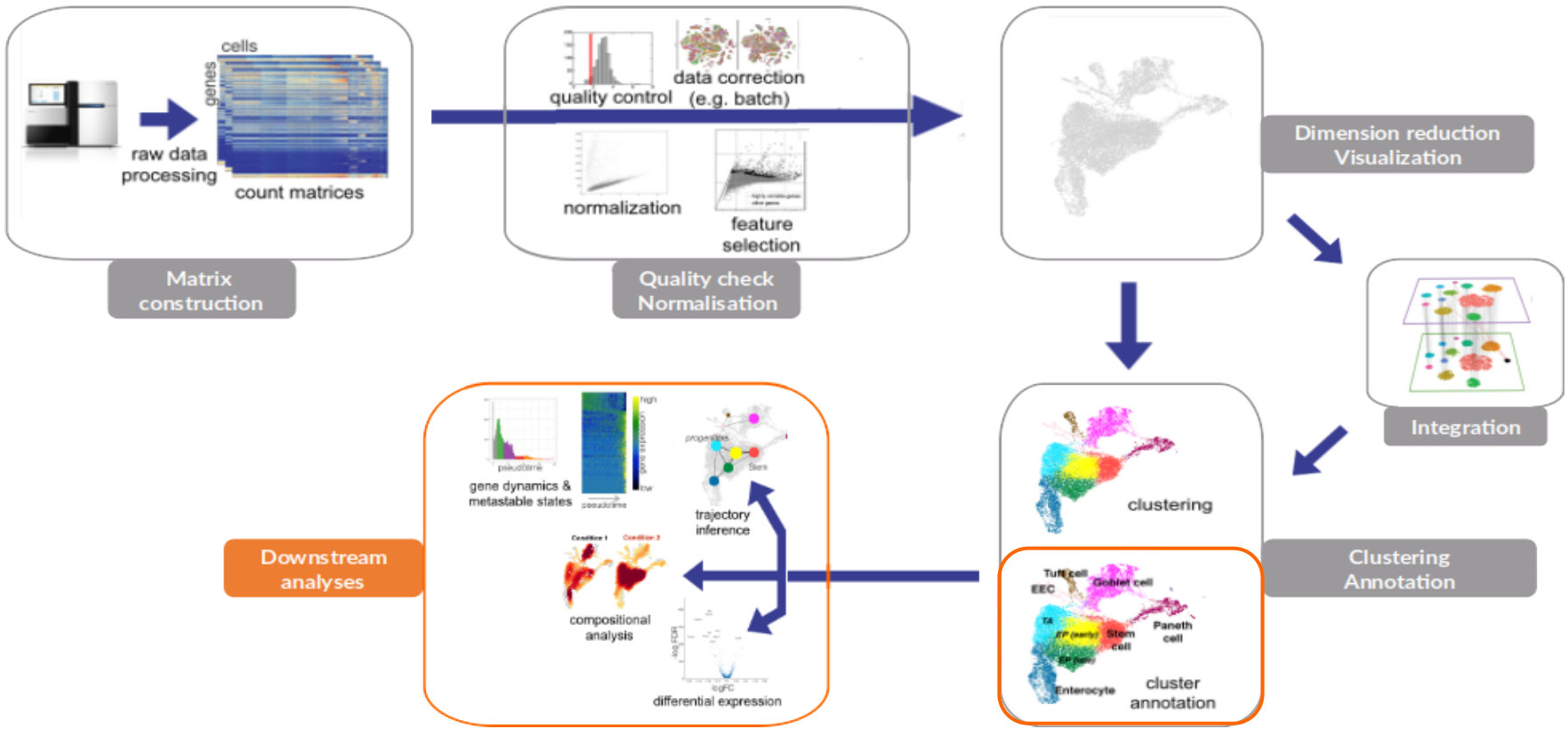
Secondary analyses

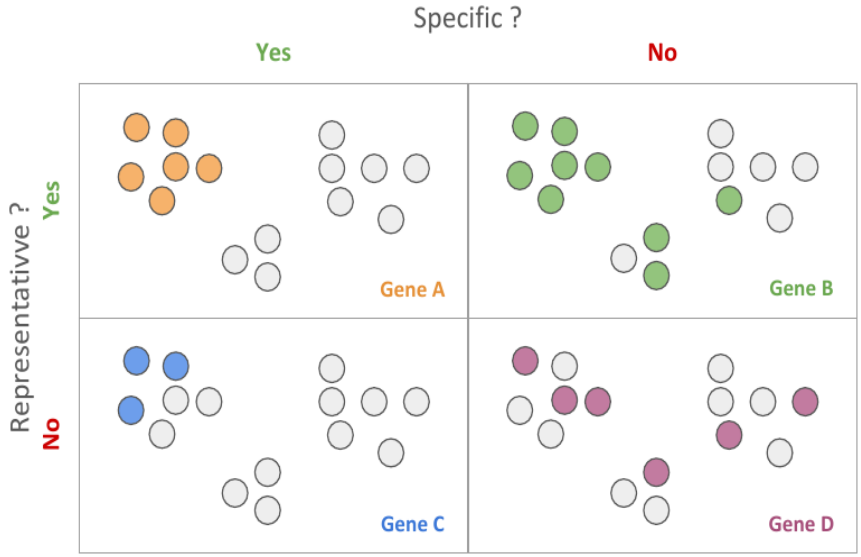
Introduction



Cell annotation

Cell annotation

Annotation relies on marker genes



Good markers are overexpressed and specific to a population

Cell annotation

Annotation strategies

Manual annotation

Using differentially expressed genes

Automatic annotation

marker-based

or

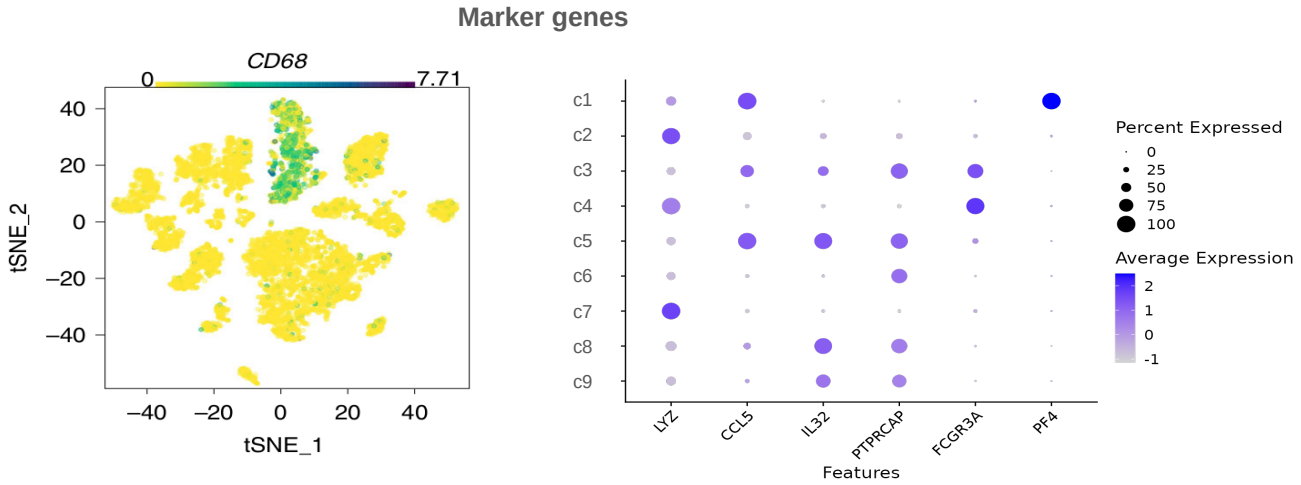
dataset-based

Cell annotation

Manual annotation
Automatic annotation

Manual annotation

- Manually review marker genes.



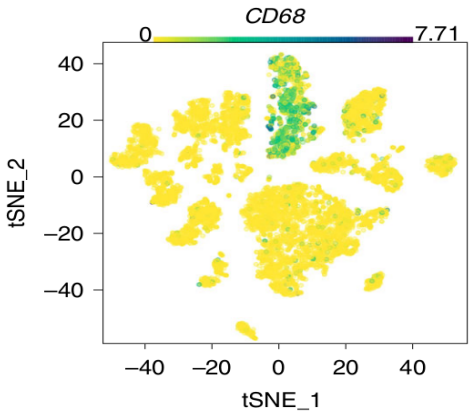
Cell annotation

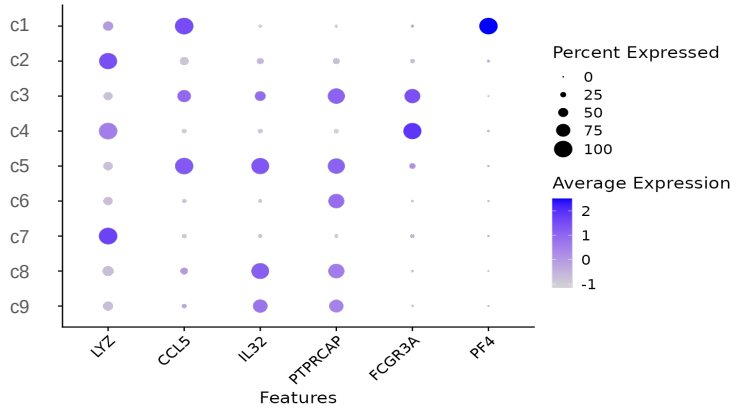
Manual annotation
Automatic annotation

Manual annotation

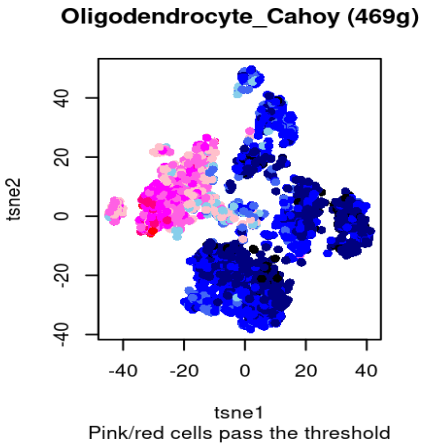
- Manually review marker genes.
- Score known signatures: AUCell, decouplR, gficf…

