Summer School 2021: Advanced topics in Single Cell Omics

RNA Velocity

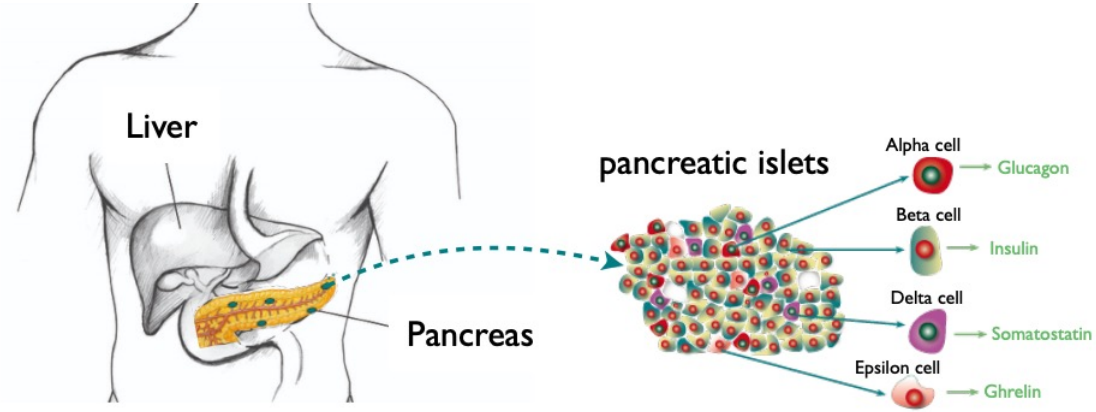
Topic 1 - Vector field representations depend on the embedding

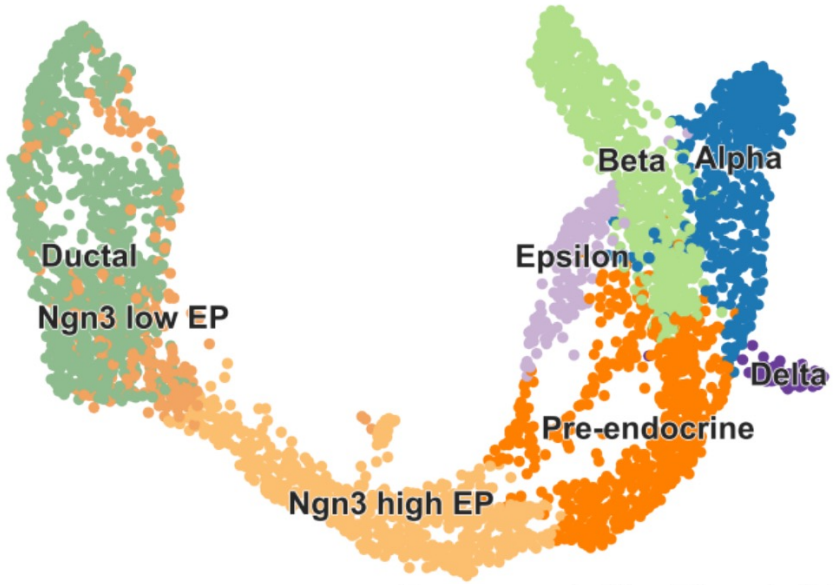
03. September 2021
Lasse Votborg Novél, Efthalia Preka, Sanna Abrahamsson, Jana Koch

Volker Bergen & Paulo Czarnewski

Development of pancreatic cells

Endocrine cells in pancreas





Bergen *et al.*, Nat. Biotech (2020)

Bastidas-Ponce *et al.*, Development (2019)

- Development of pancreatic isles highly medically relevant
- Rather well characterized cell populations & developmental stages --> ideal setting to learn about & test RNA velocity

--> Does the choice of embedding impact RNA velocity analysis ?

What makes a good embedding?

- Preserve global and local structures of the dataset
- Represent the high-dimensional vector field

Are there differences between the embeddings used for RNA velocity analysis?

Can we quantify differences?

Making sense of the data

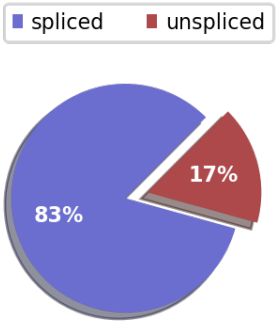
- presorted mouse Ngn3+ and epithelial progenitors at E15.5
- 10x 3' library (v2)

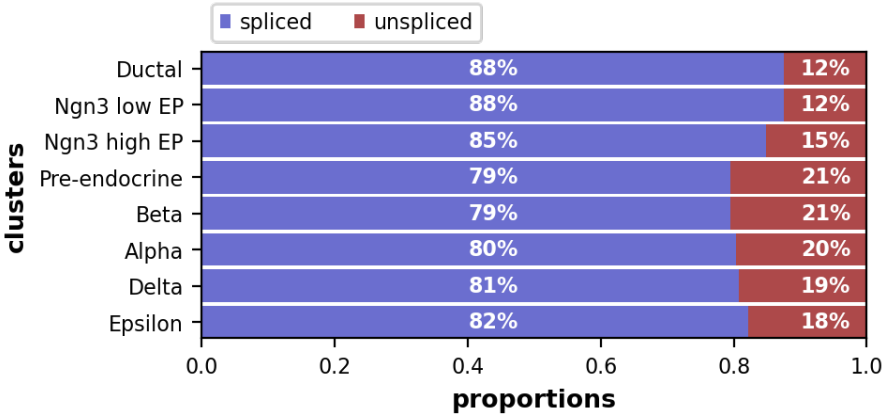
- Dataset already preprocessed:

```
scv.datasets.pancreas()
```

```
AnnData object with n_obs × n_vars = 3696 × 27998
    obs: 'clusters_coarse', 'clusters', 'S_score', 'G2M_score'
    var: 'highly_variable_genes'
    uns: 'clusters_coarse_colors', 'clusters_colors', 'day_colors', 'neighbors', 'pca'
    obsm: 'X_pca', 'X_umap'
    layers: 'spliced', 'unspliced'
    obsp: 'distances', 'connectivities'
```

