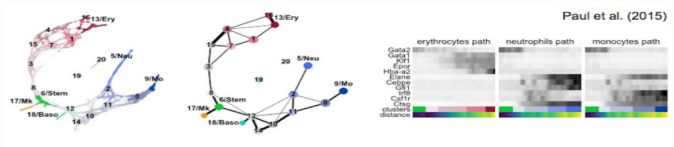
Visualization

This tutorial shows how to visually explore genes using scanpy. → tutorial: plotting/core



Trajectory inference

Get started with the following example for hematopoiesis for data of [Paul15]: → tutorial: paga-paul15



More examples for trajectory inference on complex datasets can be found in the PAGA repository [Wolf19], for instance, multi-resolut

- Same data set, similar analysis with Seurat.
- But implemented in python.
- Using jupyter notebook as a interface.

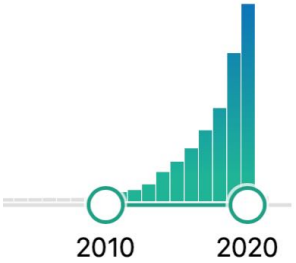
(.ipynb format can be converted to .RMD to be run in Rstudio)

New trends—deep learning methods is getting momentum in single cell analysis

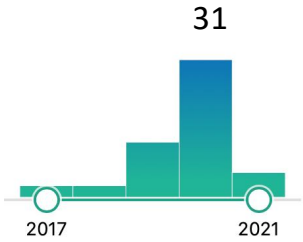
- Python packages are increasingly popular
- Single cell RNA-seq analysis is a recent development
- Single cell analysis with neural network is picking up fast