Marker genes





Signature



Cell annotation

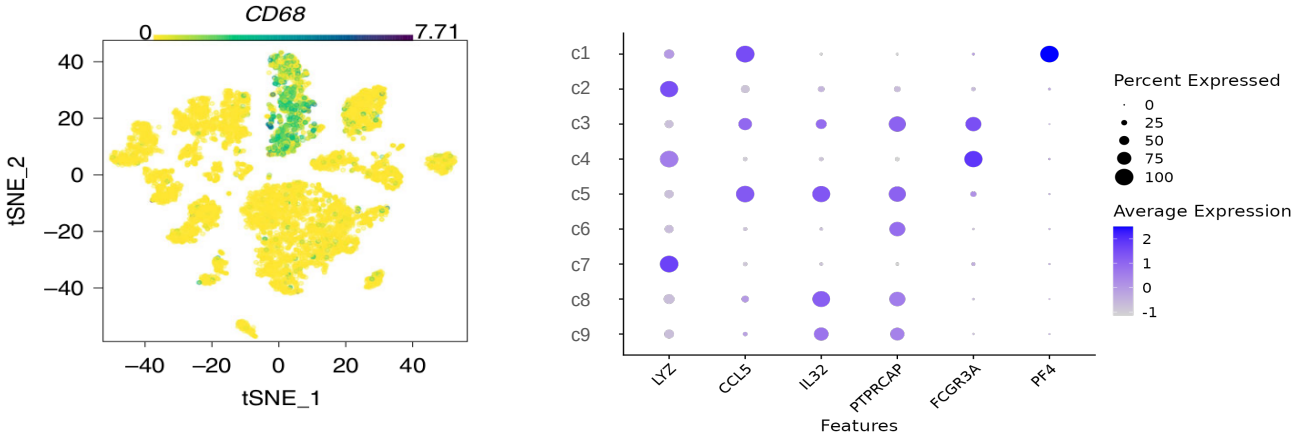
Manual annotation
Automatic annotation

Manual annotation

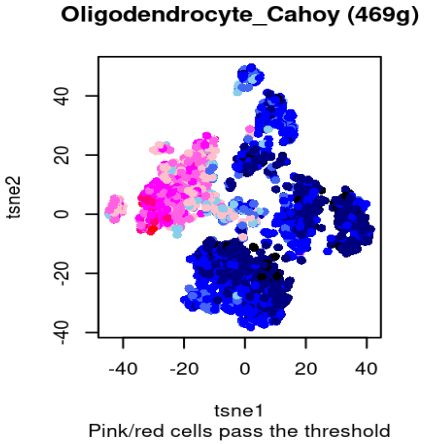
- Manually review marker genes.
- Score known signatures: AUCell, decouplR, gficf…

annotate clusters according to gene/signature expression

Marker genes



Signature



Cell annotation

Manual annotation
Automatic annotation

Manual annotation

- Easy to implement (no tool needed)
- You can define your own populations and subpopulations.

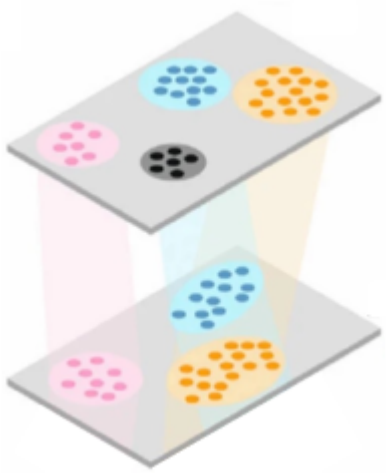
- Need *a priori* knowledge
- Time consuming
- Can be subjective
- Annotation generally at the cluster level = dependent on clustering
- “Why is my favorite gene not expressed?"

Cell annotation

Manual annotation
Automatic annotation

Automatic annotation

- Compare single cell data to a **reference** and deduce the cell labels.

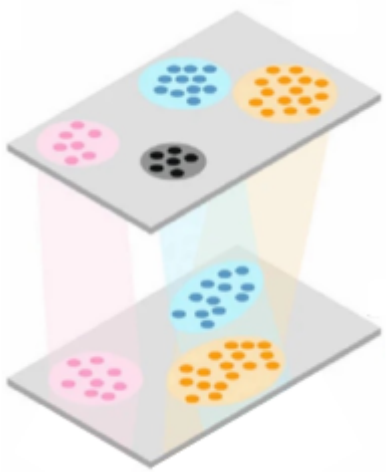


Cell annotation

Manual annotation
Automatic annotation

Automatic annotation

- Compare single cell data to a **reference** and deduce the cell labels.
- Need a tool: *singleR, CelliD, CHETAH, Azimuth (Seurat)…*

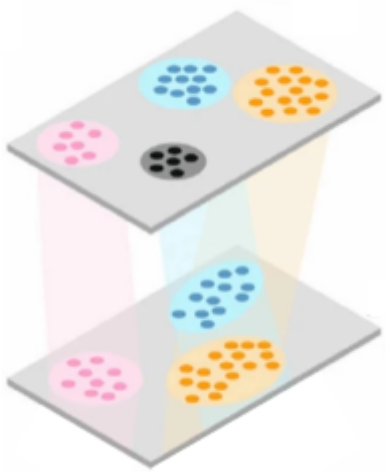


Cell annotation

Manual annotation
Automatic annotation

Automatic annotation

- Compare single cell data to a **reference** and deduce the cell labels.
- Need a tool: *singleR, CelliD, CHETAH, Azimuth (Seurat)…*
- Need a good reference. Depending on the tool:
  - Sets of marker genes
  - Whole single cell datasets (label transfer)
  - Bulk data of purified cell types



Cell annotation

Manual annotation
Automatic annotation

Automatic annotation

- Compare single cell data to a **reference** and deduce the cell labels.
- Need a tool: *singleR, CelliD, CHETAH, Azimuth (Seurat)…*
- Need a good reference. Depending on the tool:
  - Sets of marker genes
  - Whole single cell datasets (label transfer):
  - Bulk data of purified cell types
- Find references in databases, literature or use your own data…

Databases

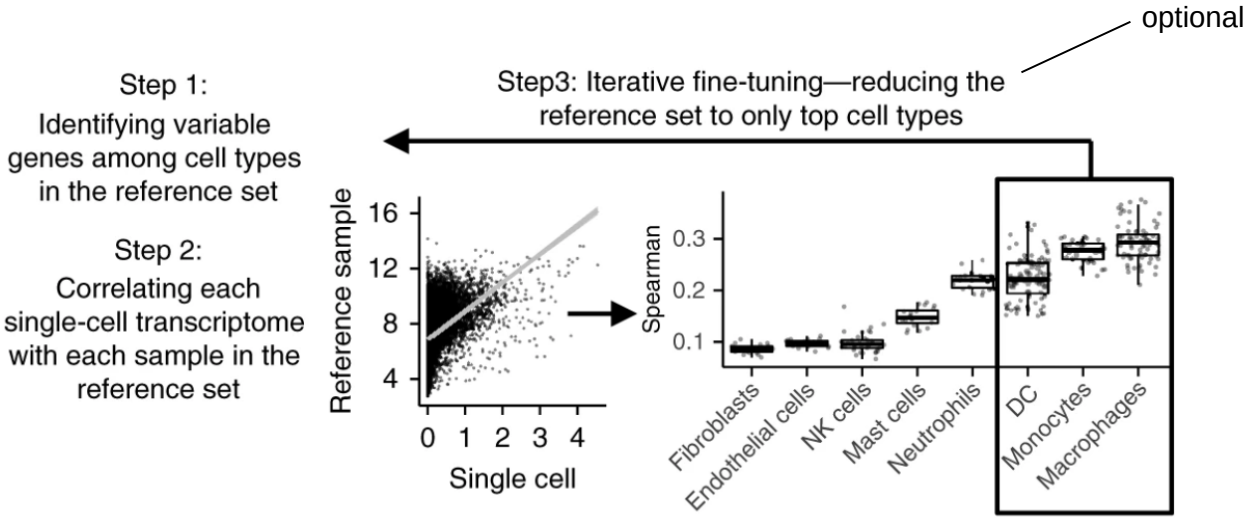
Marker genes	PanglaoDB, CellMarkers, ScSig
single cell datasets	Single Cell Expression Atlas, Tabula Muri, Immgen, Human Cancer Atlas, CancerSEA, Azimuth
bulk purified cell types	CellDex

Cell annotation

Manual annotation
Automatic annotation

Example: singleR

Principle: select reference label with the best gene expression correlation.



Aran *et al*. Nat. Immunol. 2019

Cell annotation

Manual annotation
Automatic annotation

Example: large language models

- Principle: the model identifies cell type based on list of top DEG
- R tool: GPT-Celltype for compatibility with Seurat objects
- Easy of use and allegedly better results than other automatic tools
- But a need for manual validation (AI hallucination) and GPT-4 fee

Hou & Ji, Nat. Methods 2024

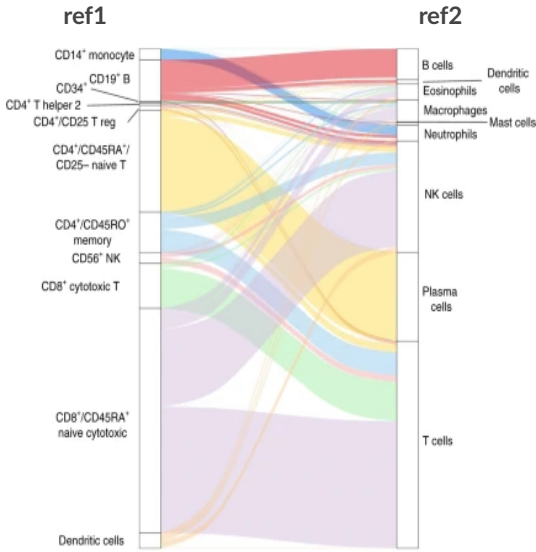
Cell annotation

Manual annotation
Automatic annotation

- Fast
- Annotation at the single cell level
- Define your own reference
- Quality scores often provided

Automatic annotation

- Need a good reference for your model. Not always easy to find.
- Incomplete, poorly matched reference data → poor results (conflicting, absent cell labels… error propagation !)
- Inconsistencies between references
- Similar cell types hard to distinguish
- Sample processing can have a huge impact on the results



Clarke *et al.*, Nat. Protoc. 2021

Cell annotation

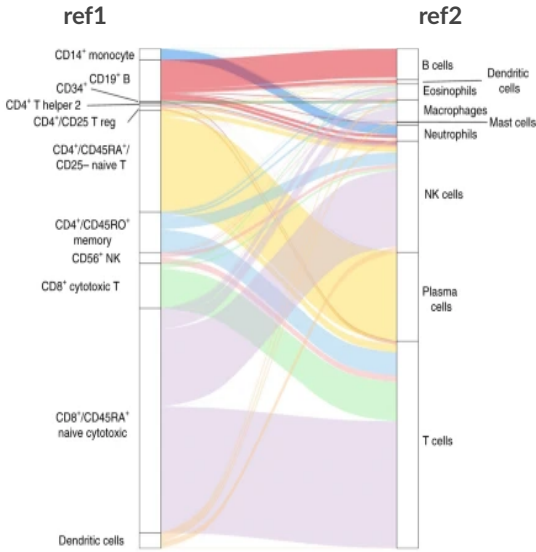
Manual annotation
Automatic annotation

- Fast
- Annotation a
- Define your own reference
- Quality scores often provided

