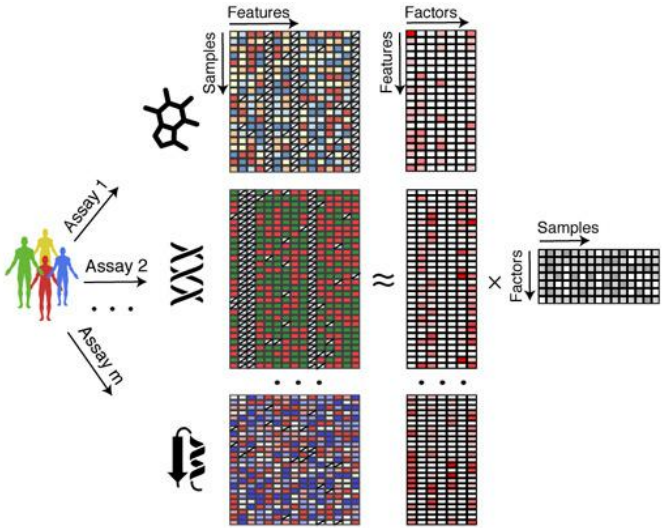
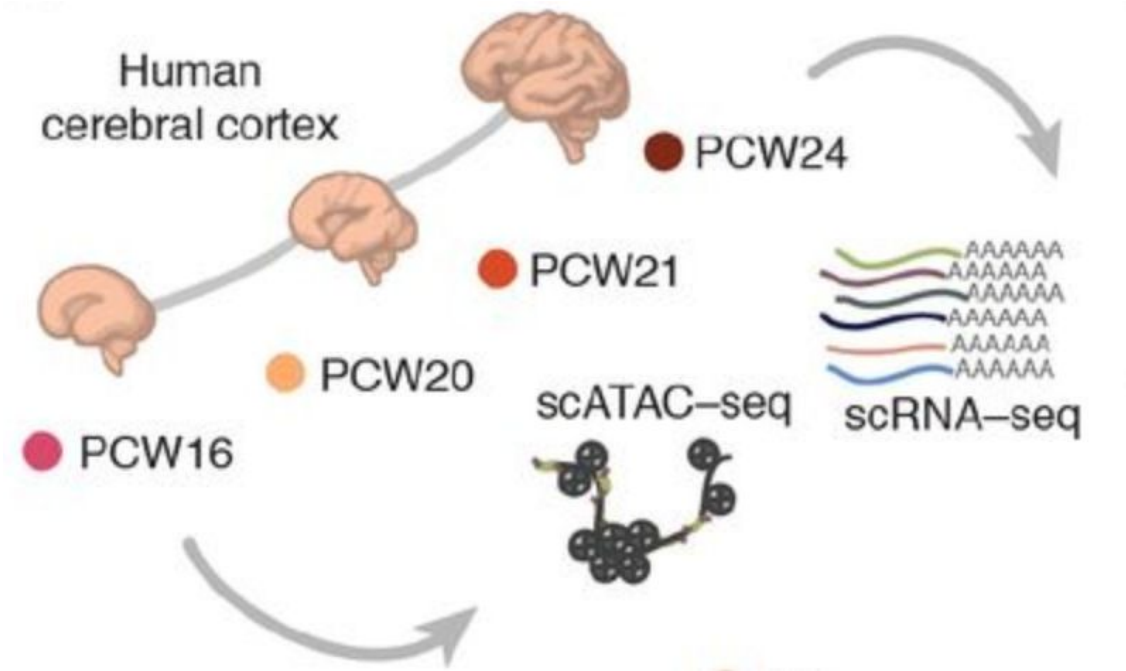
Analysis of matched multi-modal data

Multiomics group2



Dataset:

Brain cortex during development
(week 16 to week 24 post conception)



Dataset used: 10x genomics multiome (scRNA-seq and scATAC-seq) data of developing brain cortex from Trevino et al.2020, focusing on **differentiation of excitatory glutamatergic neurons.**

Aims and first impressions:

Aims:

We focused on integrating two modalities of omics data for the same subset of cells (vertical integration).

