Analysing single-cell CUT&Tag in tumors

And tools related to single-cell epigenomics analysis

25 / 10 / 2022

Pacôme Prompsy

Plan

1) Introducing single-cell epigenomics & challenges in the analysis

2) Quick comparison of scChIP-seq and scCUT&Tag

3) Best practices for single-cell histone modification analysis

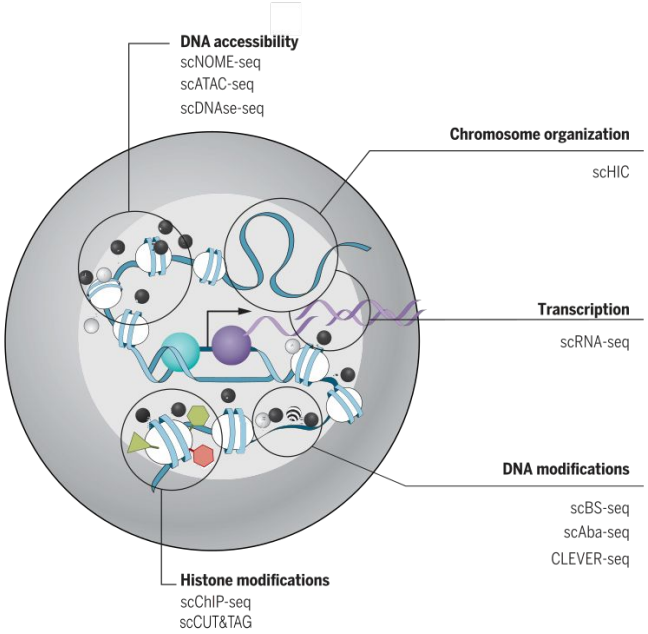
4) Analysis of single-cell CUT&Tag data of Juxtatumoral tissue of triple negative breast cancer mouse using ArchR & ChromSCape

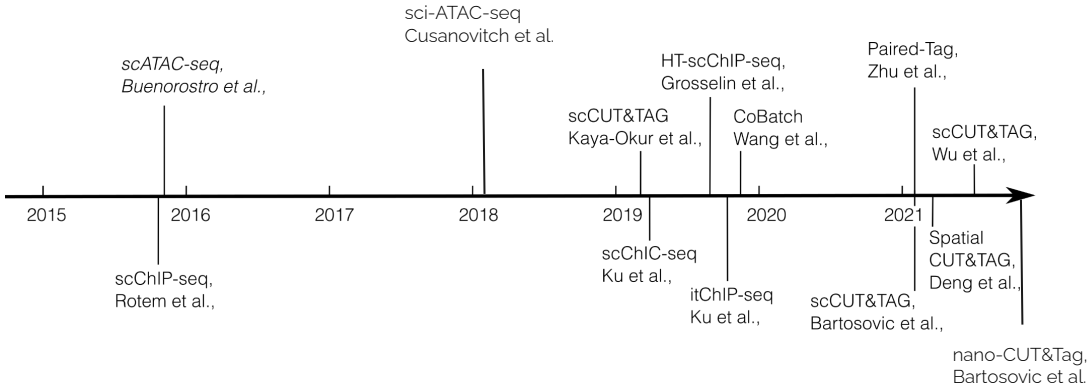
---

5) (**If enough time**) Iterative Differential Clustering

1 - Introducing single-cell epigenomics & challenges in the analysis

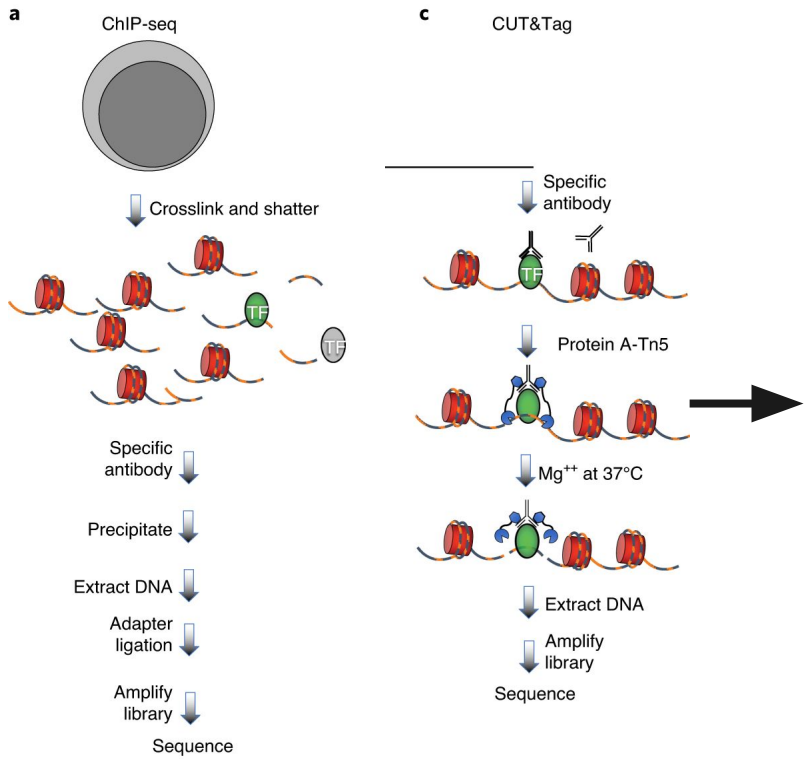
Existing Technologies





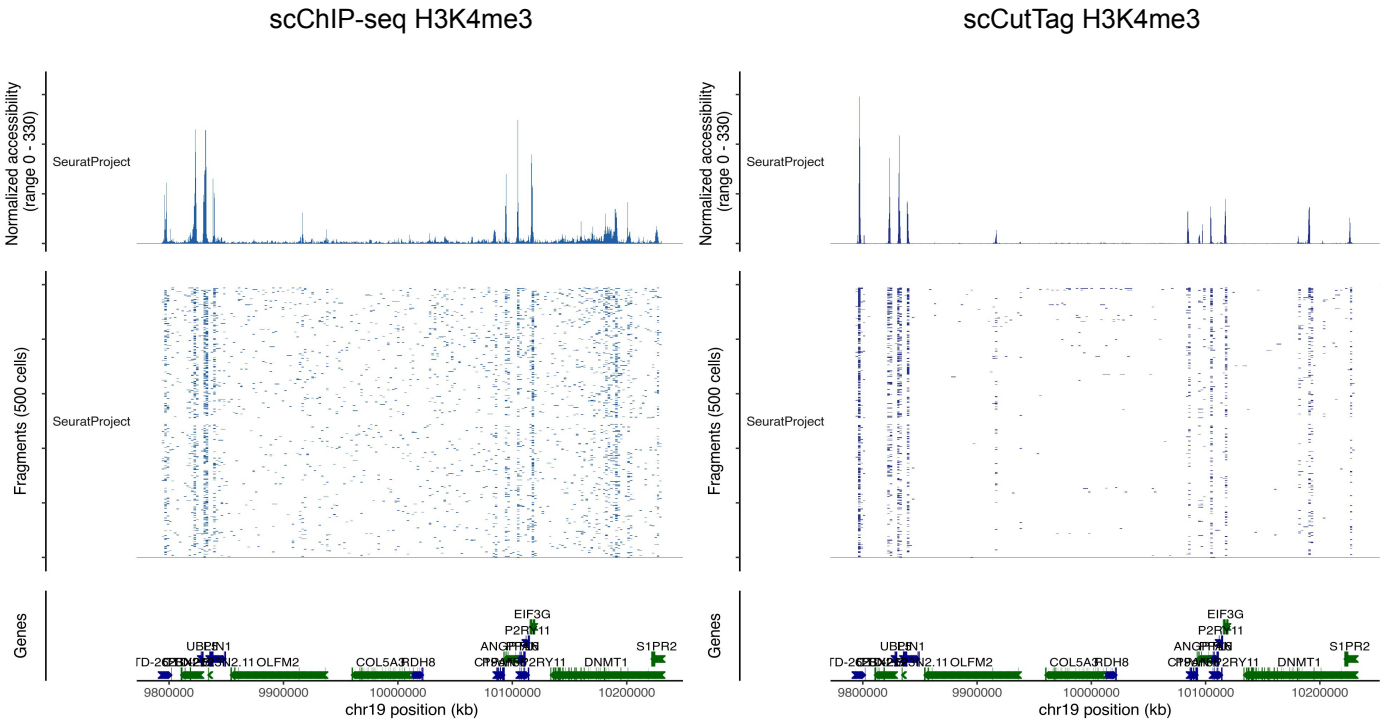
Kelsey et al., 2021

ChIP-seq vs CUT&Tag

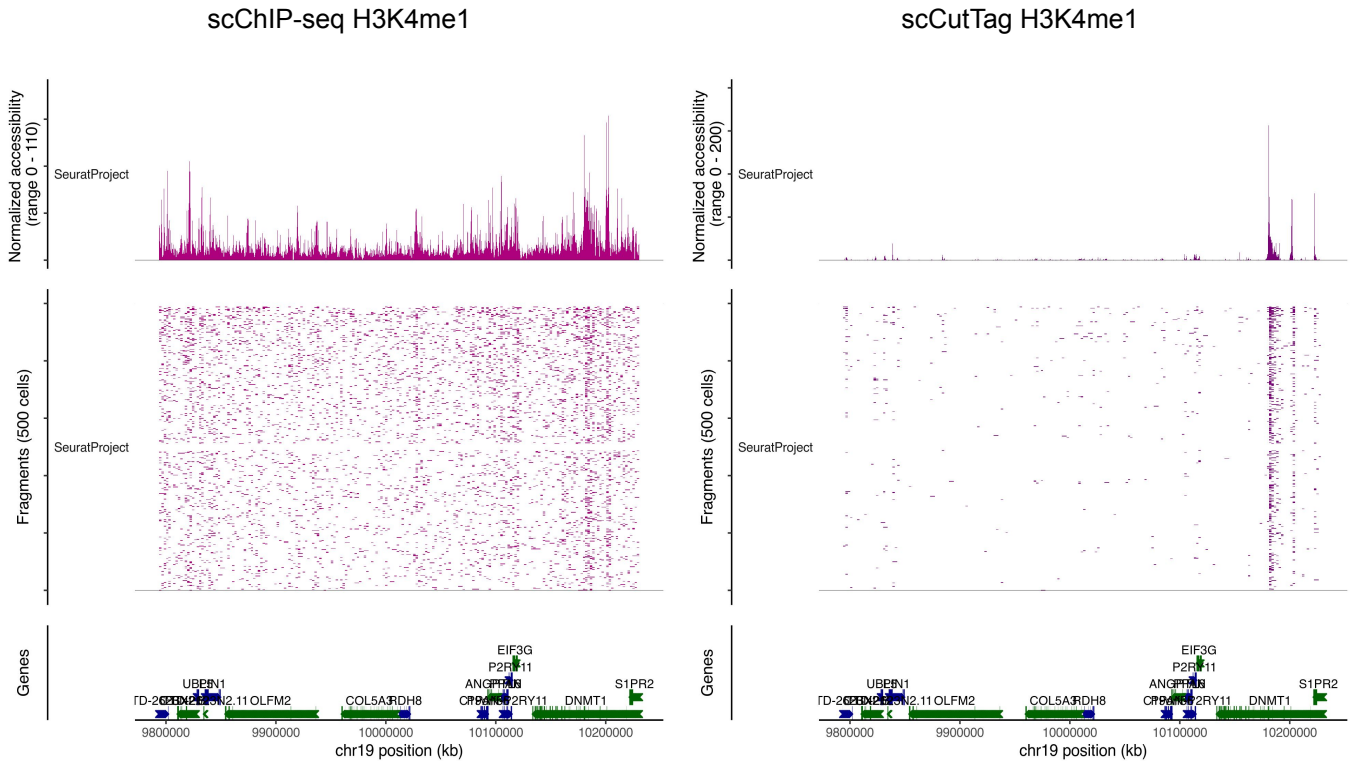


- Simpler protocol
- Uses **transposase** enzyme to both cut and add primers
- Better ratio signal vs noise
- Might be biases toward open chromatin (as coming from ATAC-seq)
- Protocol still needs to be optimized !

