# Multiomics project Diagonal integration

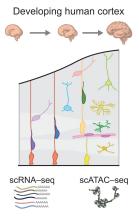
## Group 1

Anna Papiez, Gustavo Ruiz, Giada Sandrini, Giulia Protti

## Introduction

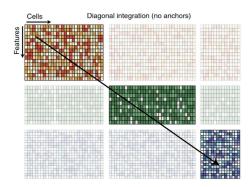
Data from different cells, comprising a single cell atlas of

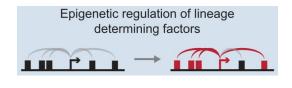
- Gene expression (scRNA-seq)
- Chromatin accessibility (scATAC-seq)



Aims of the analysis:

- 1. To perform diagonal integration of unmatched scRNA-seq and scATAC-seq data.
- 2. To associate gene expression to accessibility in the developing human cortex.





Trevino, et al., Cell, 2021

## Pre-processing of separate datasets

#### scRNA-seq data:

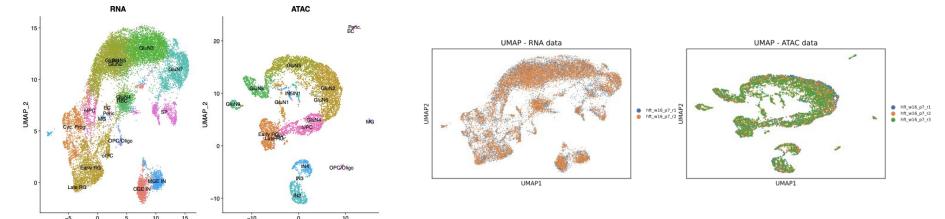
- Cells were already filtered by % of mito. and total counts.
- Filtered genes (min cells = 3).
- Normalized and log-transformed the raw counts.
- Identified highly variable genes for dimensionality reduction.
- Performed dimensionality reduction with PCA and computed a KNN graph.

UMAP

- The clusters were previously defined.

#### scATAC-seq data:

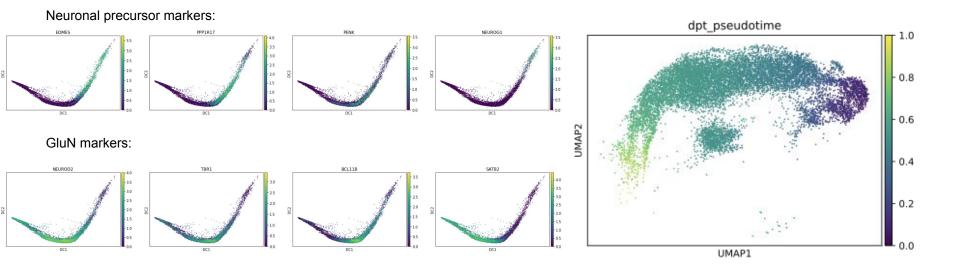
- Some QC metrics were already calculated and used for filtering cells.
- Binarized the data matrix due to sparseness.
- Filtered peaks accessible in < 10 cells.
- Performed dimensionality reduction with Latent Semantic Indexing (LSI) and computed a KNN graph.
- The clusters were previously defined.



## **Diffusion pseudotime**

Aim: estimate the order of excitatory neurons along a differentiation trajectory with diffusion pseudotime.

- Subsetted the data to include glutamatergic neurons (GluN) only.
- Performed dimensionality reduction with PCA and computed KNN graph.
- Checked the expression of marker genes along the diff. traj.
- Defined a putative root cell (max of DC1) and plotted diffusion trajectory for GluNs.



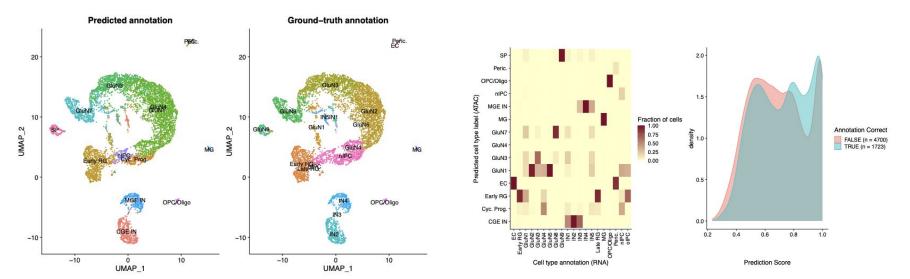
## Co-embedding

Aim: annotate cells from the ATAC dataset exploiting the labels of RNA data

How: We infer the cell type annotations for "ATAC cells" through Seurat CCA approach