We used Muon, which offers a data structure and a set of methods (eg. MOFA and WNN) to perform the integration of both modalities (here we used scRNAseq and scATACseq).

We aimed at understanding whether the joint integration performed better or worse than single modalities alone and compared different integration methods.

We aimed at identifying non-coding genomic regions where chromatin accessibility is associated with expression of genes involved in excitatory neuron development.

First impressions:

The 3 data format (eg. AnnData, Sce, Seurat), all have certain limitations when it comes to keep track of the parameters specific to the integrated object hence the efforts from the Muon team to add an extra layer on top of the AnnData objects. (memory limitations)

The main bottleneck was the heavy juggling of file formats and tools to query the different layers of the data. We lost a lot of time trying to get third party tools to work. We learned a lot !

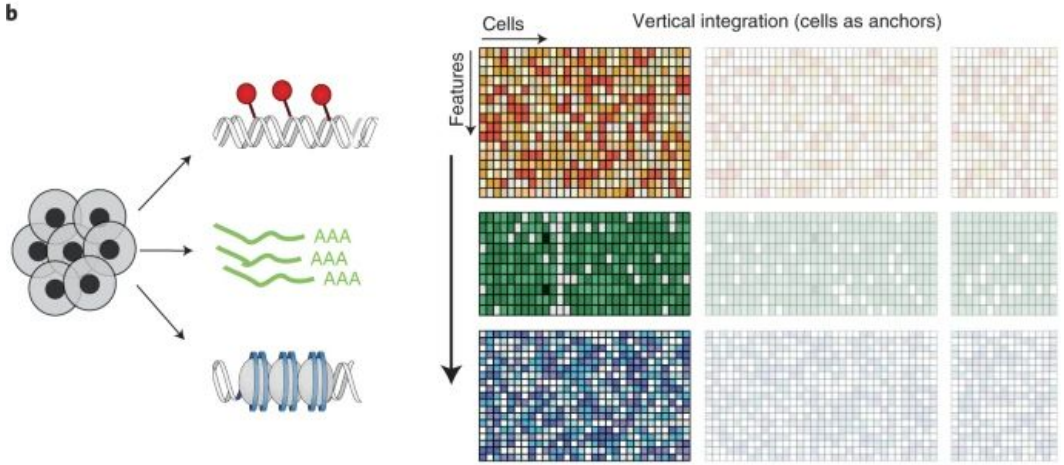
Walkthrough:

1. Preprocessing of single modalities
2. Joining modalities in a common muon object

   In order to reduce the complexity (and size) of the joint object (28000 genes x 440 000 peaks), a filtering step would be recommended (or not ?)

3. Filtering cells
   a. Filtering cells based on a pre-computed Glutamate-trajectory from the RNA modality
   b. No filtering
4. Filtering peaks
   a. ChromVar
   b. No filtering
5. Joint embedding of the two modalities
   a. Weighted nearest neighbour embedding (WNN)
   b. Multi-Omics Factor Analysis (MOFA)
6. Pseudotime inference and identification of marker genes
7. Compare different joint embeddings
8. Selecting features for chromatin-expression association
   a. Gene selection using CellRank or using marker genes identified from the pseudotime inference step
   b. Peak selection using Cicero or using peaks present in the vicinity (50kb) of the marker genes selected
9. Correlation study on the relevance of peak-gene association

Vertical integration



Strategies:

- Multi Omics Factor Analysis (MOFA) (The inferred latent factors represent the underlying principal axes of heterogeneity across the samples)
- Weighted Nearest Neighbour Analysis (defines a neighbourhood graph for the samples across different feature sets (modalities)