- Spliced and unspliced reads
- Clustering
- Cell cycle classification
- …



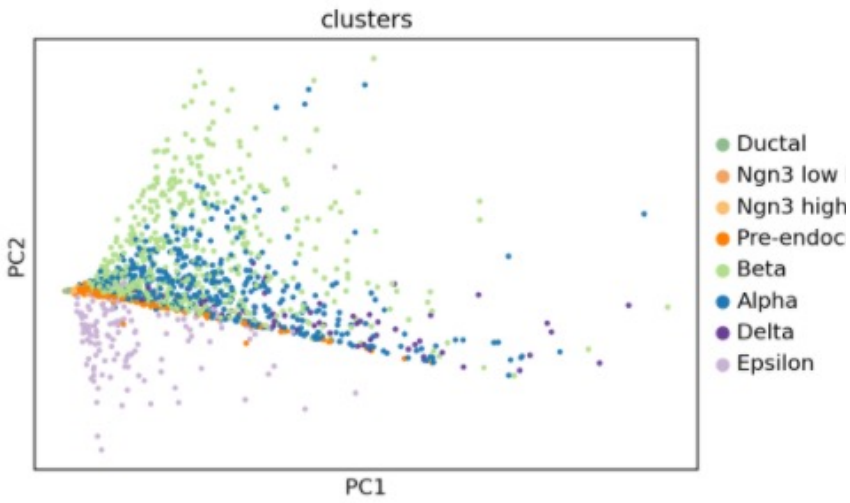


Logarithmization is important to capture the topology

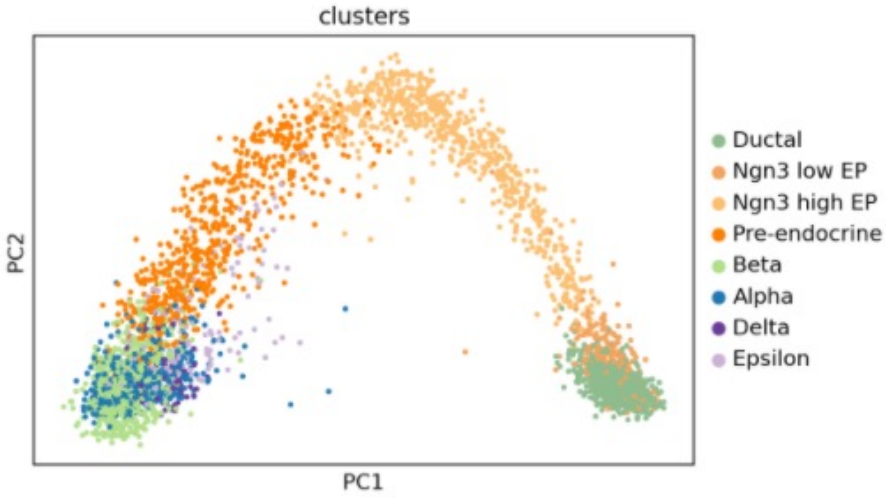
```
scv.pp.filter_and_normalize(adata, min_shared_counts=20, n_top_genes=2000)
```

```
Filtered out 20801 genes that are detected 20 counts (shared).
Normalized count data: X, spliced, unspliced.
Extracted 2000 highly variable genes.
Logarithmized X
```

Filtered and normalized data



Additionally logarithmized



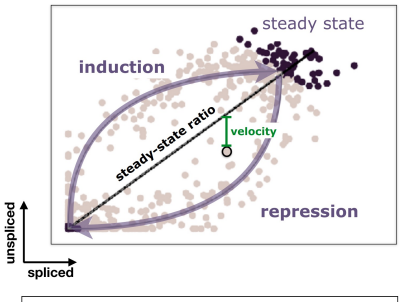
Log transformation:

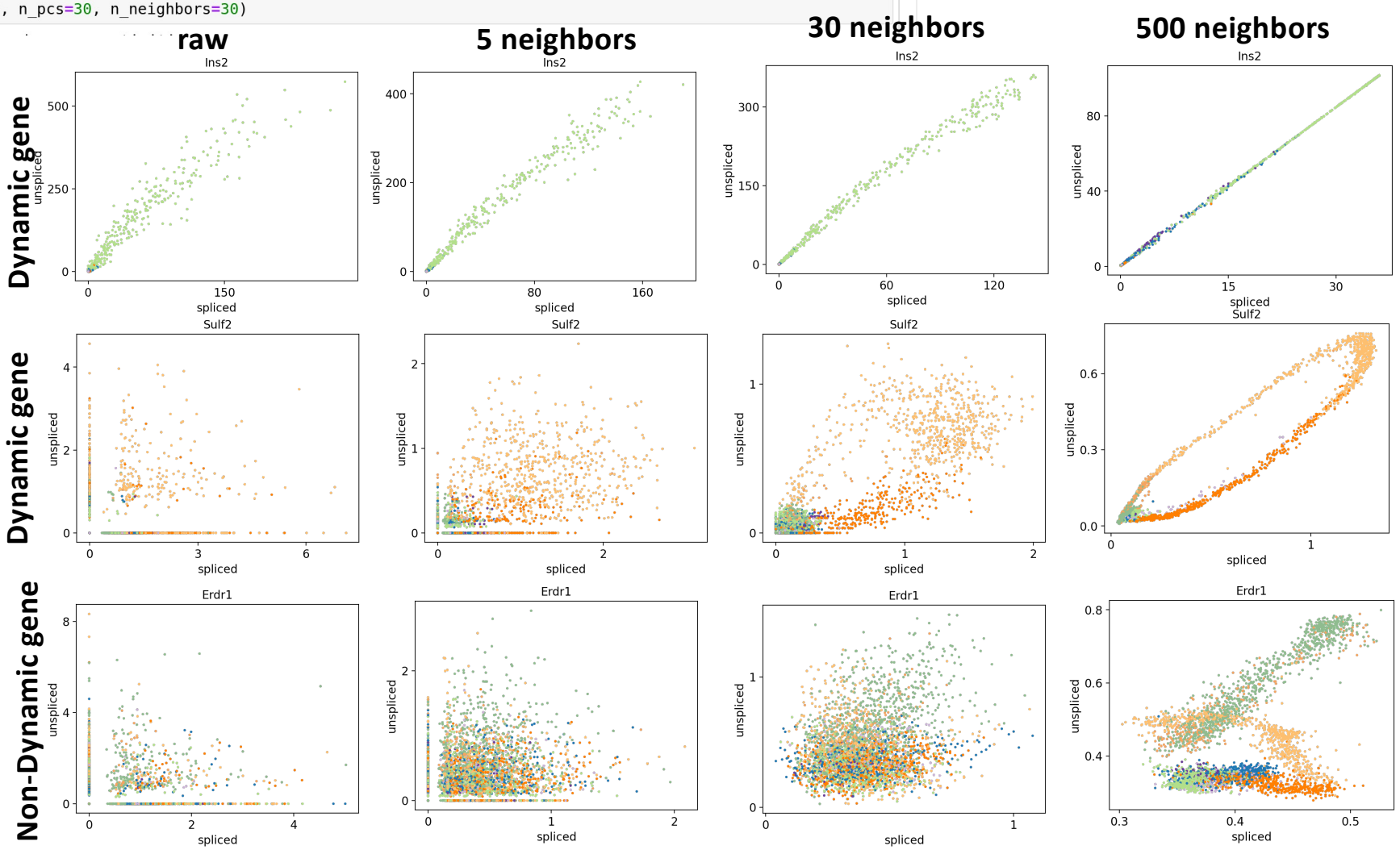
- reduces skewedness of data (important for downstream analysis tools that assume normal distribution of data

--> drastic differences for embedding

Imputation can amplify signal but can also introduce artifacts

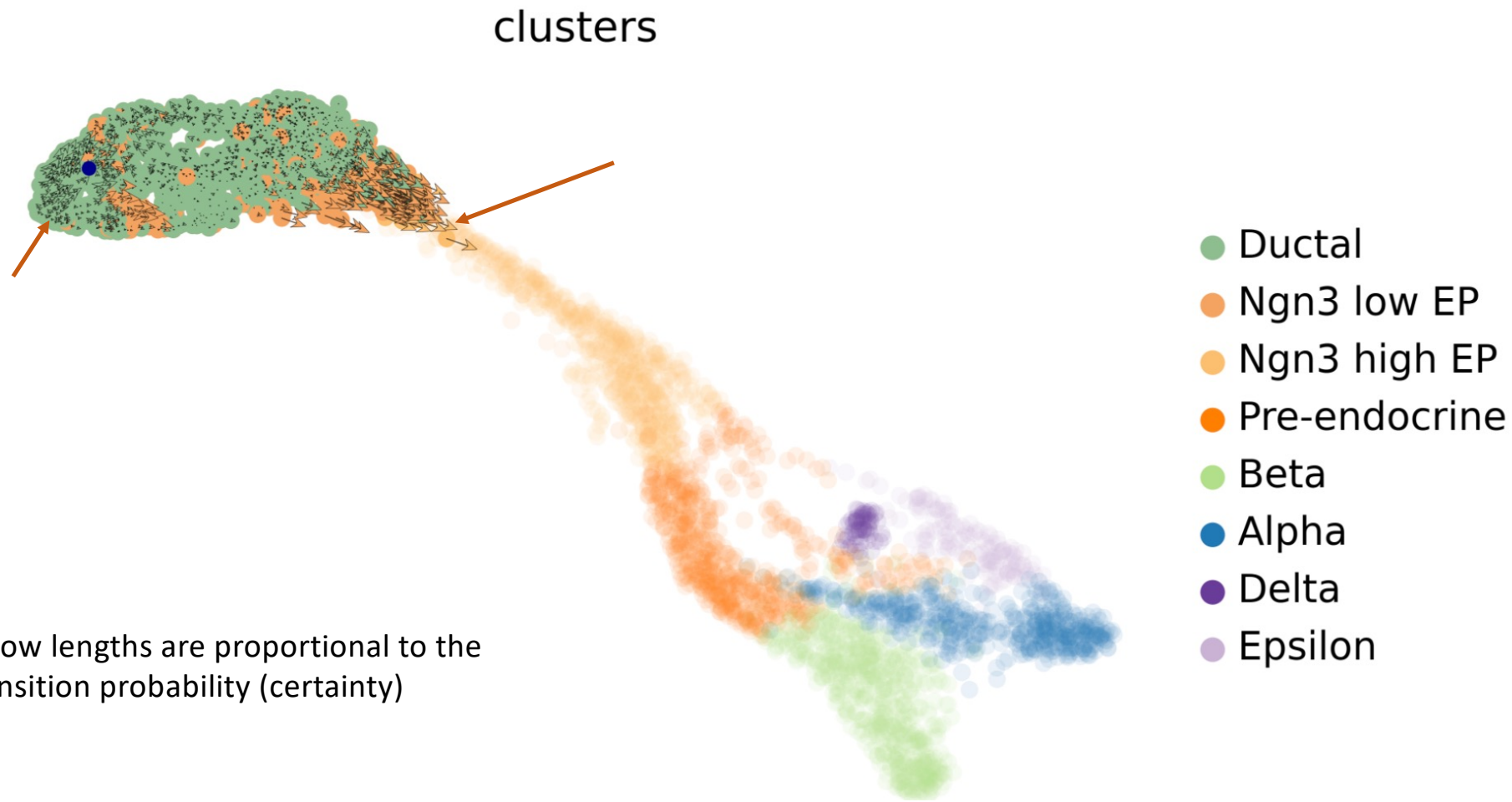
```
In [85]: scv.pp.moments(adata, n_pcs=30, n_neighbors=30)
```





