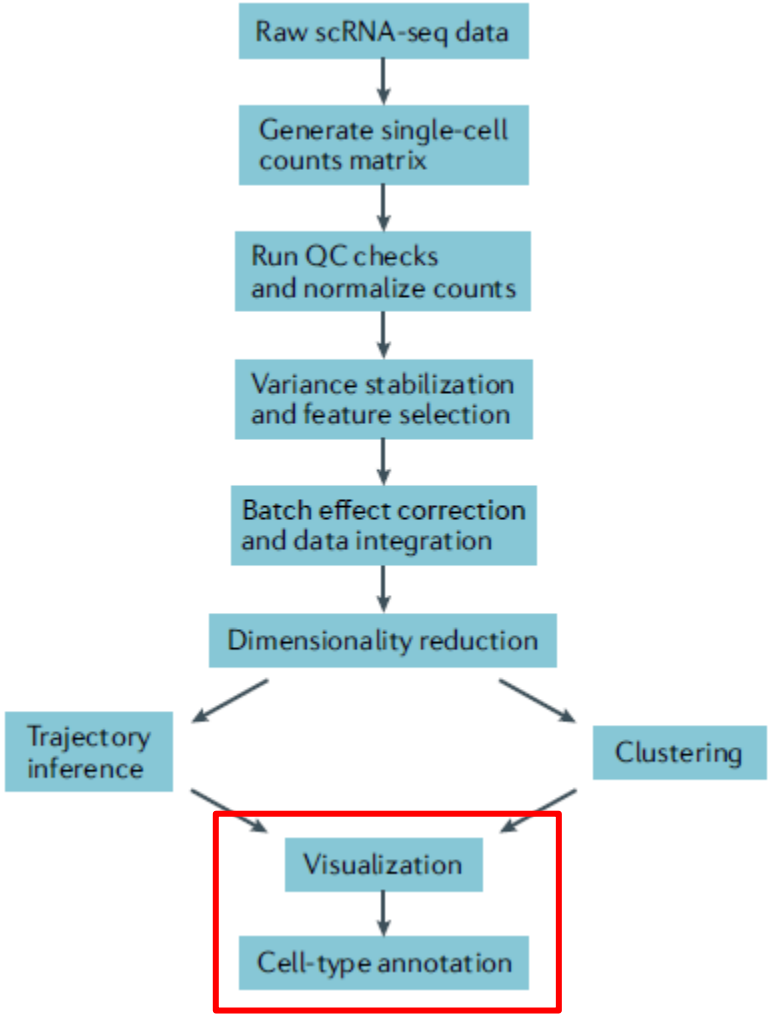
Secondary Analysis

Agnès Paquet, Nicolas Nottet

SincellTE 2022 - 01/11/2022

agnes.paquet@syneoshealth.com

Summary of what we have seen so far



Wu, Nat RevNeph 2020

Visualization tools

Cakir, NAR Gen and Bioinfo 2020

Table 1. Overview of the visualization tools and their capabilities

	ASAP	Bbrowser	cellxgene	Granatum	iSEE	Loom viewer	Loupe Cell Browser	SCope	scSVA	scVI	Single Cell Explorer	SPRING	UCSC Cell Browser
Web Sharing	✓		✓		✓	✓		✓	✓		✓	✓	✓
Interactivity	✓	✓	✓		✓		✓	✓	✓		✓	✓	✓
Docker	✓		✓		✓			✓	✓				
Cloud Support			✓	✓				✓	✓			✓	
SaaS	✓			✓				✓			✓	✓	✓
Loom	✓				✓	✓		✓	✓	✓	✓		✓
h5ad		✓	✓						✓	✓	✓		✓
SCE					✓								
Seurat		✓									✓		✓
csv/txt	✓	✓		✓					✓	✓	✓	✓	✓
Platform	Java/ R	Desktop	Python	R	R	Python	Desktop	Python	R	Python	Python	Python	Python

- Single-cell data are large, and you will have many back-forth with your collaborators
- Anticipate how you will share your results = Avoid reploting distribution for each of the collaborators favorite gene
- Sharing data and results using tools with GUI will be helpful
  - E.g.: iSEE, UCSC cell browser

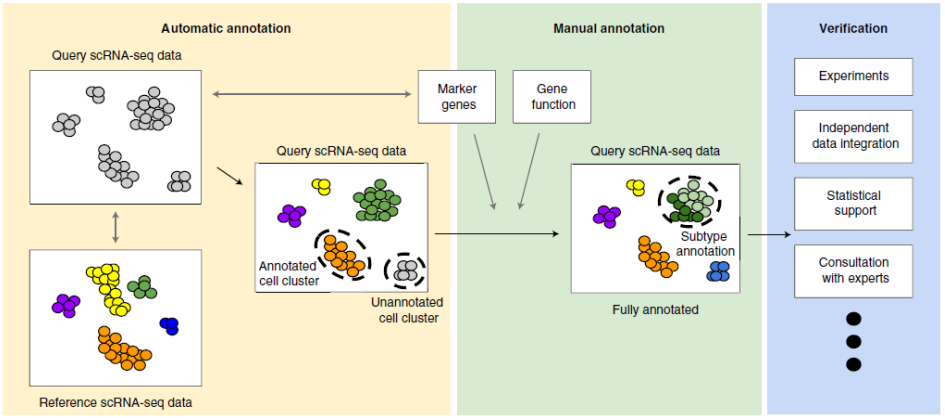
Cell Annotation workflow

- Perform a manual review of marker genes from the DE results and compare to known markers
  - Labor intensive
  - Can be subjective
  - Difficult for poorly characterized cell types / systems
  - Why is my favorite gene not expressed?