Automatic annotation

- Need a good reference for your model. Not always easy to find.
- Incomplete, poorly matched reference
  s (conflicting, absent
  ropagation !)
- Inconsistencies between references
- Similar cell types hard to distinguish
- Sample processing can have a huge impact on the results

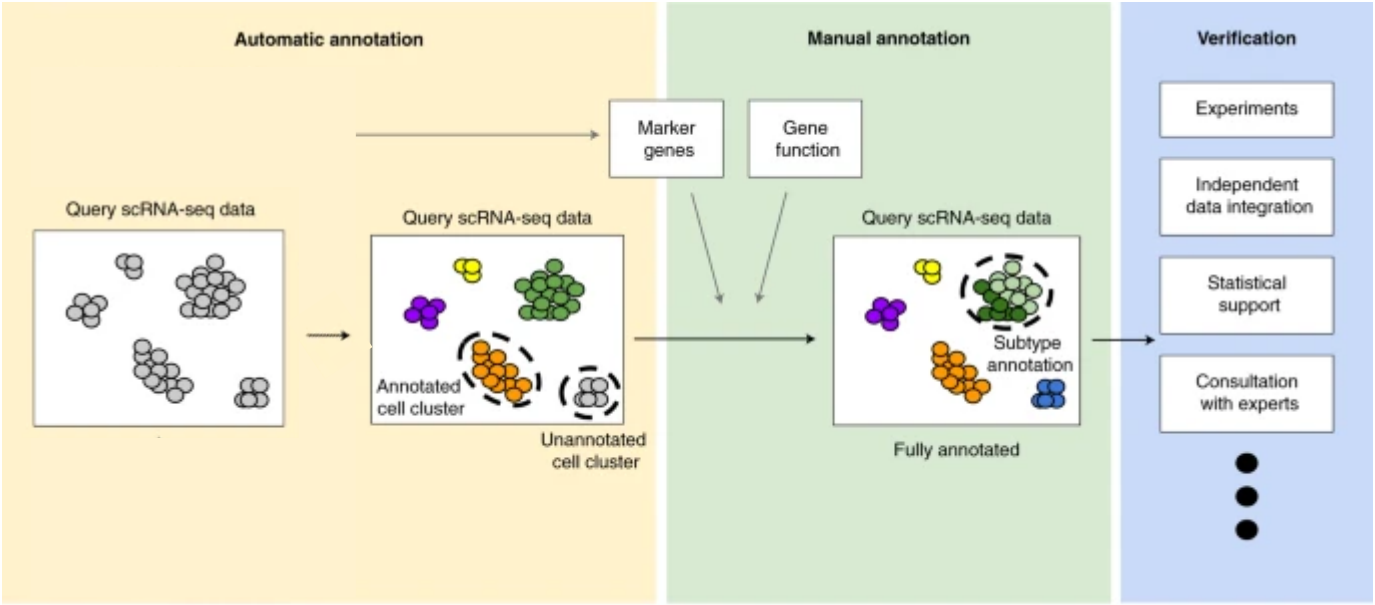
Check the results manually



Clarke *et al.*, Nat. Protoc. 2021

Cell annotation

Possible workflow



Clarke *et al.*, Nat. Protoc. 2021

Cell annotation

Some existing tools

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

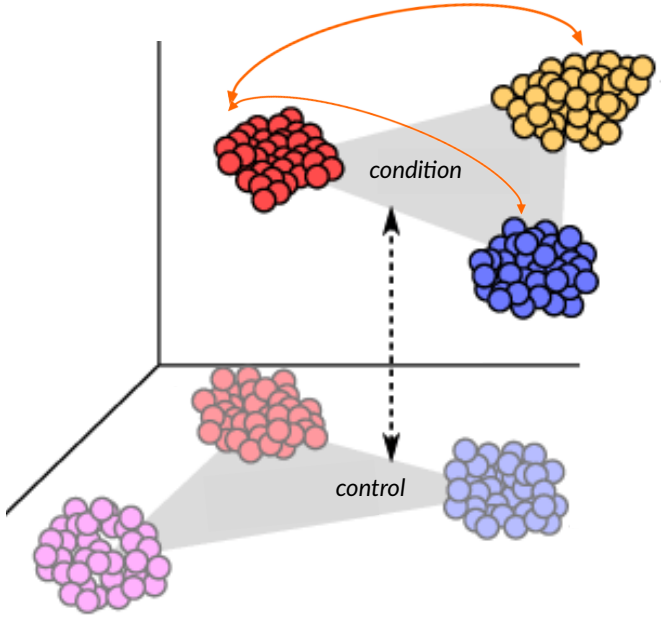
Clarke *et al.*, Nat. Protoc. 2021

Gene and function enrichment

Gene and function enrichment

What to measure

- Compare populations:
  - **1 population vs the rest: markers**

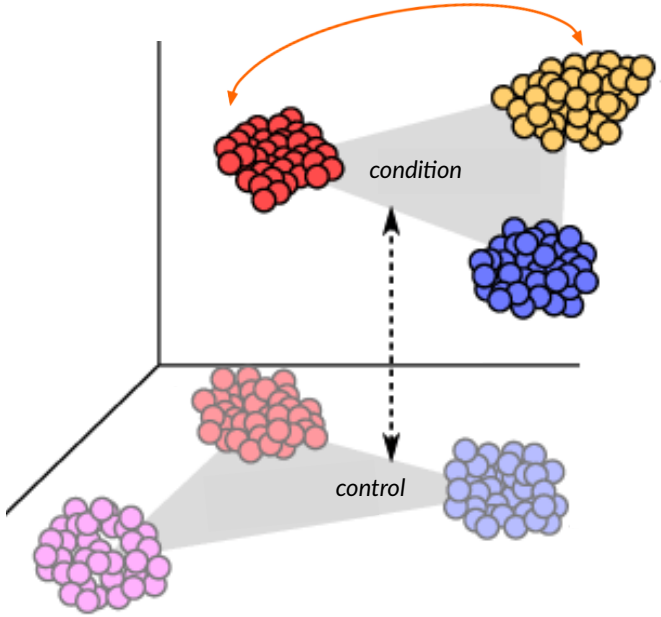


What are the specific markers of red population ?

Gene and function enrichment

What to measure

- Compare populations:
  - 1 population vs the rest: markers
  - **1 vs 1 populations**



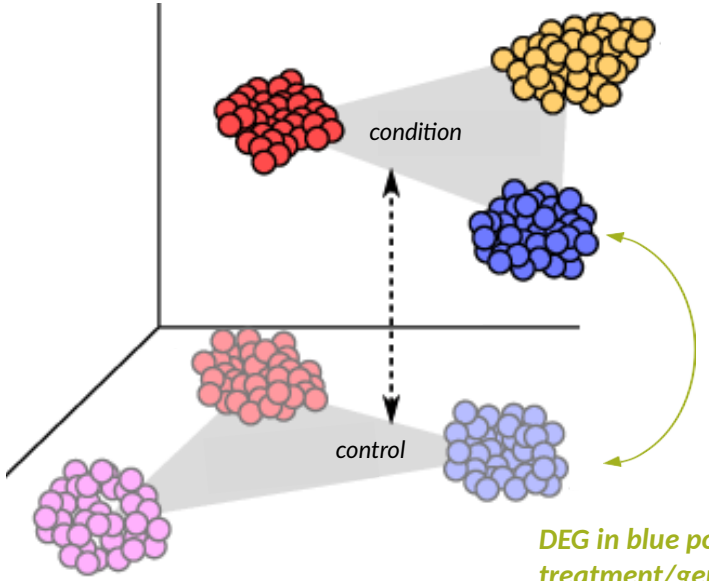
DEG between red and yellow populations ?

Gene and function enrichment

What to measure

- Compare populations:
  - 1 population vs the rest: markers
  - 1 vs 1 populations

- **Same population between different conditions**



DEG in blue population upon treatment/gene overexpression/KD... ?

Gene and function enrichment

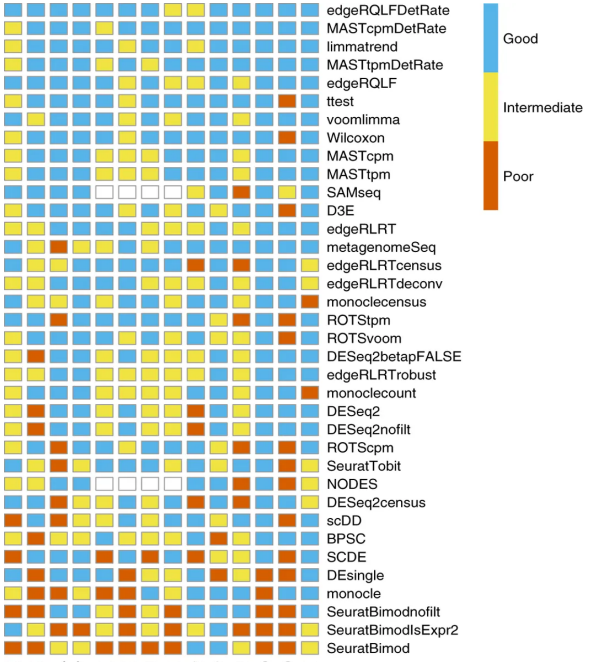
Many methods

- Many methods (over 36 in 2018)
- Relying on diverse statistical assumptions:

almost
no assumption

specific assumptions:
model of gene expression,
dropouts…

(Wilcoxon) (t-test) (GLM, bayesian inferences)

