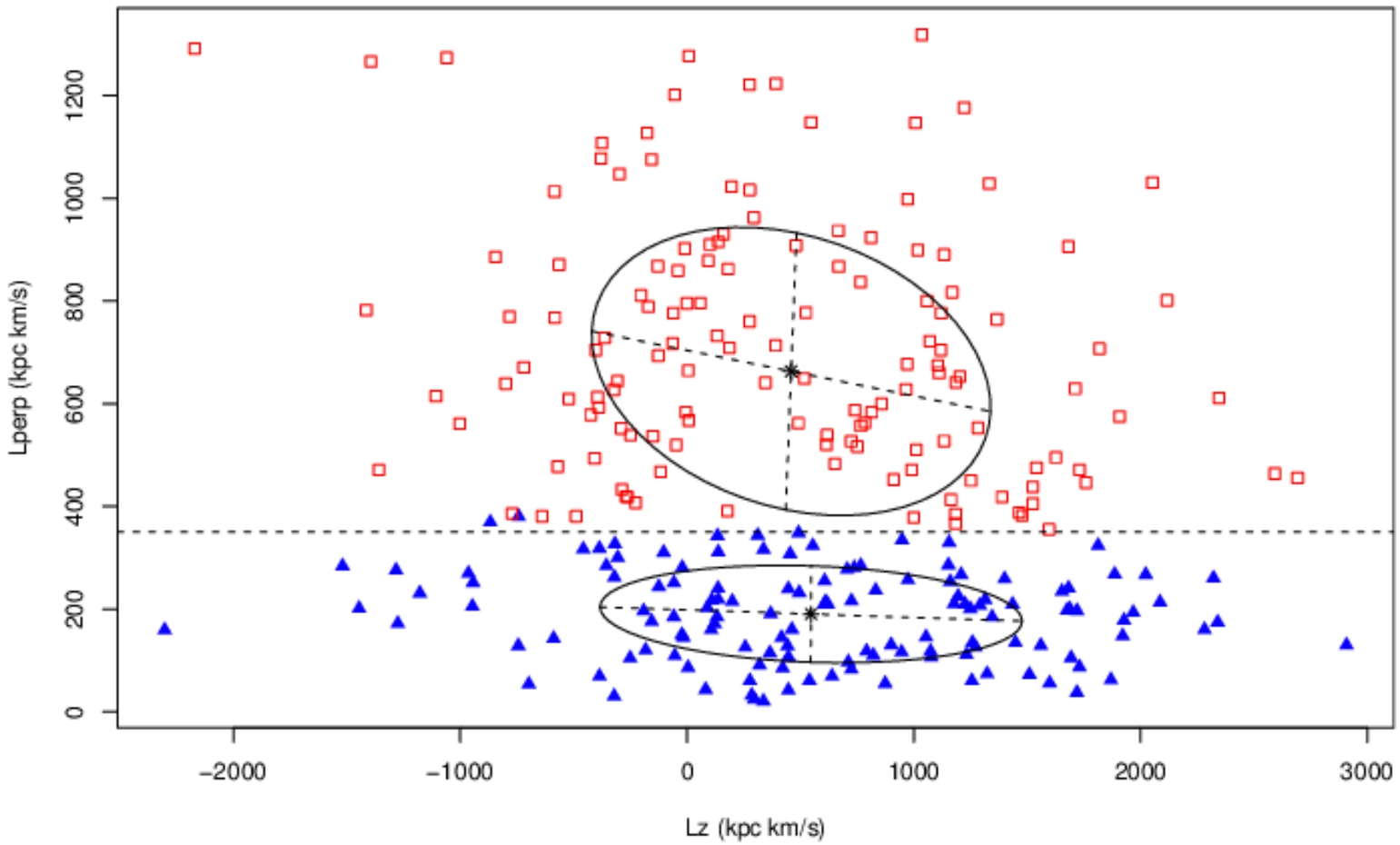
High dimensional data analysis in astronomy and biology

Lior Pachter
Caltech

June 24, 2019
Astroinformatics 2019
Caltech

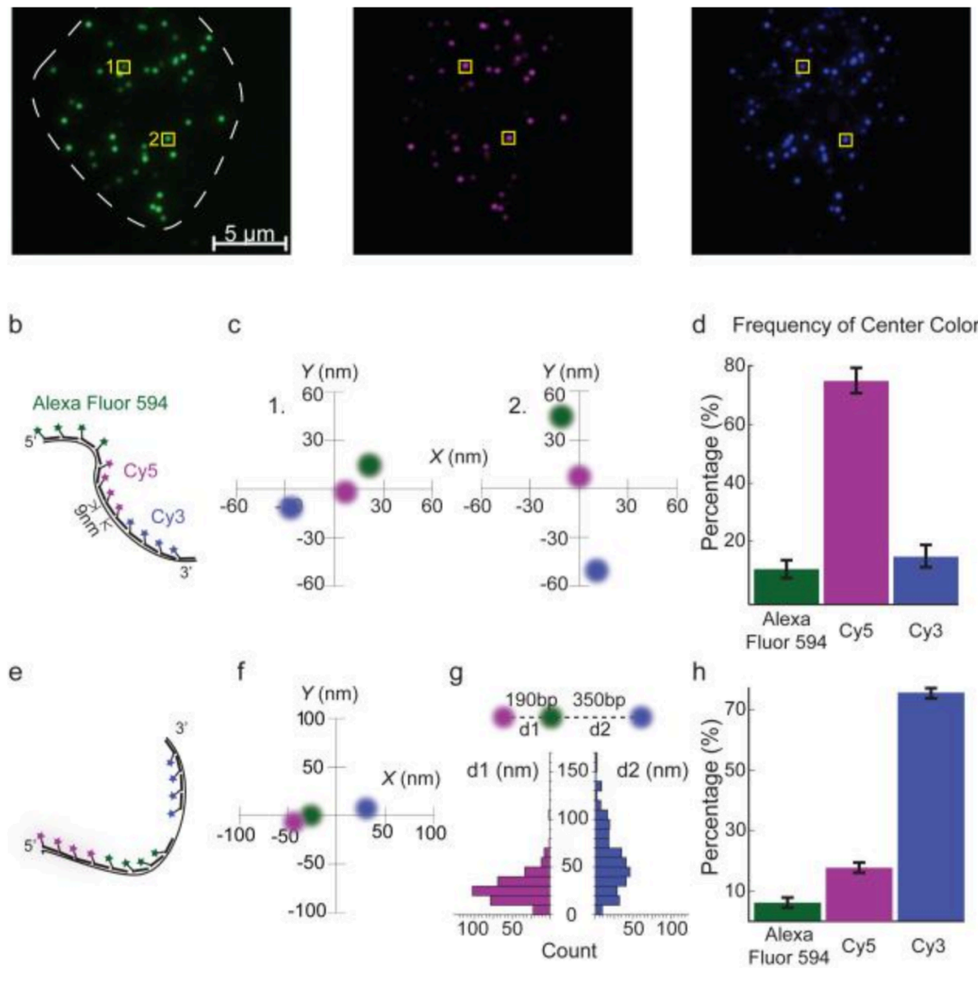
Astroinformatics from the perspective of bioinformatics