- Expertise of the biologist is often required a this step

- Use automated annotation tools
  - Many tools available.
  - Reference: Internal data or public repository data
  - Known signatures / pathways

Clarke 2021



Some existing tools

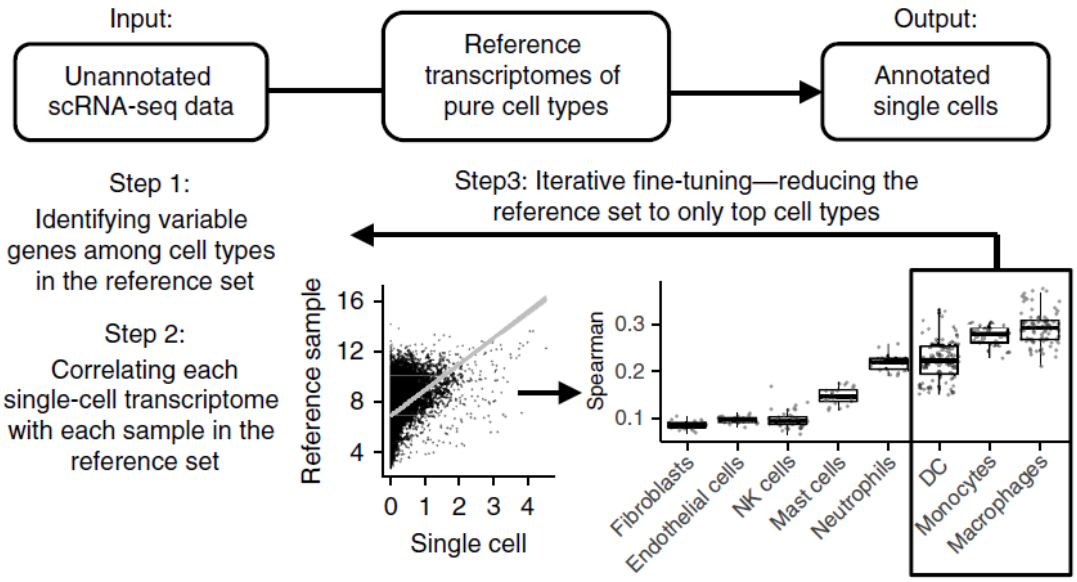
Clarke 2021

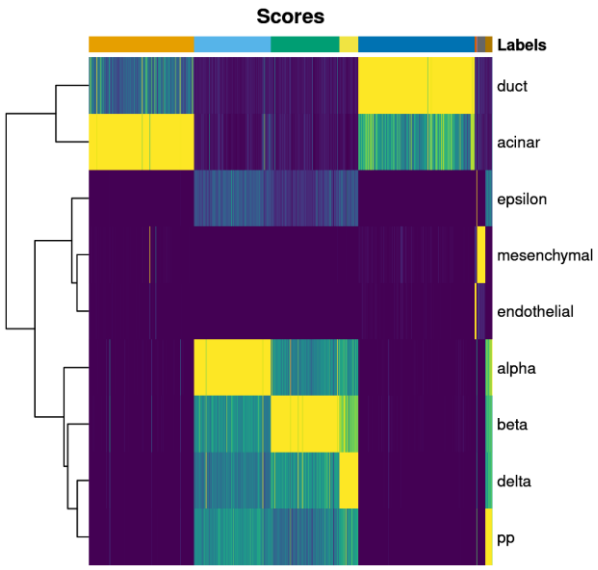
Table 2 | Summary of referenced annotation tools

Tool	Type	Language	Resolution	Approach	Allows 'None'	Notes
singleCell Net[42]	Reference based	R	Single cells	Relative-expression gene pairs + random forest	Yes, but rarely does so even when it should[33]	10–100× slower than other methods; high accuracy
scmap-cluster[41]	Reference based	R	Single cells	Consistent correlations	Yes	Fastest method available; balances false-positives and false-negatives; includes web interface for use with a large pre-built reference or custom reference set
scmap-cell[41]	Reference based	R	Single cells	Approximate nearest neighbors	Yes	Assigns individual cells to nearest neighbor cells in reference; allows mapping of cell trajectories; fast and scalable
singleR[43]	Reference based	R	Single cells	Hierarchical clustering and Spearman correlations	No	Includes a large marker reference; does not scale to data sets of ≥10,000 cells; includes web interface with marker database
Scikit-learn[102]	Reference based	Python	Multiple possible	k-nearest neighbors, support vector machine, random forest, nearest mean classifier and linear discriminant analysis	(Optional)	Expertise required for correct design and appropriate training of classifier while avoiding overtraining
AUCell[103]	Marker based	R	Single cells	Area under the curve to estimate marker gene set enrichment	Yes	Because of low detection rates at the level of single cells, it requires many markers for every cell type
SCINA[34]	Marker based	R	Single cells	Expectation maximization, Gaussian mixture model	(Optional)	Simultaneously clusters and annotates cells; robust to the inclusion of incorrect marker genes
GSEA/GSVA[36,104]	Marker based	R/Java	Clusters of cells	Enrichment test	Yes	Marker gene lists must be reformatted in GMT format. Markers must all be differentially expressed in the same direction in the cluster
Harmony[105]	Integration (Box 2)	R	Single cells	Iterative clustering and adjustment	Yes	Integrates only lower-dimensional projection of the data; seamlessly integrated into Seurat pipeline; may overcorrect data
Seurat-canonical correlation analysis[106]	Integration (Box 2)	R	Single cells	MNN anchors + canonical correlation analysis	Yes	Accuracy depends on the accuracy of MNN anchors, which are automatically-identified corresponding cells across data sets
mnnCorrect[107]	Integration (Box 2)	R	Single cells	MNN pairs + singular value decomposition	Yes	Accuracy depends on the accuracy of MNN pairs (cells matched between data sets). Referred to in Box 2
Linked inference of genomic experimental relationships (LIGER)[108]	Integration (Box 2)	R	Single cells	Non-negative matrix factorization	Yes	Allows interpretation of data set-specific and shared factors of variation. Referred to in Box 2

MNN, mutual nearest neighbor.

SingleR





Aran 2019

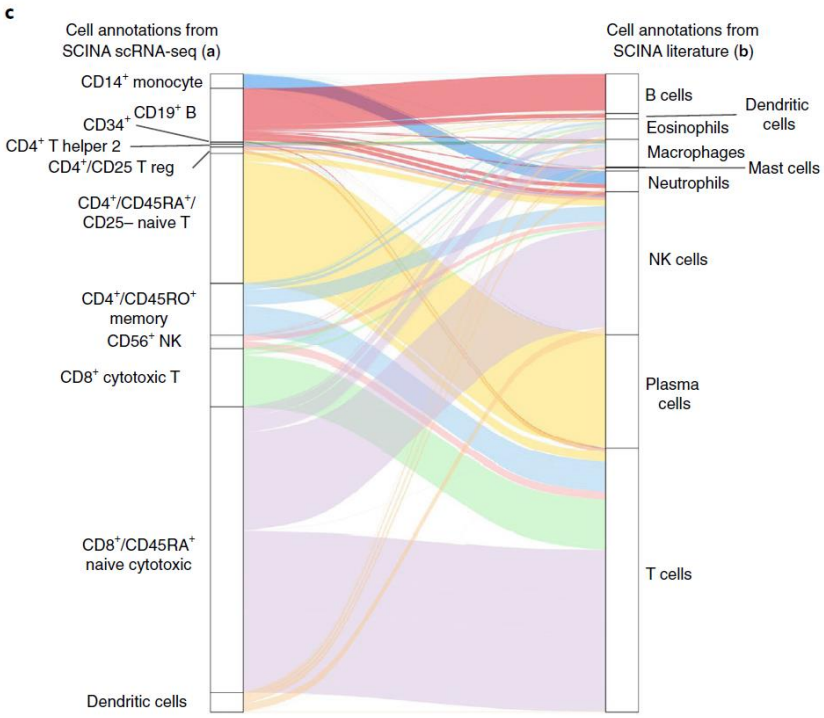
http://bioconductor.org/books/release/SingleRBook/