.--> Optimize for your dataset

Arrows of cycling vs. differentiating cells

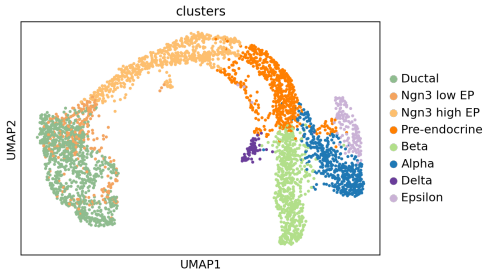


Arrow lengths are proportional to the transition probability (certainty)

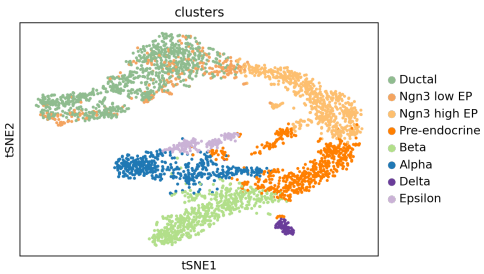
Different embeddings highlight different features of the data

*all default parameters

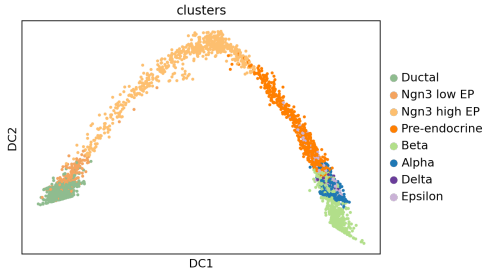
UMAP



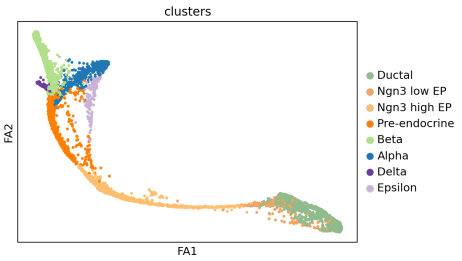