Morrison et al., 2009

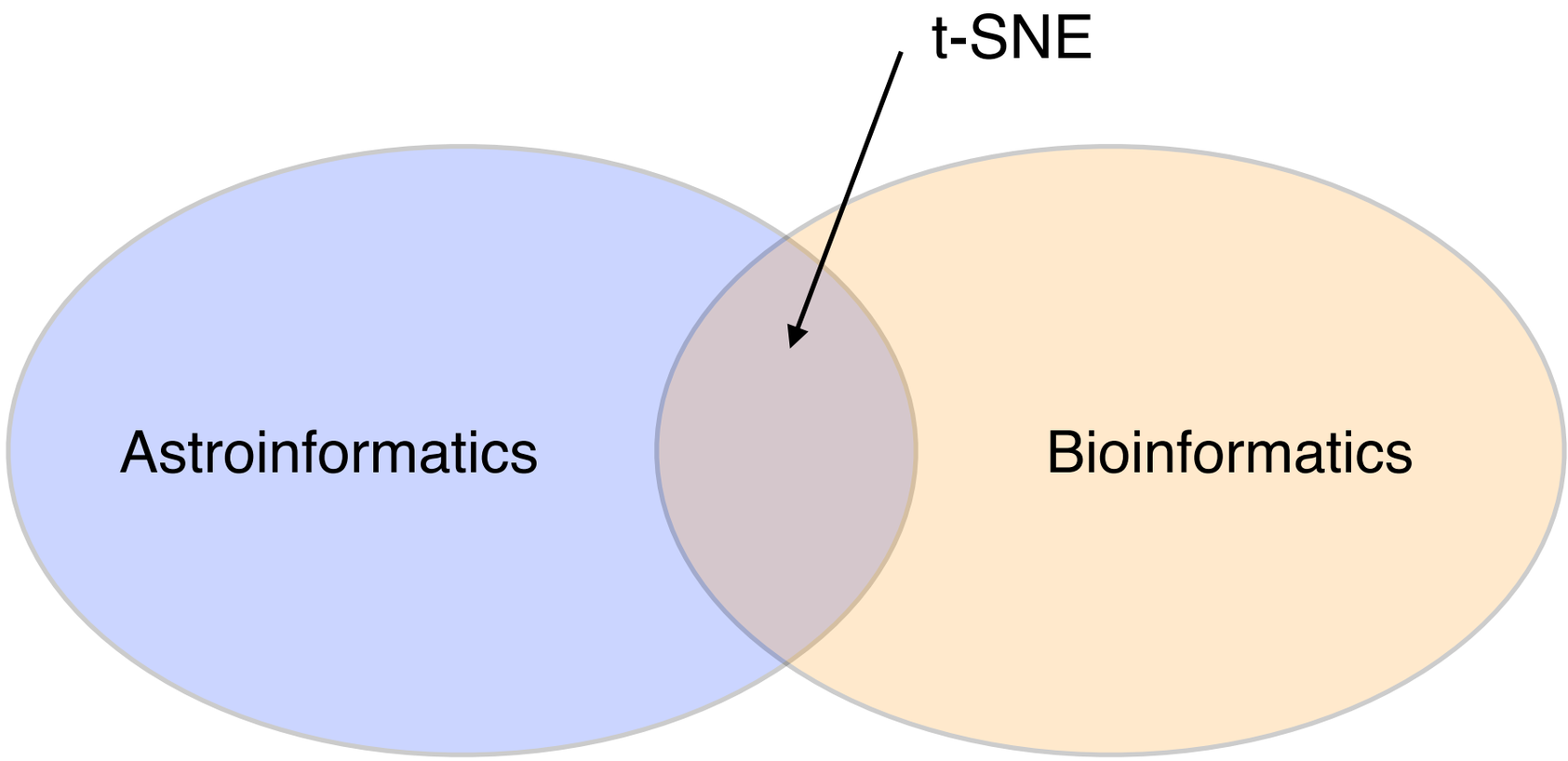
Gaussian fitting in biology

Spatial ordering of fluorophores on mRNAs can be resolved by Gaussian centroid localization



Lubeck and Cai, 2012

What else is there in the intersection?