Public Reposity Data

- Cell Atlases: HCA, Tabula Muris, Immgen, FANTOM5, Panglao DB etc
- Bulk repository: GEO, ArrayExpress

- Azimuth
  - Web based tools, with preloaded human and mouse reference sets
  - Based on transfer learning (Seurat)
  - Very easy to use and retrieve prediction results

- Caveats
  - Very large datasets can be difficult to process (100,000s of cells)
  - Cell annotations not always easy to find
  - Sample processing can have a huge impact on the data
  - **The reference may not match your data well**
    - Quality scores are often provided -> USE THEM

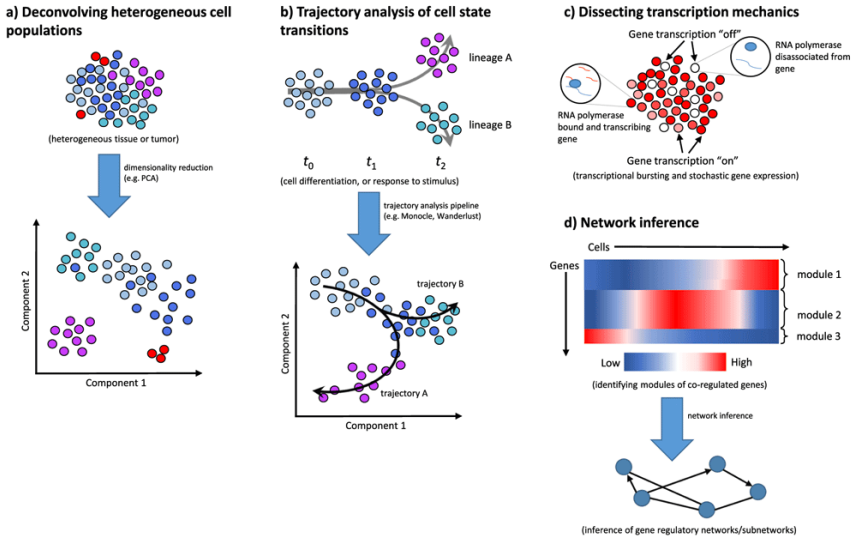


What is your question ?

1. Resolving cellular heterogeneity
   - DE analysis, clustering
   - Deconvolution of bulk data

2. Understand developmental processes and cell fate decisions
   - Trajectory inference, RNA velocity

3. Identification of co-regulated gene modules and network inference
   - Gene network inference, identification of co-regulated gene modules

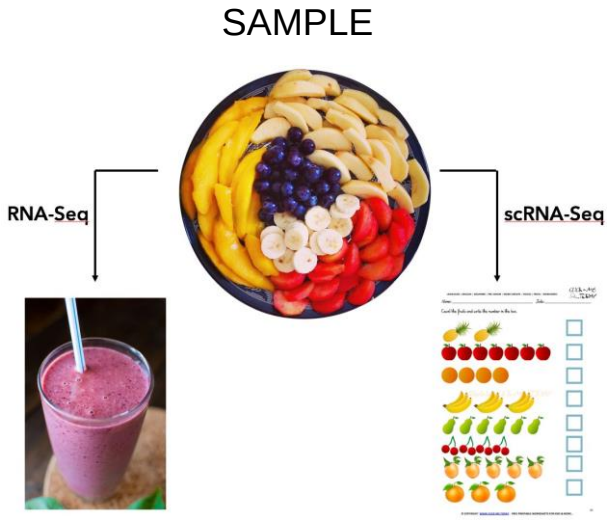


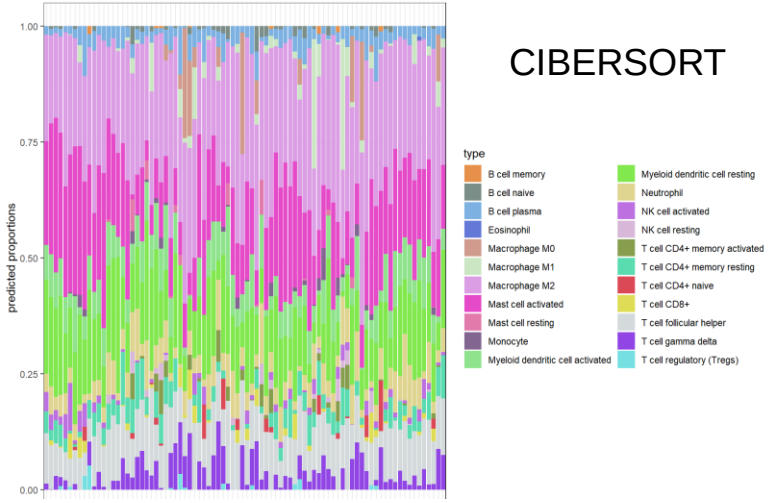
Liu S, F1000Research 2016, Haque A, Genome Medicine 2017
Zhu S, Oncotarget 2017, Griffiths JA, Molecular System Bio 2018

Overview of some advanced analysis

- Deconvolution
- Inference of cell-cell communication
- Gene network inference
- InferCNV
- Trajectory analysis
- GSVA
- RNA velocity