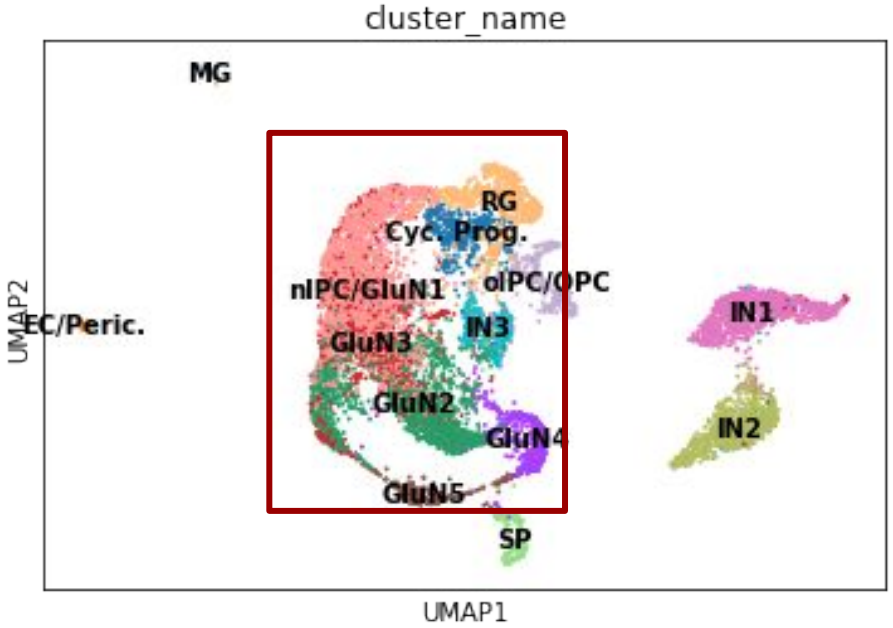
Common embedding of cells that combines information from all modalities

Q1: How much do we gain from this common embedding (for downstream analysis)?

Dataset used: scRNA-seq and scATAC-seq data of developing brain cortex from Trevino et al.2020, focusing on differentiation of excitatory glutamatergic neurons.