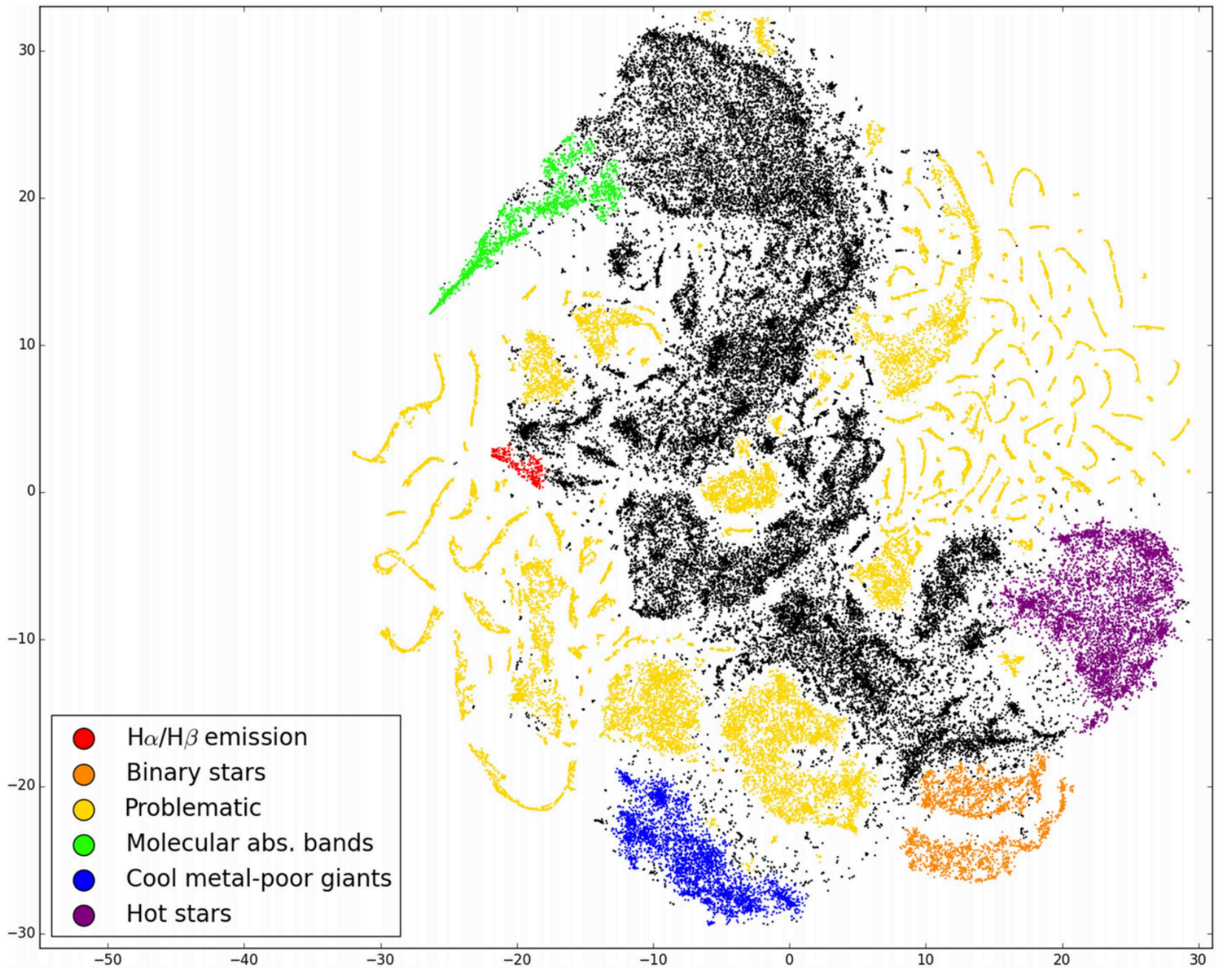
t-distributed stochastic neighborhood embedding

- Introduced by van der Maaten and Hinton in 2008.

- Minimizes the Kullback-Leibler divergence between a Gaussian distribution used to model distances in the ambient space, and a Student t-distribution modeling distances in low dimension.

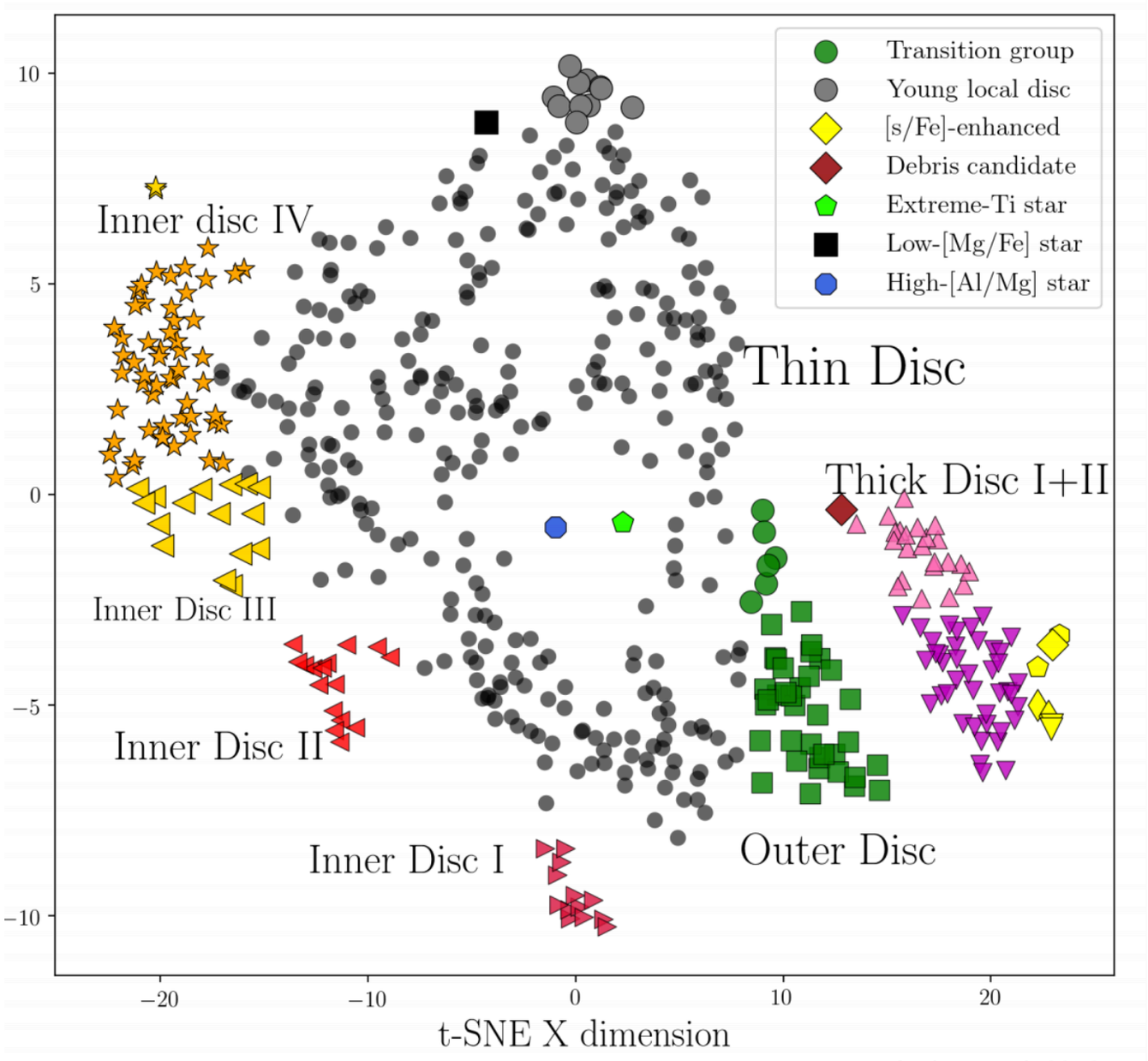
- **Theorem** (Linderman and Steinerberger, 2017): There are parameters for this algorithm that ensure rapid convergence. The algorithm behaves like spectral clustering (under some assumptions), allowing for a rigorous analysis of clustering properties.