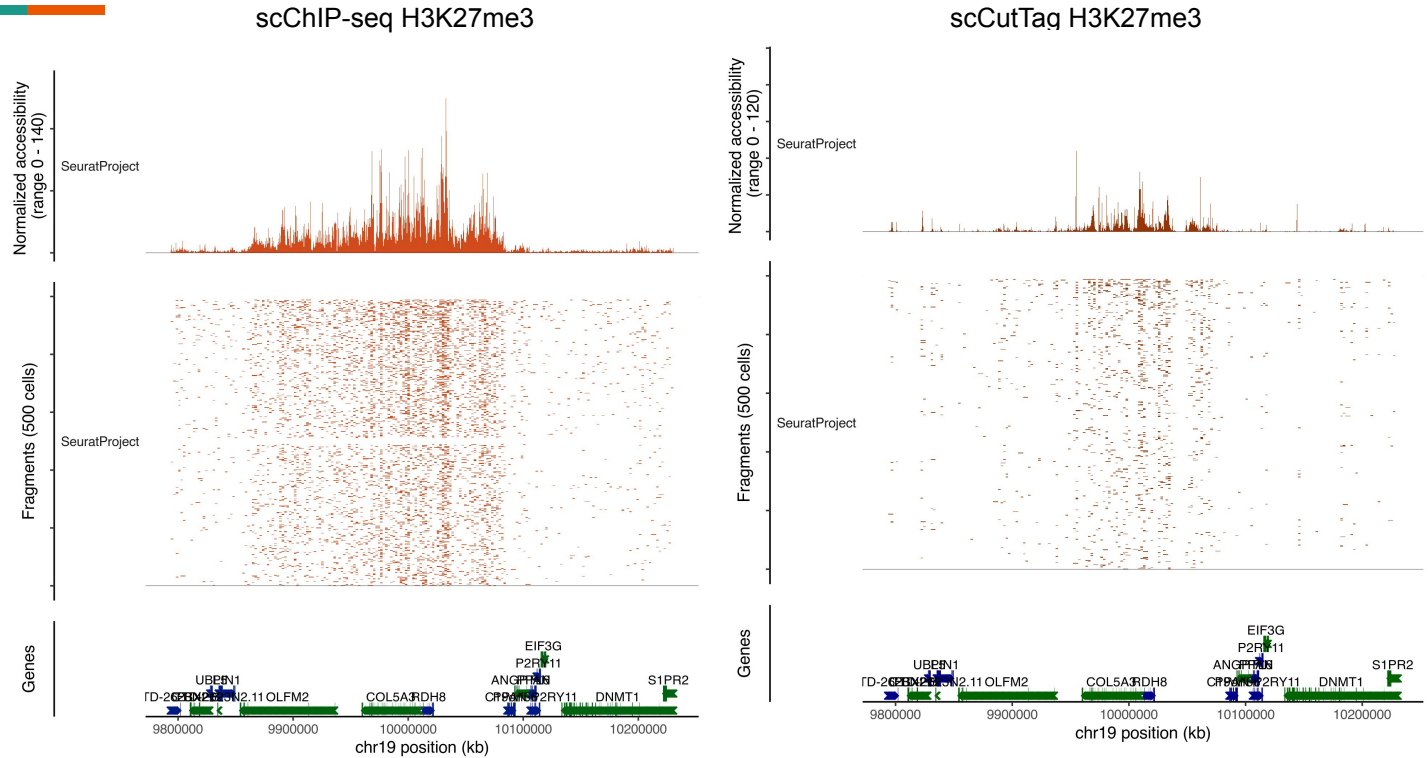
ChIP-seq vs CUT&Tag



ChIP-seq vs CUT&Tag

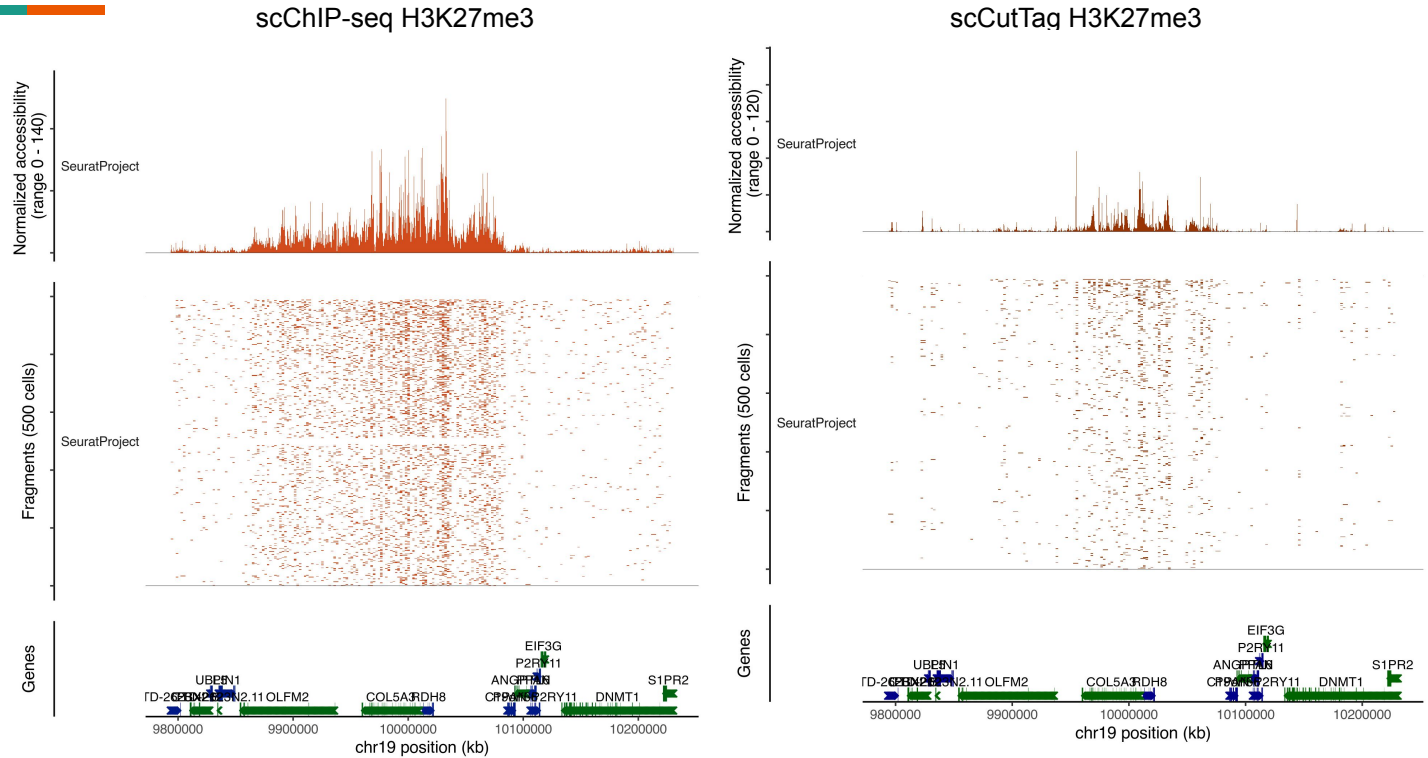


ChIP-seq vs CUT&Tag

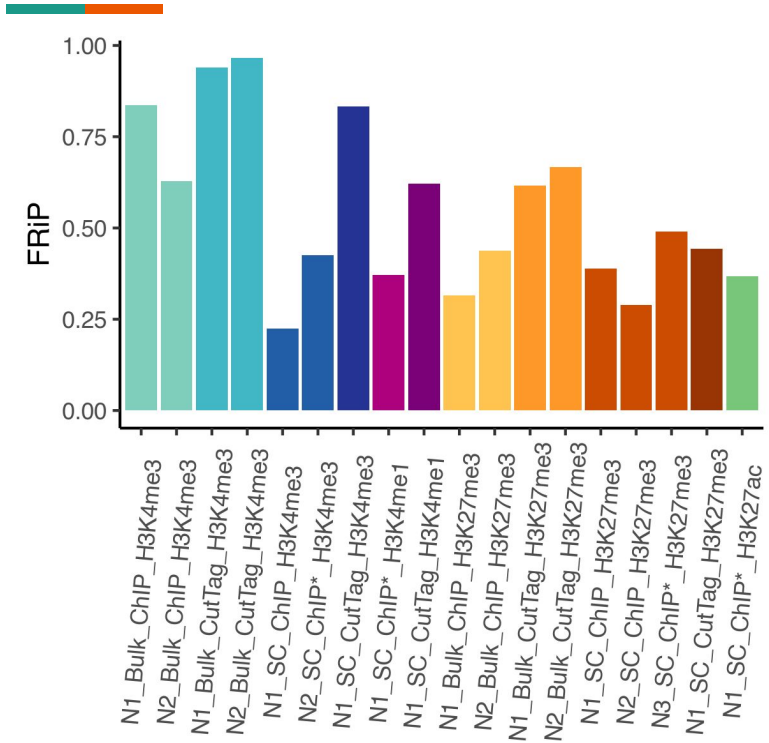


Beware of bias of the Transposase towards open chromatin

ChIP-seq vs CUT&Tag



Beware of bias of the Transposase towards open chromatin



- The FRiP is always higher in CUT&TAG compared to ChIP-seq: ~33% increase for bulk H3K4me3 ; ~2-fold increase for scH3K4me3 & H3K4me1 & bulk H3K27me3
- The FRiP is constantly higher in bulk experiments than in single-cell experiments (except scChIP H3K27me3*).
- The FRiP is higher in H3K4me3 compared to H3K4me1, H3K27ac and H3K27me3.

2 - Best practices for single-cell histone modification analysis

Challenges of scHistone (vs scRNA)

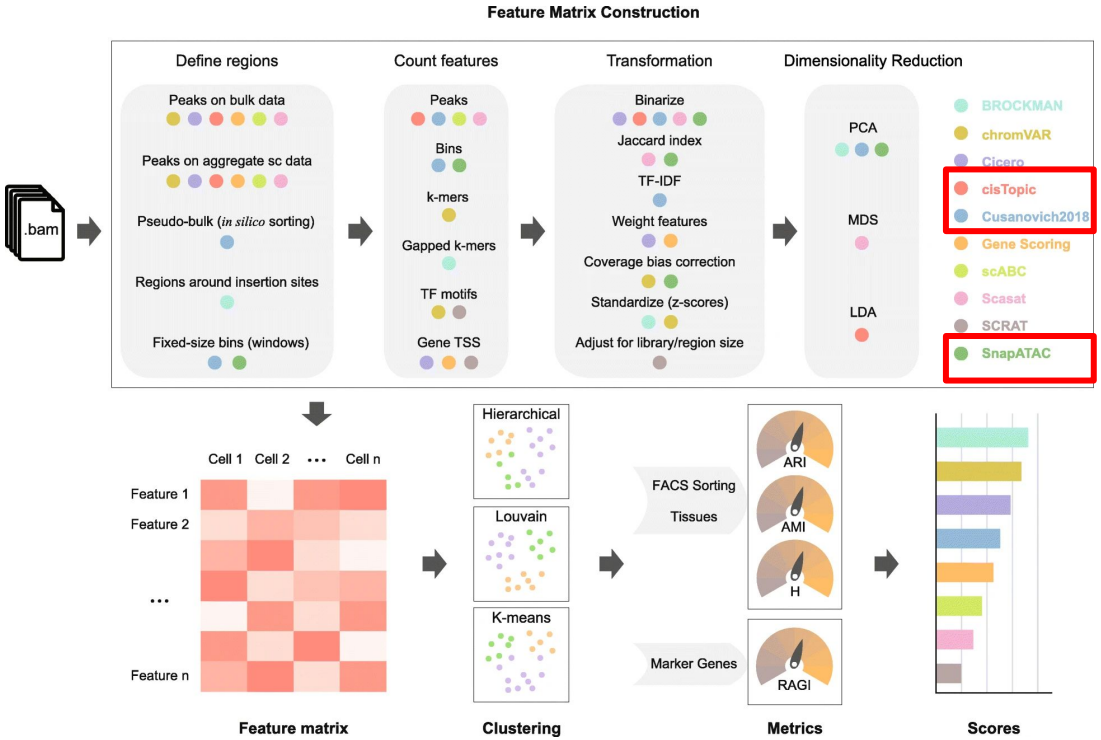
- **2** or **4** copies of DNA per cells vs 100s to 1000s of RNA per cell
  → need to **strongly amplify signal** → lots of duplicates

→ 'Read recovery' of **0.5 to 5%**, to be compared to **70-95%** in scRNA using 10X.

- Median number of reads per cell is **50x** to **100x** lower than in scRNA (**500-1000** reads/cell vs **50,000-100,000** reads/cell).

- There is no 'natural feature' (e.g. genes) to create the matrix. Features are usually one of **genomic bins**, **peaks** or **gene body + TSS**.

Existing tools (dedicated to single-cell ATAC-seq)



Plus now -

Signac - by Satija Lab (Stuart et al., 2021)

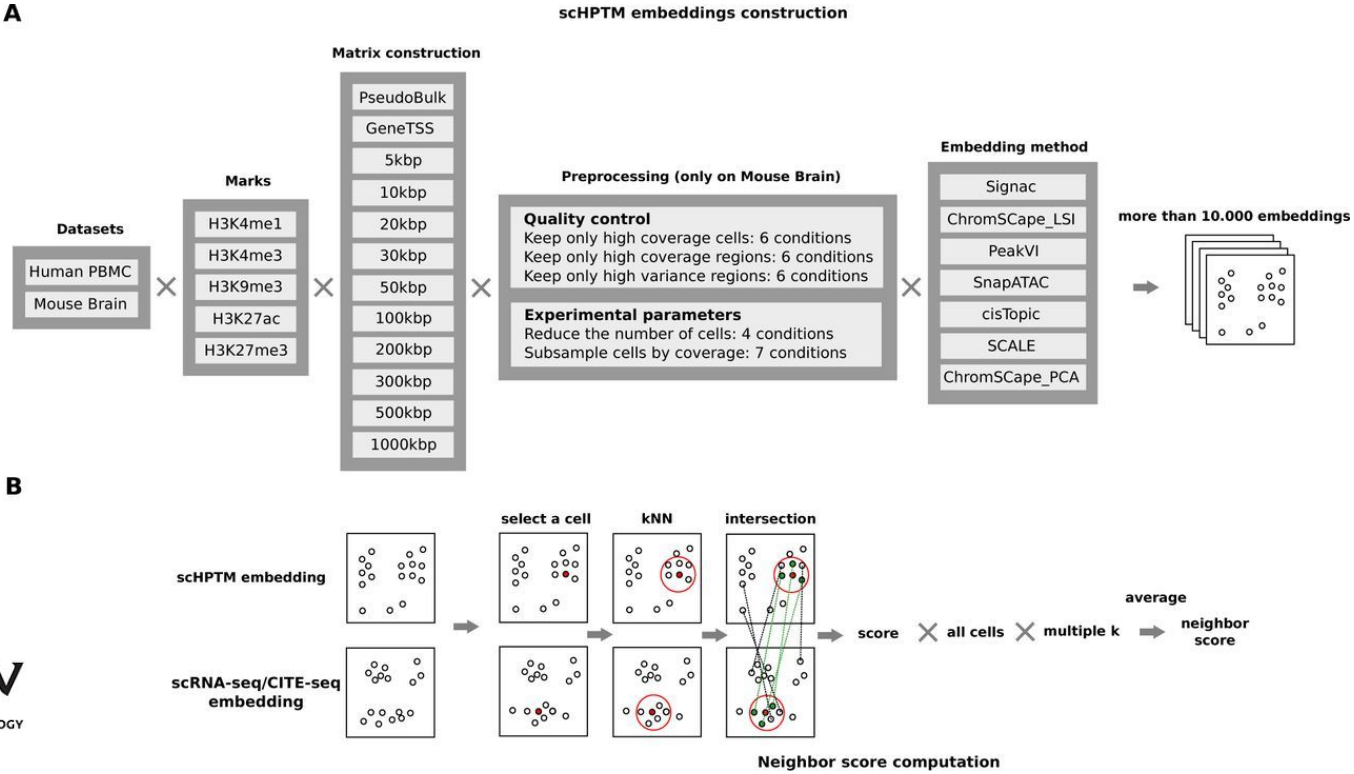
ArchR - by Greenleaf Lab (Granja et al., 2021)

ChromSCape - by our lab (Prompsy et al., 2020)

Assessment of computational methods for the analysis of single-cell ATAC-seq data, Chen et al., 2019

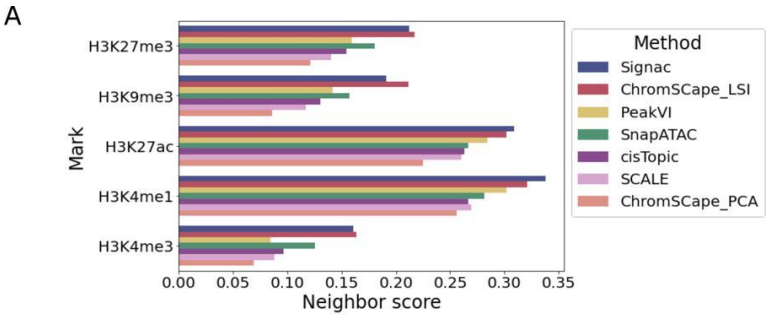
Best practices for single-cell histone modification analysis

Félix Raimundo



Raimundo et al., 2022

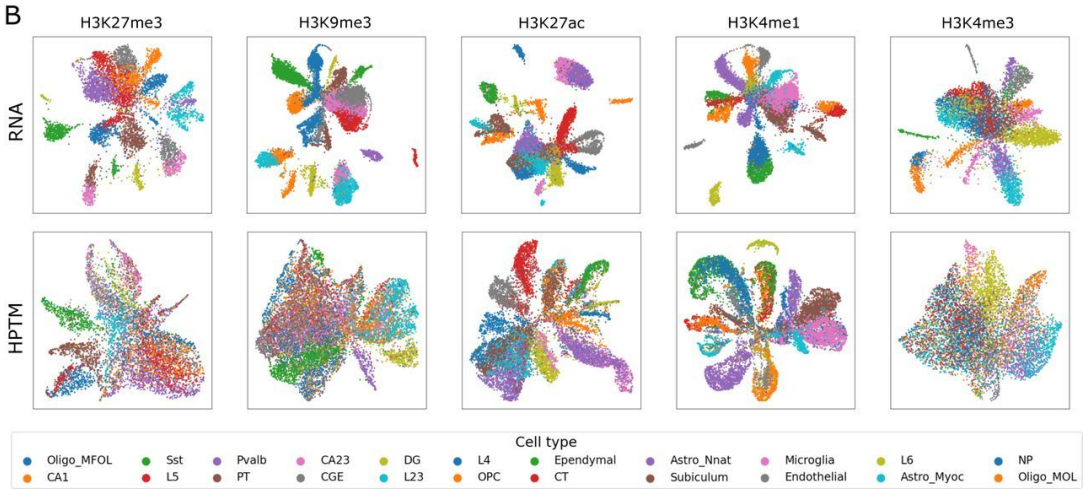
Best practices for single-cell histone modification analysis



Globally, Latent Semantic Indexing (**LSI**) generated the best embeddings.