No. of publications in Pubmed search

Single cell RNA-seq



Single cell RNA-seq + Deep learning

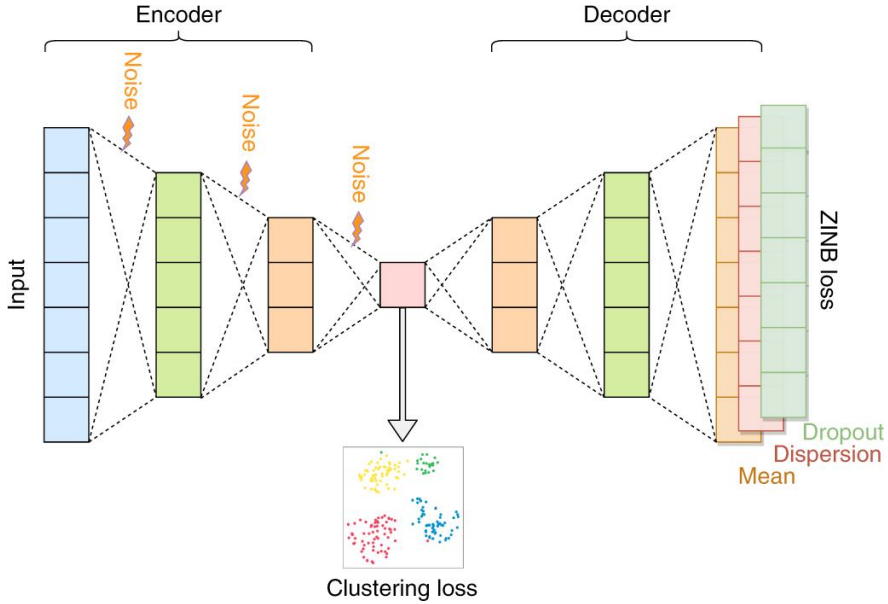


Autoencoder
Generative Adversarial Network
Transfer learning

National Institute of Allergy and Infectious Diseases

Scale up--Deep learning in single cell genomics

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

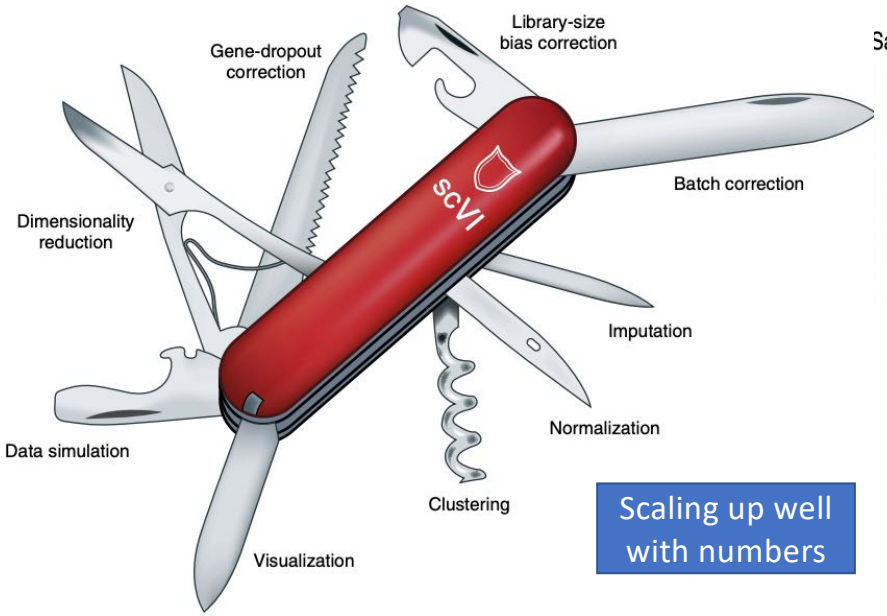


Nature Machine Intelligence, April 2019

National Institute of Allergy and Infectious Diseases

ML and single cell biology: https://www.krishnaswamylab.org/workshop

Autoencoder strategy is trendy



Scaling up well with numbers

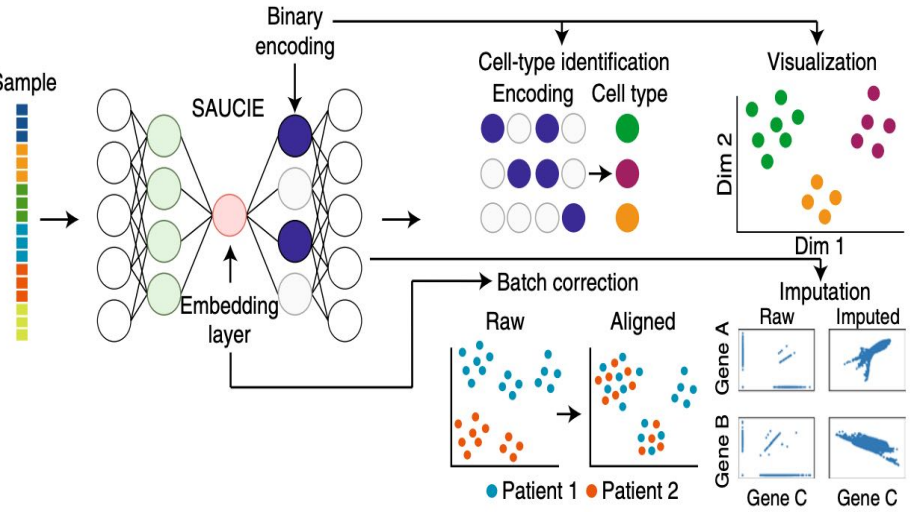
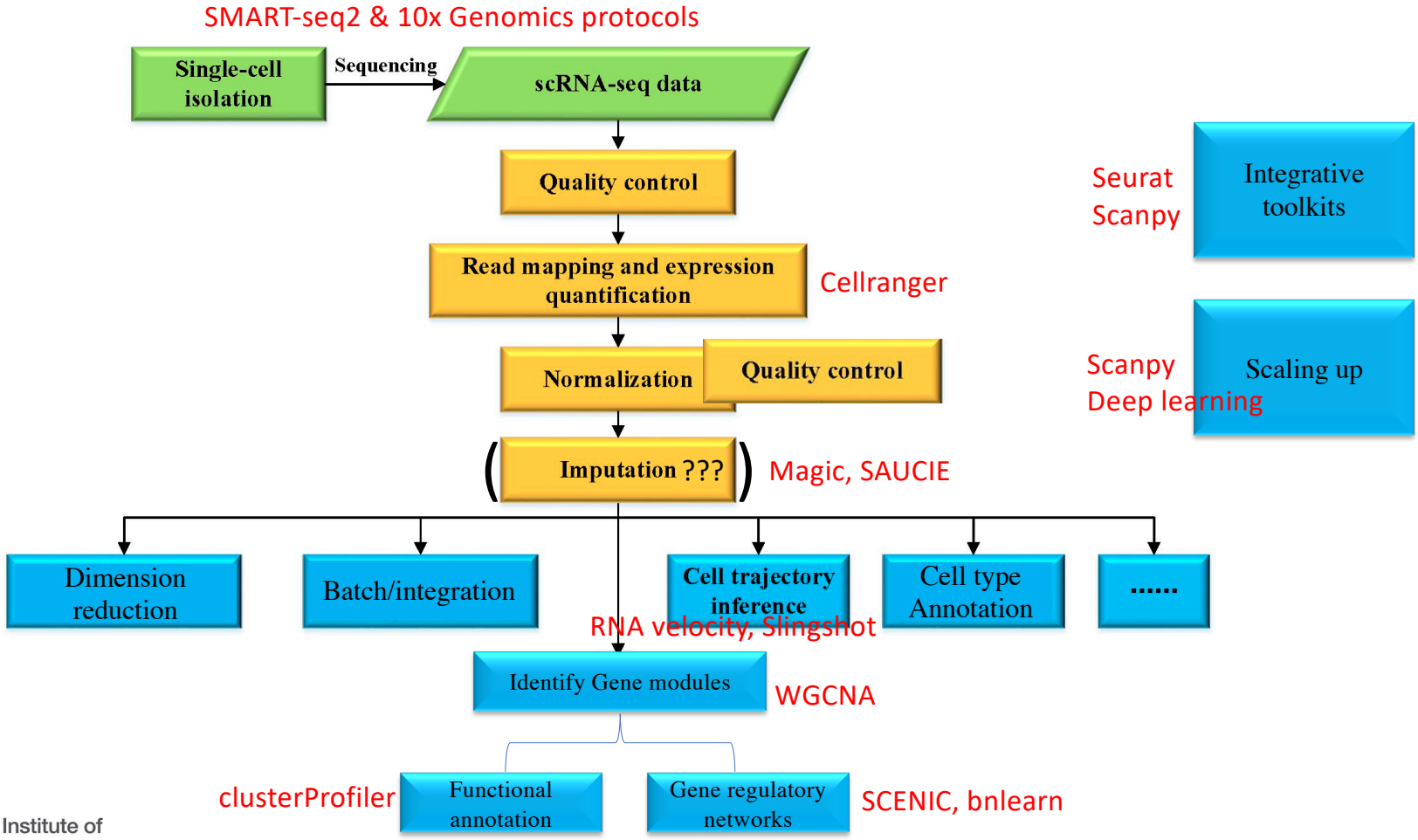