High dimensional astronomical t-SNE



Traven et al. 2017

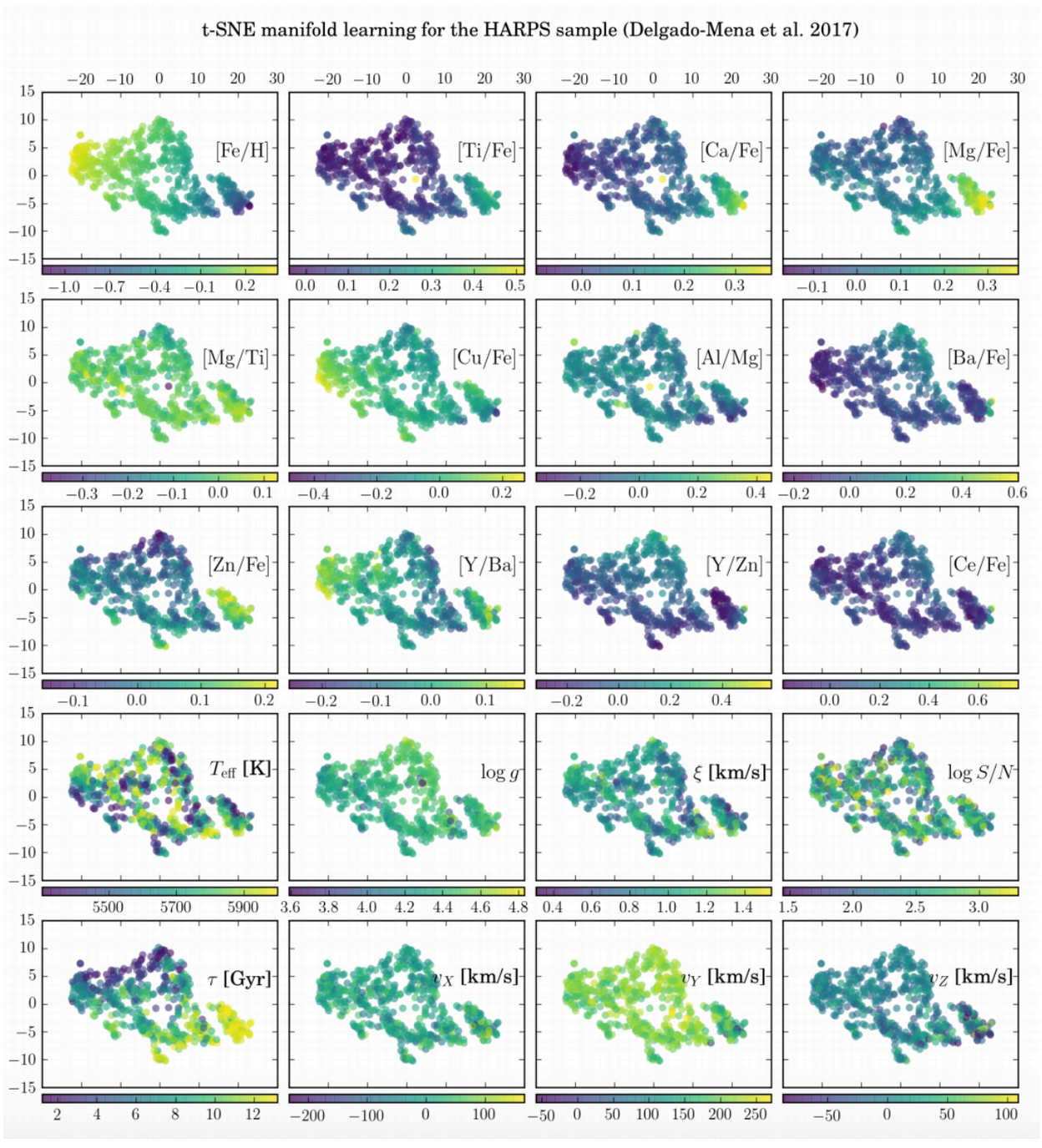
t-SNE to visualize stellar chemical abundance space