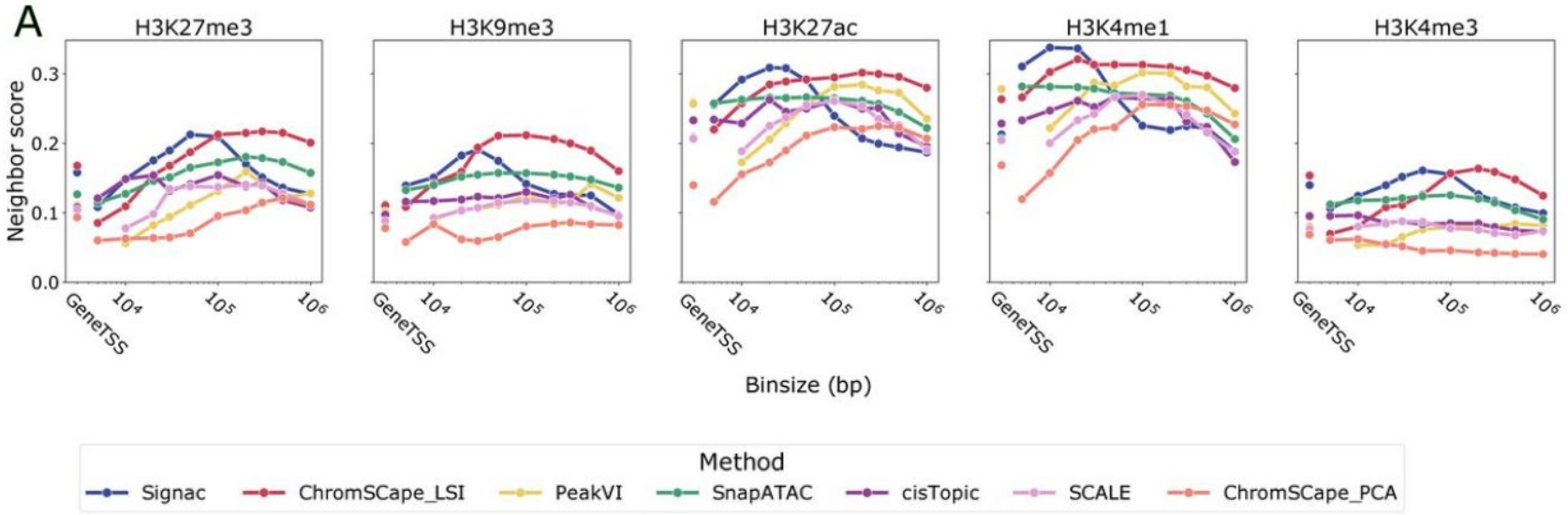
This is simply a 'TF-IDF' log-normalisation followed by a regular PCA. 1st component is removed.

First used in scATAC by Cusanovitch et al, 2018



Paired-Tag
Joint profiling of histone modifications and transcriptome in single cells from mouse brain, Zhu et al., 2021

Choice of feature is highly important

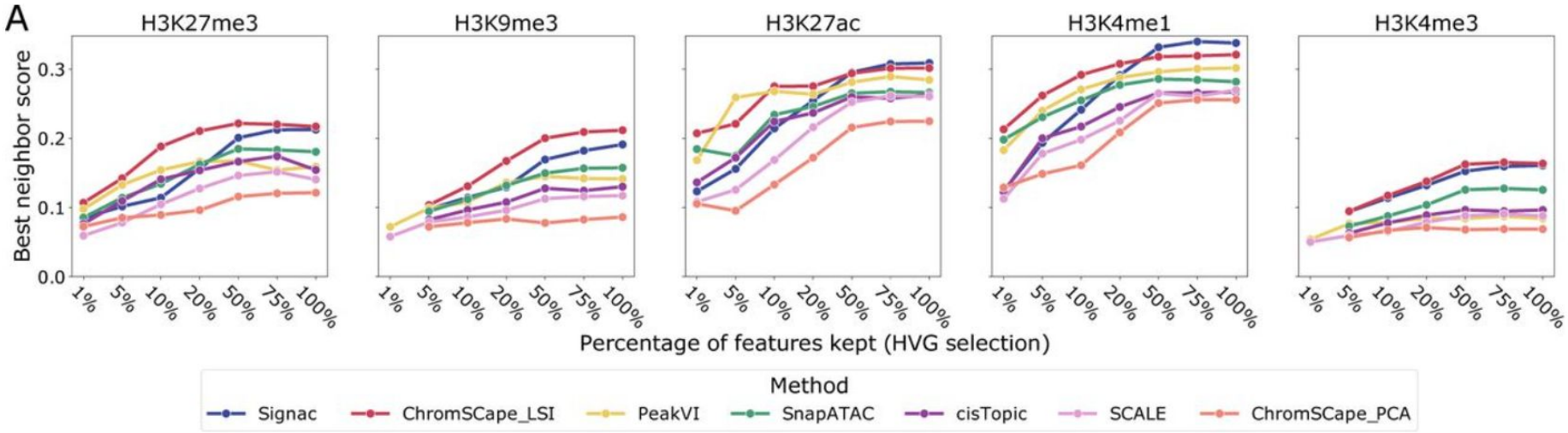


The matrix construction step **strongly impacts performances**.

Large bins (50kbp - 200kbp) perform overall best, even for active mark known to accumulate in sharp peaks.

Counting on genes and peaks constantly performed worse than genomic bins.

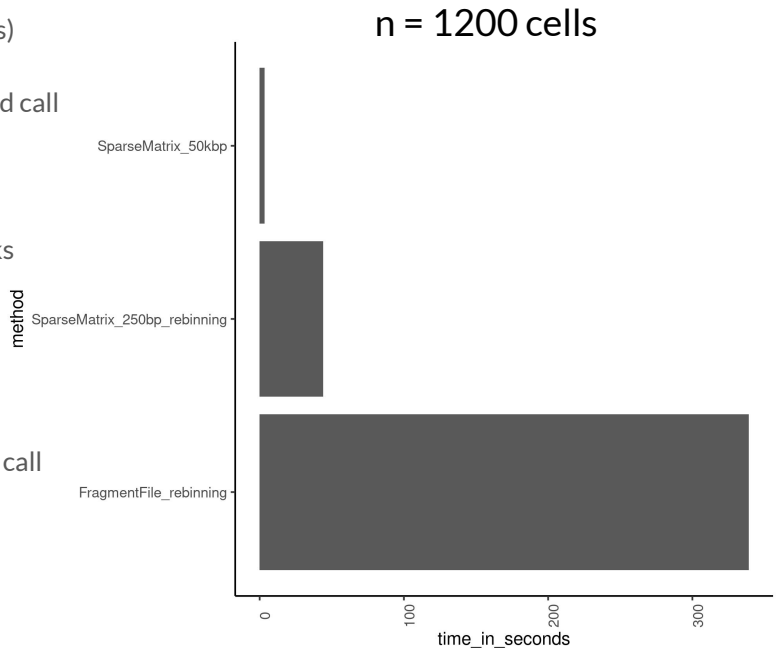
Feature selection always reduce performances



Removing features by either coverage or variance, such as in scRNA-seq always impact negatively performances.

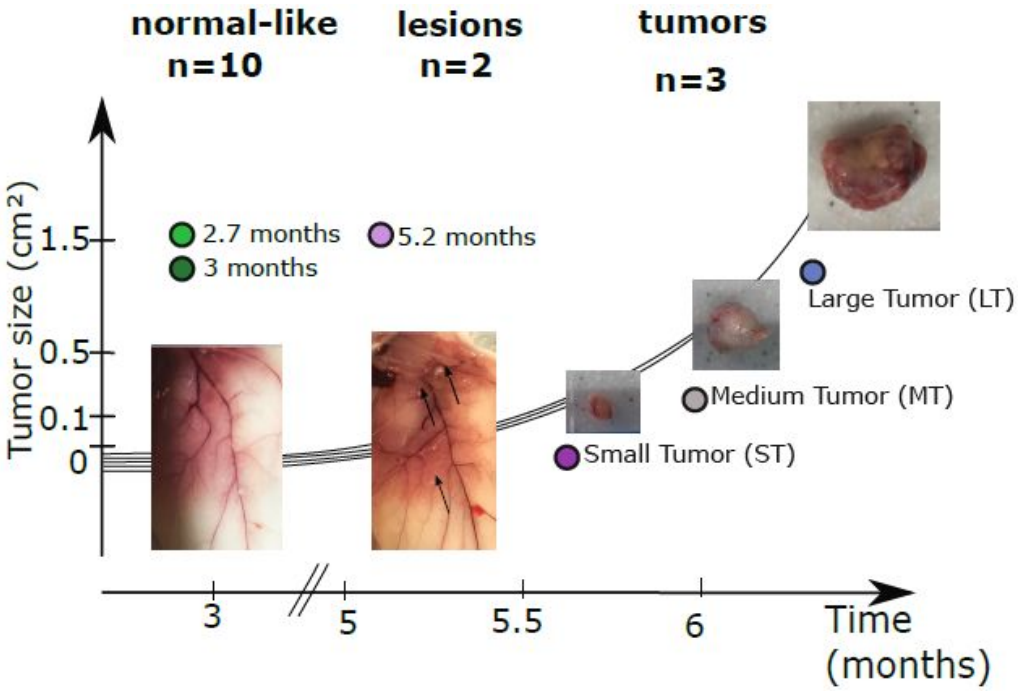
Digression of input formats and feature choice

- **Sparse Matrix Format** - Processed - 10X scRNA like (genomic bins, genes or peaks)
  - pros: light, quick to to read
  - cons: cannot re-count to recreate other features, create coverage files and call peaks
- **Fragment File Format - Raw - BED like**
  - pros: can re-count on any feature type, create coverage files and call peaks
  - cons: long time to read and re-count
- **Sparse Matrix 250-bp format - Semi Raw** - 10X scRNA like
  - \+ pros: light and fast to read, can re-bin quickly to create coverage files and call peaks
  - \+ cons: is an approximation of the real data, not precise (e.g. for TF motif analysis)



3 - Analysis of Breast Cancer Juxta-tumoral tissue in mouse

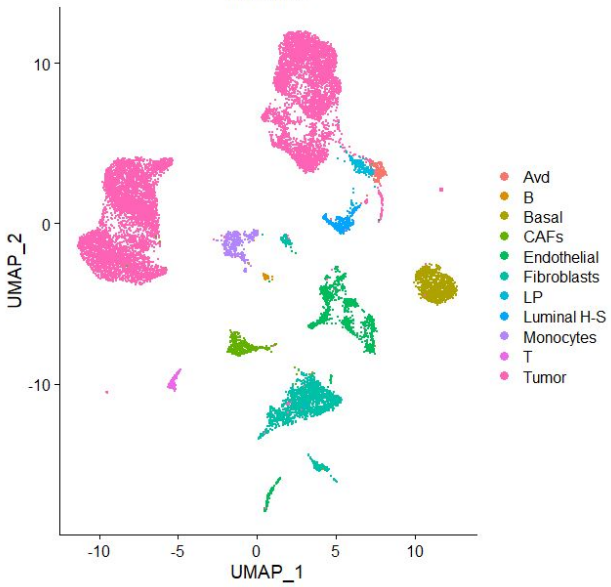
Analysis of Breast Cancer Juxtatumoral tissue in mouse

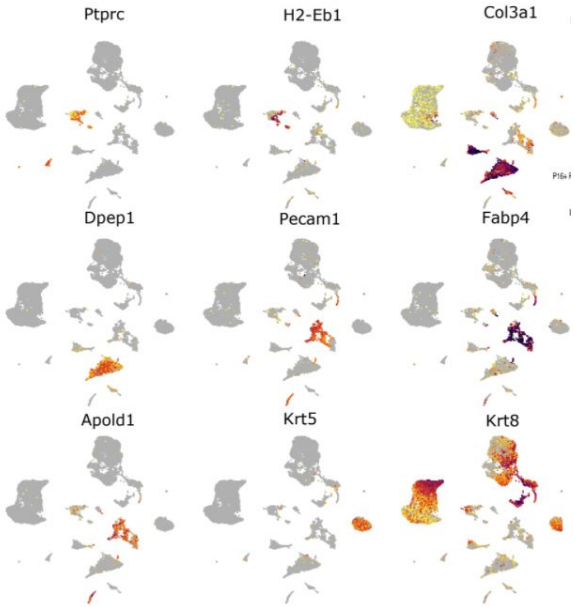


Abbreviations:
LP: Luminal Progenitor
Avd: Alveolar differentiated cells
ST: Small tumor; MT:Medium tumor; LT:Large tumor
Juxta: Juxta-tumoral : part of tissue close to the tumor site

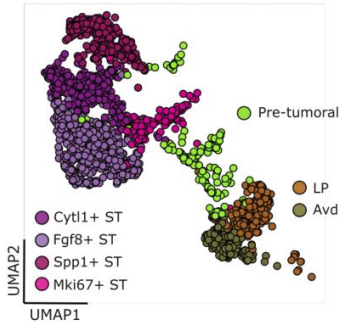
Camille Landragin

Identification of pre-tumoral LP using scRNA-seq





Identified '**pre-tumoral**' Luminal Progenitors in juxta-tumoral tissue



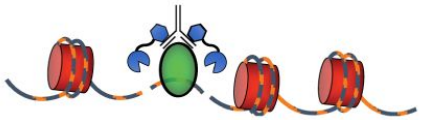
Mélissa Saichi

→ Can we find the same cell types as in scRNA-seq ?