Fig. 1 | scVI is a multifaceted tool for scRNA-seq data processing and analysis. The Bayesian deep learning and variational inference framework enables researchers to obtain scalable and accurate results across a variety of domains. Credit: Kim Caesar/Springer Nature

scVI, Nature Methods, Dec 2018
SAUCIE, Nature Methods, Nov 2019
totalVI, Nature methods, Dec 2020
SAVER-X, Nature method, Sep 2019
… …



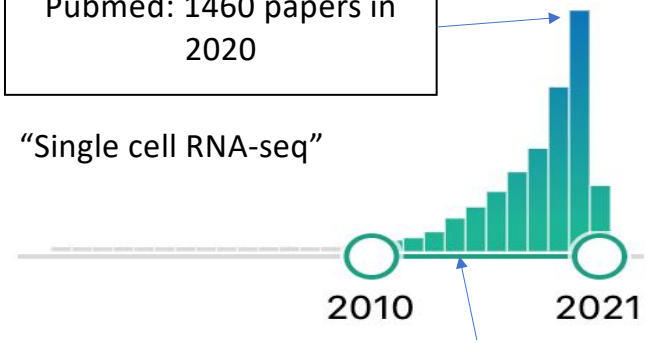
National Institute of Allergy and Infectious Diseases

https://www.frontiersin.org/files/Articles/441123/fgene-10-00317-HTML/image_m/fgene-10-00317-g001.jpg