Deconvolution - *In-silico* immunophenotyping

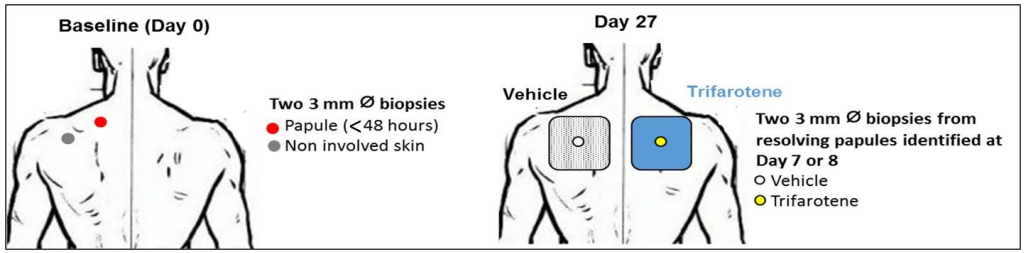


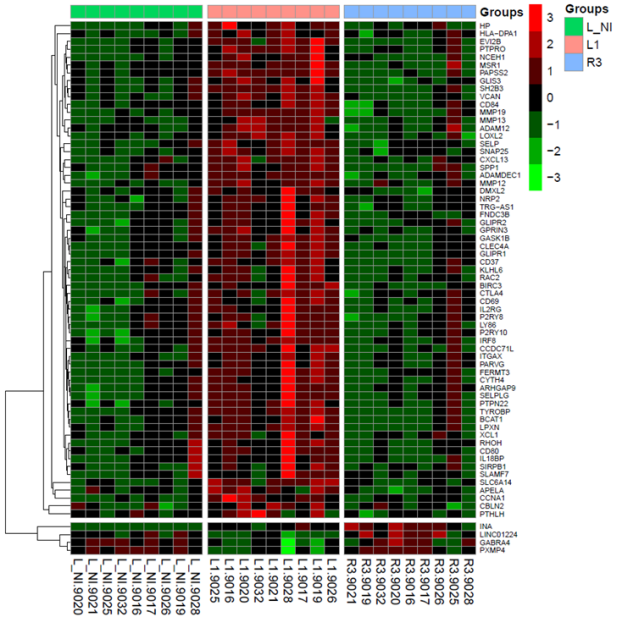


- Aim:
  - Refine bulk sample analysis based on prior knowledge about the cell types present in our sample
- Input:
  - Bulk expression data (arrays / NGS)
  - Reference cell types and their markers
  - Apply ML to estimate the proportion of each cell type in the bulk sample
- What we can do with this:-
  - Adjust for differences in cell type proportion in the differential analysis
  - Estimate the increase/decrease of specific cell types with pathology/response
- Many tools available
  - Cancer vs Immune cells infiltration
  - Very sensitive to data normalization

Example

Dreno et al, 2021



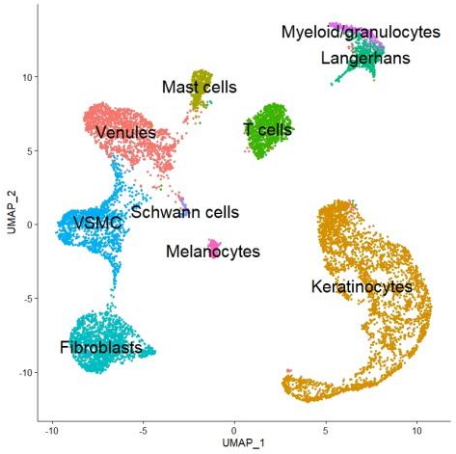


- Clinical trial data of new treatment for acne
- 4 conditions, 9 samples / condition
  - Lesional skin, non-lesional skin
  - Vehicle, Treatment
- Affymetrix arrays

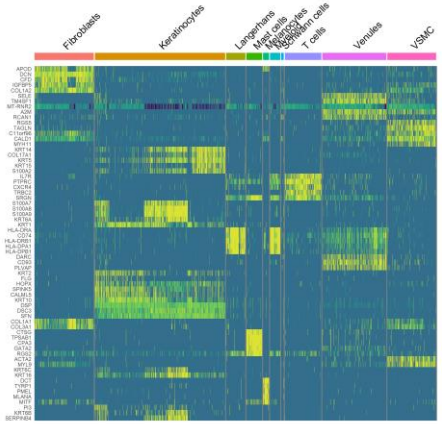
- Can we use single cell data to understand what immune cell subtypes are affected by treatment ?

Results

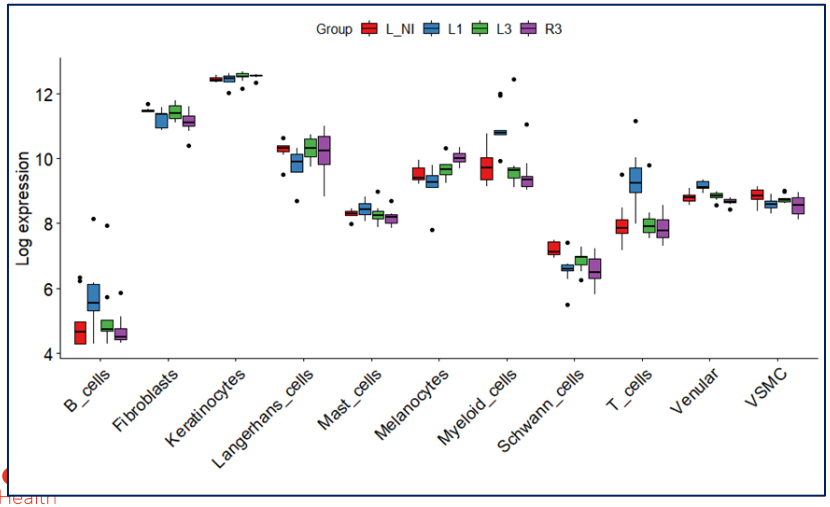
1. Analysis of acne / healthy skin scRNAseq data from GEO



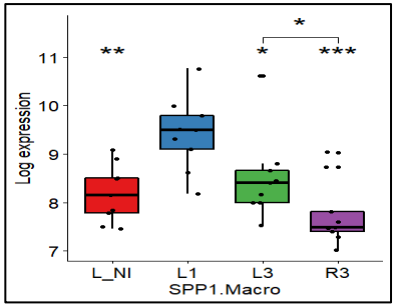
2. Definition of gene markers for each cell type present in skin + mining of literature based on markers observed in bulk DE genes



3.Deconvolution step: here, we used average expression of each cell type in our bulk samples