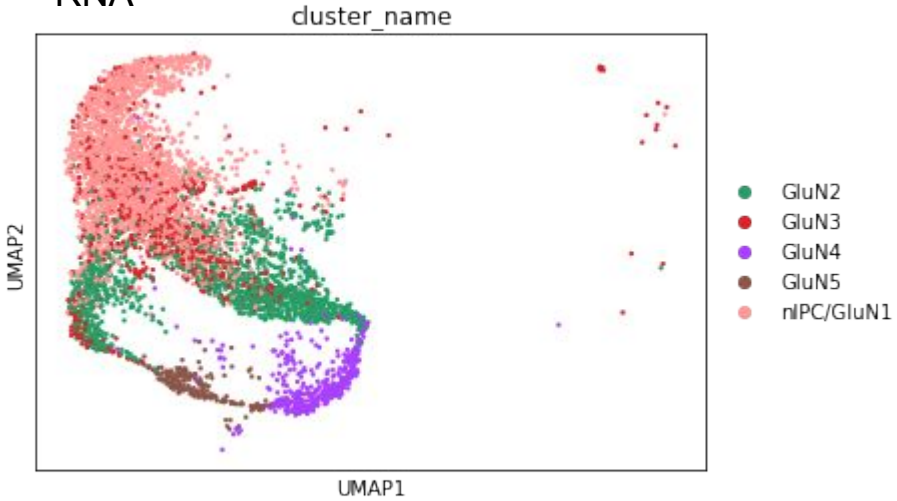
Results

Glutamatergic neuron differentiation in human brain cortex

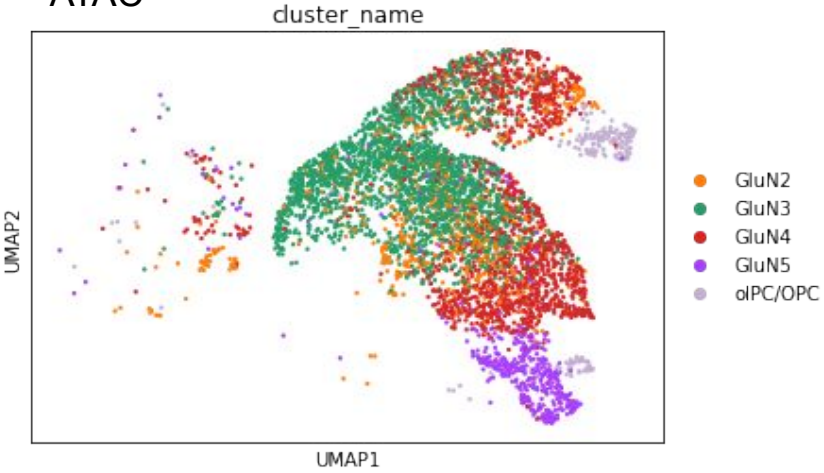


Embedding from RNA or ATAC modalities separately

RNA

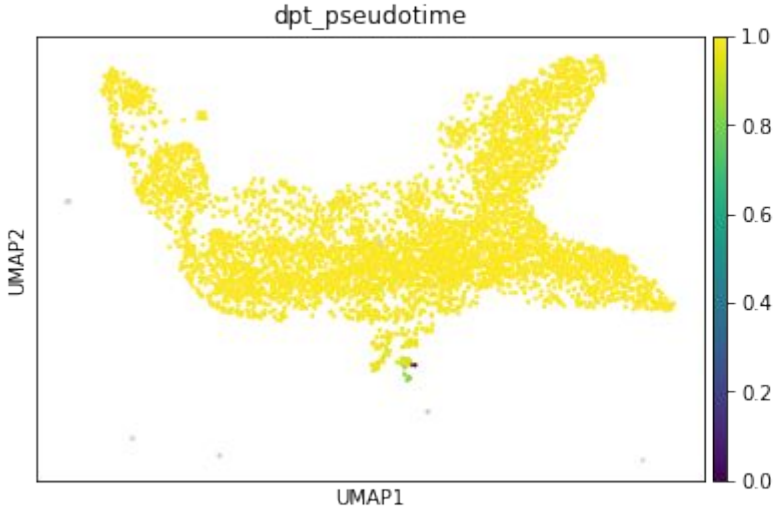


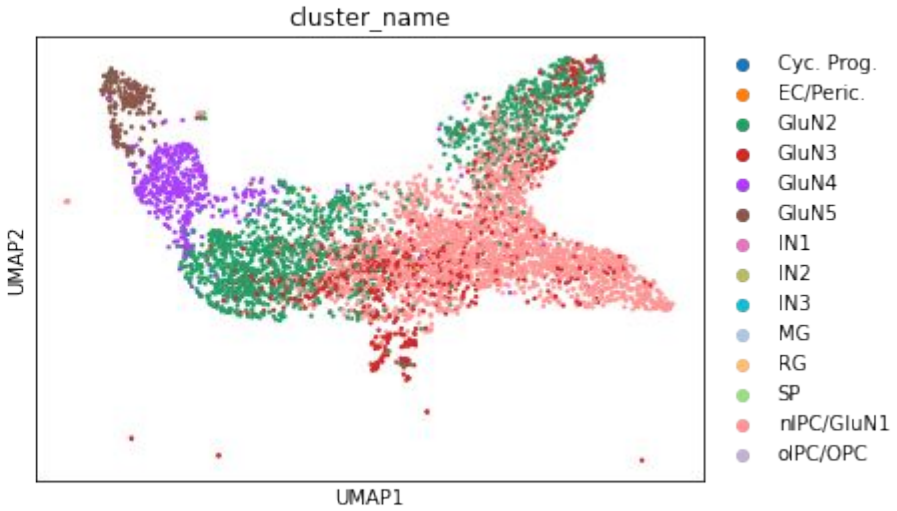
ATAC



Dense clusters forming narrow trajectories, individual outlier cells

Weighted nearest neighbour embedding (unfiltered data)