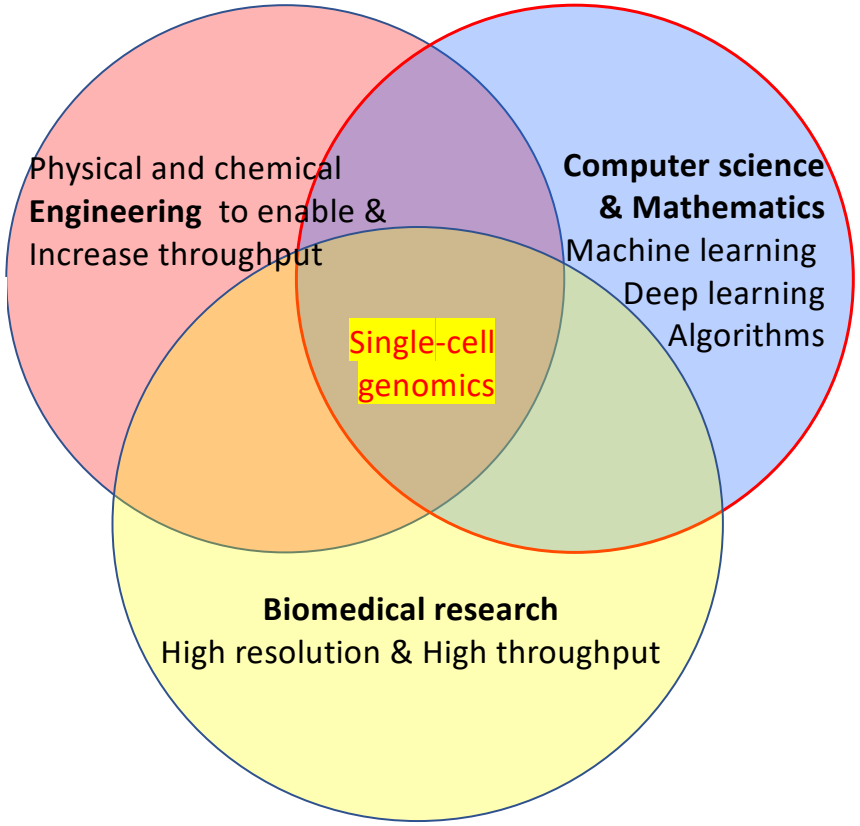
An young inter-disciplinary field with growing opportunities

Pubmed: 1460 papers in 2020

"Single cell RNA-seq"



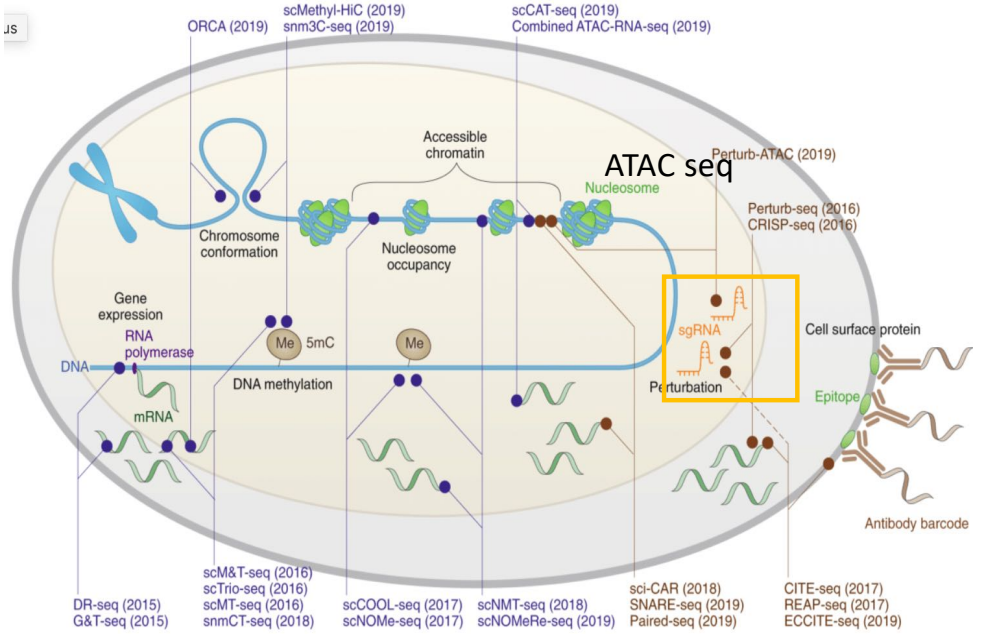
[Method of the Year 2013 Nature Methods]