tSNE



Diffmap

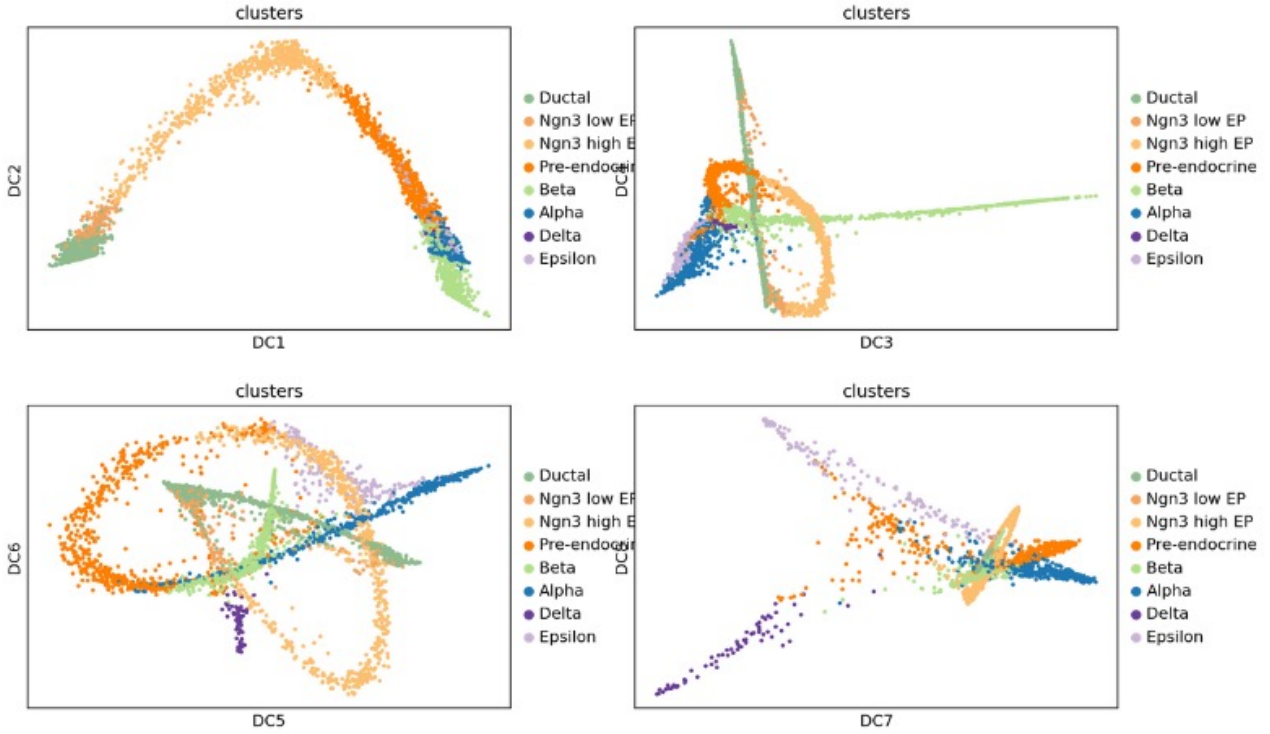


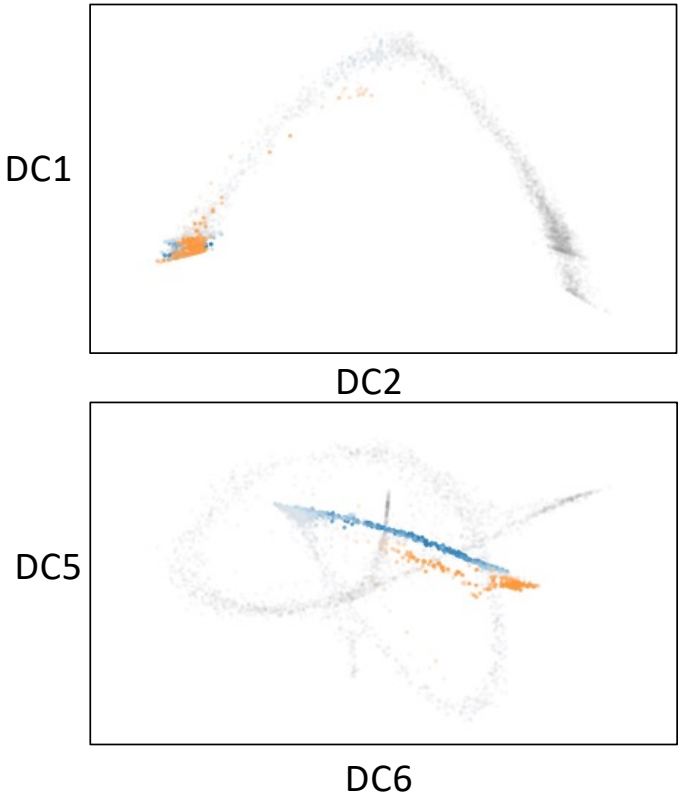
Force directed graph



Different parameters were tested in the following to assess impact on the analysis

Comprehensive view by looking at multiple components



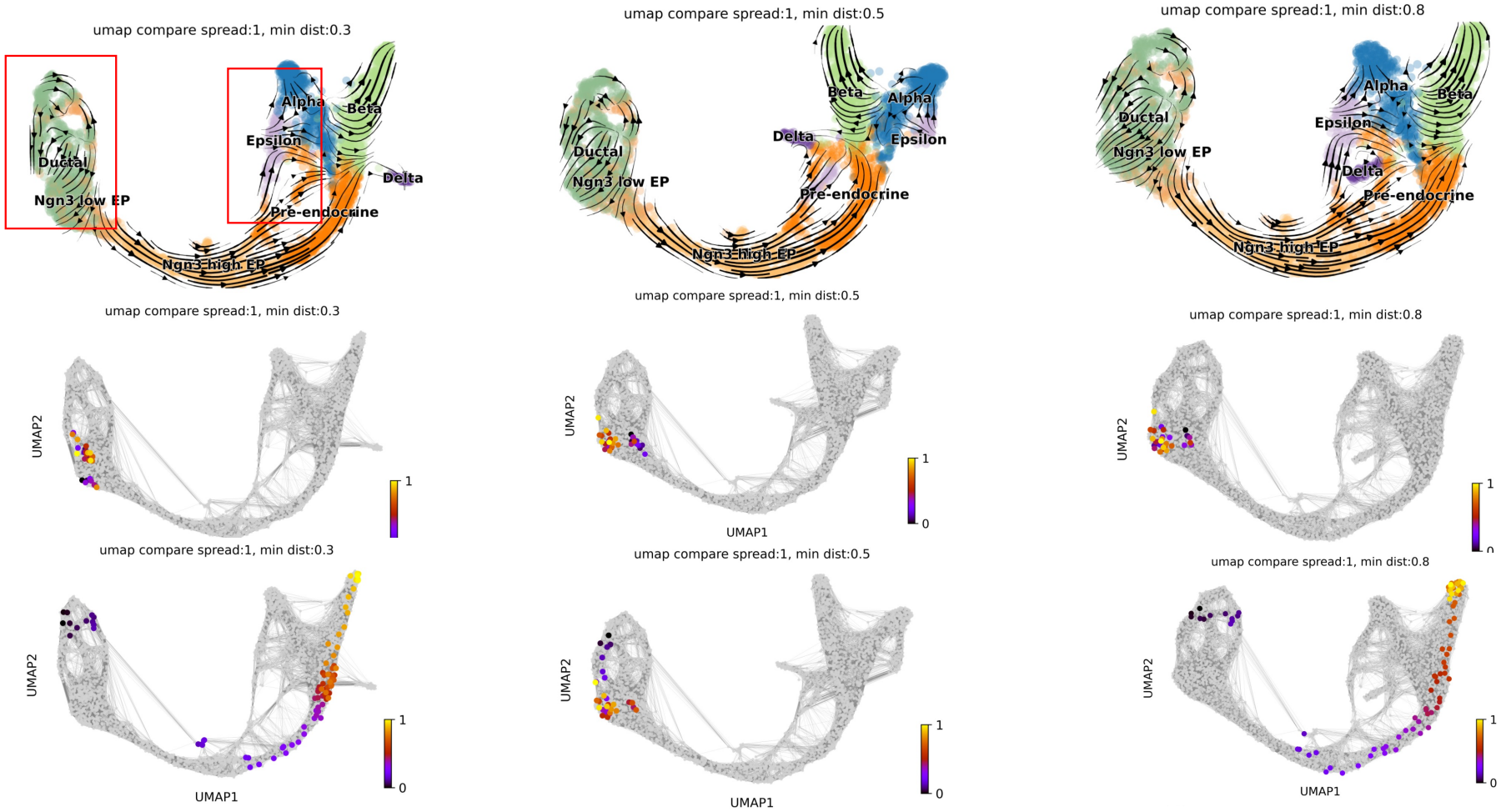


Blue = DNA Replication (s_score)
Orange = G2/ Mitosis (G2M_score)