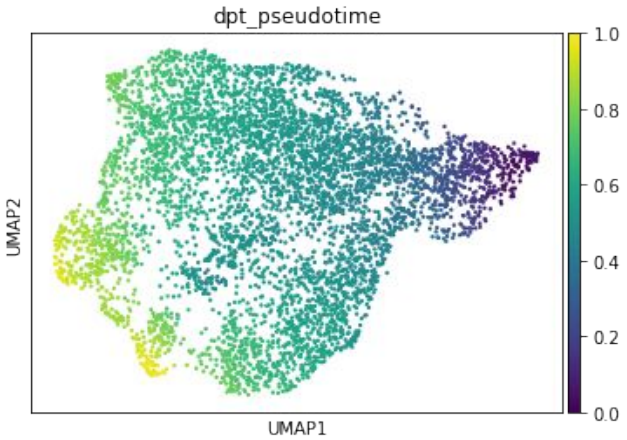


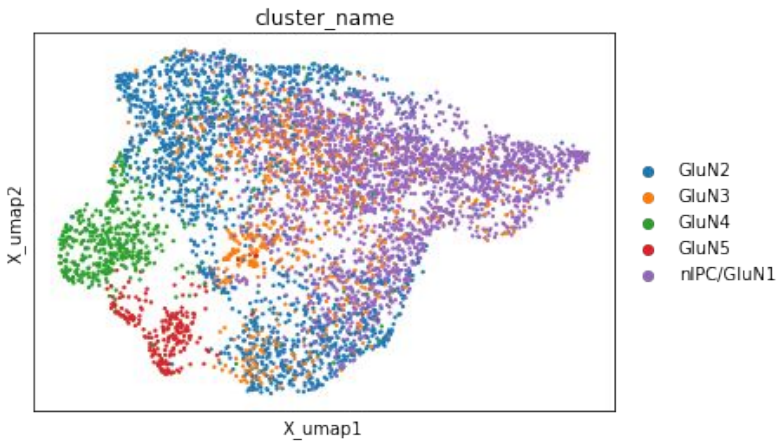


Data: joint embedding RNA_ATAC

Pseudotime: diffusion map

Weighted nearest neighbour embedding (filtered data on glutamatergic excitatory neurons (GluN)





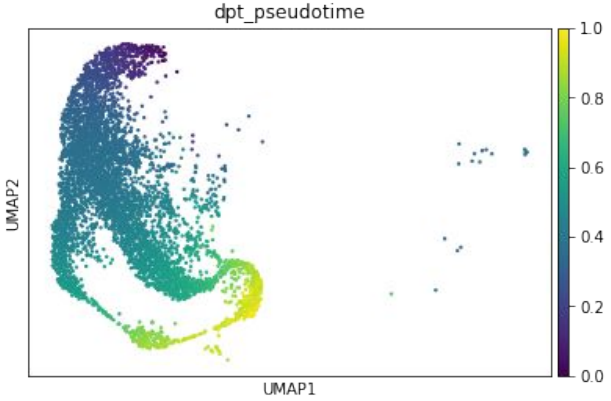
Cells are more evenly spread, no outlier cells (filtered?)

Data: joint embedding RNA_ATAC

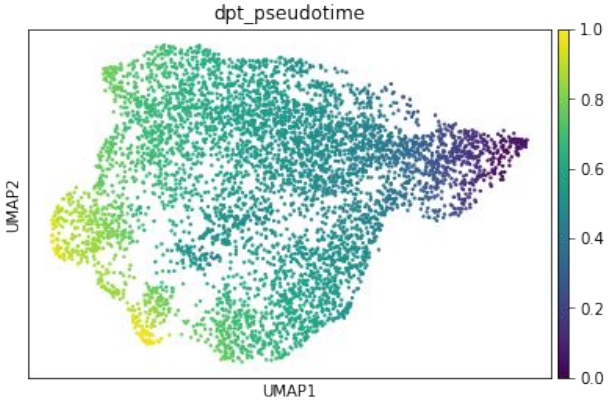
Pseudotime: diffusion map

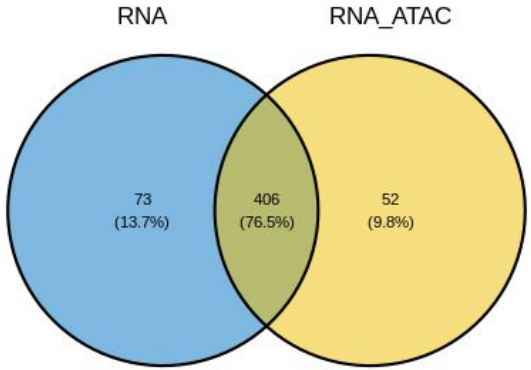
Comparing gene expression along pseudotime