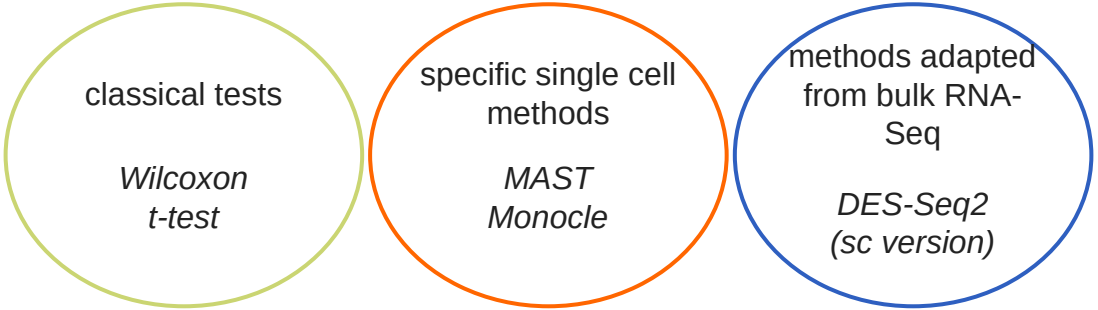


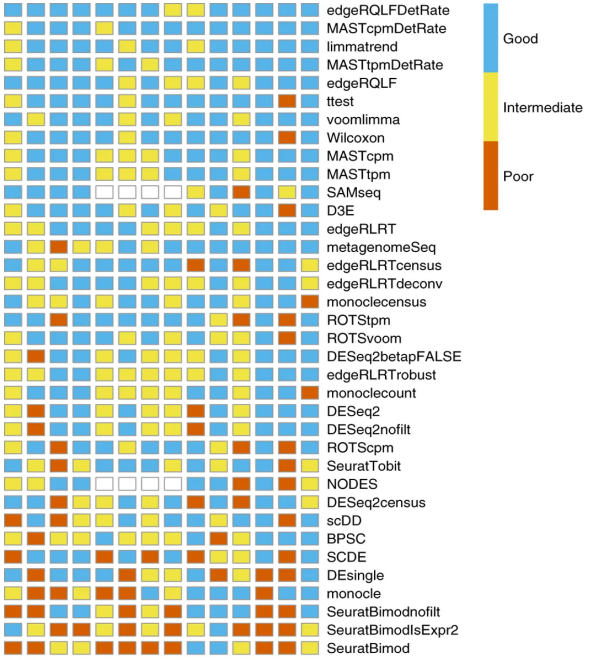
Soneson & Robinson, Nat. Methods 2018

Gene and function enrichment

Many methods

- Many methods (over 36 in 2018)
- Relying on diverse statistical assumptions
- Different origins:





Soneson & Robinson, Nat. Methods 2018

Gene and function enrichment

What about replicates

The nature of single cell data raises questions regarding replicates

- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?

Gene and function enrichment

- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?

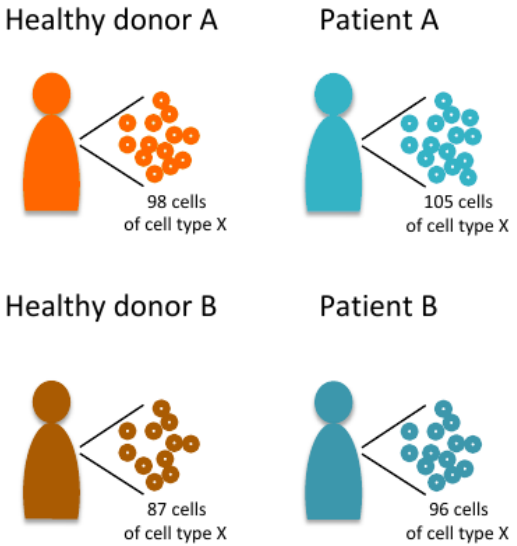
What about replicates

The nature of single cell data raises questions regarding replicates

1 replicate = 1 cell (dozens of replicate)

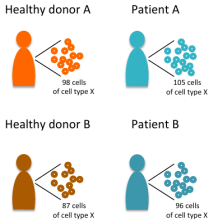
or

1 replicate = 1 sample (2 replicate/condition)



Gene and function enrichment

- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?

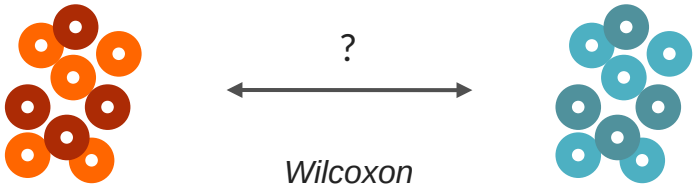


What about replicates

Based on the definition of a replicate,
methods can be classified into 2 main approaches

"single-cell methods"
(1 cell = 1 replicate)

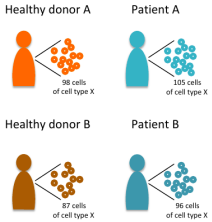
do **not** take samples into account



Wilcoxon
t-test
MAST
Monocle

Gene and function enrichment

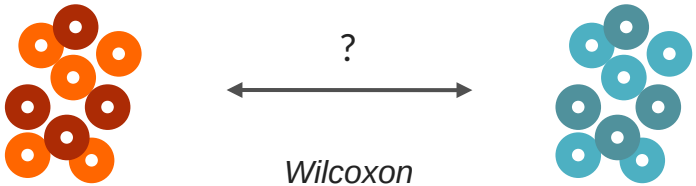
- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?



What about replicates

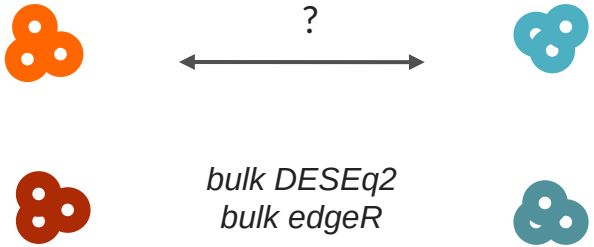
Based on the definition of a replicate,
methods can be classified into 2 main approaches

"single-cell methods"
(1 cell = 1 replicate)
do **not** take samples into account



Wilcoxon
t-test
MAST
Monocle

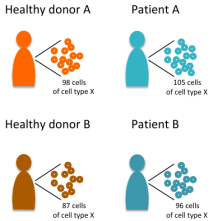
pseudobulk methods
(1 sample = 1 biological replicate)
sum cells within same replicate then bulk method



bulk DESEq2
bulk edgeR

Gene and function enrichment

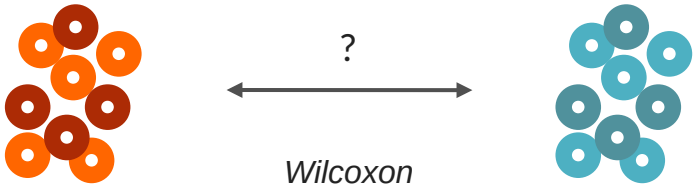
- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?



What about replicates

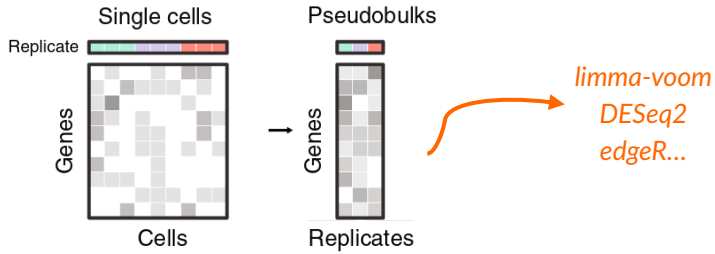
Based on the definition of a replicate,
methods can be classified into 2 main approaches

"single-cell methods"
(1 cell = 1 replicate)
do **not** take samples into account



Wilcoxon
t-test
MAST
Monocle

pseudobulk methods
(1 sample = 1 biological replicate)
sum cells within same replicate then bulk method

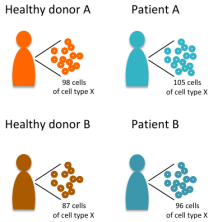


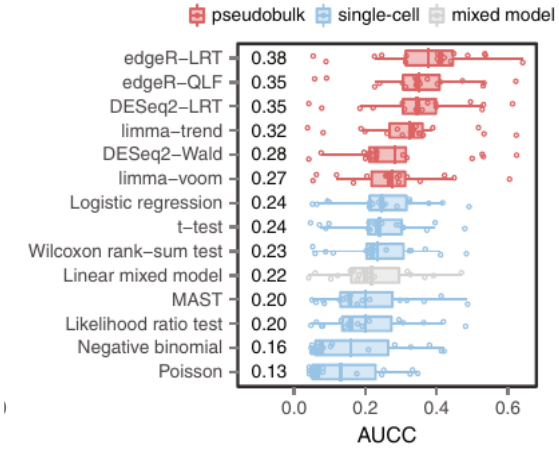
Squair *et al.*, Nat.Com 2021

Gene and function enrichment

- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?

What about replicates





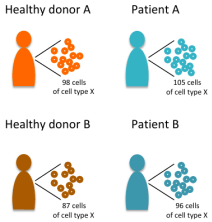
- According to Squair *et al.*, biological replicates allow:
  - less false discoveries
  - less biased towards highly expressed genes

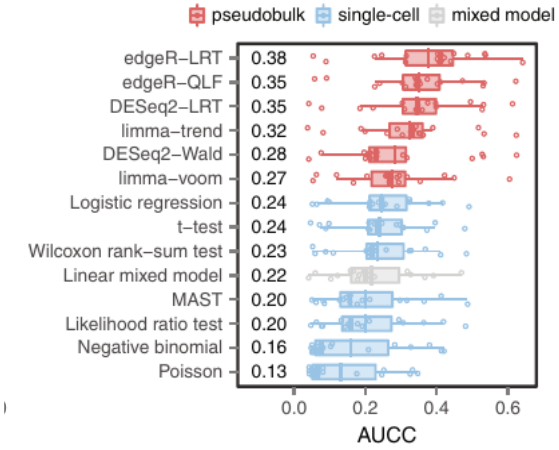
Squair *et al.*, Nat.Com 2021

Gene and function enrichment

- What is a replicate ?
- How to take replicates into account ?
- Should we use replicates ?