UMAP: lower min_dist preserves local embedding

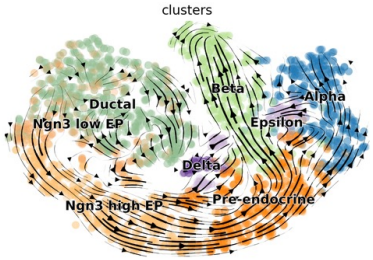
Epsilon cells endstate

Cell cycle

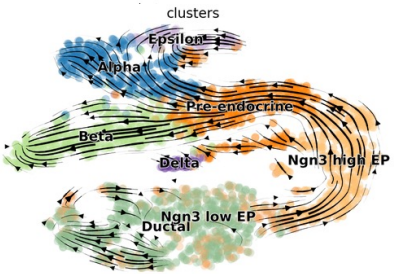


TSNE does not capture the cell cycle

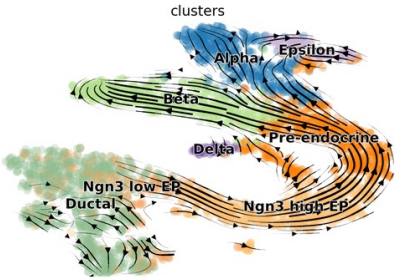
Perplexity 5



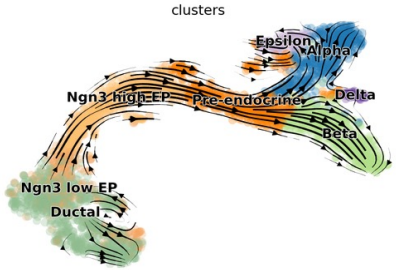
Perplexity 10



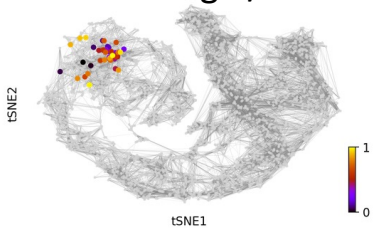
Perplexity 20

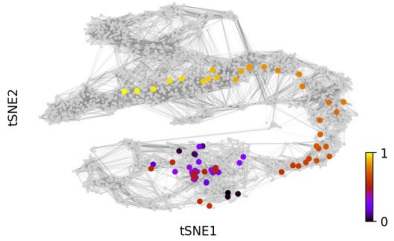


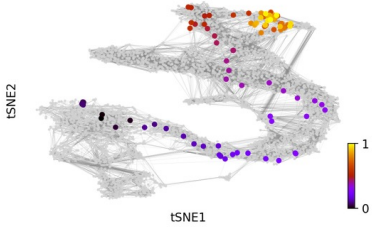
Perplexity 100

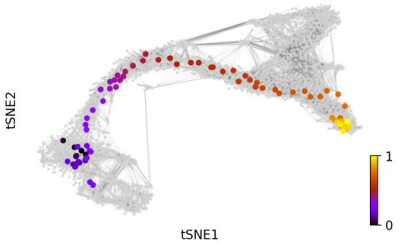


Ductal cell – neg S/G2M score