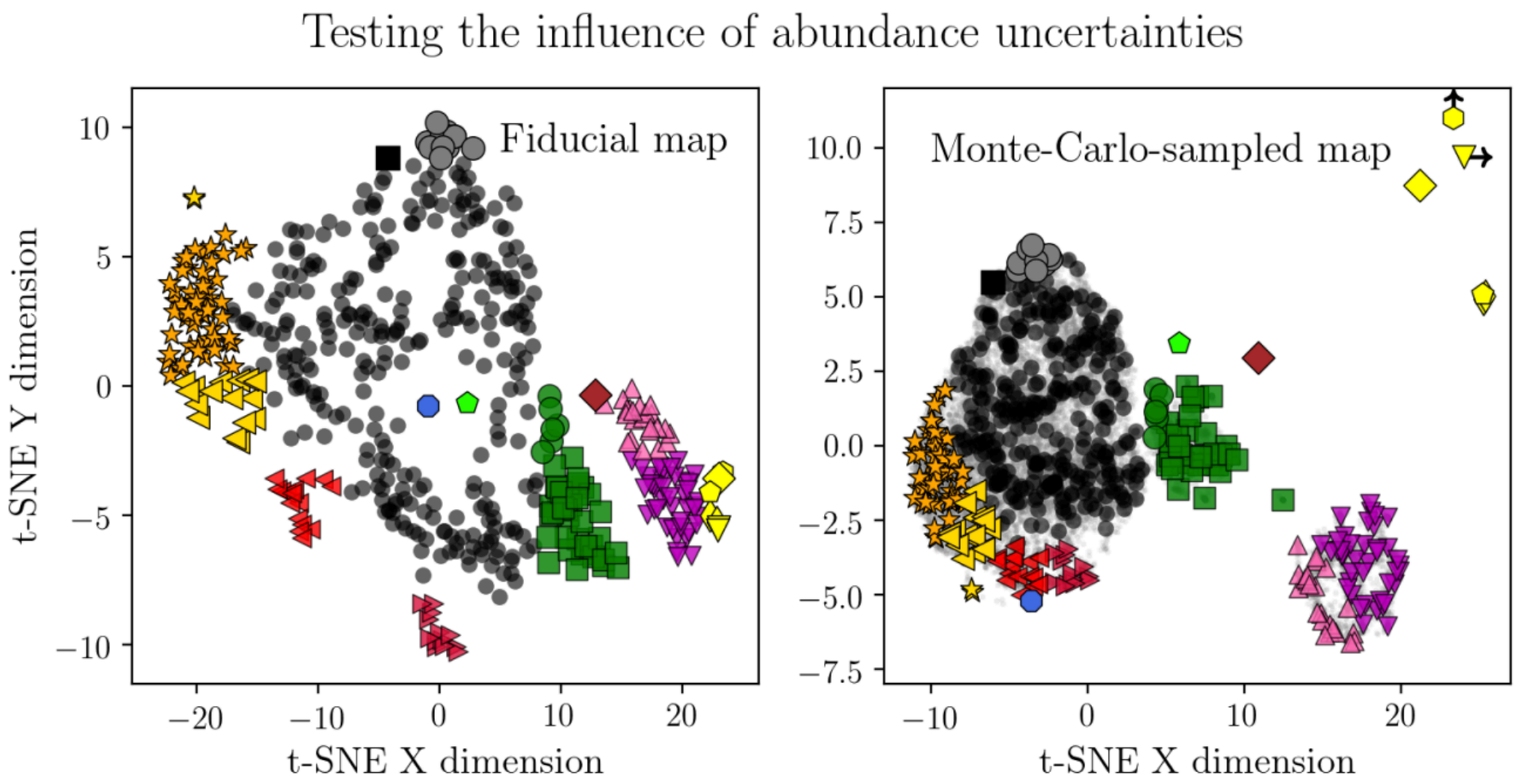
Abundances: Mg, Al, Si, Ca, TiI, Fe, Cu, Zn, Sr, Y, ZrII, Ce and Ba

Anders et al. 2018

Coloring by chemical abundances



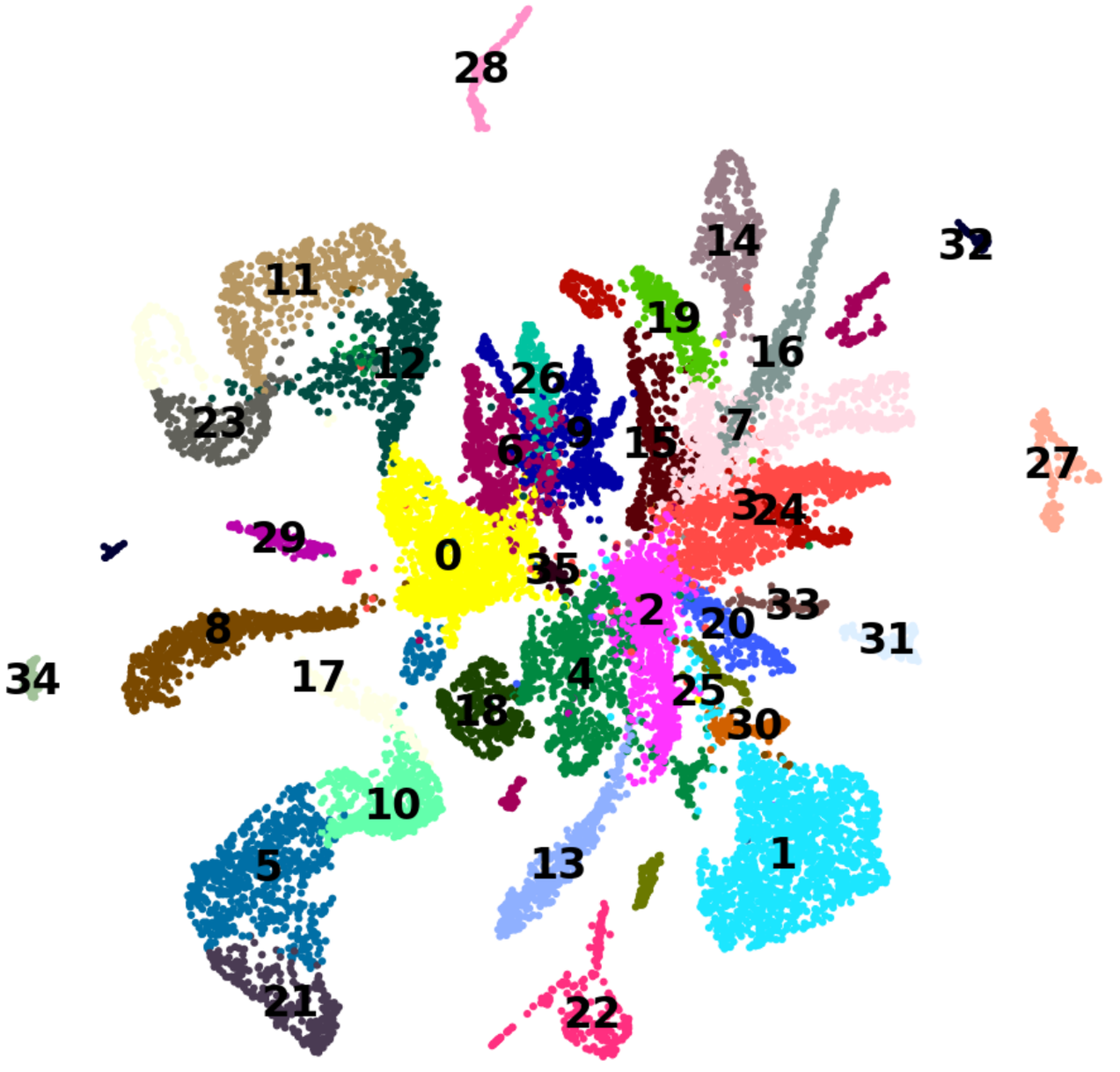
Testing robustness to abundance estimation errors