What about replicates





- According to Squair *et al.*, biological replicates allow:
  - less false discoveries
  - less biased towards highly expressed genes

- Which method to use: no consensus but:
  - pseudo-bulk methods take advantage of biological replicates
  - naive approaches would be the best second approach (even when lacking sequencing depth)

Squair *et al.*, Nat.Com 2021

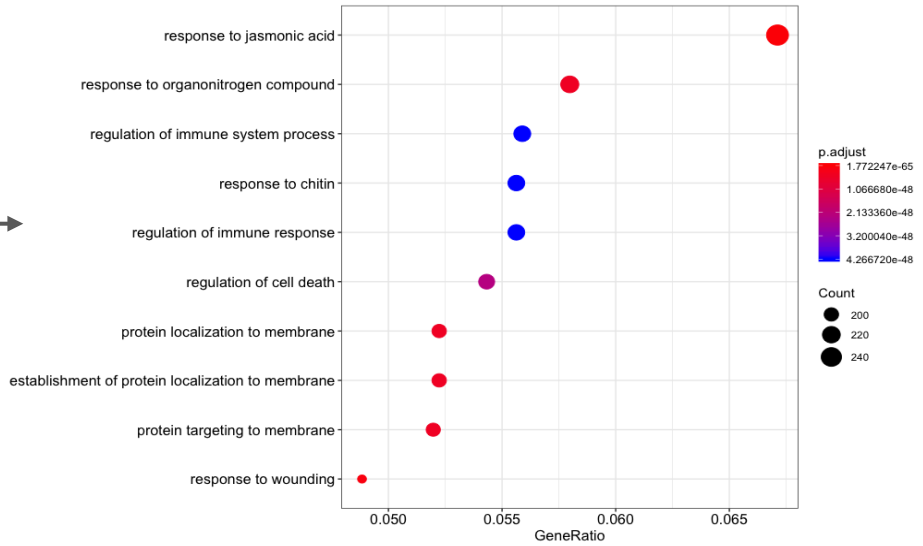
Gene and function enrichment

Functional interpretation

- Extract biologically meaningful insights from a long, hard to interpret list of genes.

“Are my DEGs more involved in cell proliferation ? Migration ? Do they belong to known pathways ?”

Gene	Fold change		
	H_2O_2	mcd1-1	pds5-1
PSO2	14.93	6.96	9.19
CST9	13.93	4.00	14.93
EAF7	13.00	2.46	9.85
MIG3	12.13	4.29	14.93
EPL1	**12.13**	**3.73**	**8.57**
NHP6A	**12.13**	**3.73**	**8.57**
SCC4	12.13	2.30	9.19
SLX1	11.31	2.83	10.56
VID21	10.56	4.59	13.00
YKU80	**10.56**	**3.25**	**9.85**
RAD59	9.19	6.50	9.19
HPR5	9.19	2.83	9.19
SIR4	9.19	2.83	7.46
EAF6	9.19	2.46	7.46
TFB3	9.19	2.14	9.19
NTG1	**8.57**	**3.48**	**7.46**
MGT1	8.57	2.64	9.19
SNF5	8.57	2.64	9.19
HMI1	8.57	2.46	6.96
RTT107	8.57	2.00	7.46
EXO1	8.00	2.64	8.00
ELC1	8.00	2.14	11.31
HEX3	8.00	2.14	6.96
DOA1	8.00	2.00	8.57
MMS4	8.00	2.00	8.00
RAD16	7.46	5.28	8.00
MEC3	7.46	2.83	6.96
PIN4	7.46	2.46	9.19
CAC2	7.46	2.00	8.57
LCD1	6.96	3.48	7.46
PAN2	6.96	2.30	12.13



Gene and function enrichment

Functional interpretation

- Rely on annotated sets of genes:

Gene and function enrichment

Functional interpretation

- Rely on annotated sets of genes:

- Databases:
  - Gene Ontology (GO): controlled, hierarchical vocabulary with fixed terms.
  - KEGG, Reactome, WikiPathway: list of pathways and high-level functions
  - MSigDB: Multiple collections of genes sets (human centered)

- Can be accessed online or with R packages (*biomaRt*, the *OrgDb* packages)

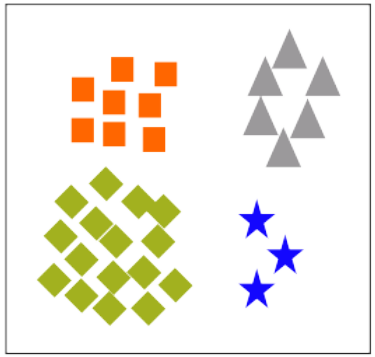
Gene and function enrichment

Functional interpretation

- **Classical methods**: over-representation analysis, GSEA… Resolution: cluster or cell type

Over Representation Analysis (ORA)

For a given gene set, are there more DE genes than what we expect by chance ?



All known genes
categorized into gene sets

Expected by chance

Observed DEG
orange genes are over-represented

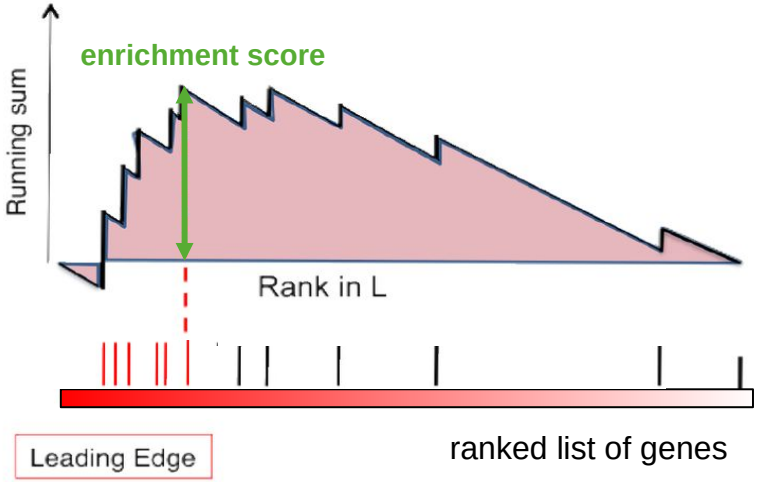
Gene and function enrichment

Functional interpretation

- **Classical methods**: over-representation analysis, GSEA… Resolution: cluster or cell type

Gene set enrichment analysis (GSEA)

- Principle:
  - rank the list of all genes: logFC, (logFC x p-value)
  - assess whether a gene set is enriched at the top or bottom of the list: enrichment score (ES), p-value
  - select high ES = low p-values



Gene and function enrichment

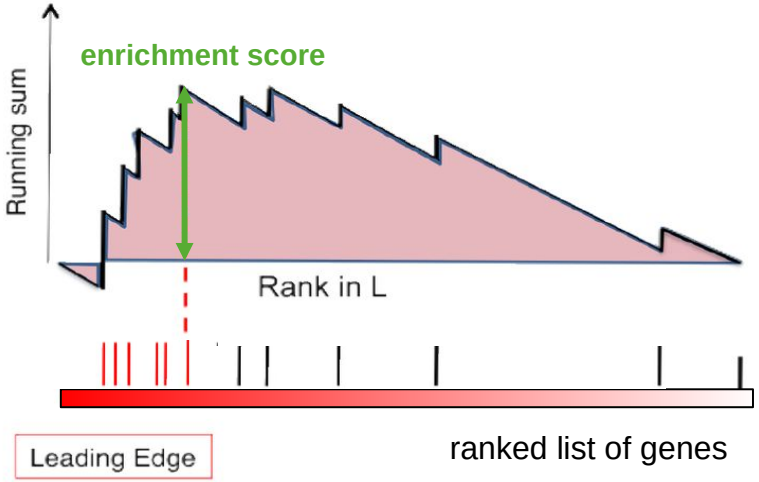
Functional interpretation

- **Classical methods**: over-representation analysis, GSEA… Resolution: cluster or cell type

Gene set enrichment analysis (GSEA)

- Principle:
  - rank the list of all genes: logFC, (logFC x p-value)
  - assess whether a gene set is enriched at the top or bottom of the list: enrichment score (ES), p-value
  - select high ES = low p-values

- GSEA focuses on coordinated differences in expression. Even not significant DEGs contribute.



Gene and function enrichment

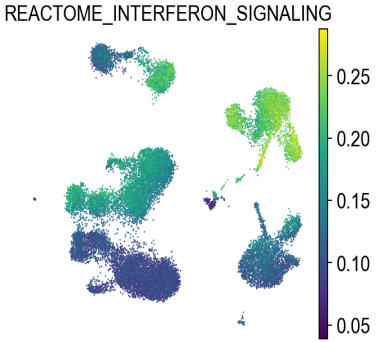
Functional interpretation

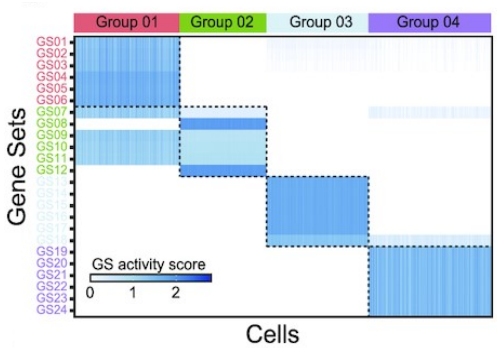
- **Classical methods**: over-representation analysis, GSEA… Resolution: cluster or cell type
- **Single cell specific methods**:
  - individual score for each cell
  - Some tools: CellID, AUCell, scGSVA, decouplR, scGSEA (in R package gficf)
  - Better than simple average or z-score because take the size of the gene set into consideration

Gene and function enrichment

Functional interpretation

- **Classical methods**: over-representation analysis, GSEA… Resolution: cluster or cell type
- **Single cell specific methods**:
  - individual score for each cell
  - Some tools: CellID, AUCell, scGSVA, decouplR, scGSEA (in R package gficf)
  - Better than simple average or z-score because the size of the gene set taken into consideration