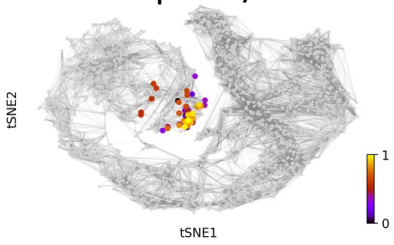


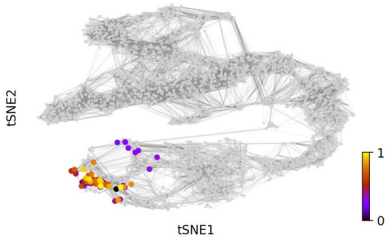


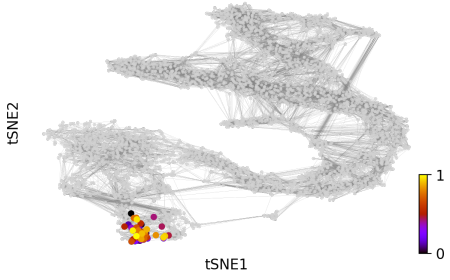


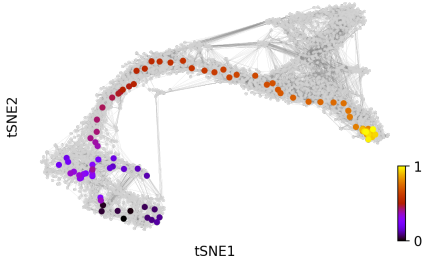


Ductal cell – pos S/G2M score

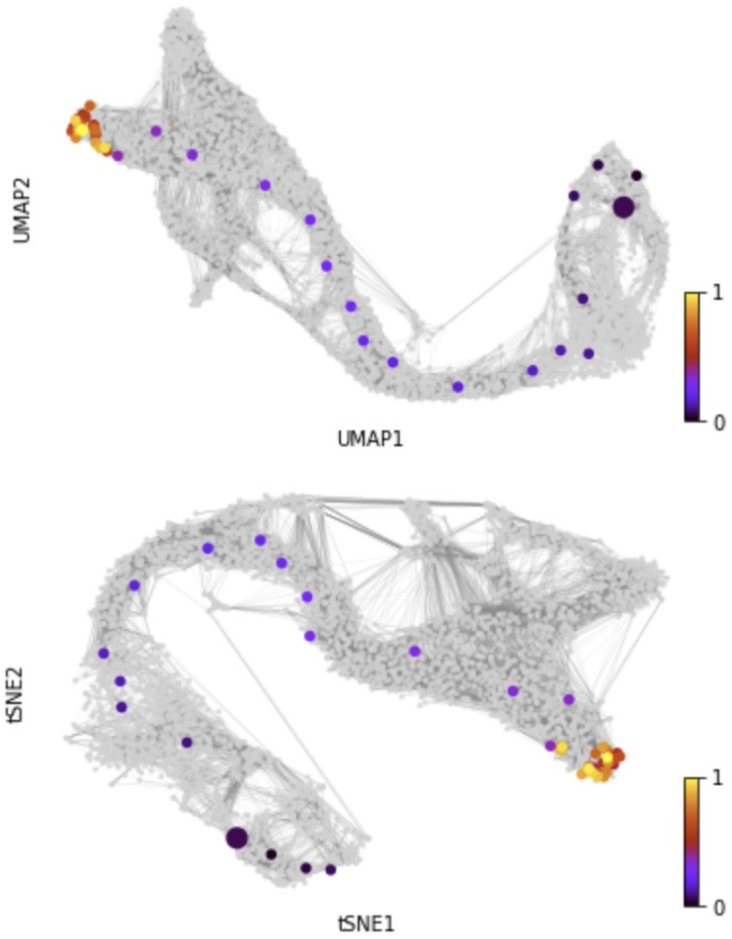








Can the vector field representation be quantified?



Transition — Length of vector

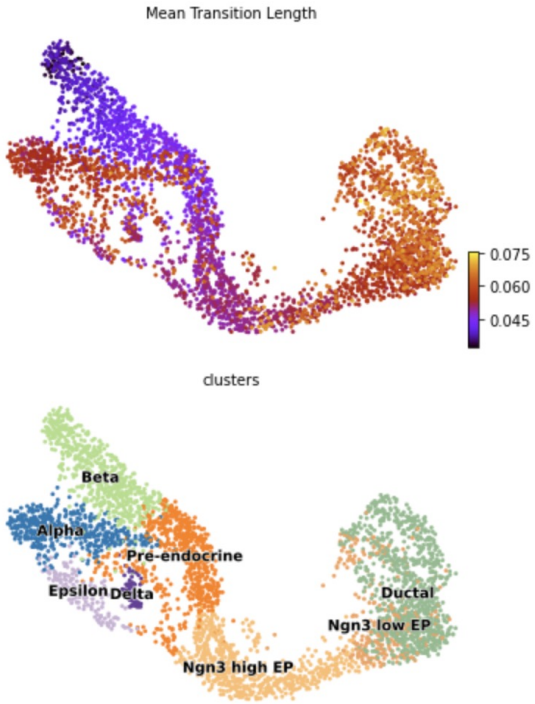
Cell — Mean of transition lengths

Embedding — Mean of cell means

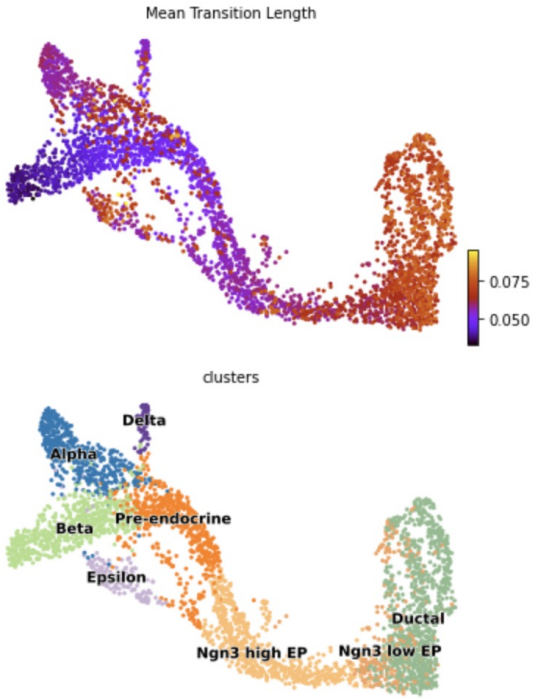
Embedding parameters change the representation