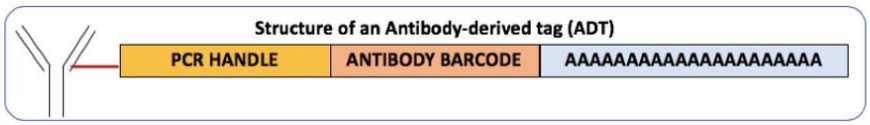
National Institute of Allergy and Infectious Diseases

Single-cell sequencing technologies: more opportunities

- **Single cell RNA-seq**
- CITE-seq/HASH-seq for surface protein antigens
- ATAC-seq to access open chromatin
- ECCITE-seq: functional screening with sgRNA
- Single cell genomics



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

Thank you!

BCBB members

Bioinformatics@niaid.nih.gov
zhuy16@nih.gov

NIH National Institute of Allergy and Infectious Diseases

Further readings

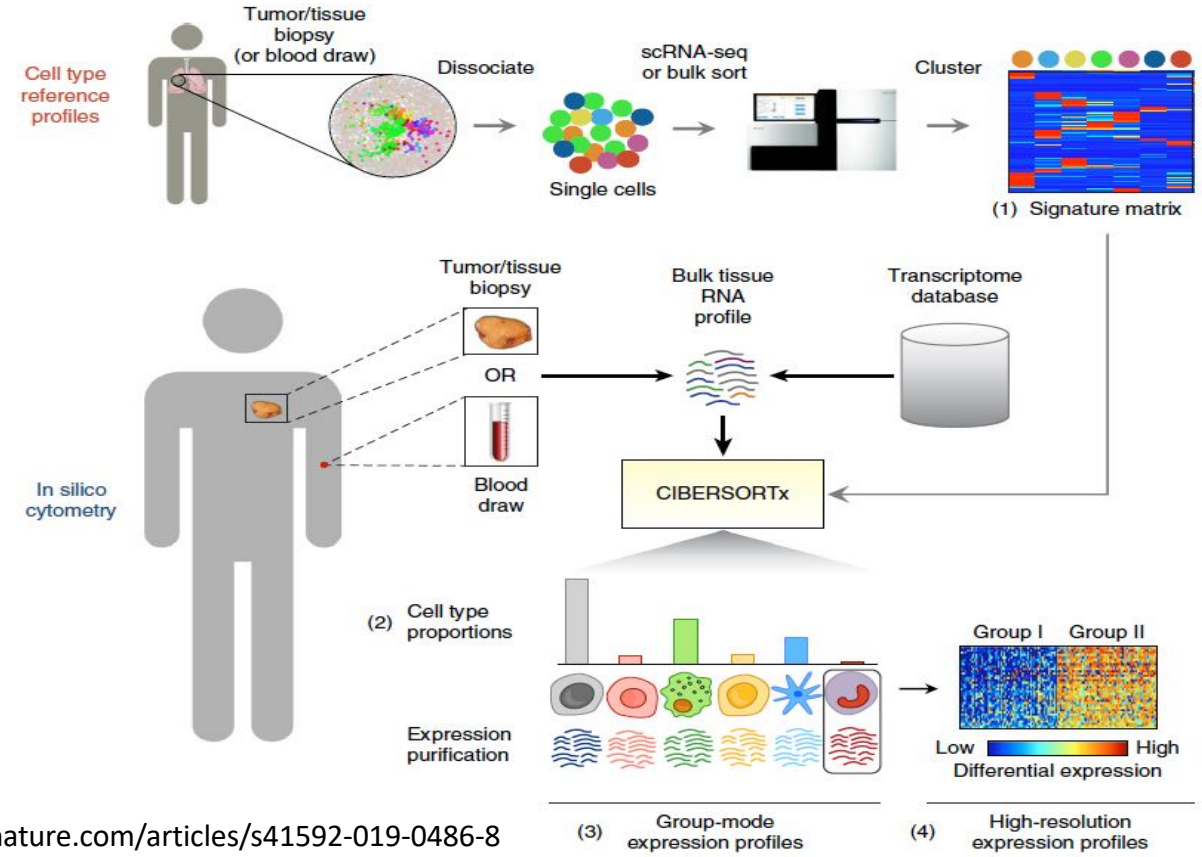
- Single cell softwares
  - https://github.com/seandavi/awesome-single-cell
- Some personal tips for learning bioinformatics
  - https://github.com/zhuy16/learning_notes
- Gene Regulatory network analysis
  - https://github.com/niaid/Gene_Regulatory_Networks
- Sean Davis's overview on scRNA-seq
  - https://figshare.com/articles/Single_Cell_Present_and_near_Future/12121674
- 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