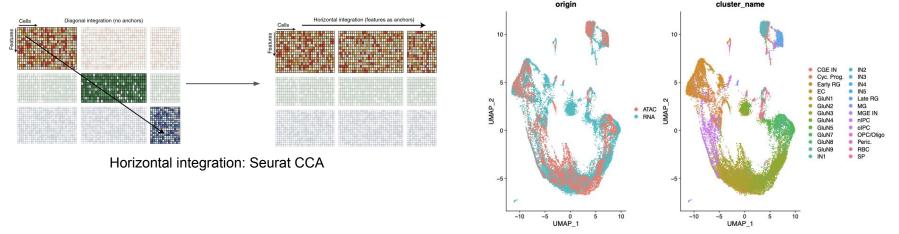
- Estimation of the gene activity based on ATAC data
- Identification of the anchors
- Transfer the metadata (cell annotations)

Comparison between the estimated cell types and the true annotation



## Co-embedding: from diagonal to horizontal configuration

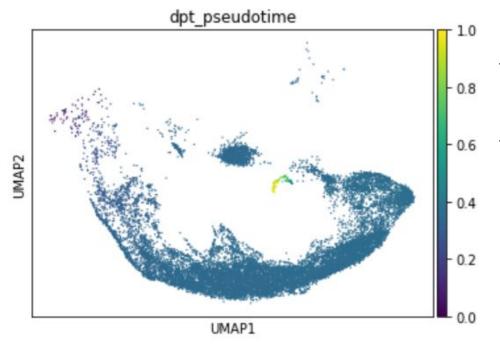
Aim: horizontally integrate the data to have a common embedding space, thus to have gene expression values also for ATAC cells



How: we impute the gene expression value for "ATAC cells" exploiting Seurat CCA approach

- Estimation of the gene activity based on ATAC data
- Identification of the anchors
- Transfer the gene expression values

## Co-embedding to define a pseudotime ordering of differentiating glutamatergic neurons from nIPCs



- Once we have a common embedding, we can use standard similarity-based trajectory inference methods to order the excitatory neurons in pseudotime.
- We used the Diffusion Pseudotime implementation in sc.tl.dpt.

## Selecting features for chromatin accessibility-expression associations

After having identified a common embedding and a common pseudotime axis, we need to **select the features** that we will use to **associate gene expression to chromatin accessibility**.

Why the feature selection step?

- 1. Test all the peaks against all the genes is **computationally expensive + multiple test burden**
- 2. Long range interactions on the genome are not very common (no sense to test for associations between genes and chromatin regions that are extremely far apart e.g. on different chromosomes)

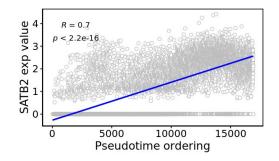


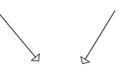


Aim: Select a subset of interesting genes that seem to have a dynamic behaviour in the differentiation trajectory

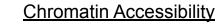
How: We **correlated** the log-normalized gene expression to the value of pseudotime (Spearman correlation)

SATB2: known marker of glun differentiation - highest correlation value





List of gene-peak pairs





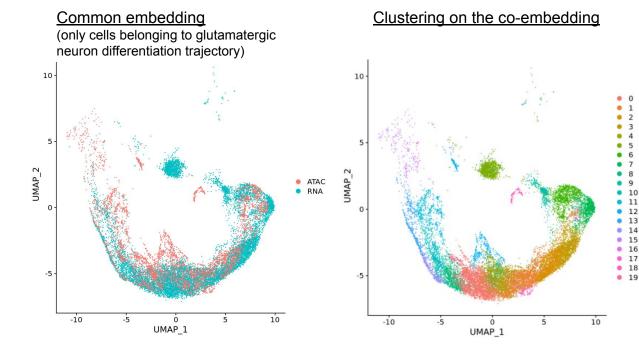
Aim: Select a subset of peaks that could probably be involved in gene expression regulation

How: We subset the possible gene-region pairs to regions within a certain range of the gene (100000 base pairs)

## Aggregating expression/accessibility profiles from multiple cells

Why this step?

- 1. To associate gene expression to accessibility (last step) we need the same cell-level unit
- 2. To deal with the high sparsity of the scATAC profiles
- 3. To prioritize the most robust associations
- 4. To reduce the computational burden of testing for associations



#### For each cluster, summarize expression/ATAC counts (AggregateExpression function from Seurat package)

#### **RNA** assay

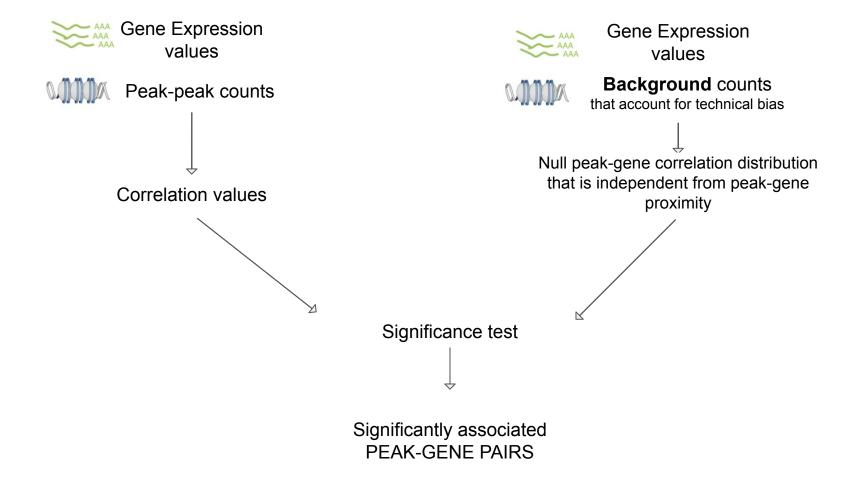
	Cluster 1	Cluster 2	Cluster	Cluster M
Gene1	Exp value			
Gene2				
GeneN				