RNA



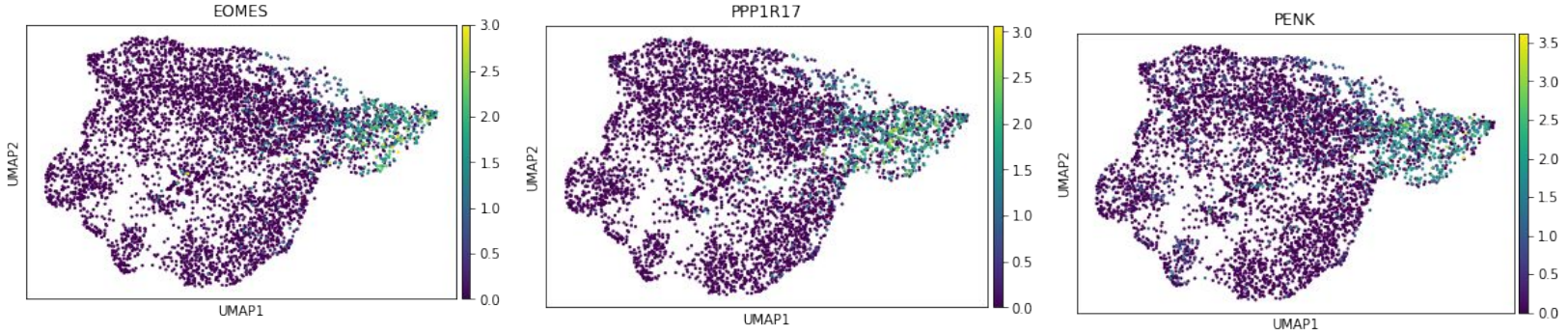
RNA+ATAC





Gene overlap between the two trajectories

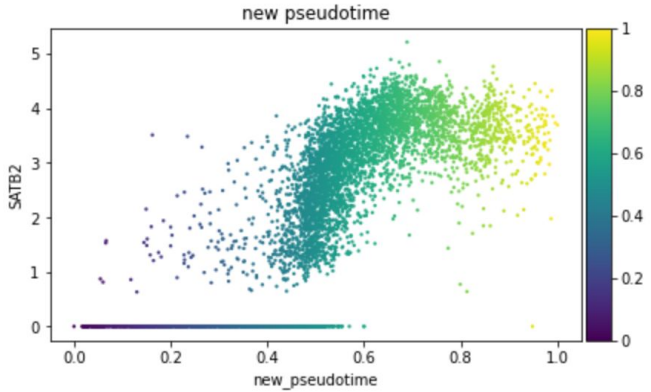
Mapping unique genes RNA+ATAC

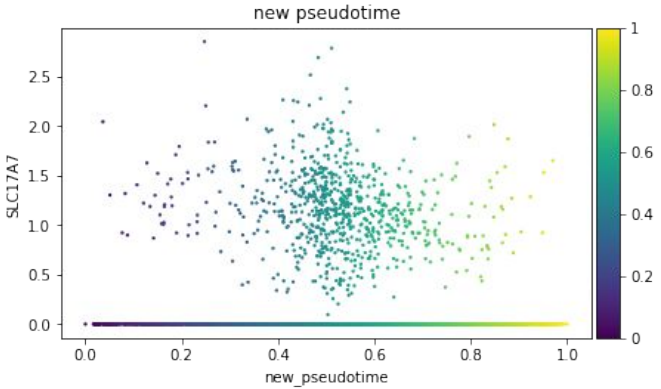



```
marker_genes = {
    "nIPC":['EOMES', 'PPP1R17', 'PENK', 'NEUROG1', 'NEUROG2'],
    "GluN":['NEUROD2', 'TBR1', 'BCL11B', 'SATB2', 'SLC17A7']
            }
```