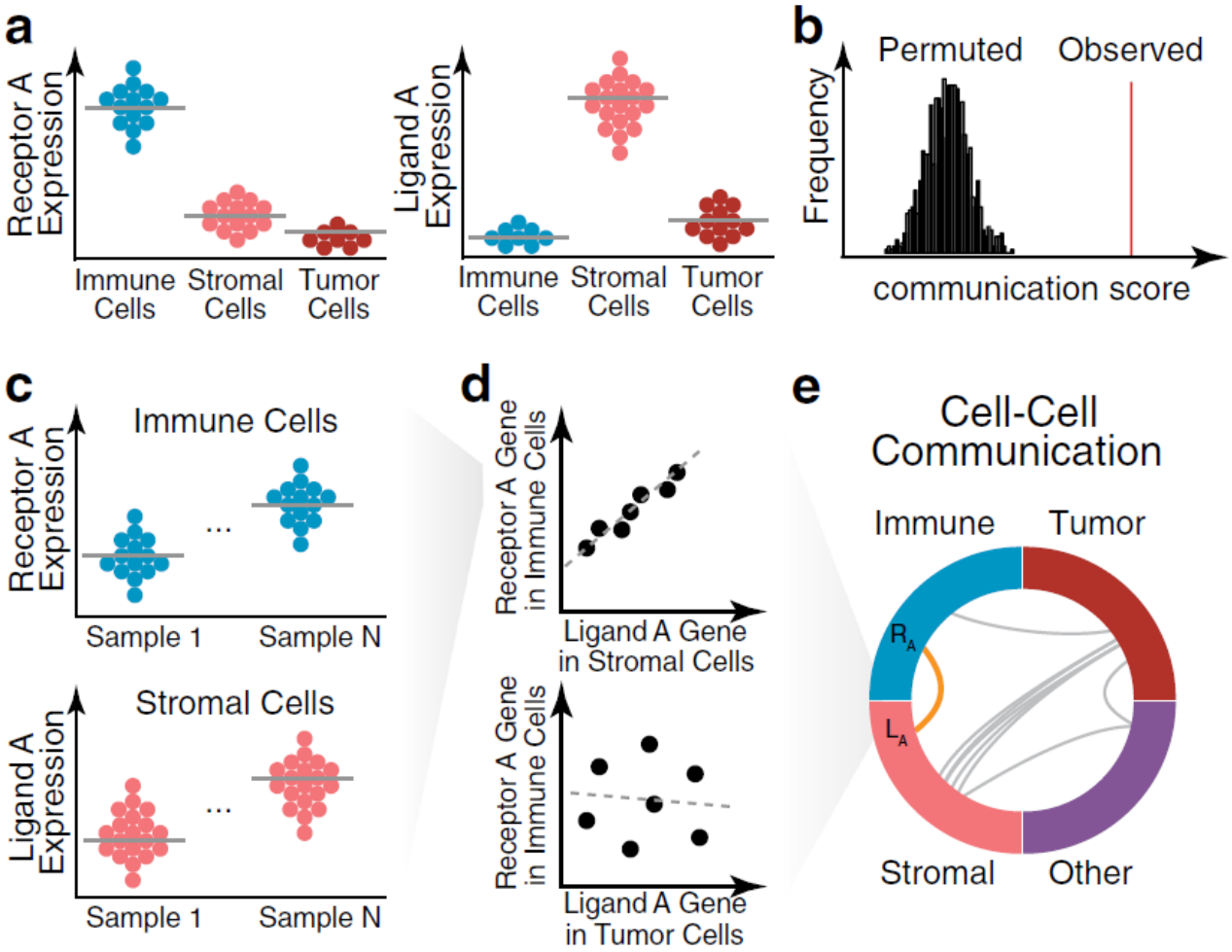


4.Refined analysis of macrophage subtypes show that the treatment is downregulating SPP1+ macrophages



Dreno et al, 2021

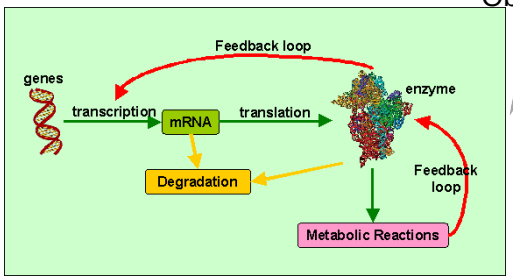
Inference of Cell-Cell communications



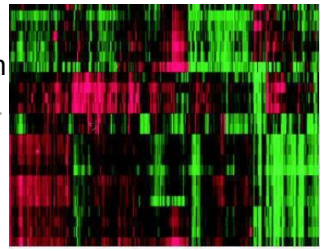
Fan, EMM 2020

Gene Regulatory Network

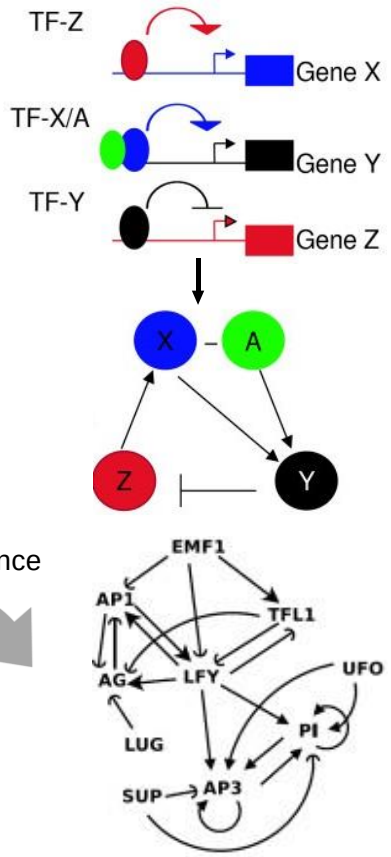
- Gene regulatory networks (GRNs) are the on-off switches on a cell operating at the gene level
- Two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition
- It can be inferred from correlations in gene expression data, time-series gene expression data, and/or gene knock-out experiments…