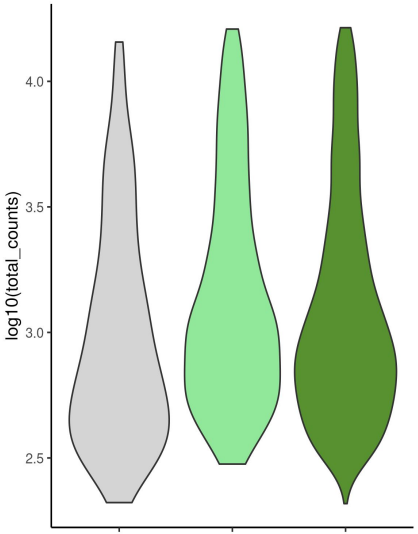
→ Can we identify a 'LP pre-tumoral' population in our juxta tissue, based on H3K4me1 landscape only ?

scH3K4me1 of Juxta and Control tissues

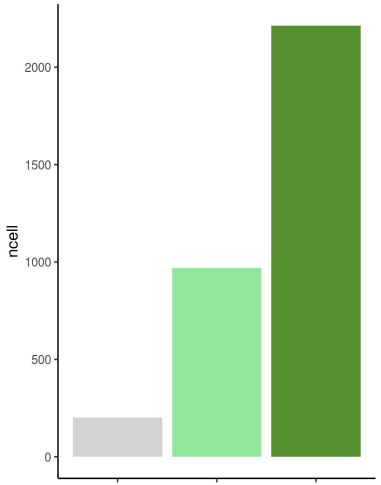
scCUT&Tag + 10X



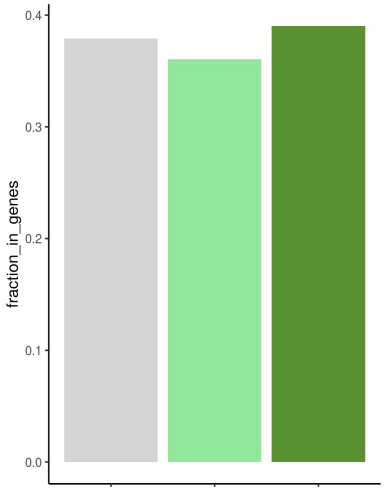
log10 total reads



Number of cells



Fraction of reads around genes
(genebody + 1000bp)

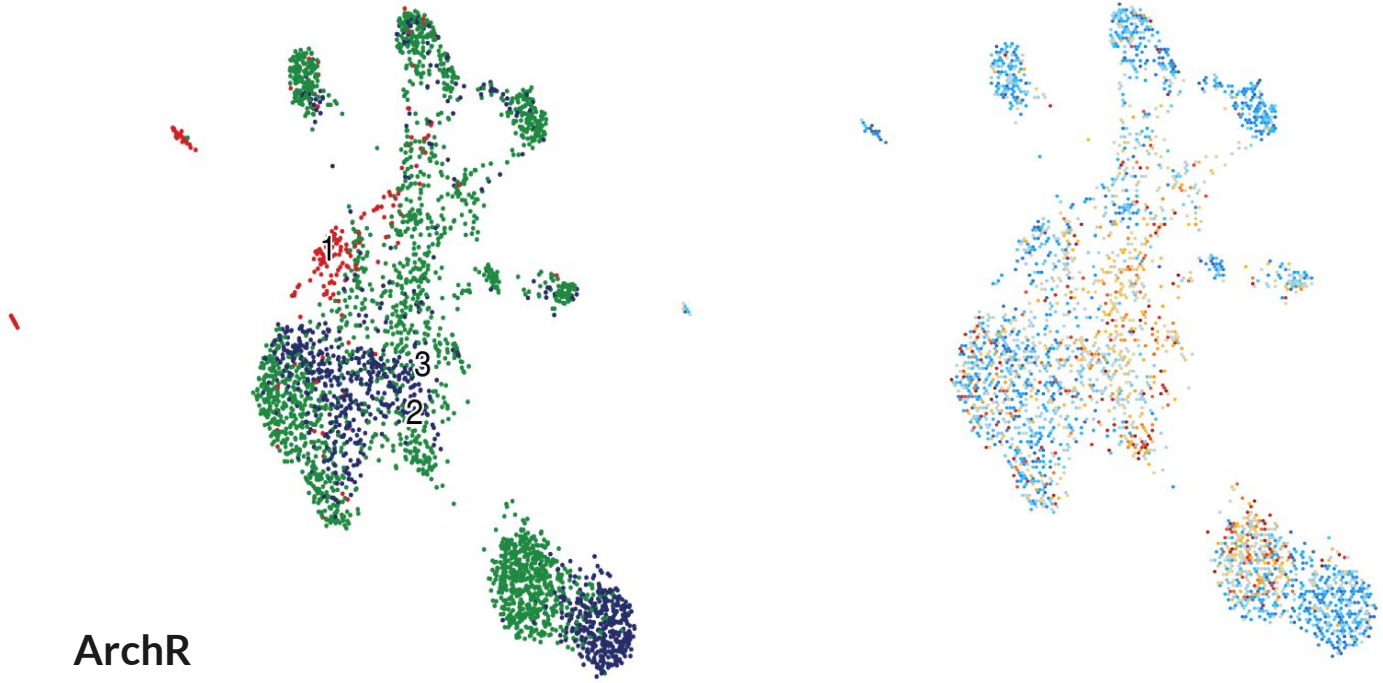


Genomic bins of 20kbp - No feature removal - Cells > 200 unique fragments & remove top 5%

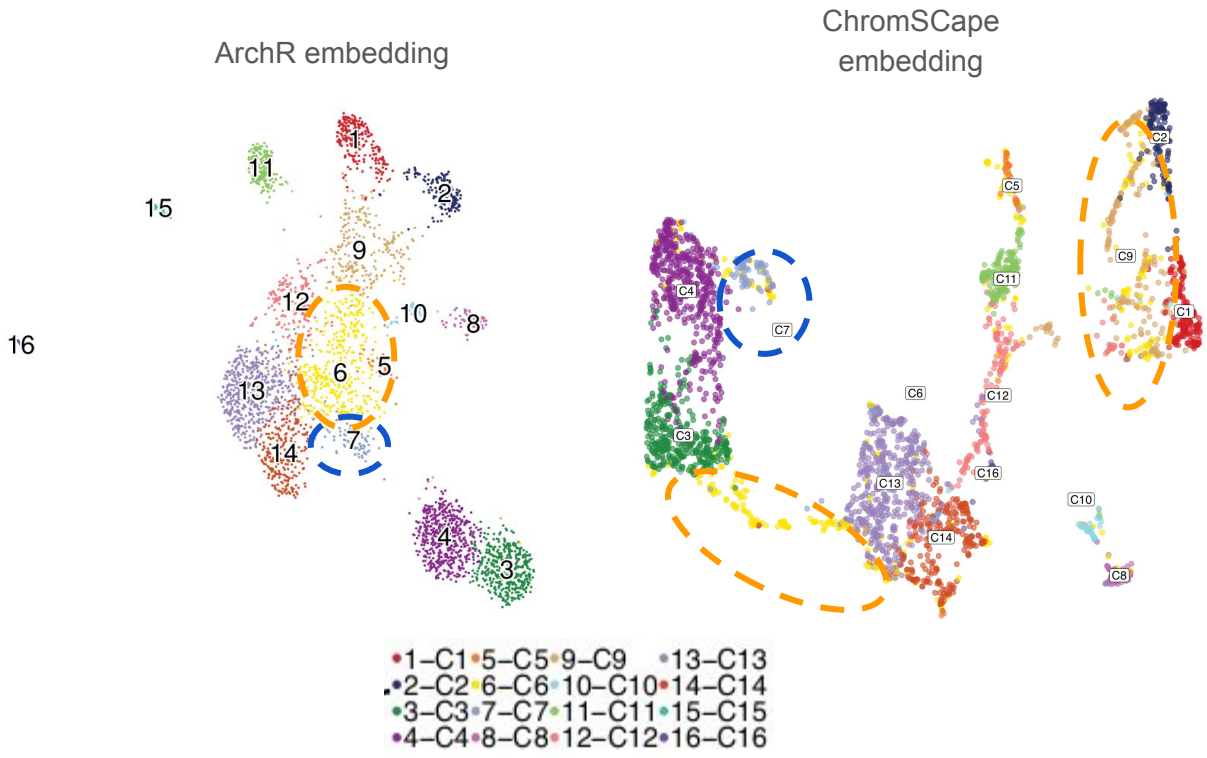
ChromSCape - UMAPs

ChromSCape

ArchR - UMAPs


C_7412_mm_m02y22_H3K4me1_250
J_7583_mm_m06y22_H3K4me1_1_250
J_7584_mm_m06y22_H3K4me1_250
1
3
2
ArchR

Identifying a ‘strange’ cluster

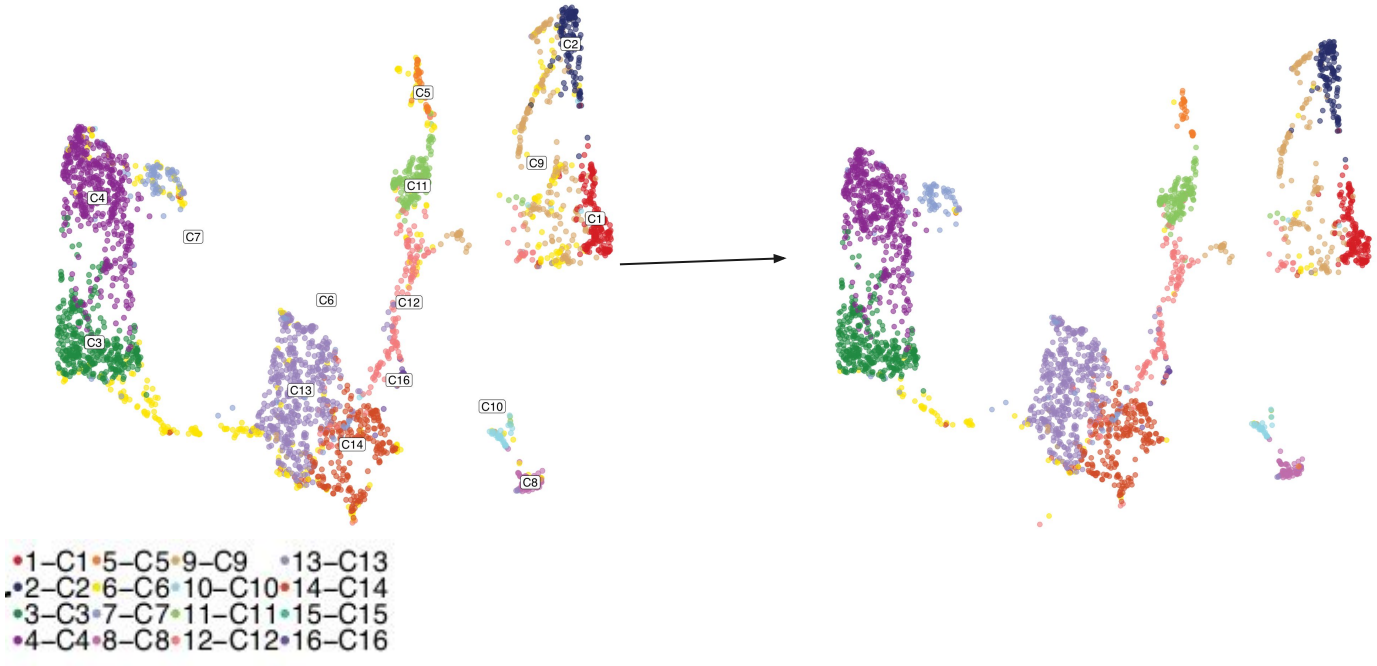


Cluster 6 seems is highly covered and clusters in between main cell groups in ChromSCape embedding

Removing highly covered cells remove specifically cluster 6

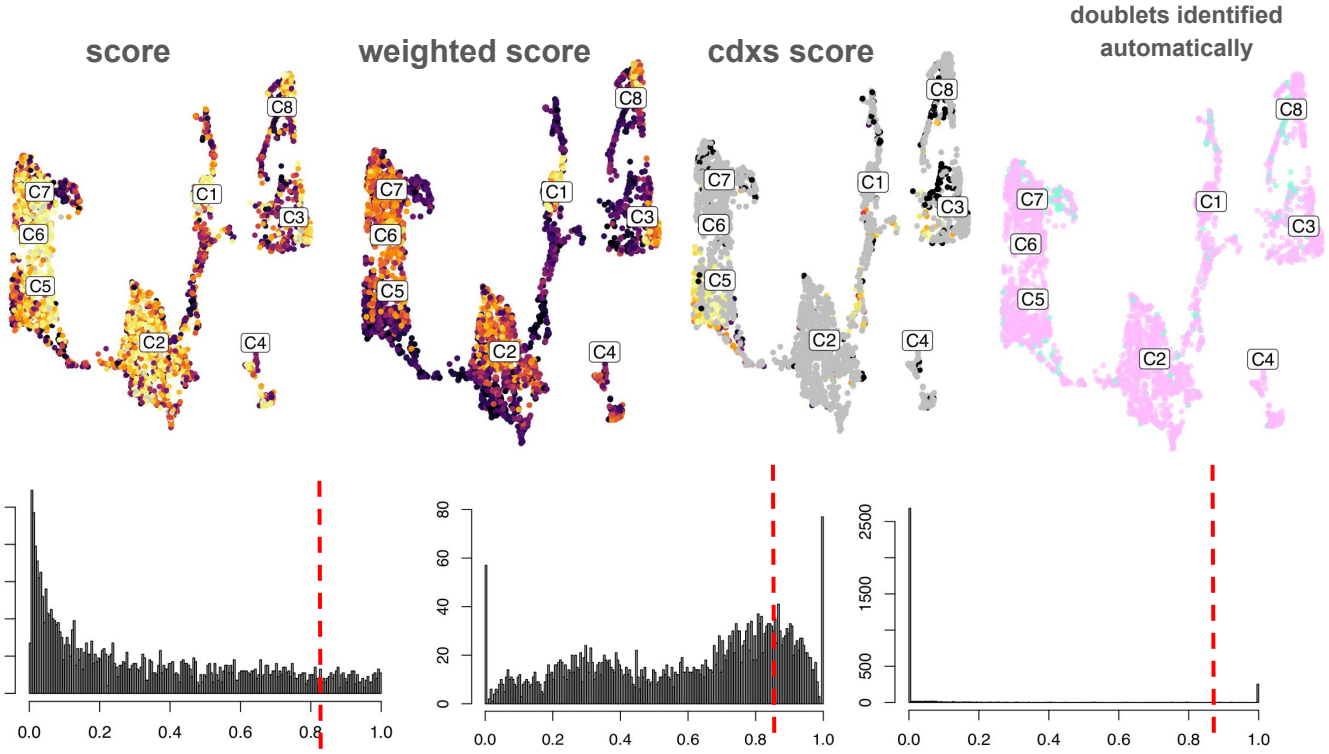
Before top 15% covered cell removal

After top 15% covered cell removal

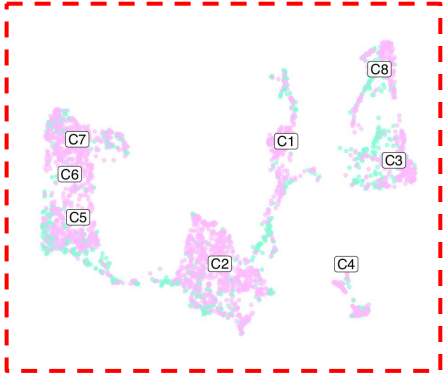


Removing additionally the top 15% of cells removes in priority cells from cluster 6 (80% vs ~8.5% in average)