Others aspects not touched

- Scoring or regression out the the cell cycle related gene changes
- Using single cell RNA-seq data to do deconvolution on bulk RNA-seq.
  - Sybersortx
- Using maps to project new data to reference cell types.
- Finding out microbial reads from single cell RNA-seq data.
- Single bacteria sequencing
- Estimate copy number variation from tumors.
- Estimate SNP mutations from single cell data.

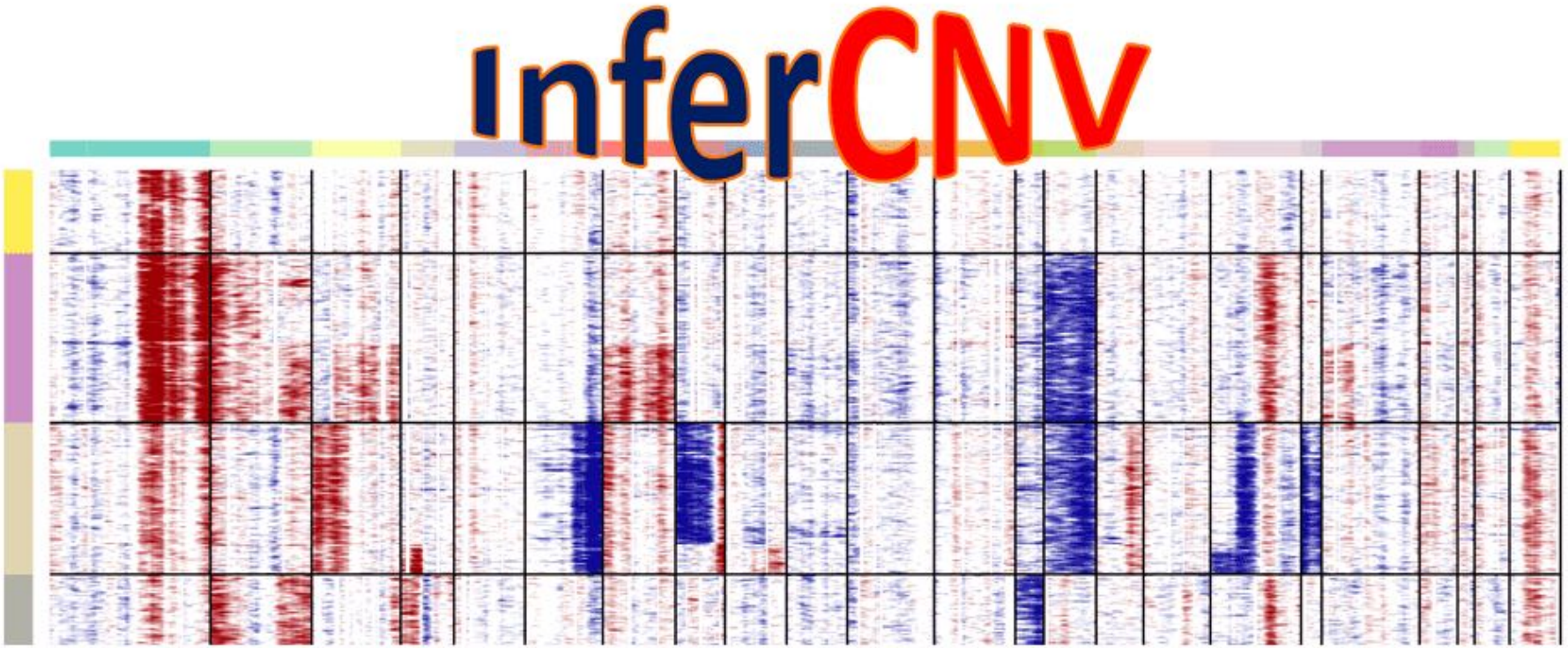
Eleven grand challenges in single-cell data science
Lähnemann et al. Genome Biology (2020) 21:31