Traven et al. 2019

t-SNEs are also used in biology… **so what**?

with Jase Gehring
Brady Weissbourd, David Anderson



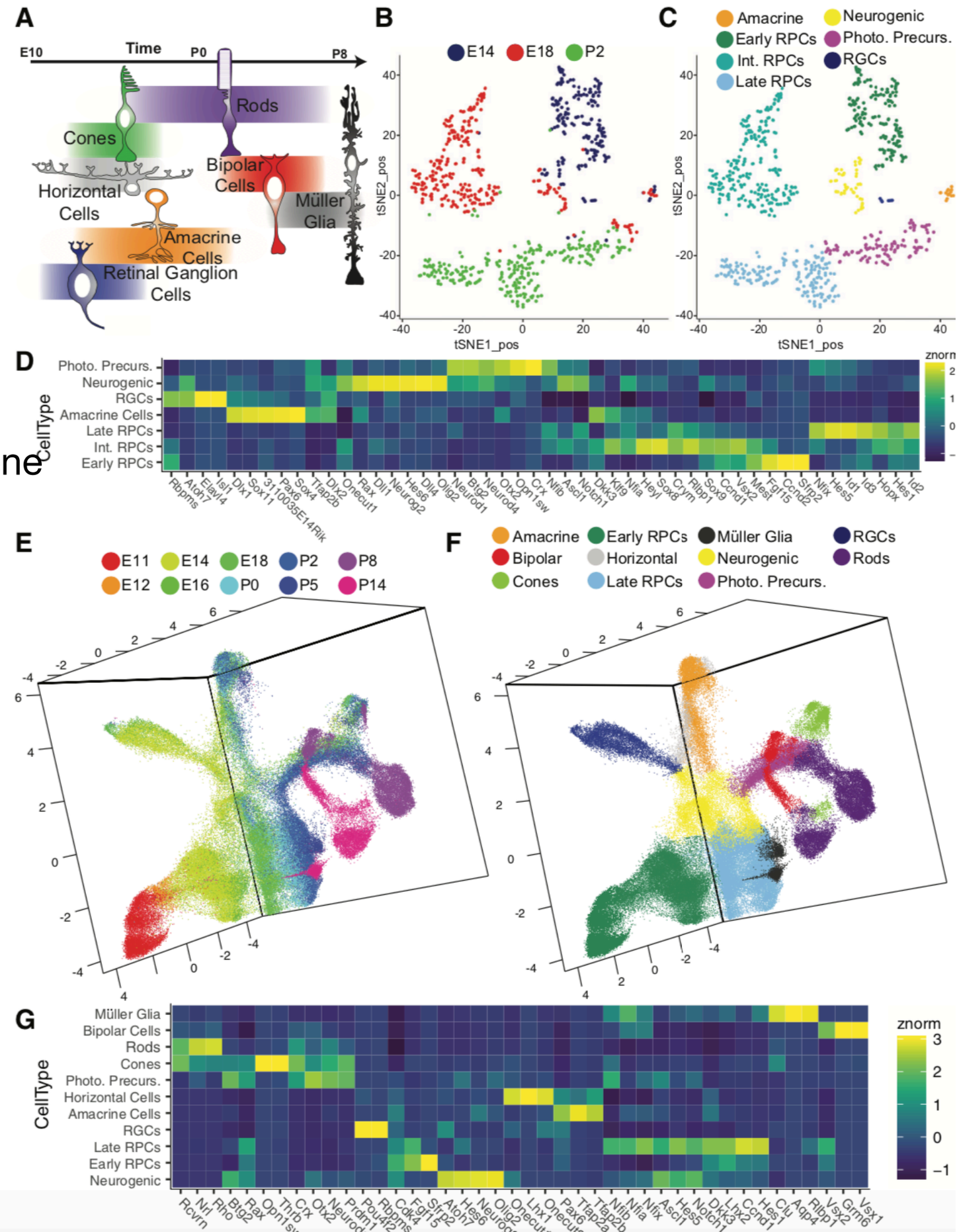
0 Neural stem cells	12 Early cnidocytes	24 Blank
1 Blank	13 Bell rim endoderm	25 Neuronal subtype
2 Blank	14 Blank	26 RFamide neurons
3 Blank	15 Blank	27 Neuronal subtype
4 Blank	16 Blank	28 Bioluminescent neurons
5 Conodopine-2	17 Late cnidocytes 1	29 Gastrovascular endoderm
6 SynTag5 neurons	18 Blank	30 Blank
7 Canal endoderm	19 Blank	31 pp9 neurons
8 Muscle	20 Blank	32 Blank
9 GLWamide neurons	21 Conodopine-3	33 Blank
10 Conodopine-1	22 Neuronal subtype	34 Blank
11 Mid Cnidocytes	23 Late cnidocytes 2	35 Oocytes

Biological t-SNE:

- A: Time, Space and Type
- B.C. t-SNE are clumpy
- D.G. Chemical/Frequency => Gene
- E. F. UMAP

(McInnes, Healy, Melville 2018)

Clark et al. 2019



The standard “t-SNE workflow”