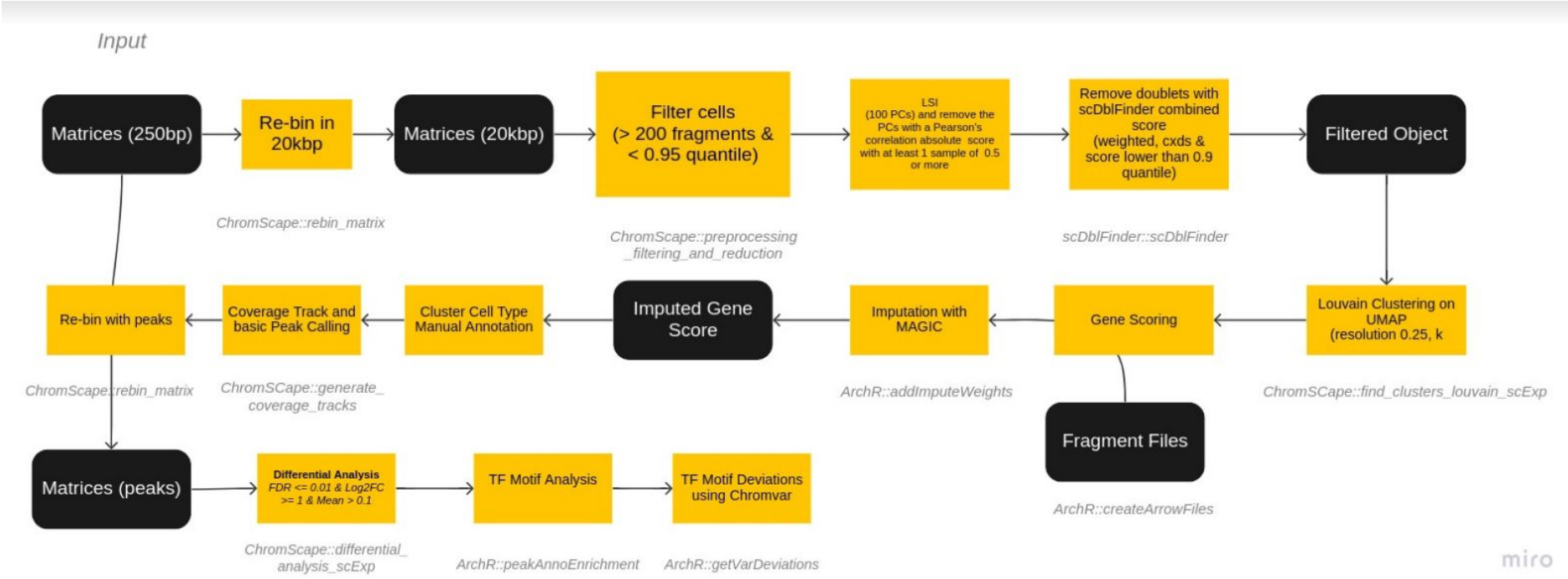
scDblFinder



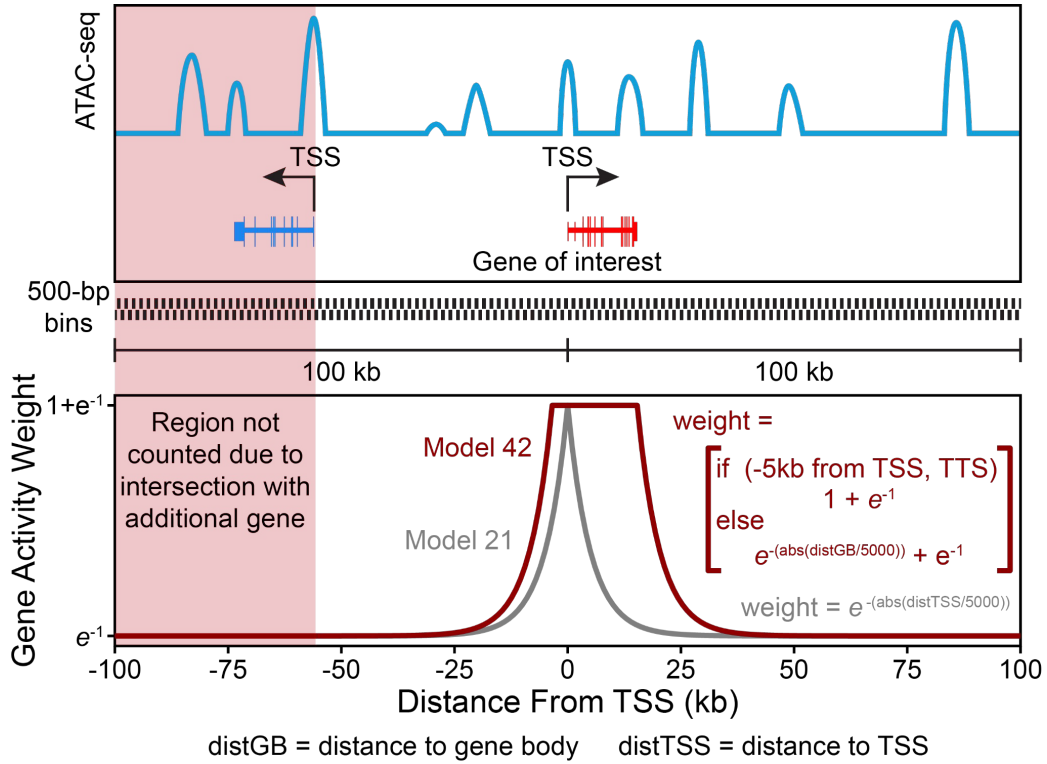
doublets identified
weighted score < q85 & cdxs score < q85



Final ‘Roadmap’ of the scH3K4me1 analysis

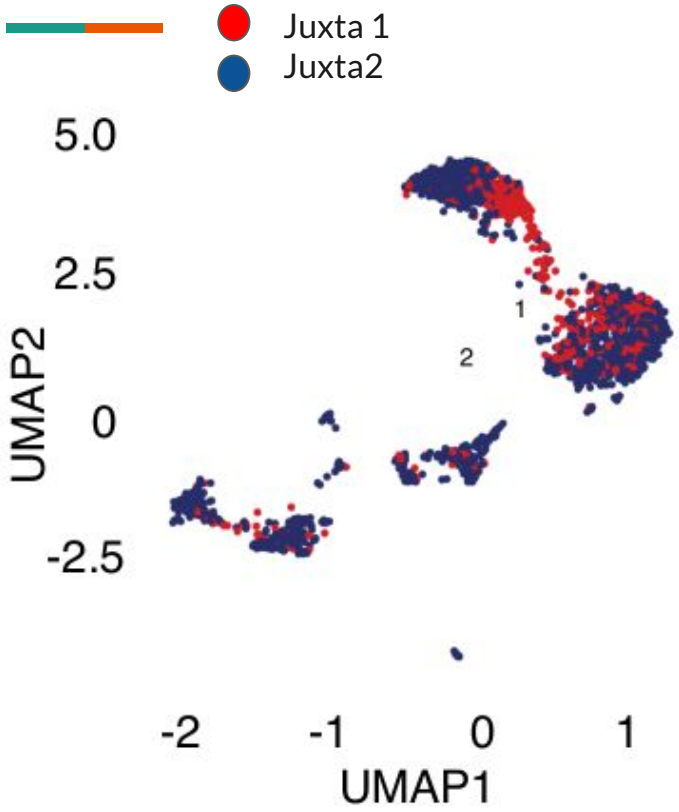


Calculating Gene Scores with ArchR

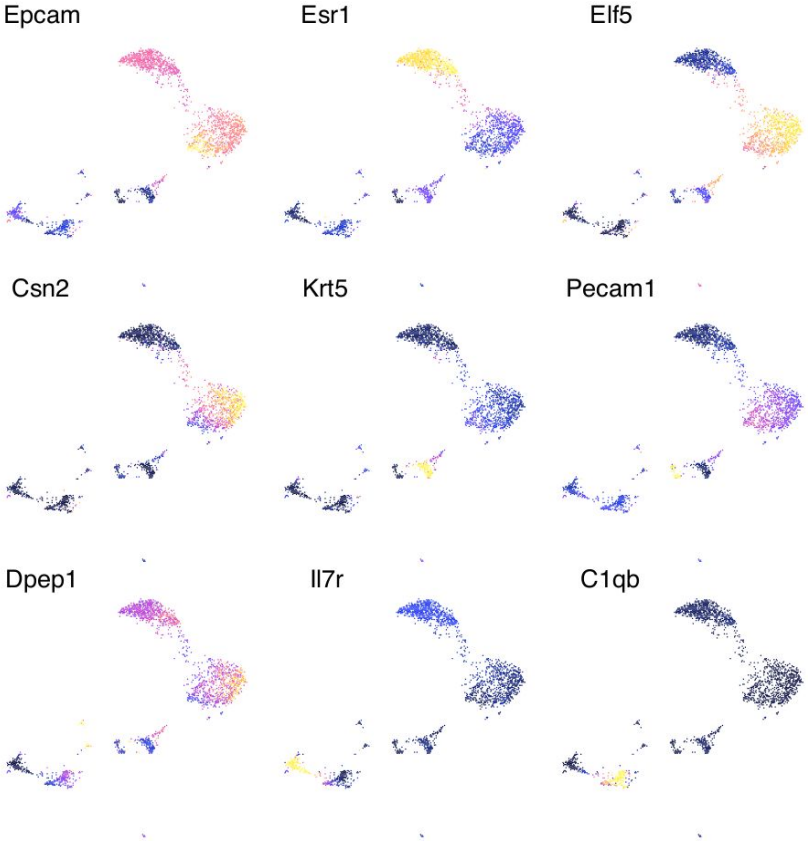


→ ArchR has a specific model to count reads around genes

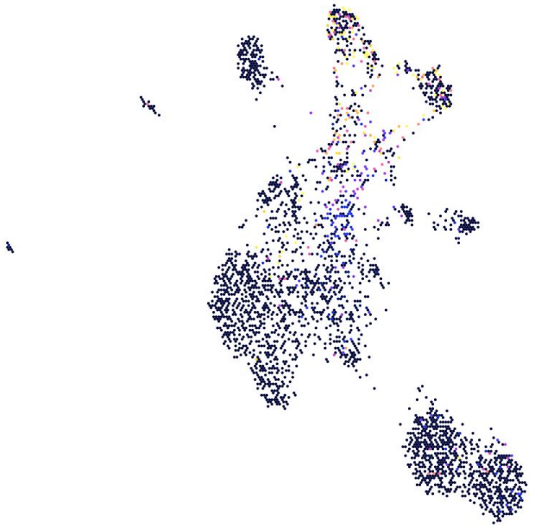
H3K4me1 of Juxta tissues



MAGIC Imputation on GeneScoring

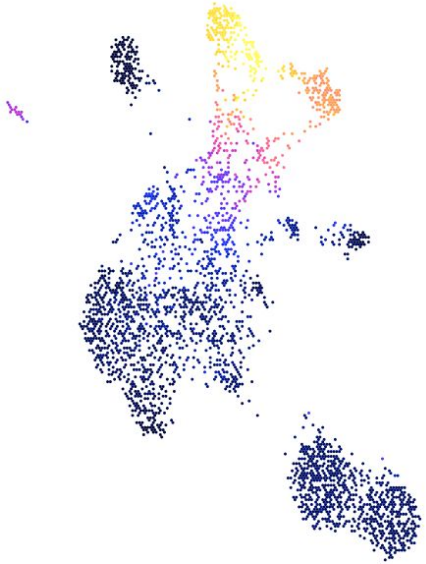


Comparison MAGIC vs Coverage



Ptprc
Before MAGIC

MAGIC



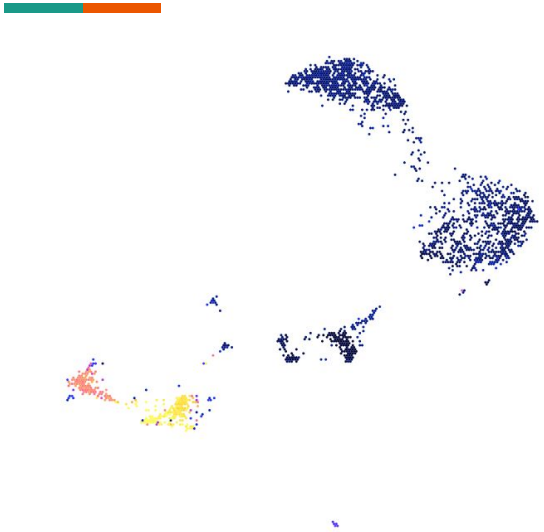
Ptprc
After MAGIC

!!!

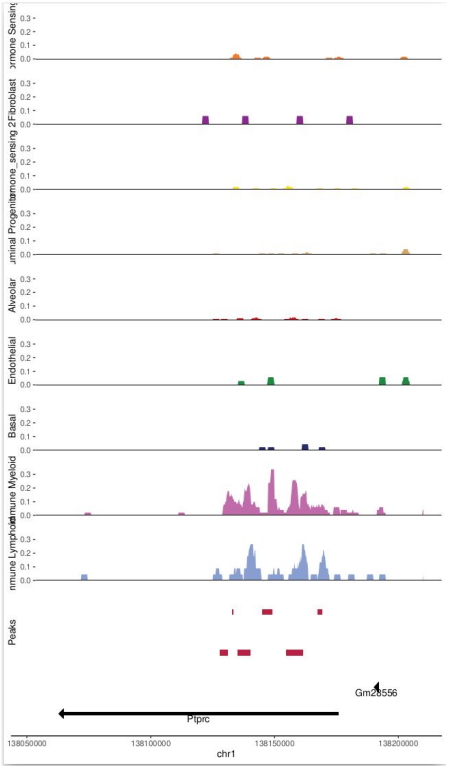
We must be careful when visualizing
UMAP with MAGIC imputation !
Always check coverage !

!!!

Comparison MAGIC vs Coverage



Ptprc
After MAGIC

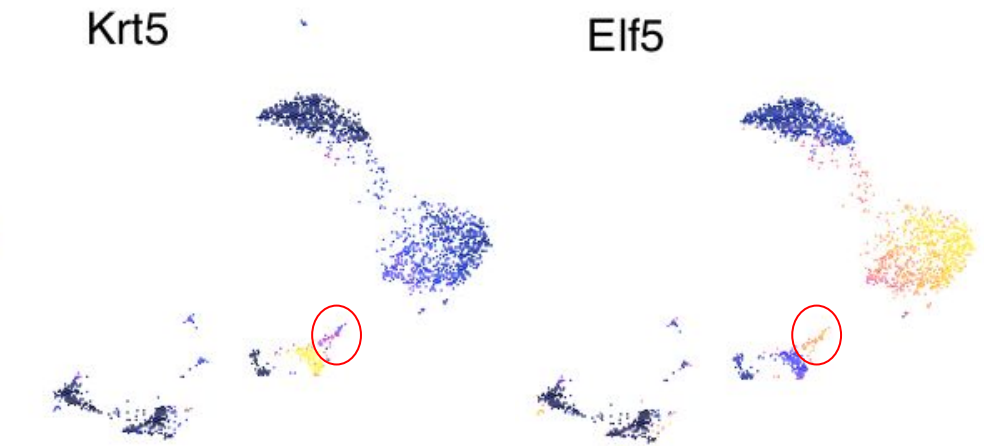


!!!