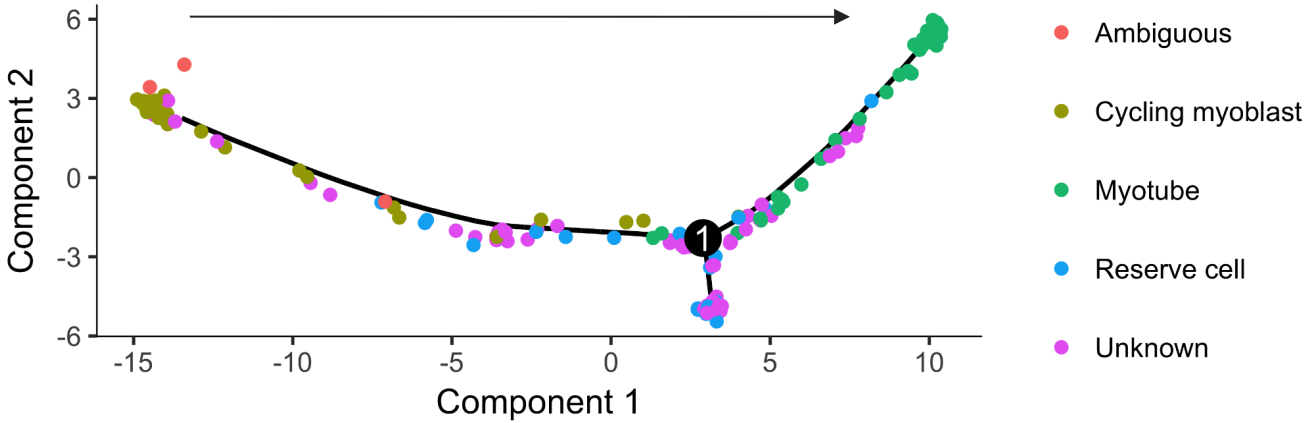


Trajectory analysis

Trajectory analysis

Dynamic processes

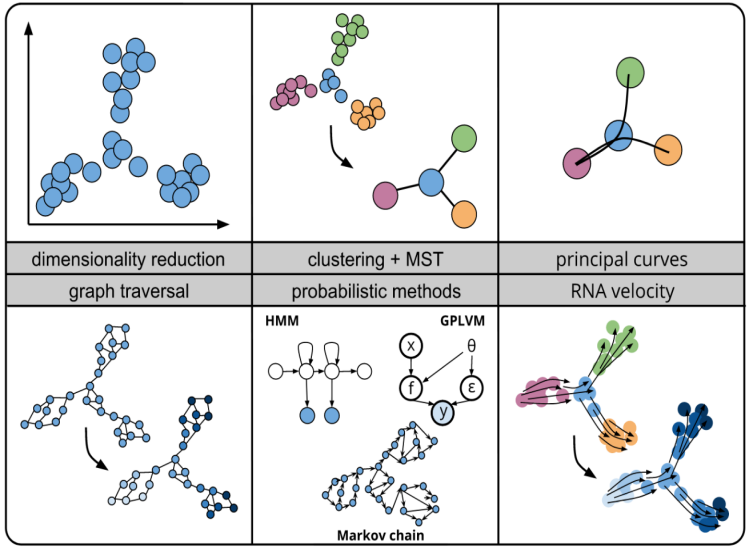
- Study dynamic processes: development, differentiation, immune response…
- Single cell sample heterogeneity: *different cell states present in the same sample* ⇒ order them.



Trajectory analysis

Dynamic processes

- Principle:
  - work in reduced dimension
  - infer trajectories in pseudo-time
  - Many approaches (probabilistic, cluster-based, graph-based)

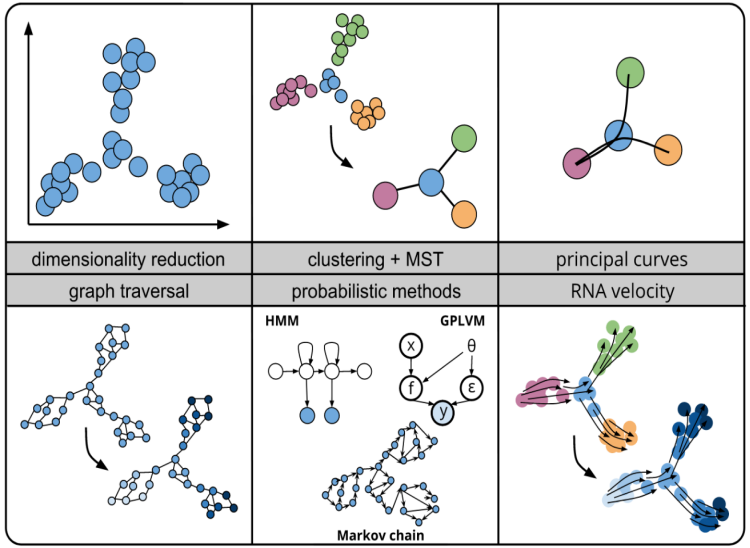


Deconinck *et al.*, Cur. Opin. Syst. Biol 2021

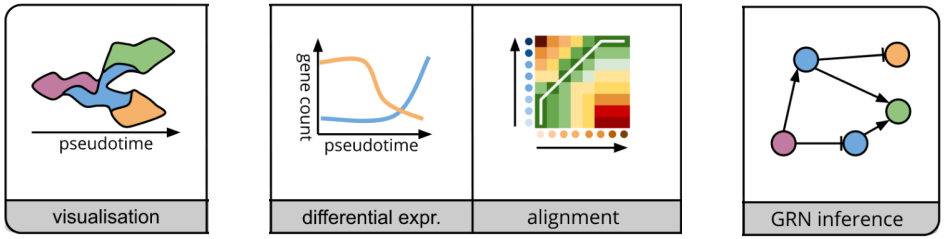
Trajectory analysis

Dynamic processes

- Principle:
  - work in reduced dimension
  - infer trajectories in pseudo-time
  - Many approaches (probabilistic, cluster-based, graph-based)



- Downstream analyses:

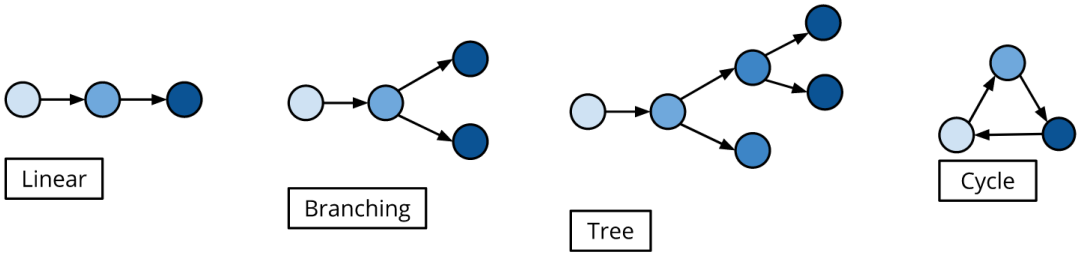


Deconinck *et al.*, Cur. Opin. Syst. Biol 2021

Trajectory analysis

Limits

- Topology:
  - A biological process can be

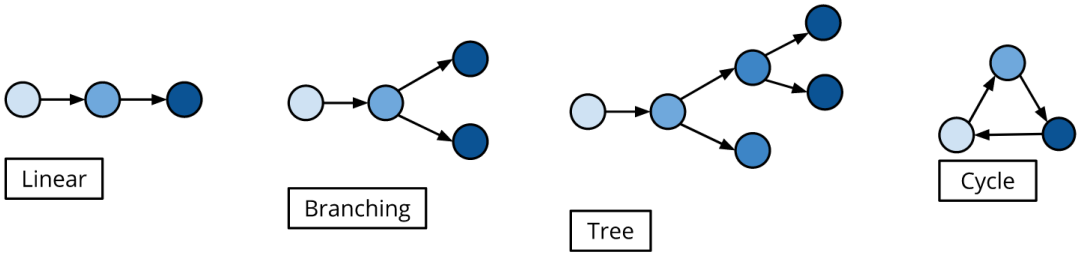


Deconinck *et al.*, Cur. Opin. Syst. Biol 2021

Trajectory analysis

Limits

- Topology:
  - A biological process can be



  - All methods cannot not correctly infer all topologies

    ⇒ Need *a priori* topology knowledge

    ⇒ Then choose a corresponding method

    ⇒ If no prior knowledge, compare several inferences

Deconinck *et al.*, Cur. Opin. Syst. Biol 2021

Trajectory analysis

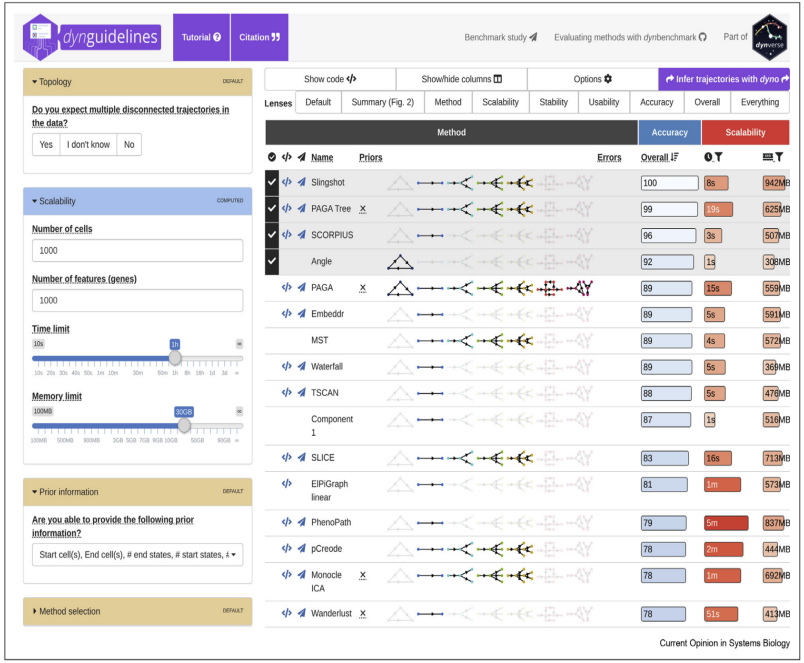
Limits

- Topology:
  - A biological process can be
  - All methods cannot not correctly infer all topologies

    ⇒ Need *a priori* topology knowledge

    ⇒ Then choose a corresponding method

    ⇒ If no prior knowledge, compare several inferences
  - Package **dynverse** implements most methods and helps choosing



Deconinck *et al.*, Cur. Opin. Syst. Biol 2021

Trajectory analysis

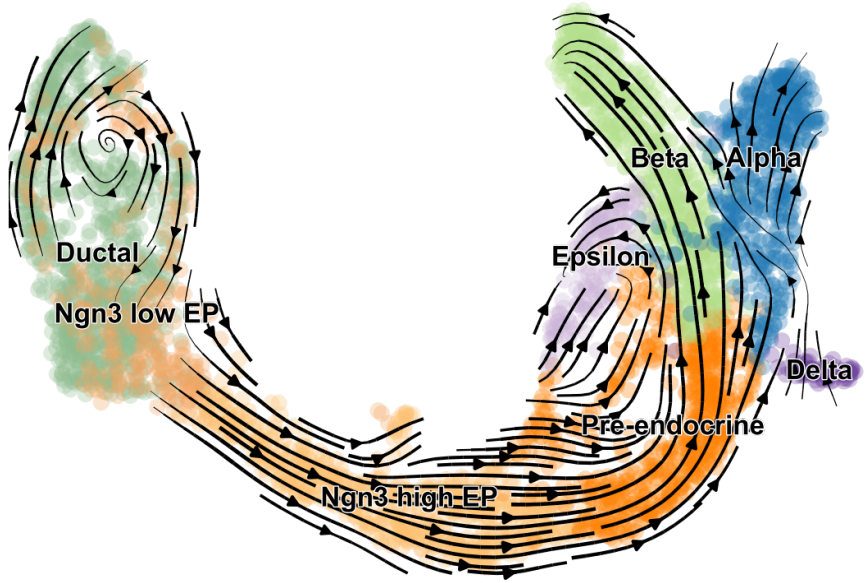
Limits

- Topology:

- Only based on gene expression similarity ⇒ add information:
  - RNA maturation: RNA velocity: scVelo, CellRank
  - other modalities
  - time points

Trajectory analysis

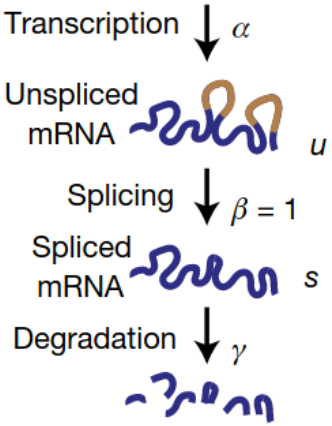
RNA velocity

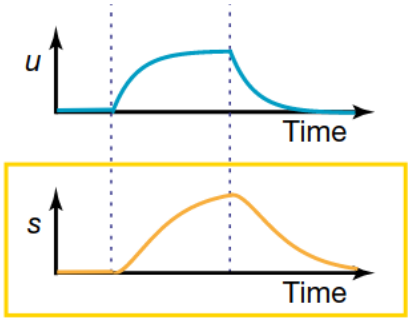


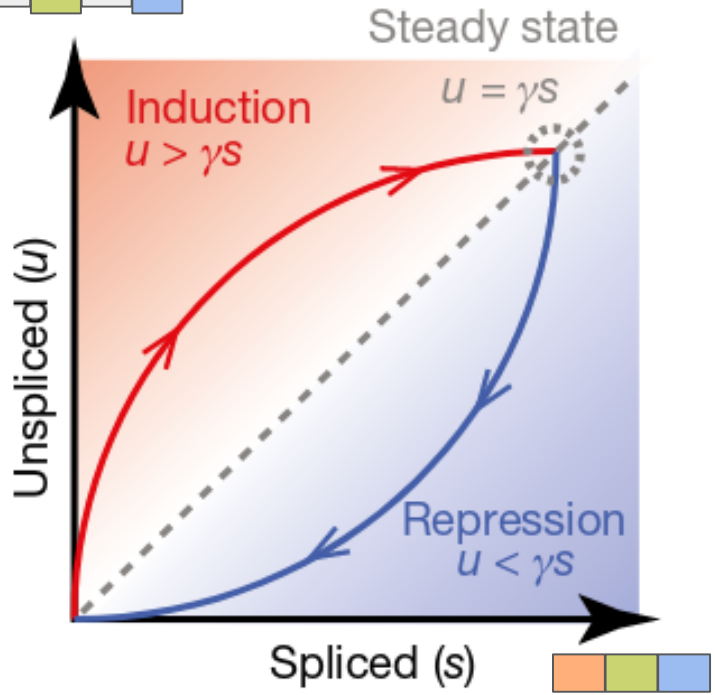
Trajectory analysis

RNA velocity

- Levels of immature and mature RNA change upon activation/repression.





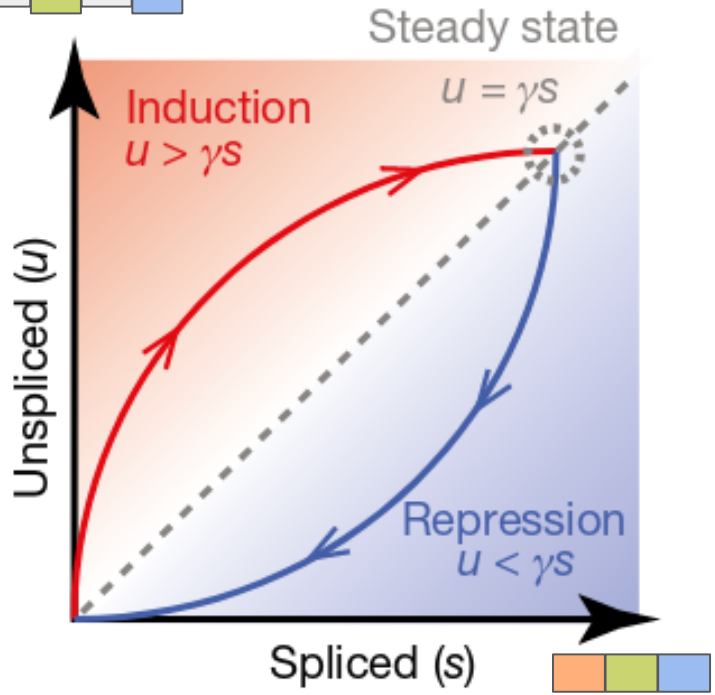


La Manno *et al.*, Nature 2018

Trajectory analysis

RNA velocity

- Levels of immature and mature RNA change upon activation/repression.
- Modelized as differential equations



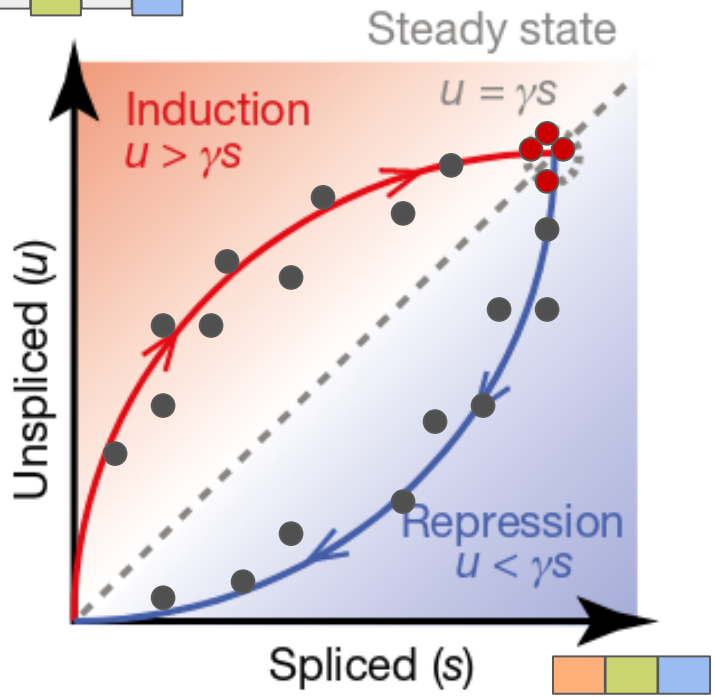
$$\frac{ds}{dt} = u - \gamma s \qquad \frac{du}{dt} = \alpha - \beta u$$

La Manno *et al.*, Nature 2018

Trajectory analysis

RNA velocity

- Levels of immature and mature RNA change upon activation/repression.
- Modelized as differential equations
- Single cells used to fit model



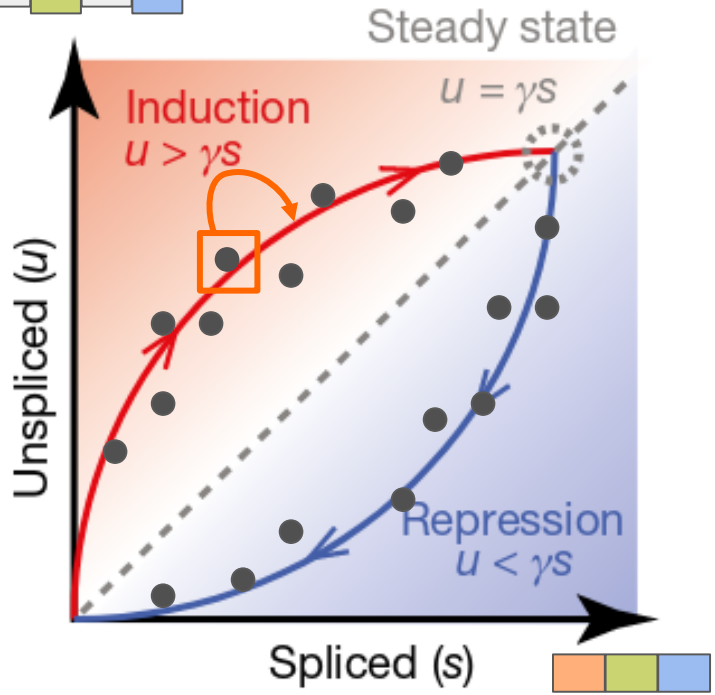
$$\frac{ds}{dt} = u - \gamma s \qquad \frac{du}{dt} = \alpha - \beta u$$

La Manno *et al.*, Nature 2018

Trajectory analysis

RNA velocity

- Levels of immature and mature RNA change upon activation/repression.
- Modelized as differential equations
- Single cells used to fit model
- For each cell, predict genes expression at t+1
  ⇒ vector of gene expression at t+1
  (i. e. next cell state)

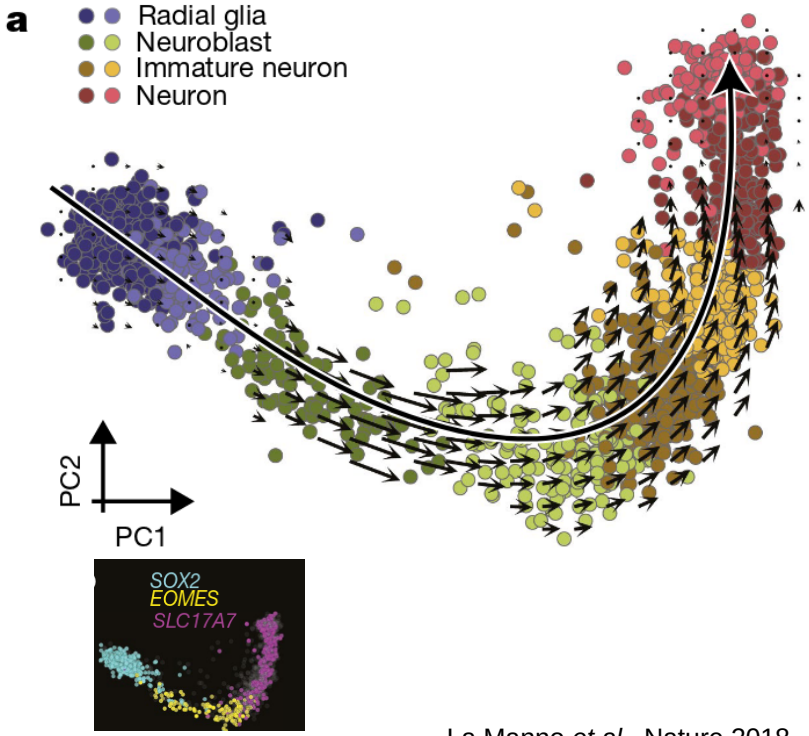


La Manno *et al.*, Nature 2018

Trajectory analysis

RNA velocity

- Levels of immature and mature RNA change upon activation/repression.
- Modelized as differential equations
- Single cells used to fit model
- For each cell, predict genes expression at t+1
  ⇒ vector of gene expression at t+1
  (i. e. next cell state)