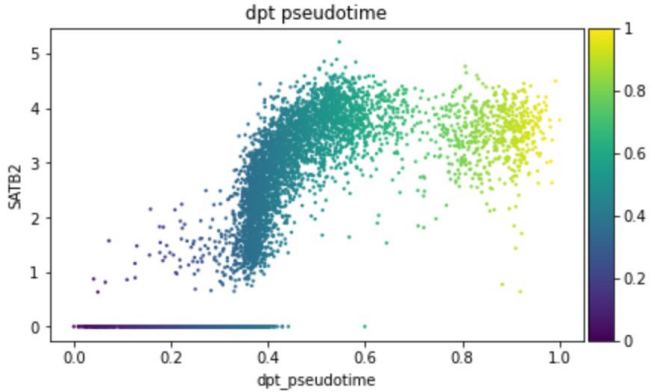
Glutamatergic neuron specific marker gene expression along pseudotime

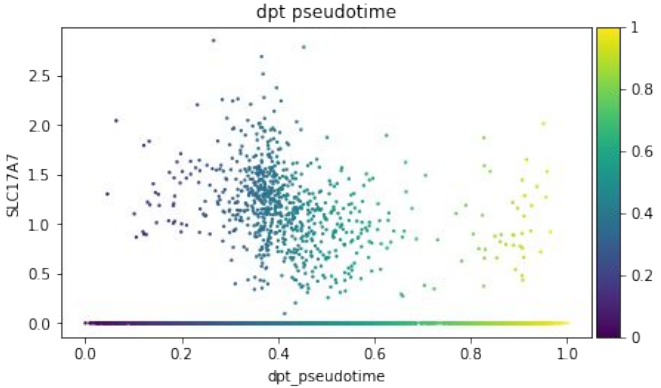
RNA+ATAC





RNA



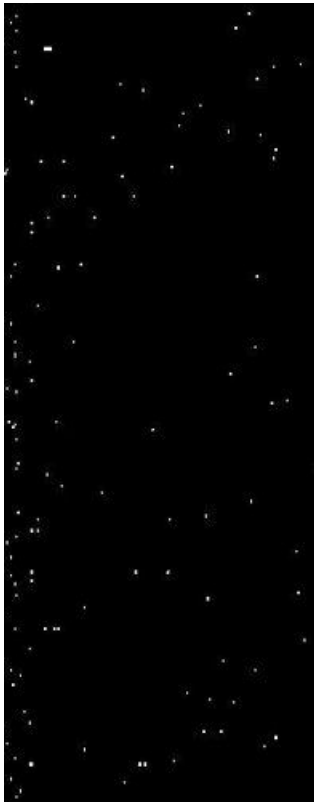


New pseudotime: pseudotime in joint embedding RNA+ATAC

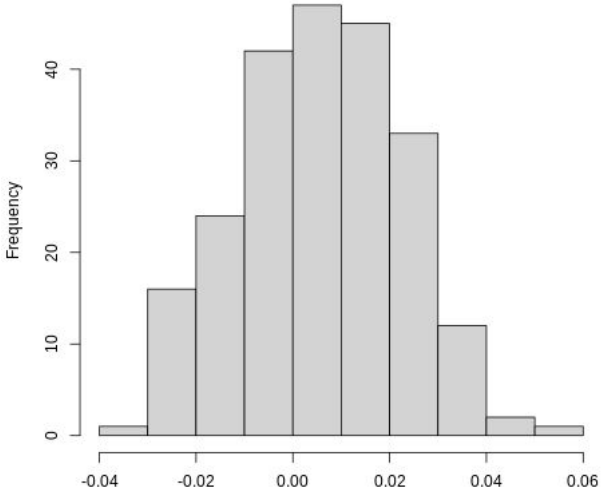
dpt_pseudotime: pseudotime in RNA embedding