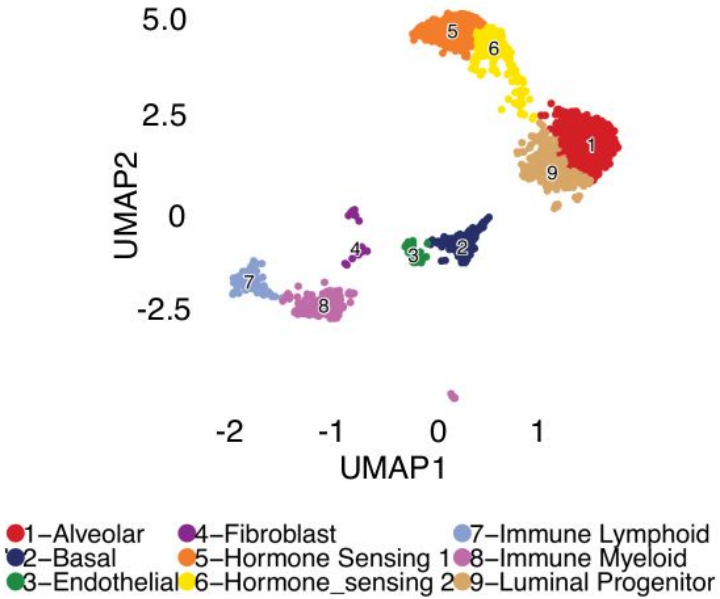
We must be careful when visualizing
UMAP with MAGIC imputation !
Always check coverage !

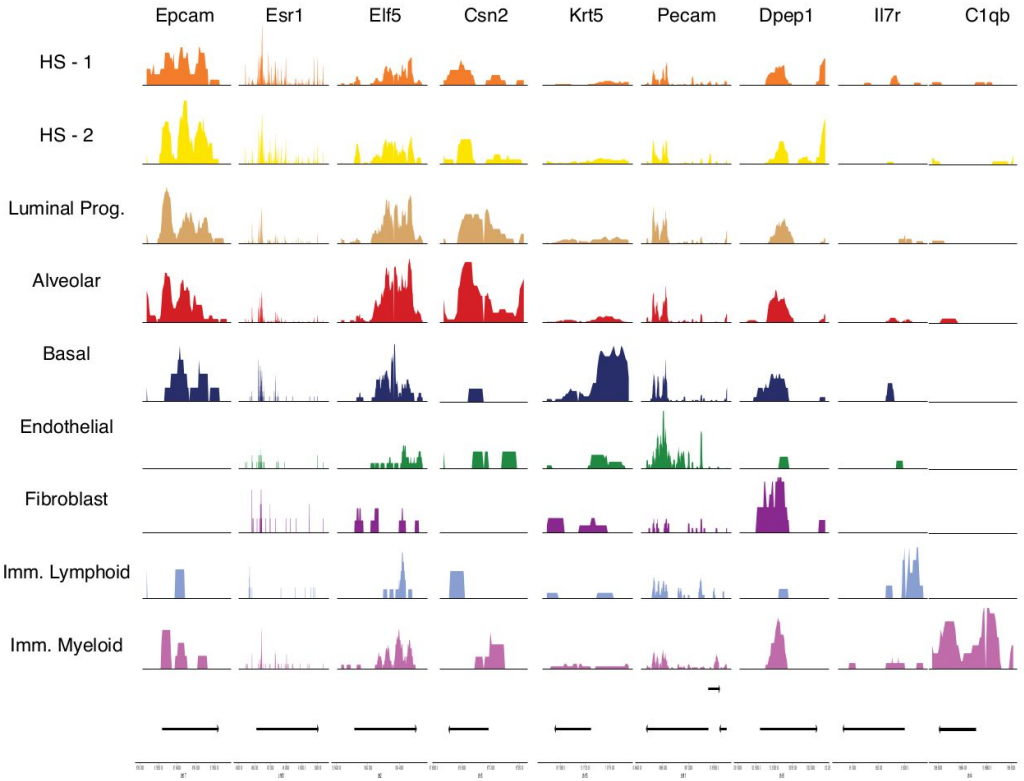
!!!

A potentially interesting population between LP and basal



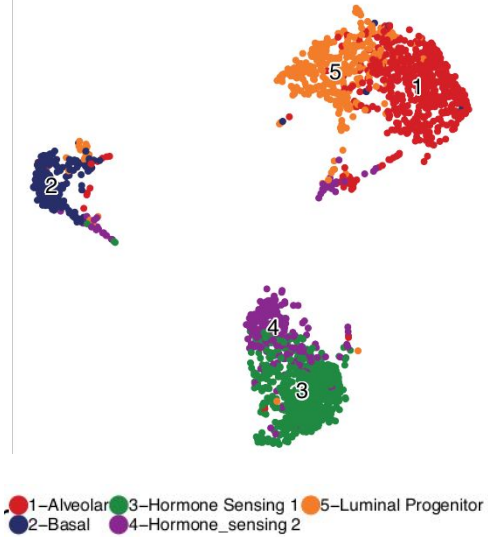
H3K4me1 of Juxta tissues



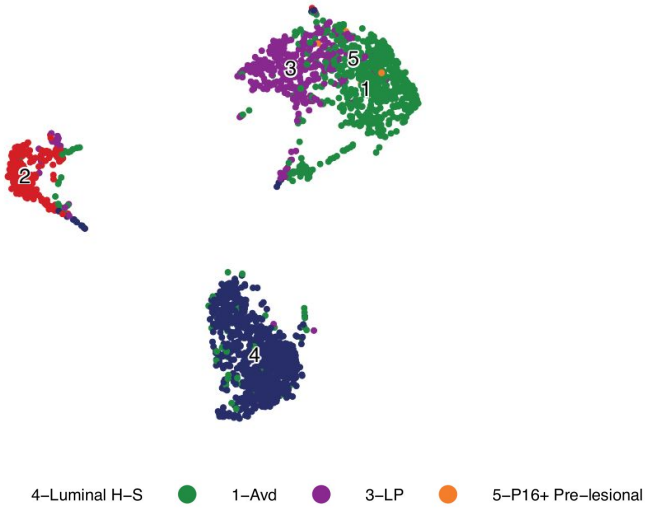


Label transfer of scRNA annotation on the epithelial compartment using ArchR

Annotation of Epithelial Cells

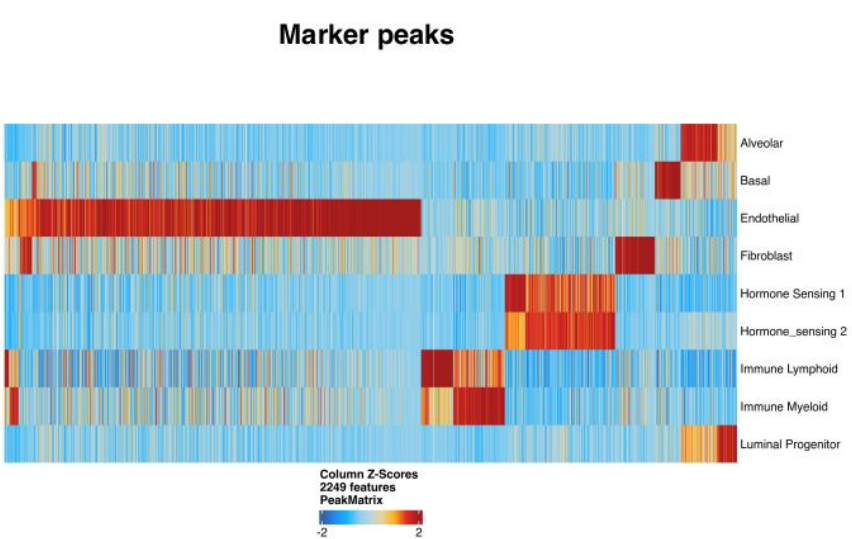


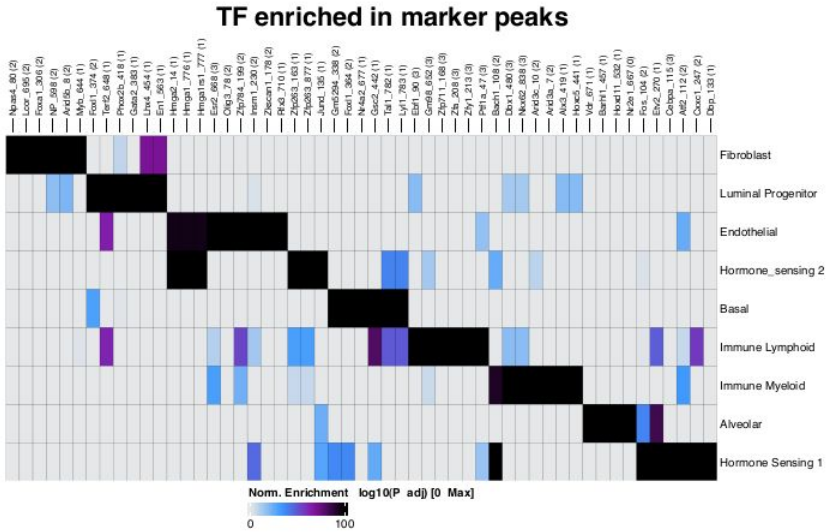
Label Transfer from scRNA



Best ways of integrating

H3K4me1 of Juxta tissues





Problem → size of peaks might be too large to enrich relevant TF

Conclusions

- The H3K4me1 landscape defines with good granularity the cell types in juxta-tumoral tissue

- Using different tools can be a generally good idea for exploration of a new dataset for example to identify doublets

- To overcome sparsity one can use larger bins, imputation and cluster coverage

4 - Iterative Differential Clustering

Biologically relevant unsupervised clustering

Tool in development, looking for feedback !

What is a ‘cell cluster’ ?

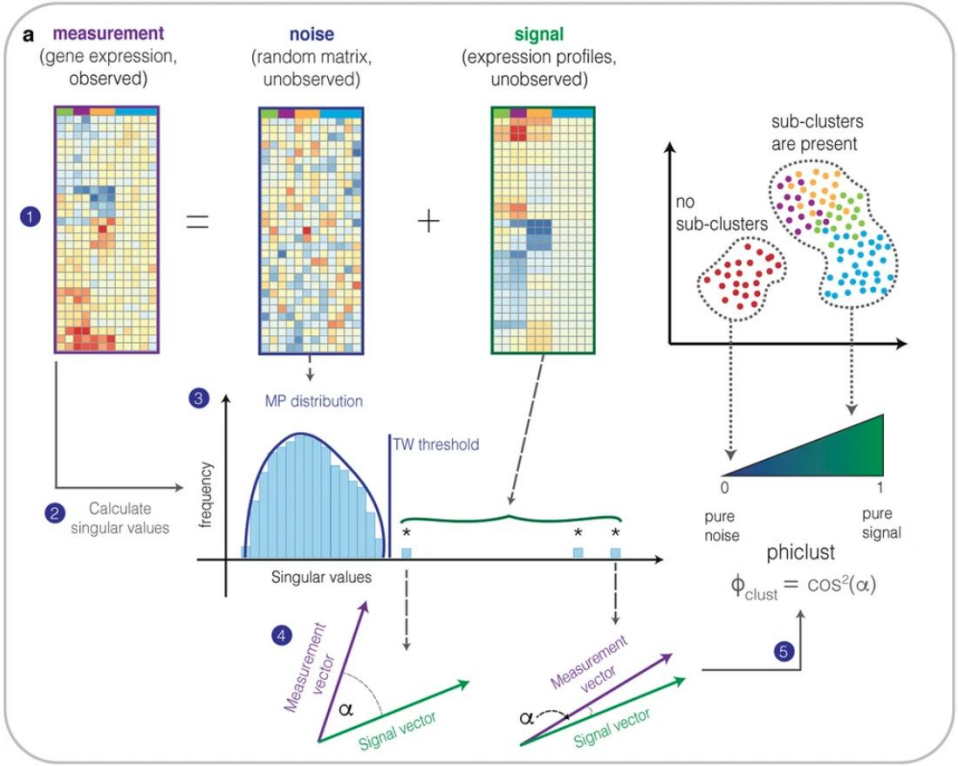
What is a ‘cluster’ ? Phiclust

Phiclust: *a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations (Mircea et al., 2022)*

→ Gives a score cell clusters with non-random substructure using random matrix theory and PCA

→ The score enable users to decide if the cluster can be safely divided into additional subclusters or not

cons: ressource intensive (time x memory)



What is a ‘cluster’ ? TooManyCells / TooManyPeaks

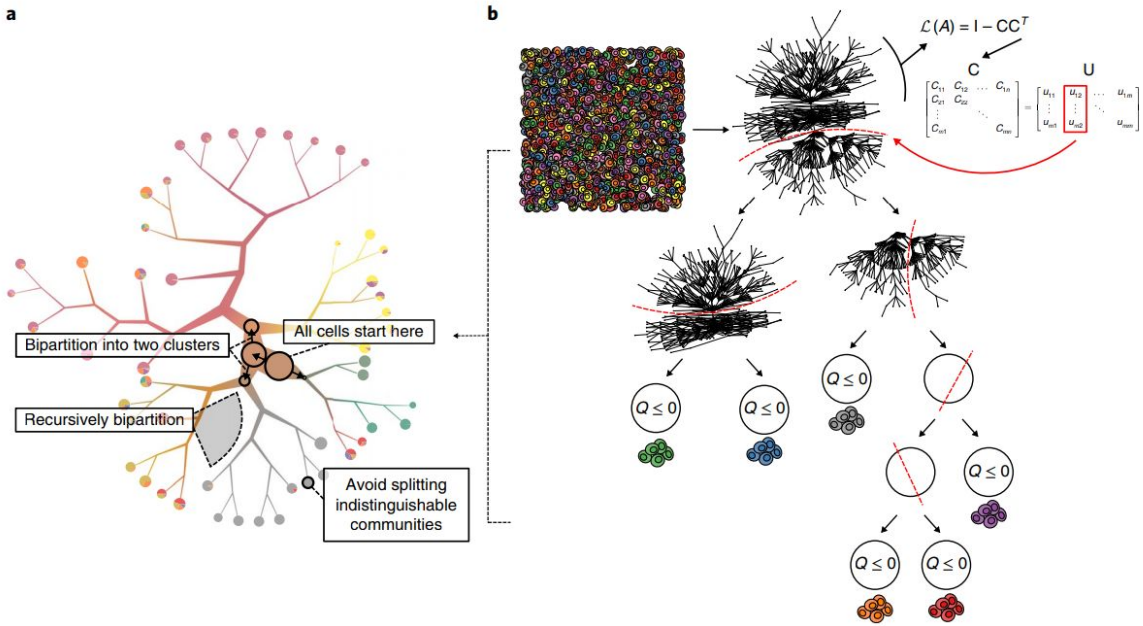
TooManyCells *identifies and visualizes relationships of single-cell clades (Schwartz et al., 2020)*

→Hierarchical bipartition spectral clustering, runs LSI at each step.