UMAP

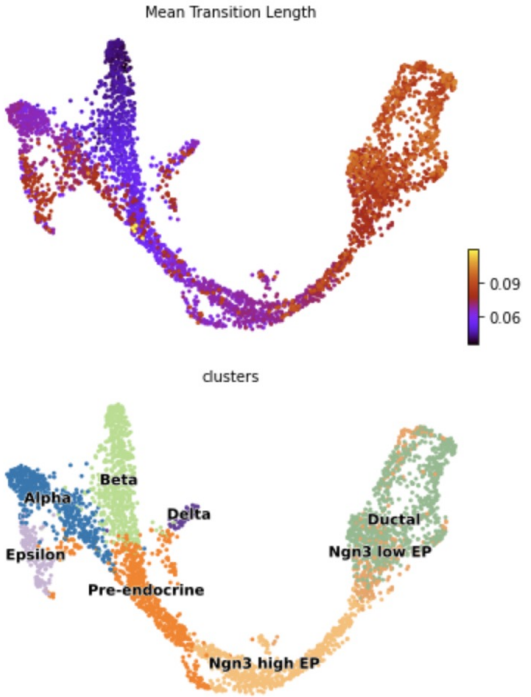
min_distance=0.5
spread=0.5



min_distance=0.5
spread=1



min_distance=0.5
spread=2

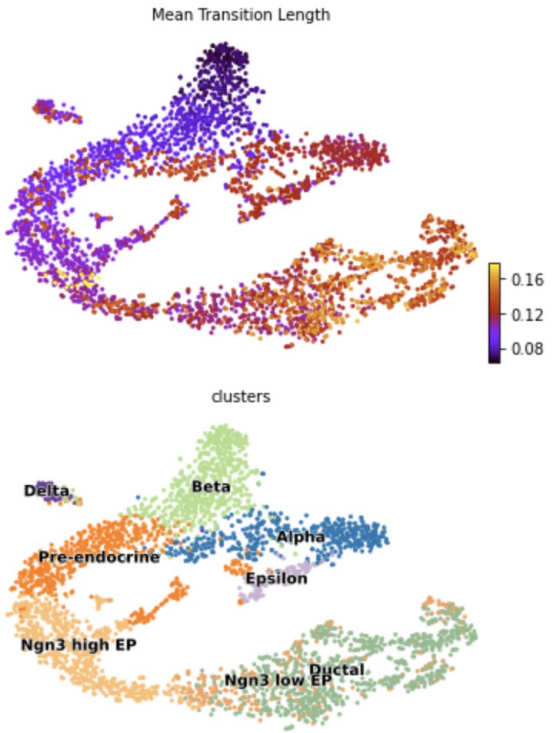


Mean transition length: 0.053 0.059 0.070

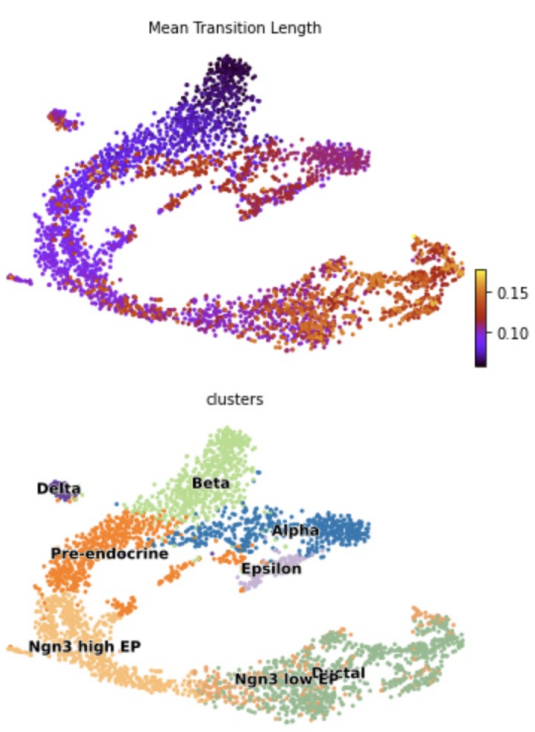
Embedding parameters change the representation

tSNE

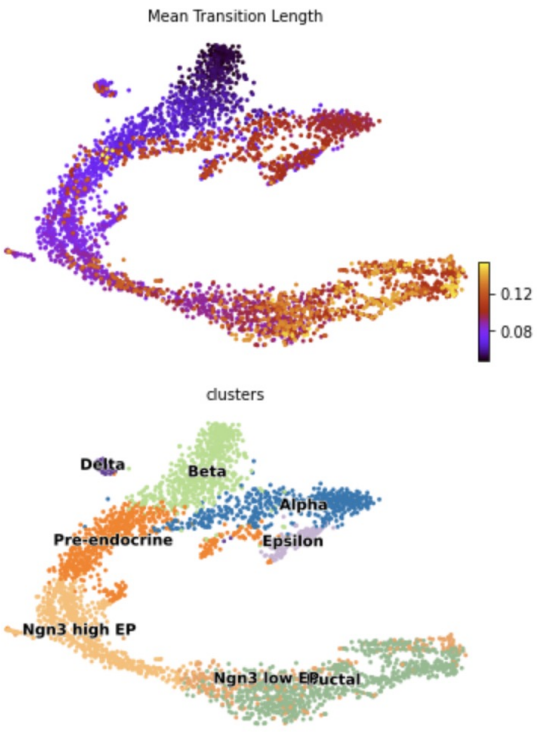
Perplexity=30



Perplexity=50



Perplexity=100



Mean transition length: 0.111 0.102 0.090

Is transition length a good quantification measure?

Embedding	Configuration	Mean of transition lengths
UMAP	min_dist=0.1 ; spread=0.1	0.017
	min_dist=0.1 ; spread=0.5	0.027
	min_dist=0.3 ; spread=0.5	0.035
	min_dist=0.5 ; spread=0.5	0.053
	min_dist=0.7 ; spread=0.5	0.059
	min_dist=0.5 ; spread=1	0.059
	min_dist=0.5 ; spread=2	0.070
tSNE	perplexity=10	0.116
	perplexity=30	0.111
	perplexity=50	0.102
	perplexity=100	0.090
	perplexity=150	0.080
	perplexity=300	0.091

Conclusion

i. Log-norm & imputation are important for the representation of the data

ii. Choice of embedding configuration may impact biological conclusion

iii. For a comprehensive overview we recommend looking at more than just your favourite TNSE, and also multiple dimensions (diffusion map).

iv. Using the cell transition/connectivity graph, we can highlight where topology might not have been preserved.

v. Metrics such as mean transition length may be used to find the optimal embedding parameter set.

Thank you for your attention!!