#### ATAC assav

11

	Cluster 1	Cluster 2	Cluster 	Cluster M
Peak1	ATAC count			
Peak2				
PeakN				

### Associating gene expression to accessibility



## Functional pathway analysis

Reactome gene sets

Over-representation analysis of genes with significant correlations to accessible peaks.

#### REACTOME NERVOUS SYSTEM DEVELOPMENT REACTOME NEURONAL SYSTEM REACTOME L1CAM INTERACTIONS REACTOME NEUROTRANSMITTER RECEPTORS AND POSTSYNAPTIC SIGNAL TRANSMISSION GOBP CELL MORPHOGENESIS INVOLVED IN DIFFERENTIATION REACTOME TRANSMISSION ACROSS CHEMICAL SYNAPSES GOBPICELL JUNCTION ORGANIZATION GOBP SYNAPSE ORGANIZATION REACTOME ASSEMBLY AND CELL SURFACE PRESENTATION OF NMDA RECEPTORS SOBP CELL MORPHOGENESIS INVOLVED IN NEURON DIFFERENTIATION REACTOME PROTEIN PROTEIN INTERACTIONS AT SYNAPSES GOBP CELL PART MORPHOGENESIS Count GOBP HEAD DEVELOPMENT REACTOME ACTIVATION OF NMDA RECEPTORS AND POSTSYNAPTIC EVENTS Coun GOBP CELLULAR COMPONENT MORPHOGENESIS REACTOME MUSCLE CONTRACTION GOBP REGULATION OF CELL PROJECTION ORGANIZATION GORP AXON DEVELOPMENT REACTOME NEGATIVE REGULATION OF NMDA RECEPTOR MEDIATED NEURONAL TRANSMISSION p.adjust CORP. REGULATION OF CATION TRANSMEMBRANE TRANSPORT REACTOME RAS ACTIVATION UPON CA2 INFLUX THROUGH NMDA RECEPTOR GOBP REGULATION OF NEURON PROJECTION DEVELOPMENT 0.005 GOBP REGULATION OF TRANSPORTER ACTIVITY 0.010 REACTOME UNBLOCKING OF NMDA RECEPTORS GLUTAMATE BINDING AND ACTIVATION GOBP FOREBRAIN DEVELOPMENT 0.015 REACTOME LONG TERM POTENTIATION GOBP SYNAPSE ASSEMBLY 0.020 GORP NEURON PROJECTION GUIDANCE REACTOME CREB1 PHOSPHORYLATION THROUGH NMDA RECEPTOR MEDIATED ACTIVATION OF RAS SIGNALING GOBP REGULATION OF CATION CHANNEL ACTIVITY REACTOME CARDIAC CONDUCTION GOBP REGULATION OF SYNAPSE STRUCTURE OR ACTIVITY GORD REGULATION OF NEUROTRANSMITTER RECEPTOR ACTIVITY REACTOME SIGNALING BY NTRKS GOBP NEURON MIGRATION REACTOME INTERACTION BETWEEN L1 AND ANKYRINS -GOBP REGULATION OF NMDA RECEPTOR ACTIVITY 0.10 0.15 0.20 REACTOME NEUREXINS AND NEUROLIGINS -GeneRatio REACTOME REDUCTION OF CYTOSOLIC CA LEVELS REACTOME CRMPS IN SEMA3A SIGNALING 0.1 0.2 GeneRatio

GO: Biological process gene sets

### Summary of analysis steps

- Preprocessing of RNA-seq and ATAC-seq data separately
- Co-embedding from diagonal to horizontal integration (CCA)
- Order excitatory neurons along a differentiation trajectory (dpt)
- Feature selection:
  - Genes based on correlation to the trajectory
  - Peaks subset the possible gene-region pairs to regions within 100000 bp of the gene
- Aggregate expression/accessibility profiles (computational intensity reduction and dealing with ATAC sparsity)
- Identify significant gene expression to accessibility associations

### Challenges

- ATAC-seq data is large lots of memory trouble
- Deciding which method and parameters to choose for:
  - Co-embedding (e.g. how to count summarize ATAC-seq signal over genes)
  - Feature selection (variable genes/correlation to pseudotime, chromVAR/Cicero...)
  - Aggregation (clustering, subsampling)
- Interoperability between AnnData, SCE, and Seurat

## Thank you for your kind attention

and

## Thanks to Emma and Charlotte