→ Uses ‘Newman–Girvan’ modularity (Q) as a stopping criteria instead of an optimization parameter

→ Allows detection of rare populations not found by other algorithlms

cons: tends to produce high number of very small clusters

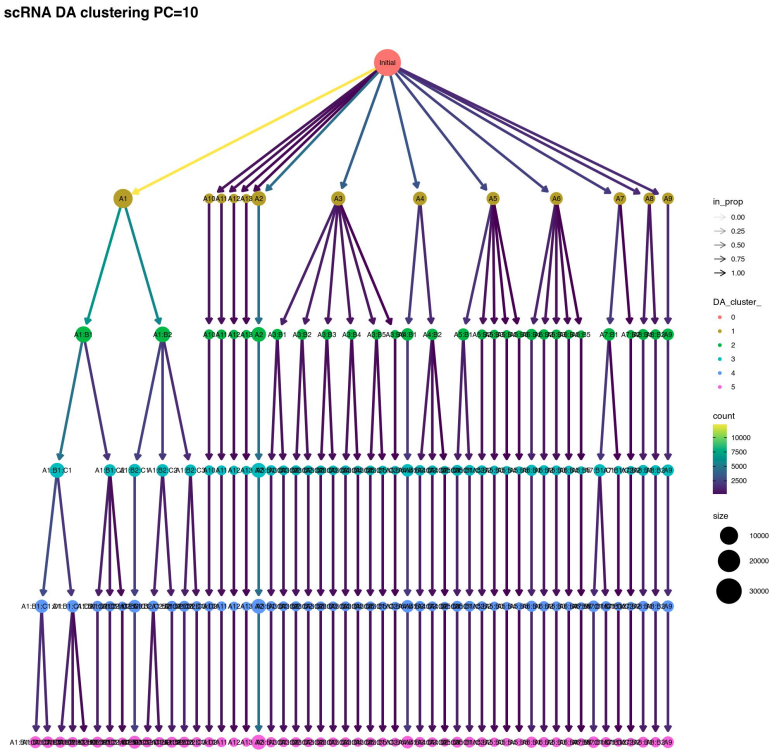


What is a ‘cluster’ ? Clustree

Clustree *enables visualization of clusters at different resolutions.*

→ If the clusters are stable when changing the resolution, this should mean that the clusters are stable

cons: Do not indicate the number of clusters to choose.



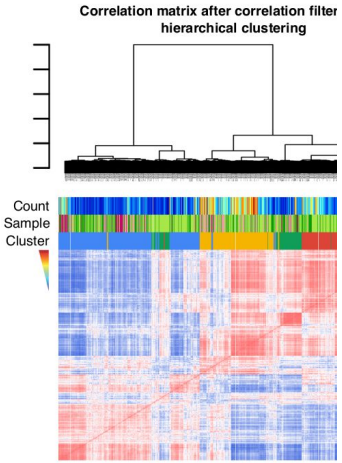
What is a ‘cluster’ ?

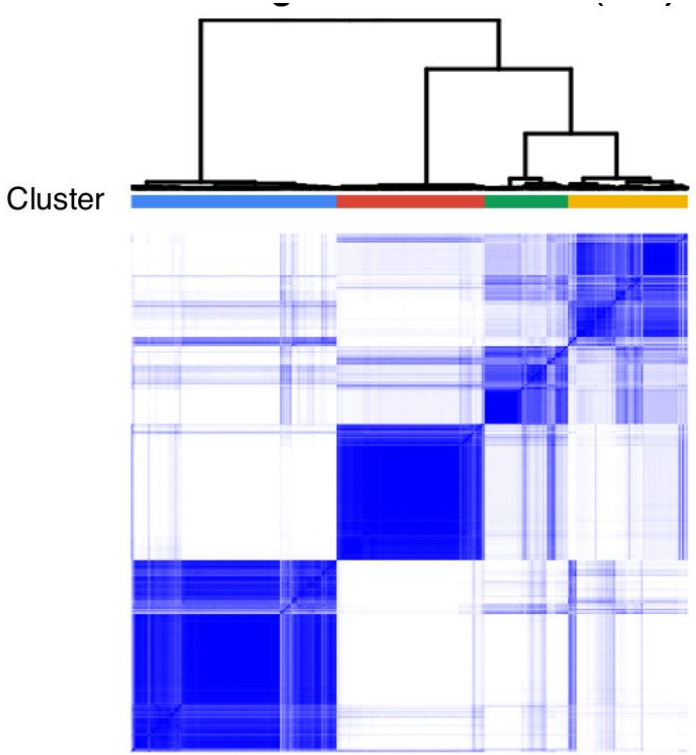
ConsensusClusterPlus *or* ***SC3*** *enables to bootstrap clustering to assess*

→ If the clusters are stable when subsampling the cells, this should mean that the clusters are stable

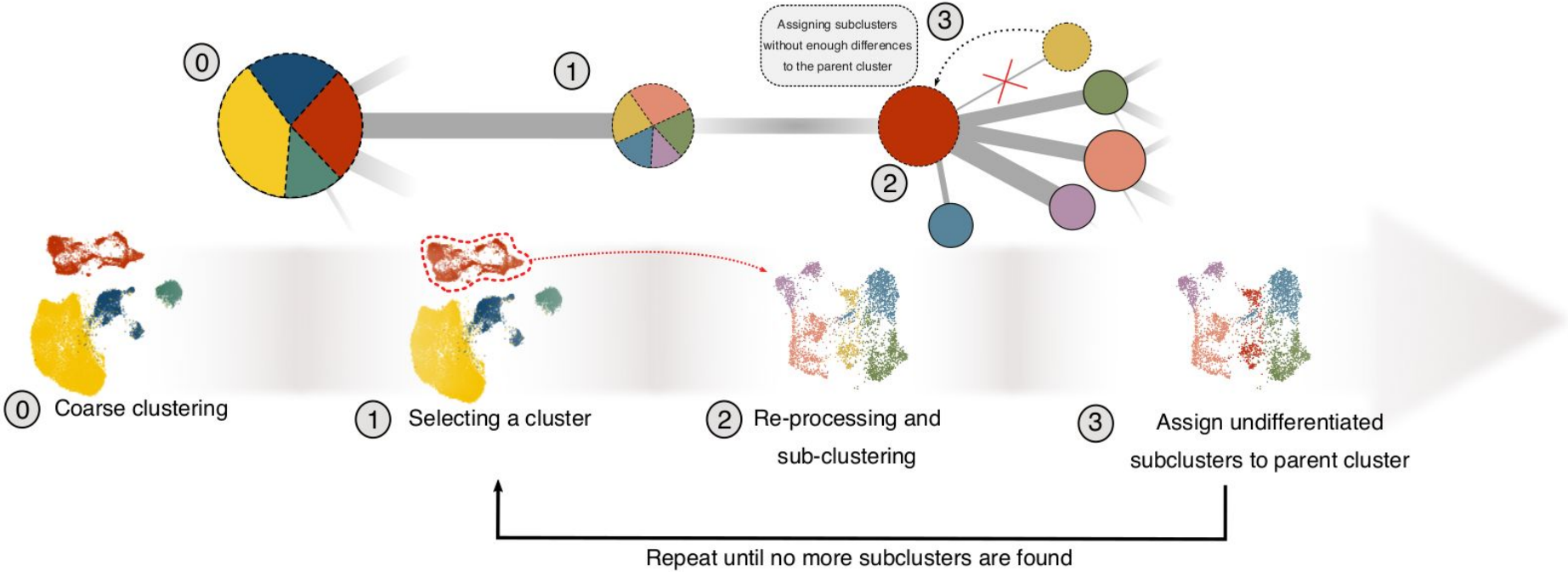
cons: ressource intensive (time x memory)

Consensus Clustering



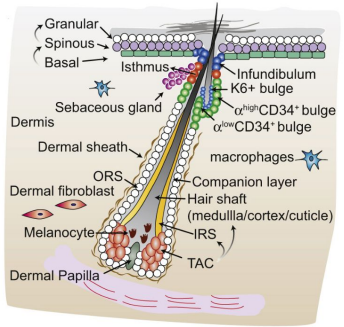


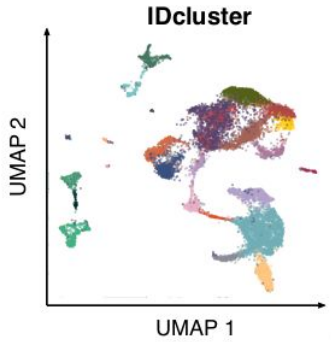
What is a ‘cluster’ ? Iterative Differential Clustering (IDclust)

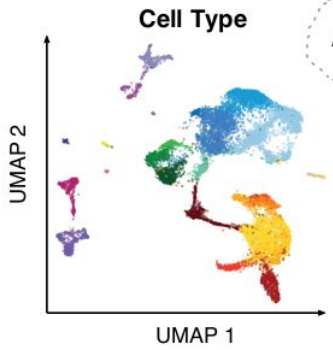


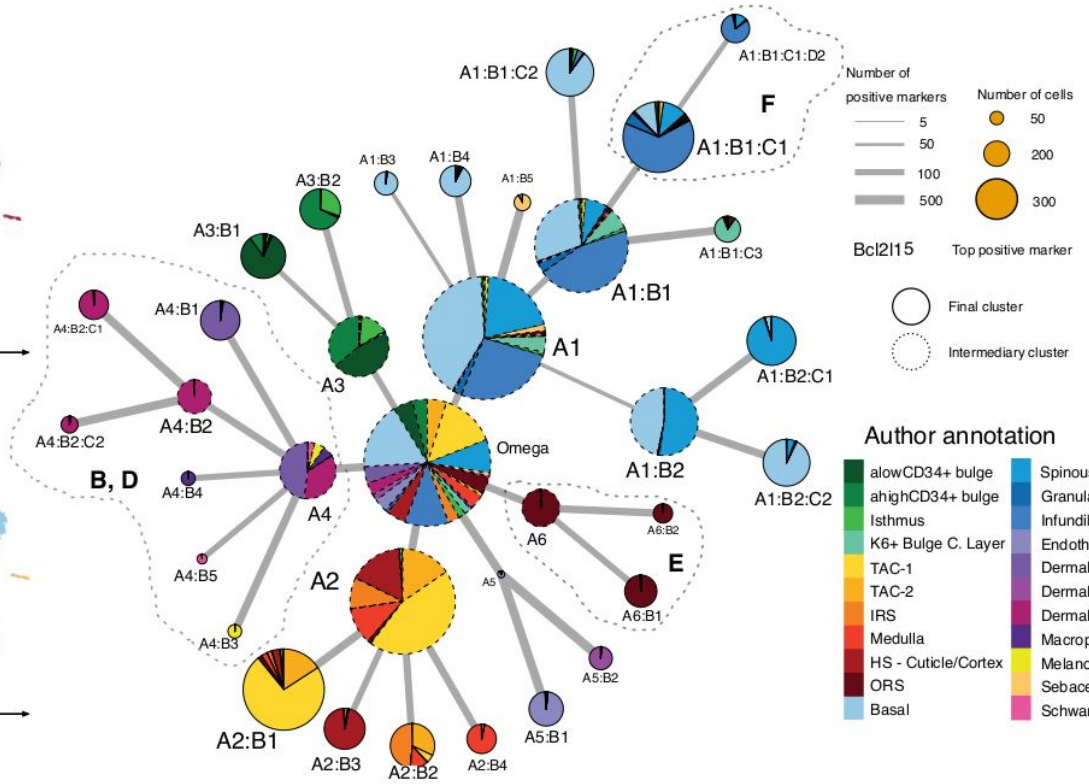
From abstract hyperparameters to biologically relevant parameters: logFC, q.value and % activation.