Seurat v3.0 Command List

Compiled: 2019-04-04

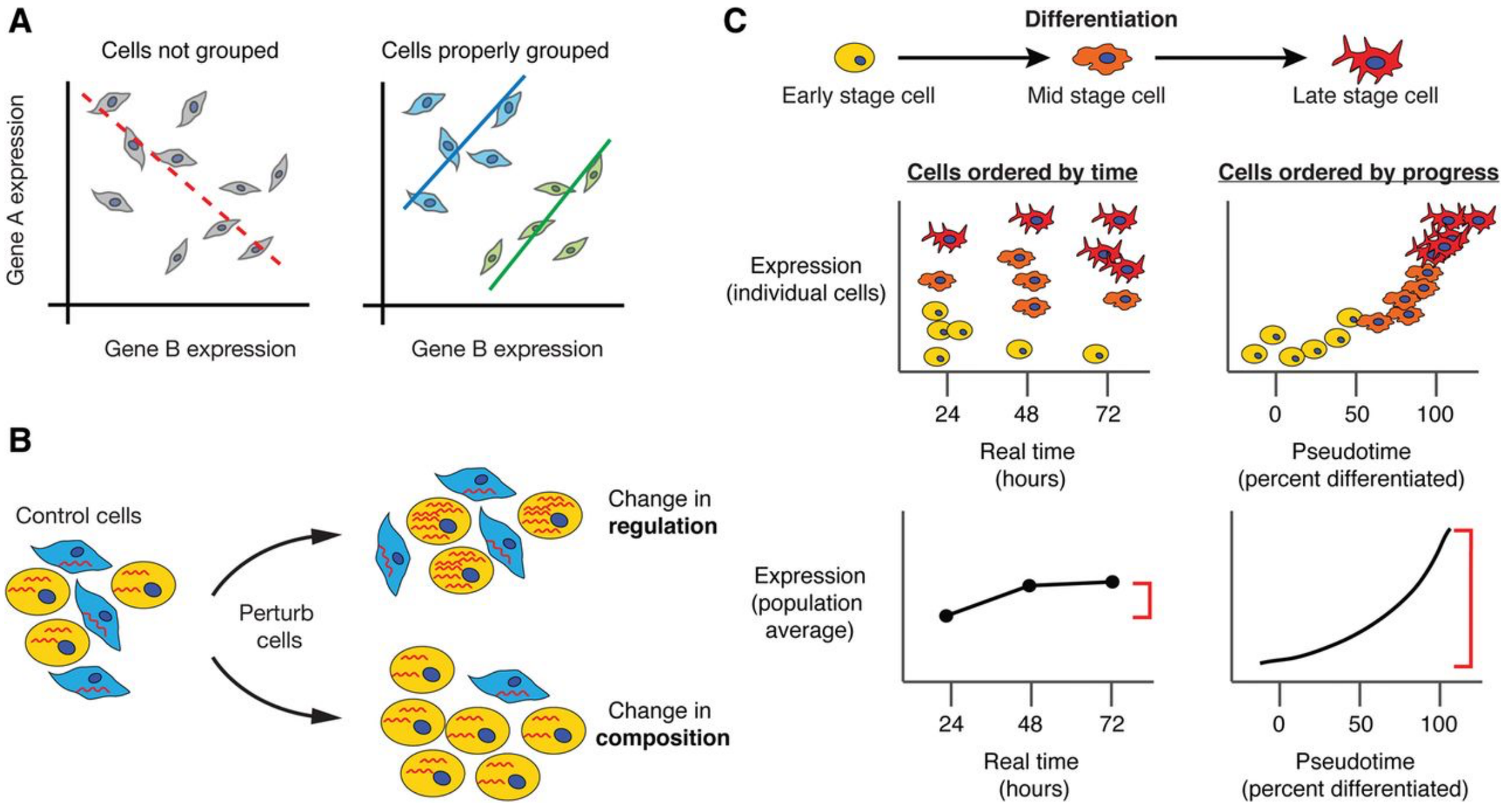
Seurat Standard Worflow | `Seurat` Object Interaction | Data Access | Visualization in Seurat v3.0

Multi-Assay Features | Seurat v2.X vs v3.X

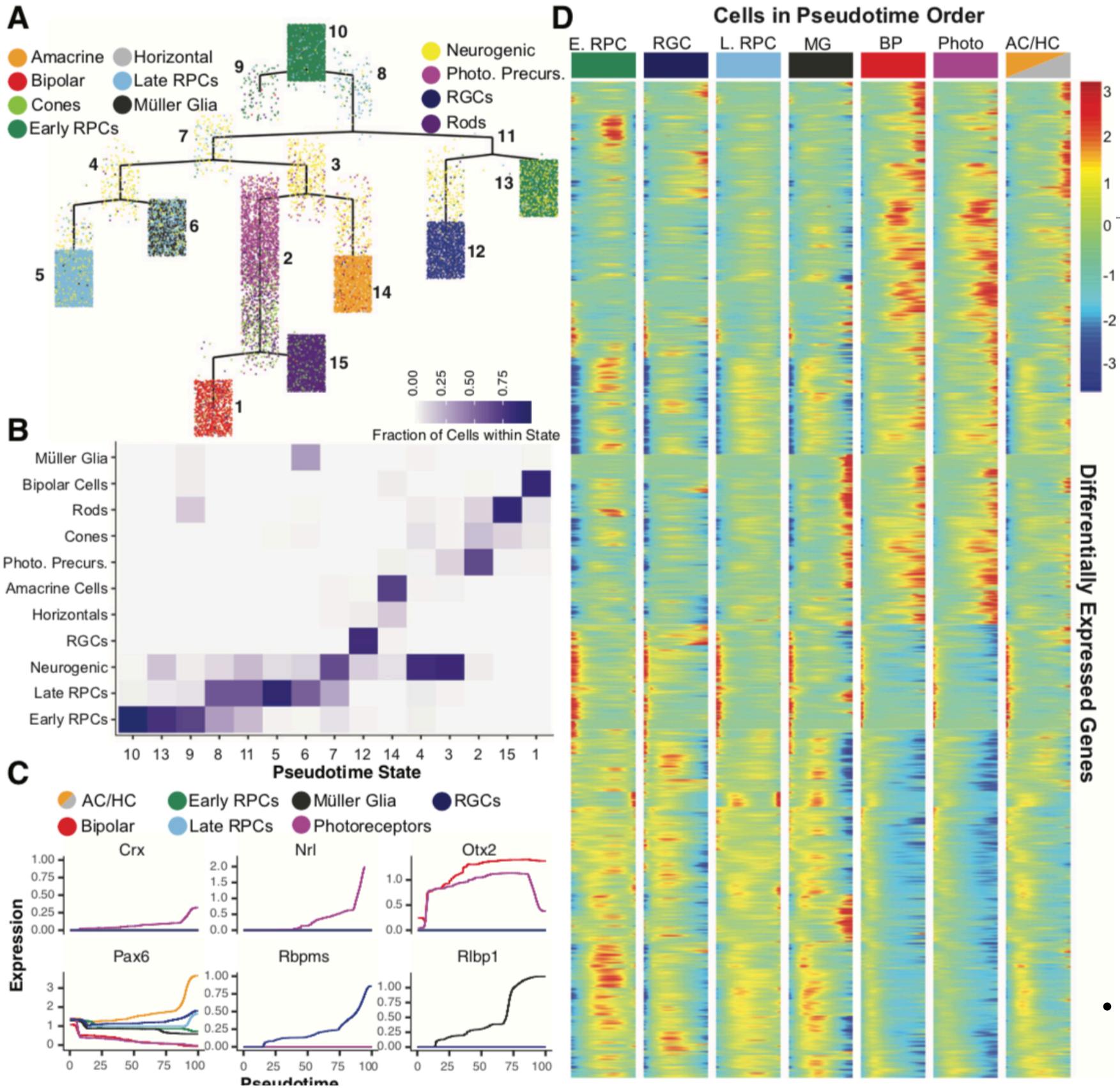
The standard Seurat workflow takes raw single-cell expression data and aims to find clusters within the data. For full details, please read our tutorial. This process consists of data normalization and variable feature selection, data scaling, a PCA on variable features, construction of a shared-nearest-neighbors graph, and clustering using a modularity optimizer. Finally, we use a t-SNE to visualize our clusters in a two-dimensional space.

```
pbmc.counts <- Read10X(data.dir = "~/Downloads/pbmc3k/filtered_gene_bc_matrices/hg19/")
pbmc <- CreateSeuratObject(counts = pbmc.counts)
pbmc <- NormalizeData(object = pbmc)
pbmc <- FindVariableFeatures(object = pbmc)
pbmc <- ScaleData(object = pbmc)
pbmc <- RunPCA(object = pbmc)
pbmc <- FindNeighbors(object = pbmc)
pbmc <- FindClusters(object = pbmc)
pbmc <- RunTSNE(object = pbmc)
DimPlot(object = pbmc, reduction = "tsne")
```