La Manno *et al*., Nature 2018

Trajectory analysis

Summary

- Reconstitution of dynamic processes
- Insights about transient states
- Additional information can improve results (RNA velocity)

- *a prior* topology knowledge needed or risk of invalid results (topology)

  → not completely data driven

  → not completely exploratory

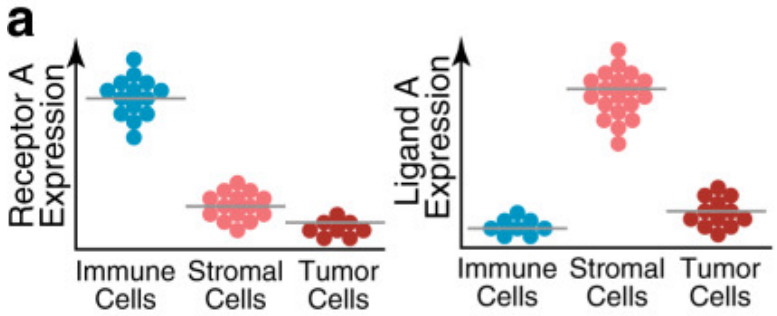
- Developed to study processes in the range of a few hours. Can it be extended to longer processes ?

Overview of advanced analyses

Overview of advanced analyses

Cell Cell Communication

- Evaluate receptor - ligand expression across cell types
- Guided by databases of know receptor-ligand interactions



PhoneDB

Fan *et al.*, Exp. Mol. Med 2020

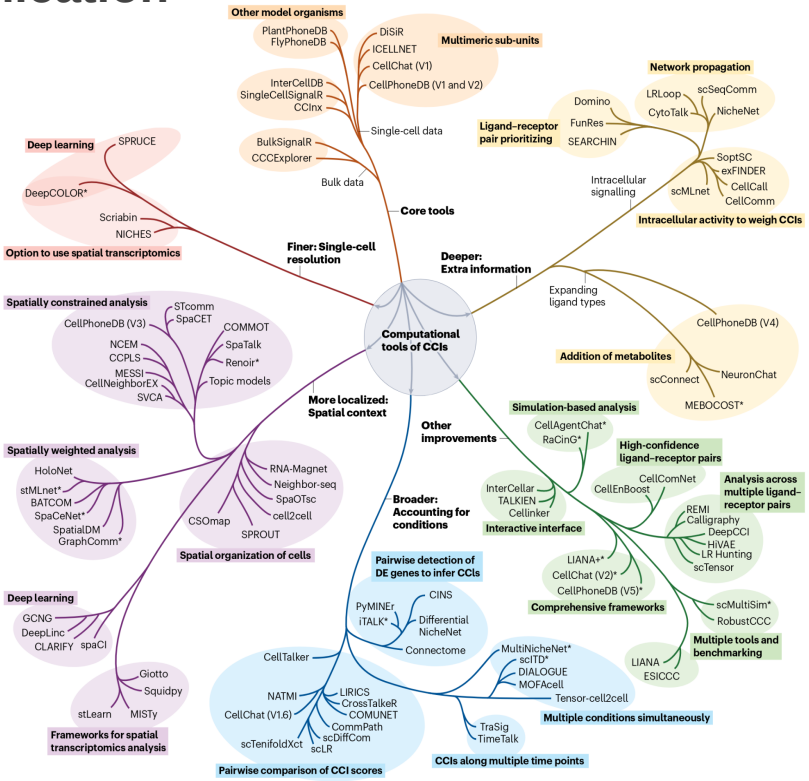
Overview of advanced analyses

Cell Cell Communication

Diversification of the methods:

- Computational:
  - add information: e.g.: intracellular pathways activity
  - robustness: replicates, conditions, time-points
  - spatial context (spatial transcriptomics)

- Experimental:
  - isolating doublets
  - track barcode diffusion

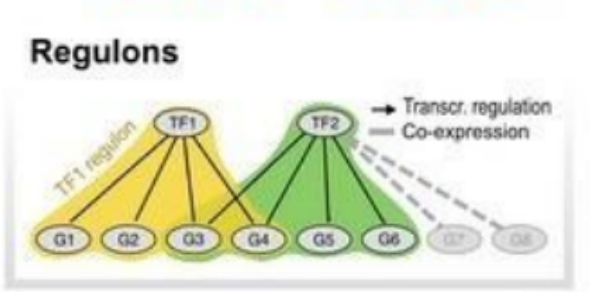


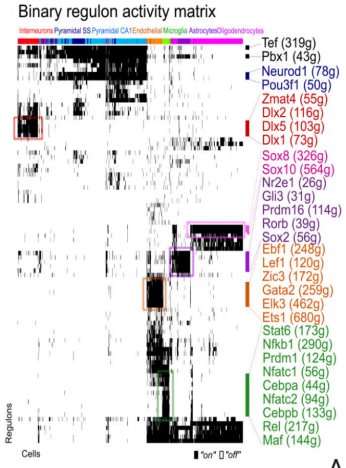
Armingol et *al.*, Nat Rev Genet 2024

Overview of advanced analyses

Gene regulatory networks inference

- GNR: set of interactions governing gene expression and cell functions
- scRNA-Seq: focus on transcription factors + cis taget elements
- Used to characterize and understand cell types or states



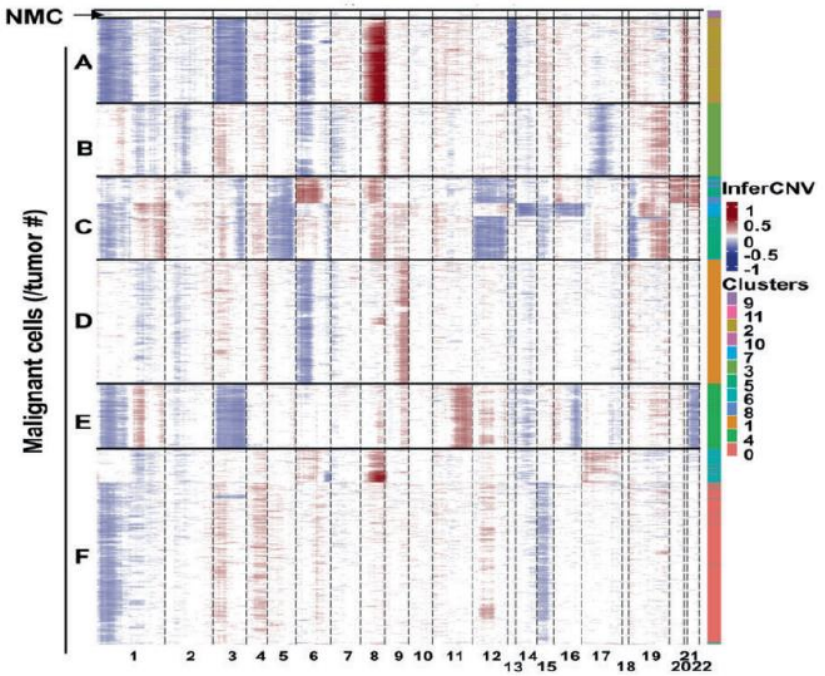


Aibar *et al.*, Nat Methods 2017

Overview of advanced analyses

copy number variations

- R package inferCNV
- identify large scale CNVs
- Principle: compare gene expression between samples and a set of reference “normal” cells.



Dong *et al.*, Front Genet 2020

References

1. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019 Feb;20(2):163–72.
2. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. SCENIC: Single-cell regulatory network inference and clustering. Nat Methods. 2017 Nov;14(11):1083–6.
3. Badia-i-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, Taus P, Dugourd A, Holland CH, Ramirez Flores RO, Saez-Rodriguez J. decoupleR: ensemble of computational methods to infer biological activities from omics data. Kuijjer ML, editor. Bioinformatics Advances. 2022 Jan 10;2(1):vbac016.
4. Franchini M, Pellecchia S, Viscido G, Gambardella G. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. NAR Genomics and Bioinformatics. 2023 Jan 10;5(1):lqad024.
5. Akira C, Loredana M, Emmanuelle S, Antonio R. Cell-ID: gene signature extraction and cell identity recognition at individual cell level [Internet]. 2020 [cited 2024 Jun 14]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.07.23.215525
6. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat Methods [Internet]. 2024 Mar 25 [cited 2024 Apr 11]; Available from: https://www.nature.com/articles/s41592-024-02235-4
7. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, Bader GD. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc. 2021 Jun;16(6):2749–64.
8. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018 Apr;15(4):255–61.
9. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, Hudelle R, Qaiser T, Matson KJE, Barraud Q, Levine AJ, La Manno G, Skinnider MA, Courtine G. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021 Sep 28;12(1):5692.
10. Deconinck L, Cannoodt R, Saelens W, Deplancke B, Saeys Y. Recent advances in trajectory inference from single-cell omics data. Current Opinion in Systems Biology. 2021 Sep 1;27:100344.
11. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, van Bruggen D, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. RNA velocity of single cells. Nature. 2018 Aug;560(7719):494–8.
12. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller HB, Pe'er D, Theis FJ. CellRank for directed single-cell fate mapping. Nat Methods. 2022 Feb;19(2):159–70.
13. Dong X, Zhang L, Hao X, Wang T, Vijg J. SCCNV: A Software Tool for Identifying Copy Number Variation From Single-Cell Whole-Genome Sequencing. Front Genet [Internet]. 2020 Nov 16 [cited 2024 Jun 14];11. Available from: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.505441/full
14. Armingol E, Baghdassarian HM, Lewis NE. The diversification of methods for studying cell–cell interactions and communication. Nat Rev Genet. 2024 Jan 18;1–20.

Aknowledgements

Some parts of this course are inspired or adapted from

Agnes Paquet presentation from the course
SincellTE 2022 / Single-Cell : Transcriptomics, Spatial and Multi-Omics

Thibault Dayris presentation from The AVIESAN – IFB – Inserm course
Traitement des données de génomique obtenues par séquençage à haut débit

Lorette Noiret presentation from the Cancéropôle Ile de France
MOOC NGS & Cancer – Single Cell