SHARE-seq of mouse skin - clustering scRNA



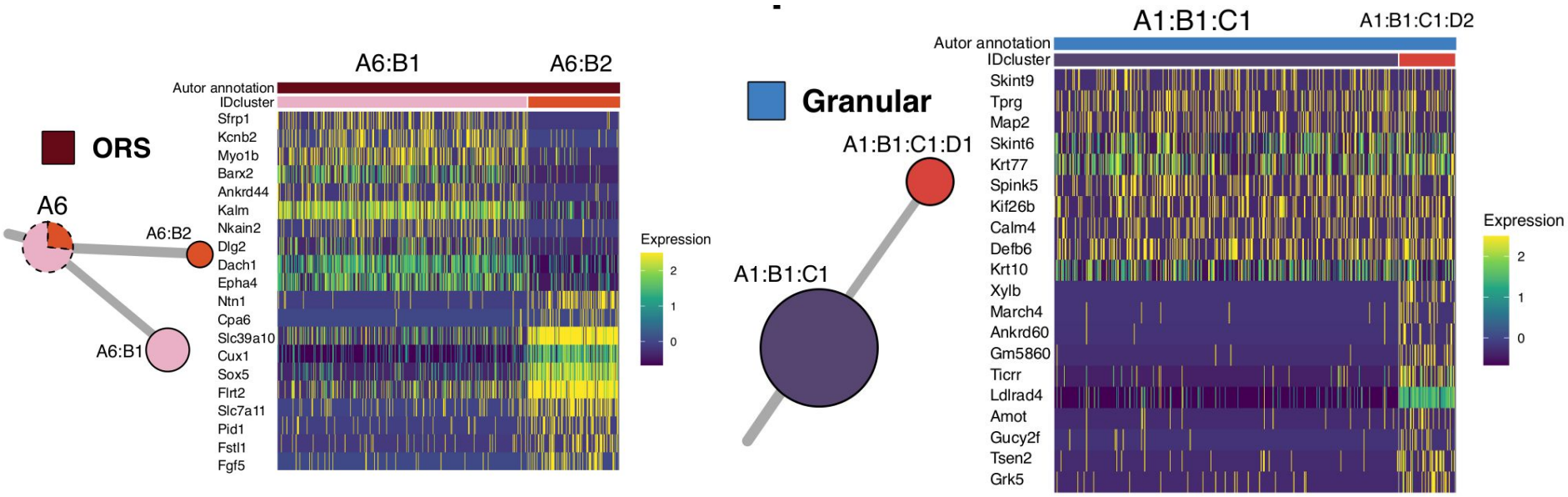




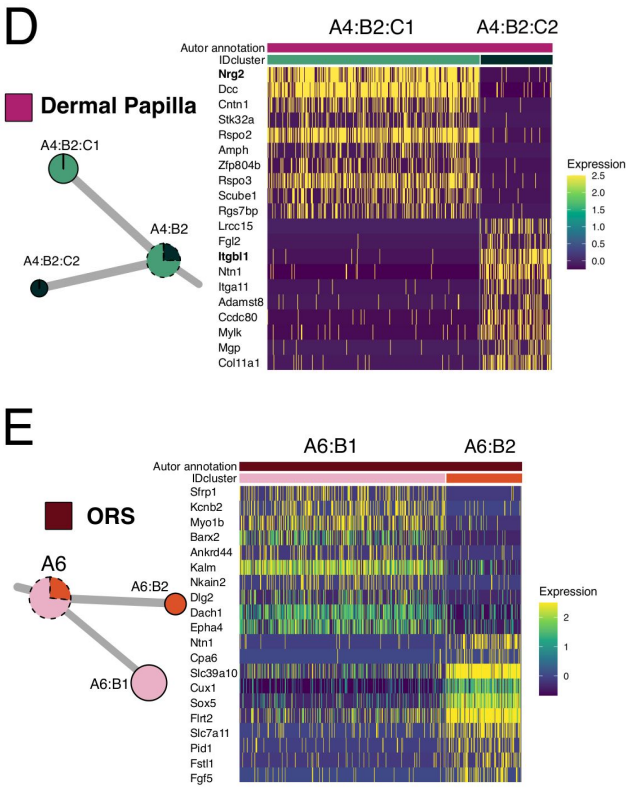


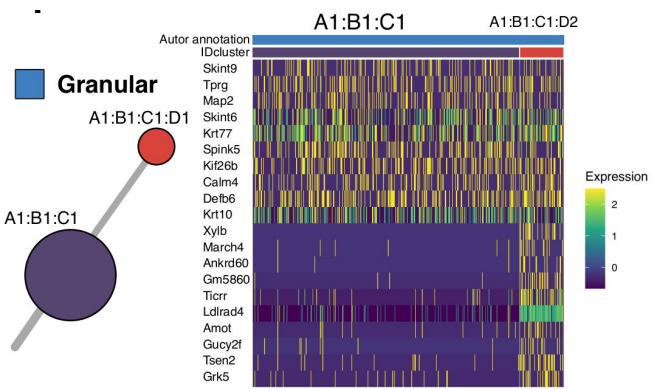
Ma & al.,
SHARE-seq

Different hierarchies of clusters

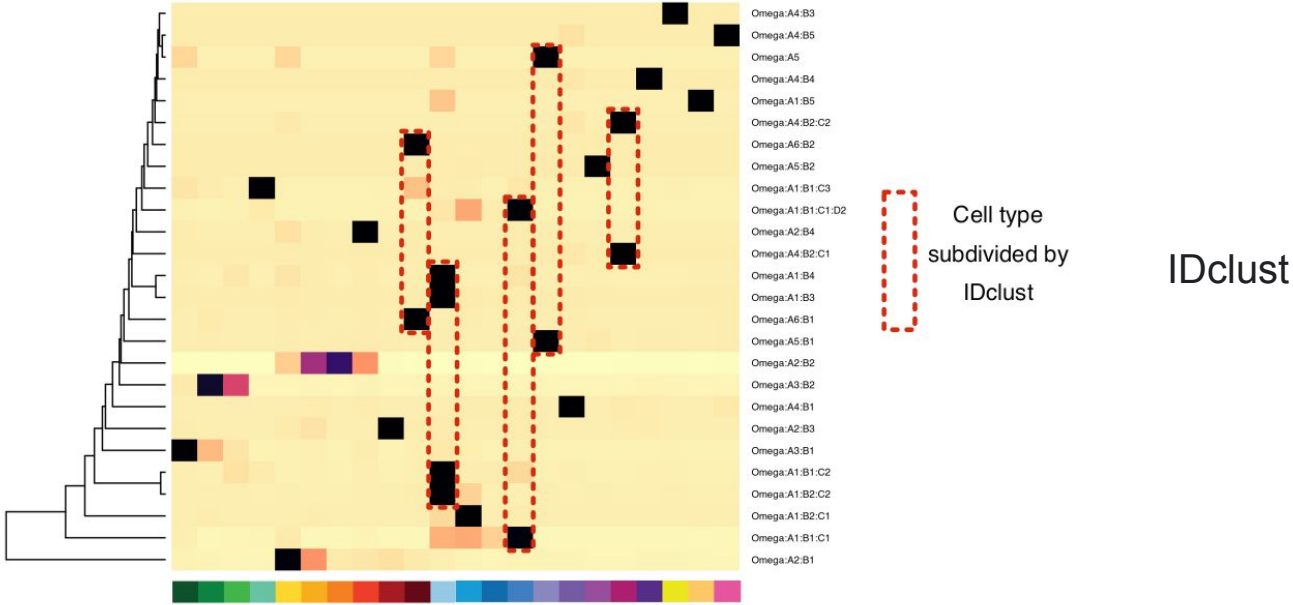


Different hierarchies of clusters

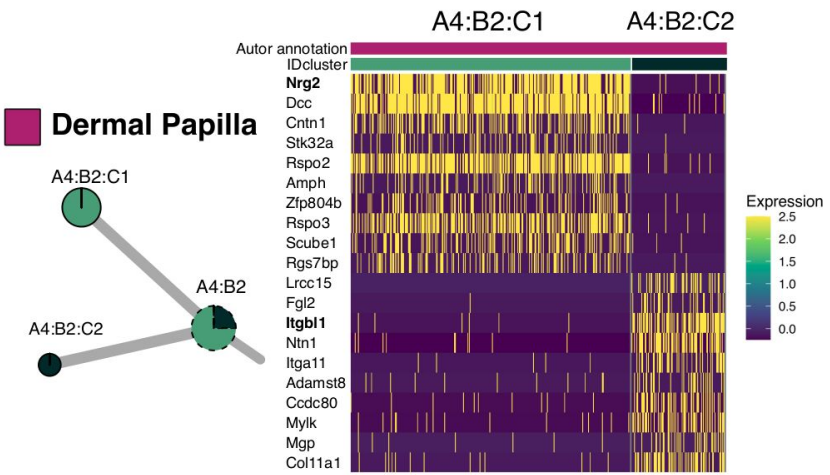


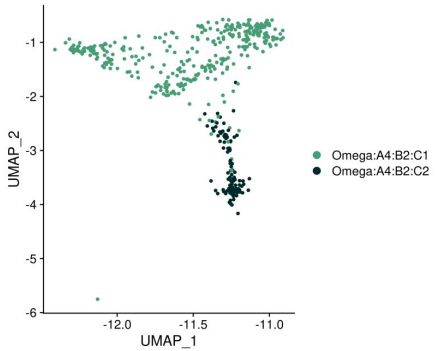


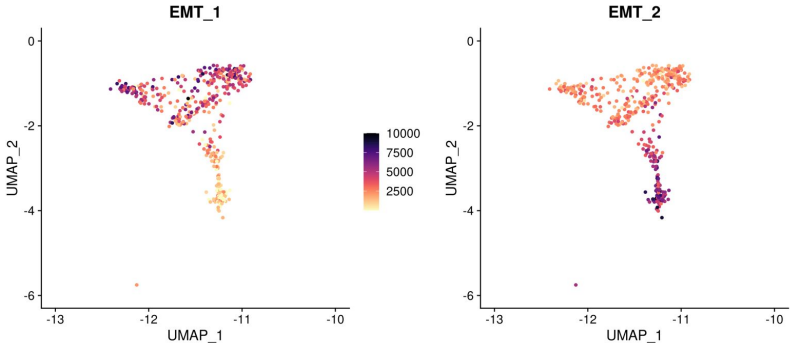
Dividing some annotated clusters into multiple



Two different EMT submodules activated in Dermal Papilla

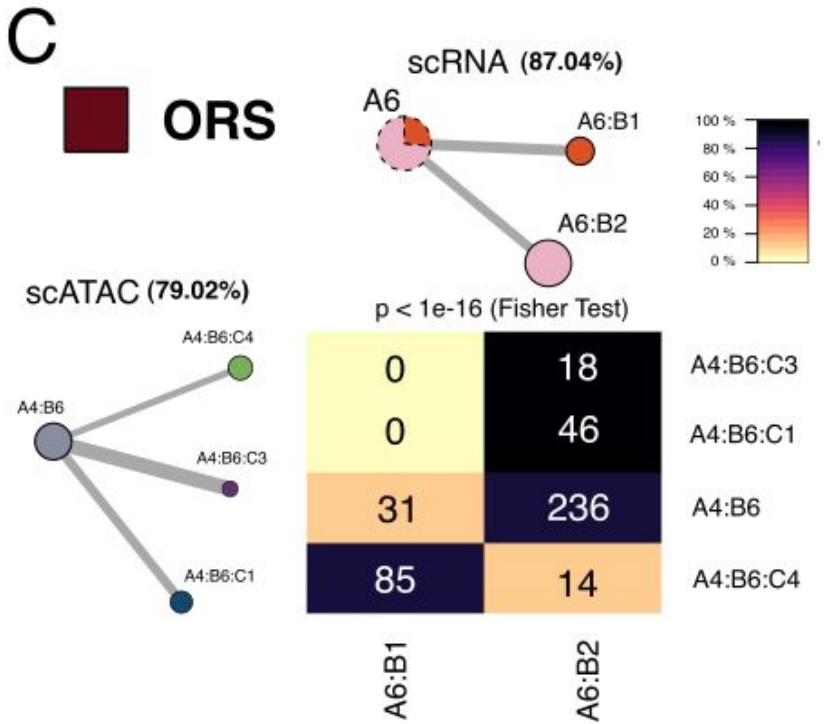




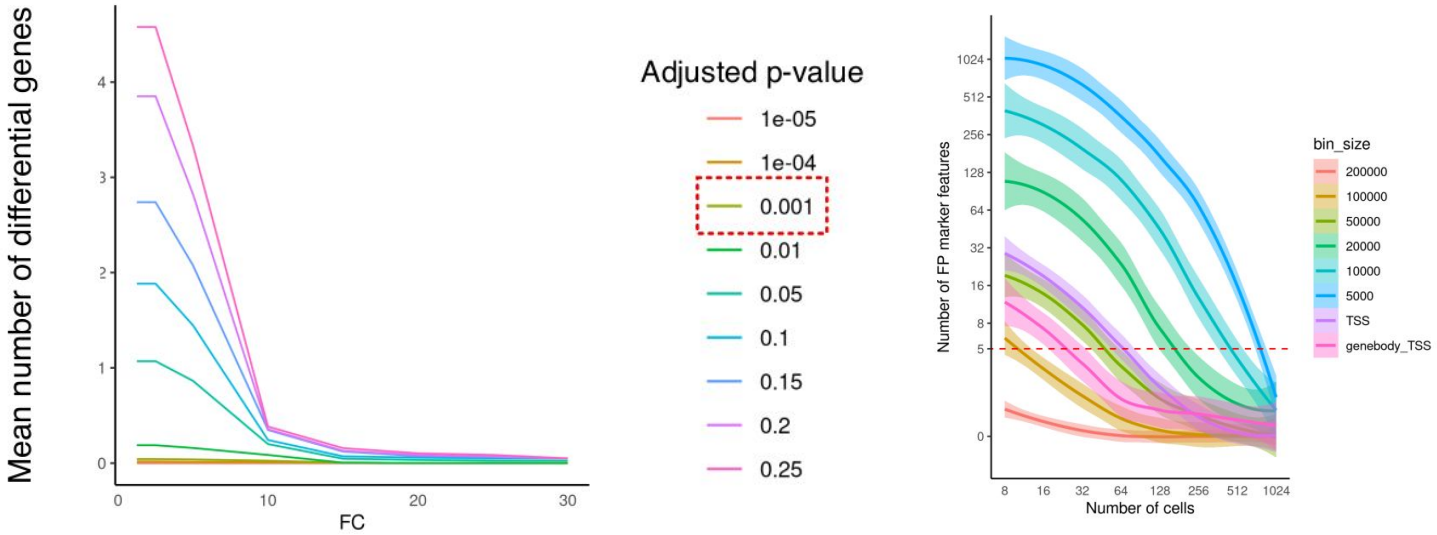


Using MAYA by Yuna Landais

IDclust can identify new consistent clusters in both modality (RNA & ATAC)



Benchmarking the differential analysis parameters (epigenomics)



- Running Differential analysis below 50 cells in the cluster produces high number of false positives

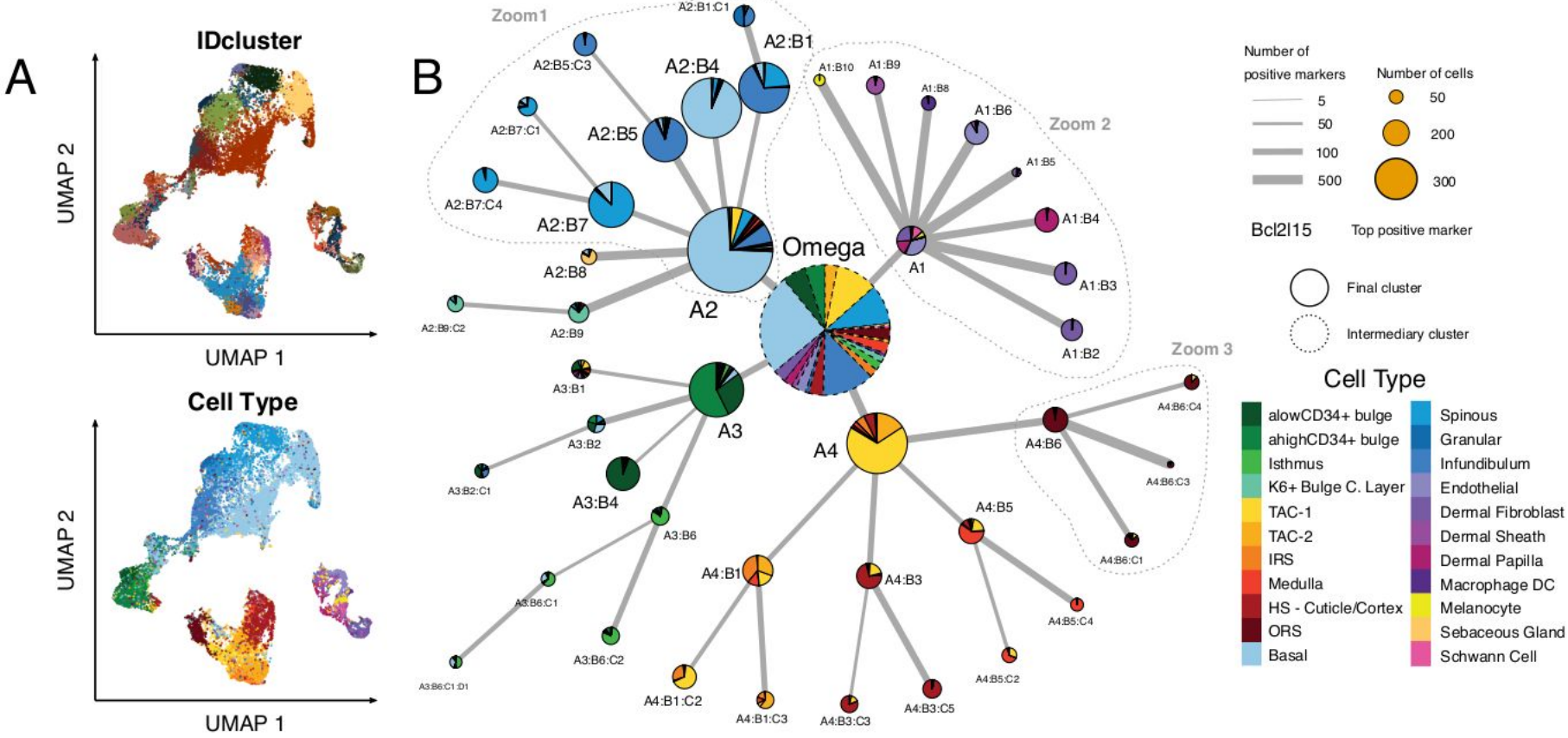
Conclusions

- IDclust is a method that finds cluster based on their differences
- IDclust can find different kind of hierarchies in the clusters

Limitations

- The framework is ‘as good’ as the differential analysis (e.g. limits in small clusters)
- The inner clustering parameters can still have an impact on the final results
- Effects such as cell cycle can influence the latest partition of the data

SHARE-seq - ATAC



PairedTag - H3K4me1