Observation



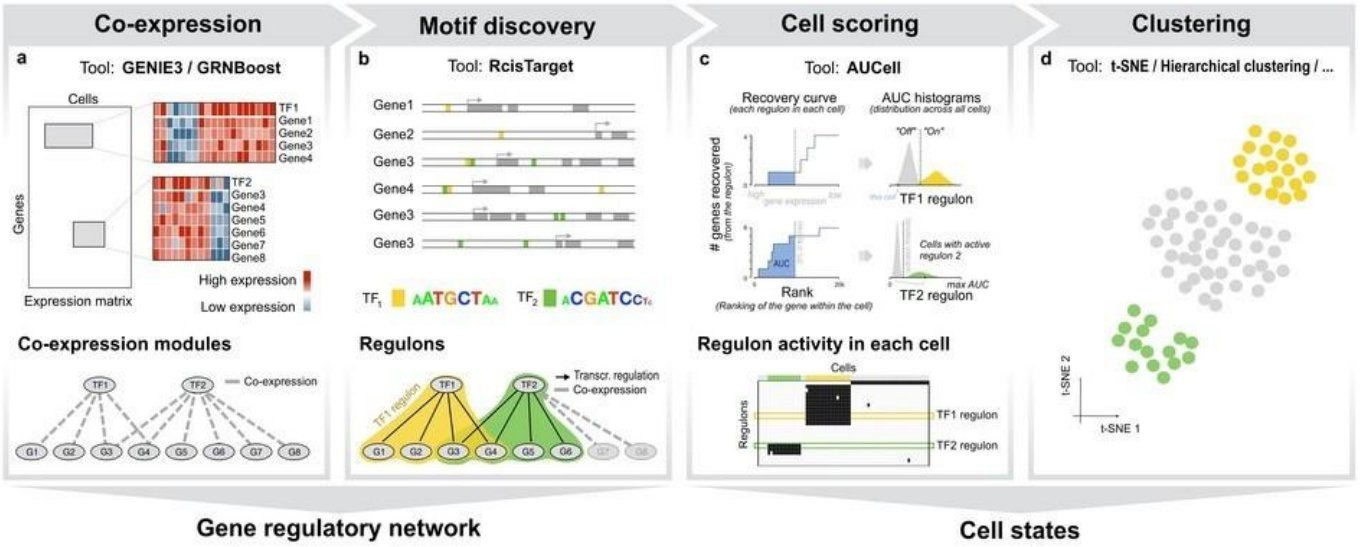
Inference



https://www.cs.purdue.edu/homes/ayg/TALKS/STC_CHICAGO10/Introductory_material/regulatory_networks.ppt

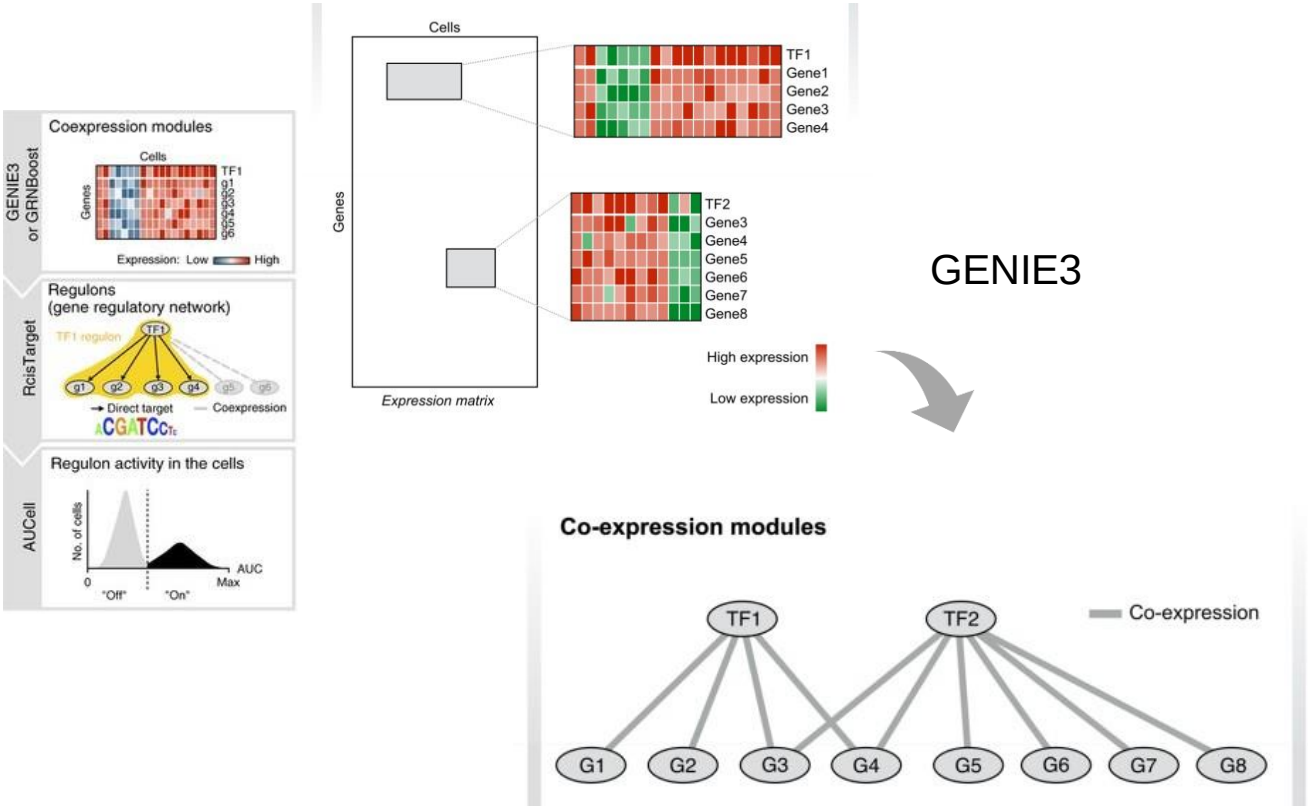
SCENIC: single-cell regulatory network inference and clustering

- SCENIC is a tool to simultaneously reconstruct gene regulatory networks and identify stable cell states from single-cell RNA-seq data. The gene regulatory network is inferred based on co-expression and DNA motif analysis, and then the network activity is analyzed in each cell to identify the recurrent cellular states.



Aibar et al. (2017) Nature Methods.

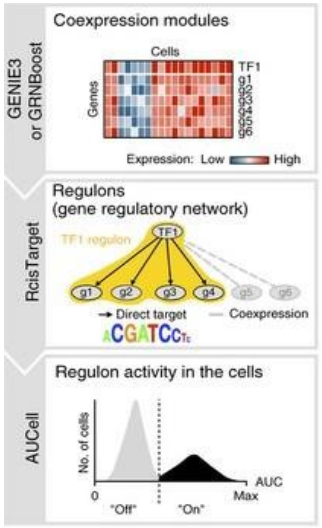
Step1. TF-based co-expression network

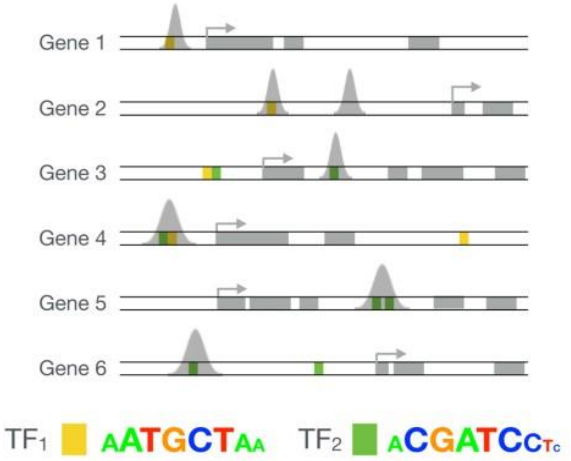


Aibar et al. (2017) Nature Methods.