scATAC-seq specific challenges

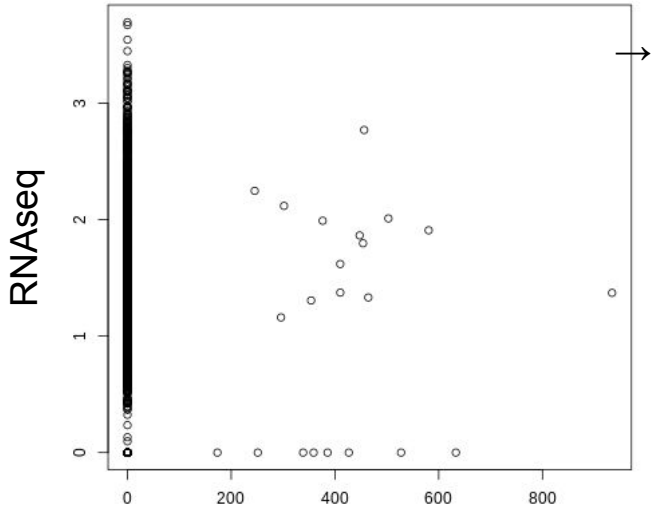
- Many features (>400,000) but extremely sparse
- Not feasible for many tools i.e. MOFA (exceeded memory)
- Different filtering approaches
  - Mapping statistics difficult (e.g. area that is always open is not interesting)
  - Available tools, i.e. ChromVAR did not run
    - futile time consuming installation attempts
  - Provided with motif matches (by Emma)
    - Selected genes that are targets of variable TFs (in scRNAseq data)
      - Still comprise ~55% of all peaks
    - Filtering peaks by number of motifs that are distant from gene of interest.
    - Merging peaks on proximal genes
      - Still too sparse for peaks



Correlation of chromatin peaks with genes from the joint embedding



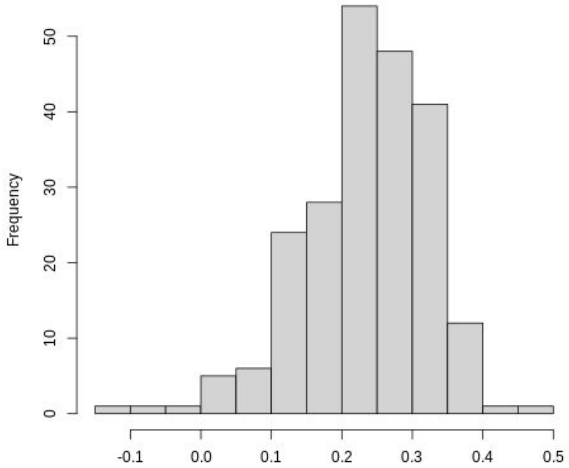
Correlation of ATAC and RNA



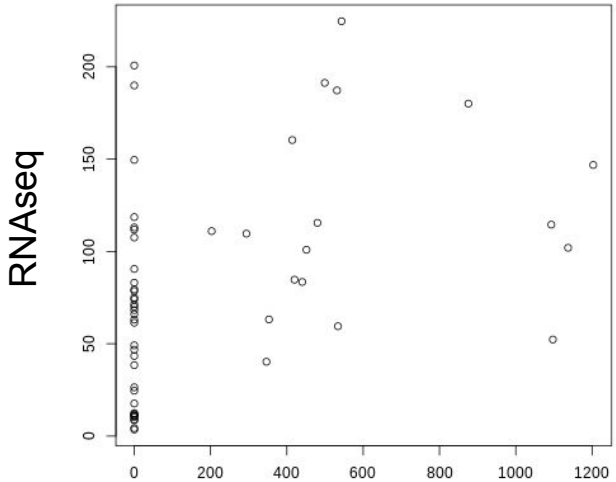
Merged peaks

→ meta cells needed!

Correlation of chromatin peaks with genes from the joint embedding



Correlation of ATAC and RNA



Merged peaks

Conclusions

- **There is no standard for how to handle the joint-modalities objects**
- **Size scales up very quickly when integrating multiple modalities (Memory issues)**
- **The contribution to the joint embedding of the different modalities may vary upon modalities and dataset**
- **The analysis pipeline will vary according to the modalities. Current implementations may not easily deal with the joint-embedding objects**

Acknowledgements

- SciLifeLab-SIB Summer School (organizers)

Mentors

- Charlotte Soneson
- Emma Dann

Thank you for your attention