Use single cell data to deconvolute bulk-RNA-seq--CybersortX



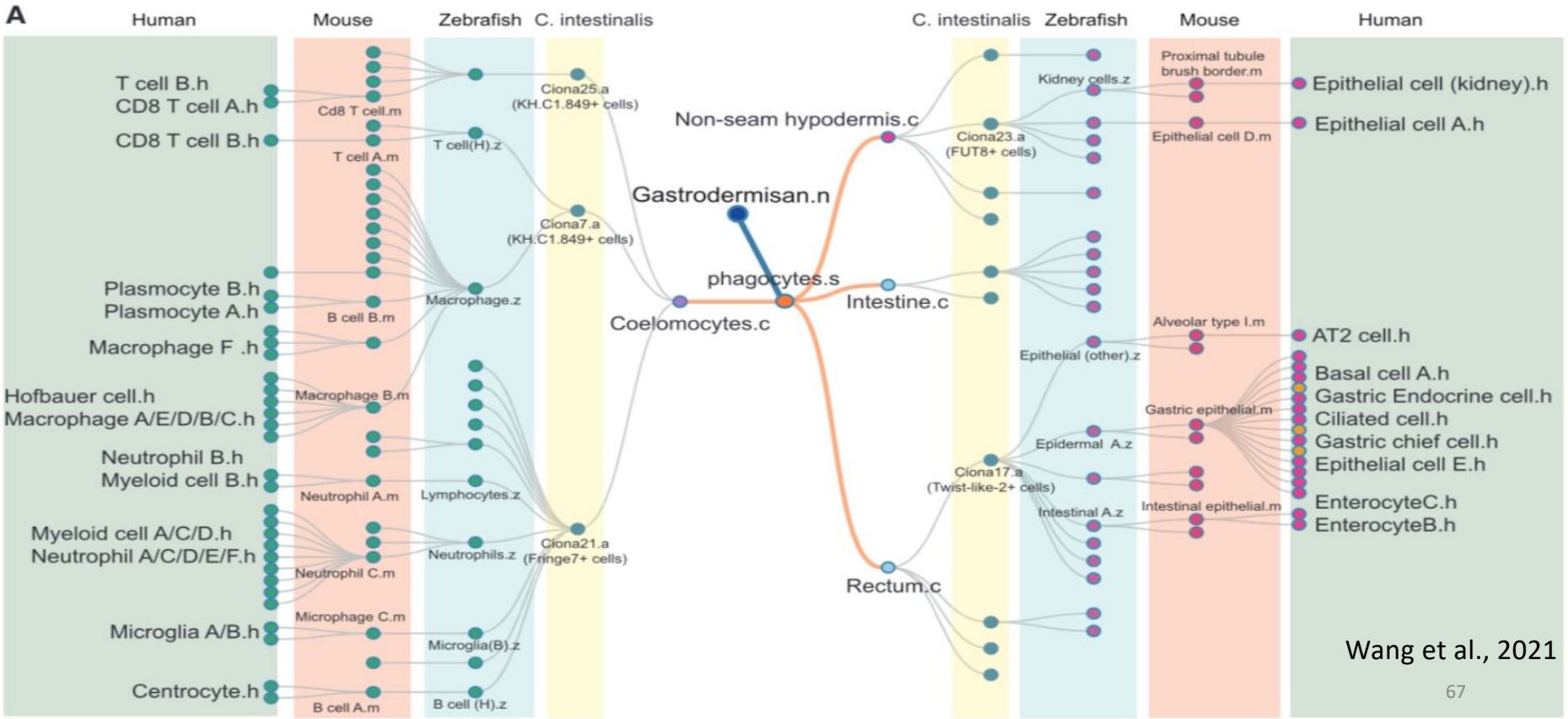
https://www.nature.com/articles/s41592-019-0486-8

Inferring copy number variations in tumor samples



https://rpubs.com/bman/418918

To study evolution of cell types in lung



Wang et al., 2021