Step2. Gene regulatory network

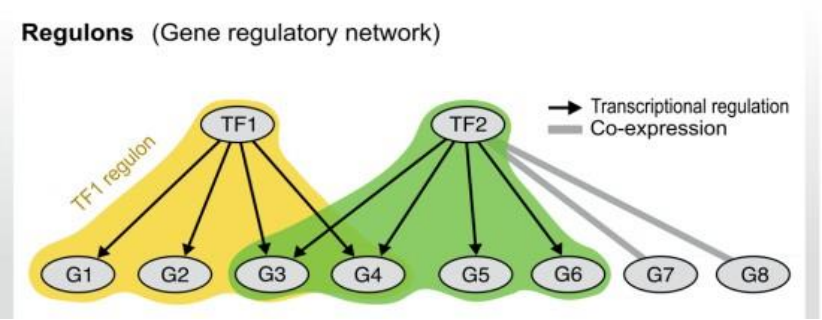
SCENIC





RcisTarget

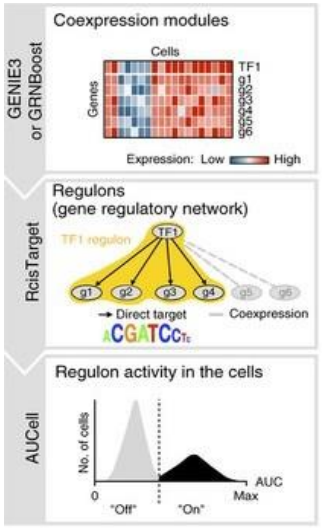
cis-regulatory sequence analysis

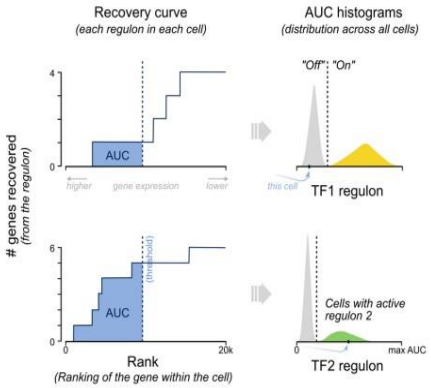


Aibar et al. (2017) Nature Methods.

Step3. Activity of the network in each cell

SCENIC

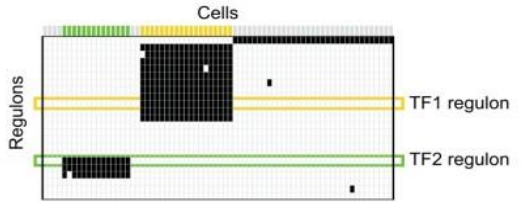




AUCell

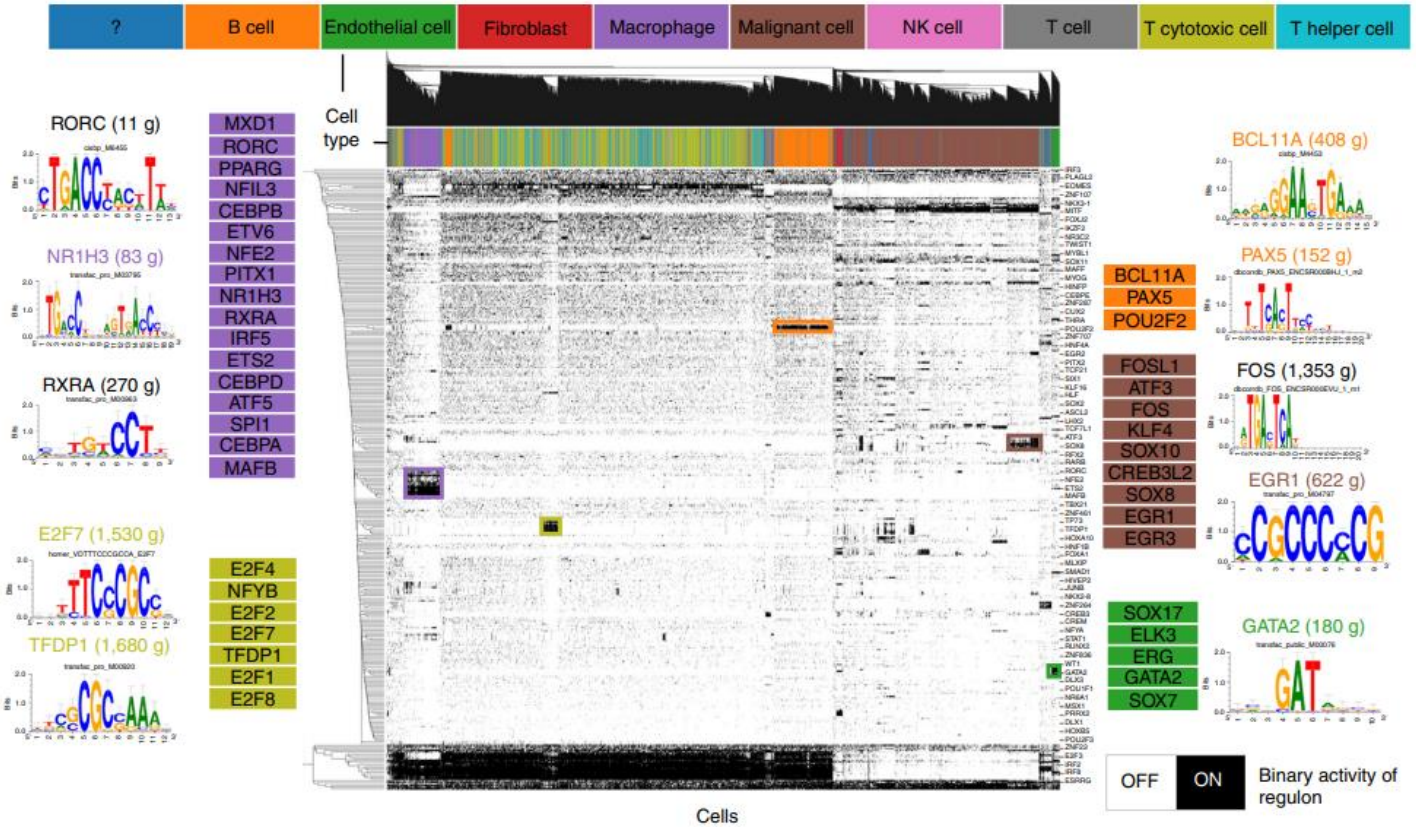
Identifying cells with active gene-sets

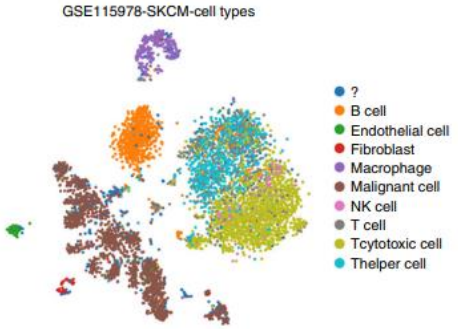
Regulon activity matrix (Network activity in each cell)



Aibar et al. (2017) Nature Methods.

SCENIC Results

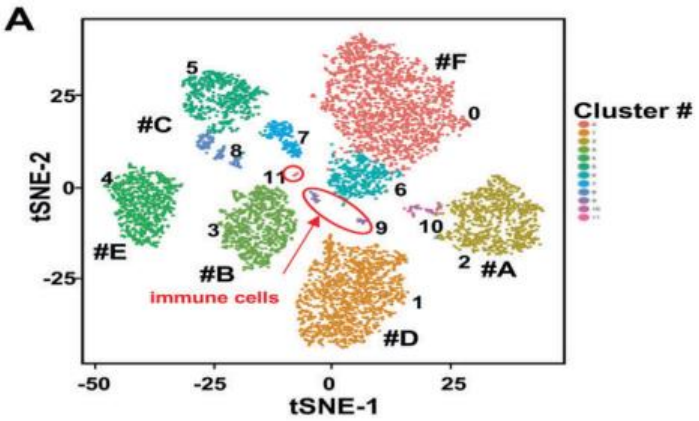


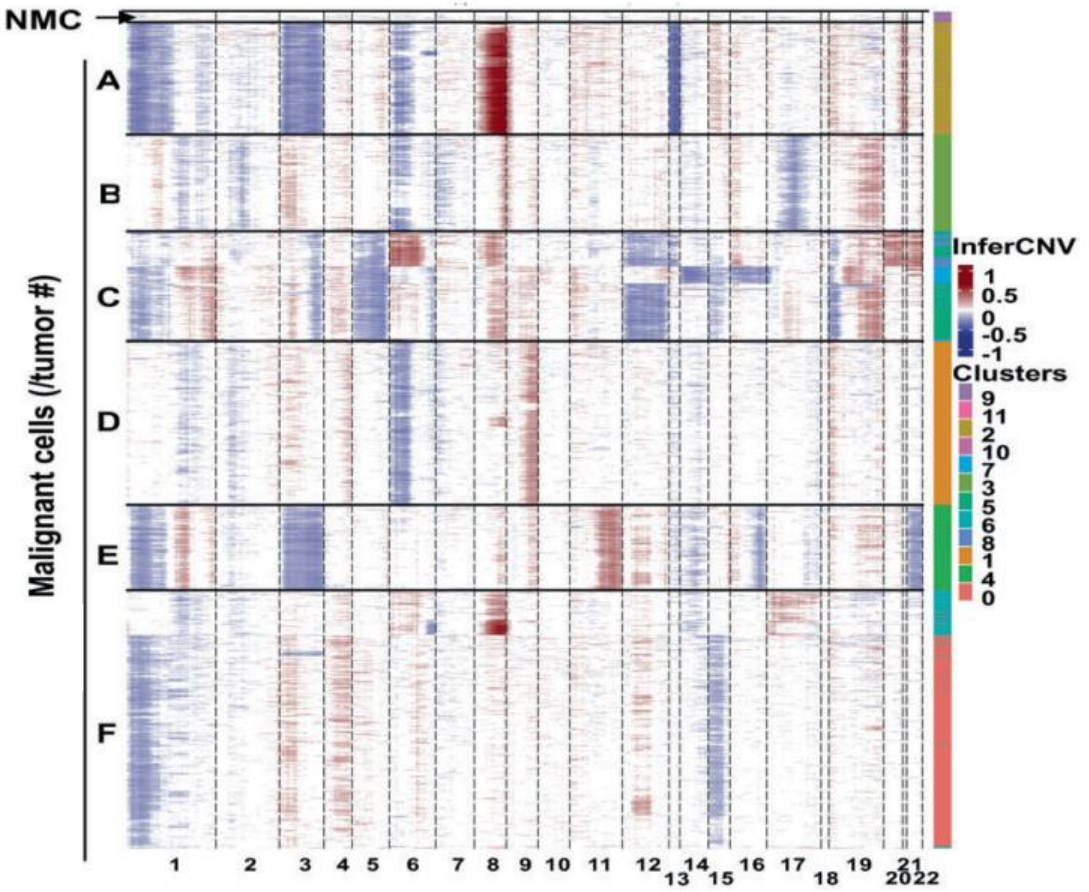


Van de Sande *et al.* (2020) Nature Protocol

InferCNV

- InferCNV is a tool used to explore tumor single cell RNA-Seq data to identify evidence for somatic large-scale chromosomal copy number alterations, such as gains or deletions of entire chromosomes or large segments of chromosomes.
- This is done by exploring expression intensity of genes across positions of tumor genome in comparison to a set of reference 'normal' cells.





Pandiani *et al.* (2021) Cell death and differentiation

References

- Yan Wu and Kun Zhang, Tools for the analysis of high- dimensional single- cell RNA sequencing data, Nat Rev Nephrology
- Huang Q et al, Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data, Genomics Proteomics Bioinformatics 19 (2021) 267
- Clarke Z, Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods, Nature Protocols 2021
- azimuth.hubmapconsortium.org
- http://bioconductor.org/books/release/SingleRBook/
- Aran D, et al, Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. 2019
-
- Cakir B**,** Comparison of visualization tools for single-cell RNAseq data, NAR Genomics and Bioinformatics, 2020
- Cobos FA, Benchmarking of cell type deconvolution pipelines for transcriptomics data, Nat Comm 2020

- Fan J et al, Single-cell transcriptomics in cancer: computational challenges and opportunities, Experimental & Molecular Medicine 2020

- Aibar S et al, SCENIC: single-cell regulatory network inference and clustering, Nat Method 2017
- https://scenic.aertslab.org/

- inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

Thank you