Why single-cell?

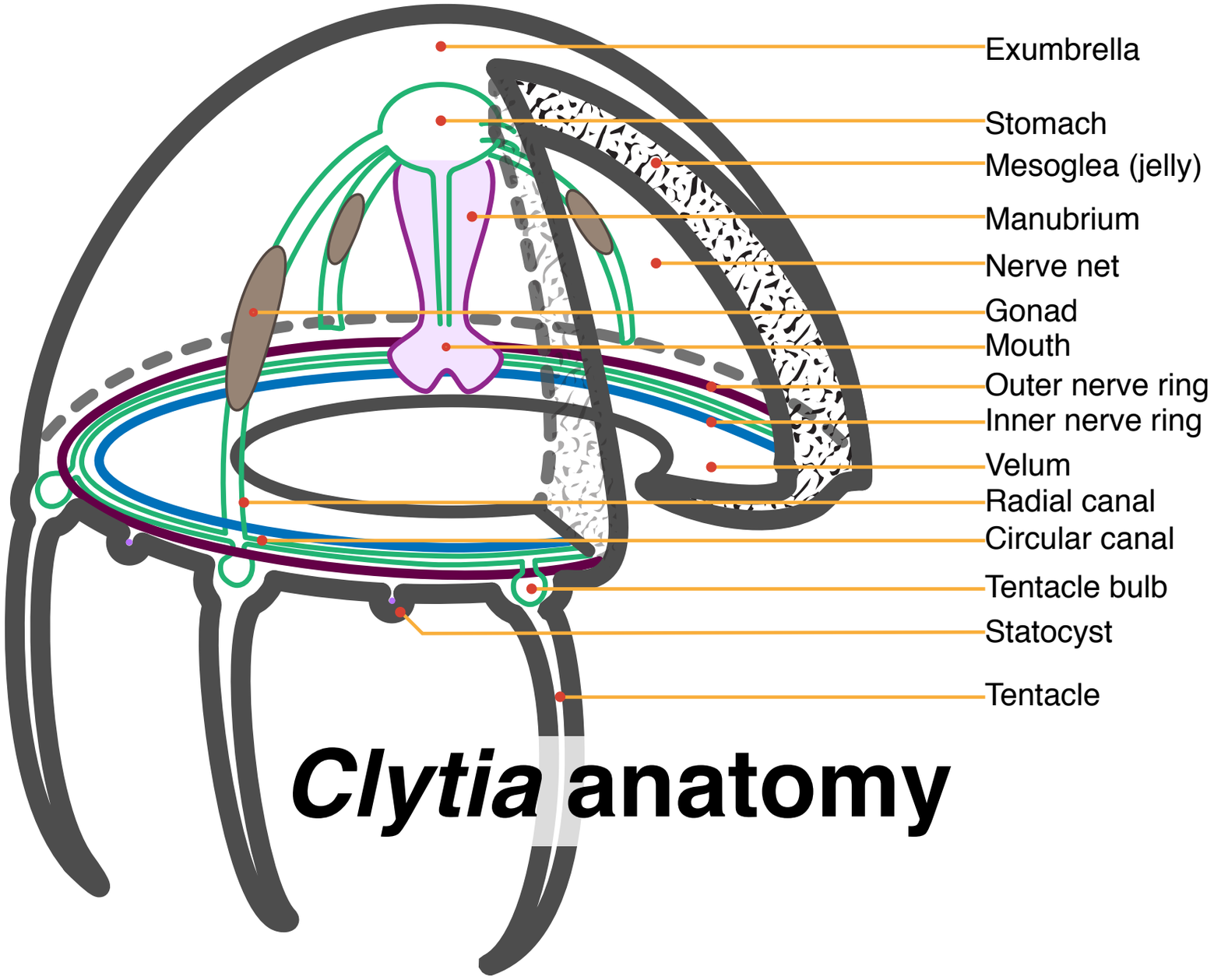


Trapnell, 2015



- Clark et al. 2019

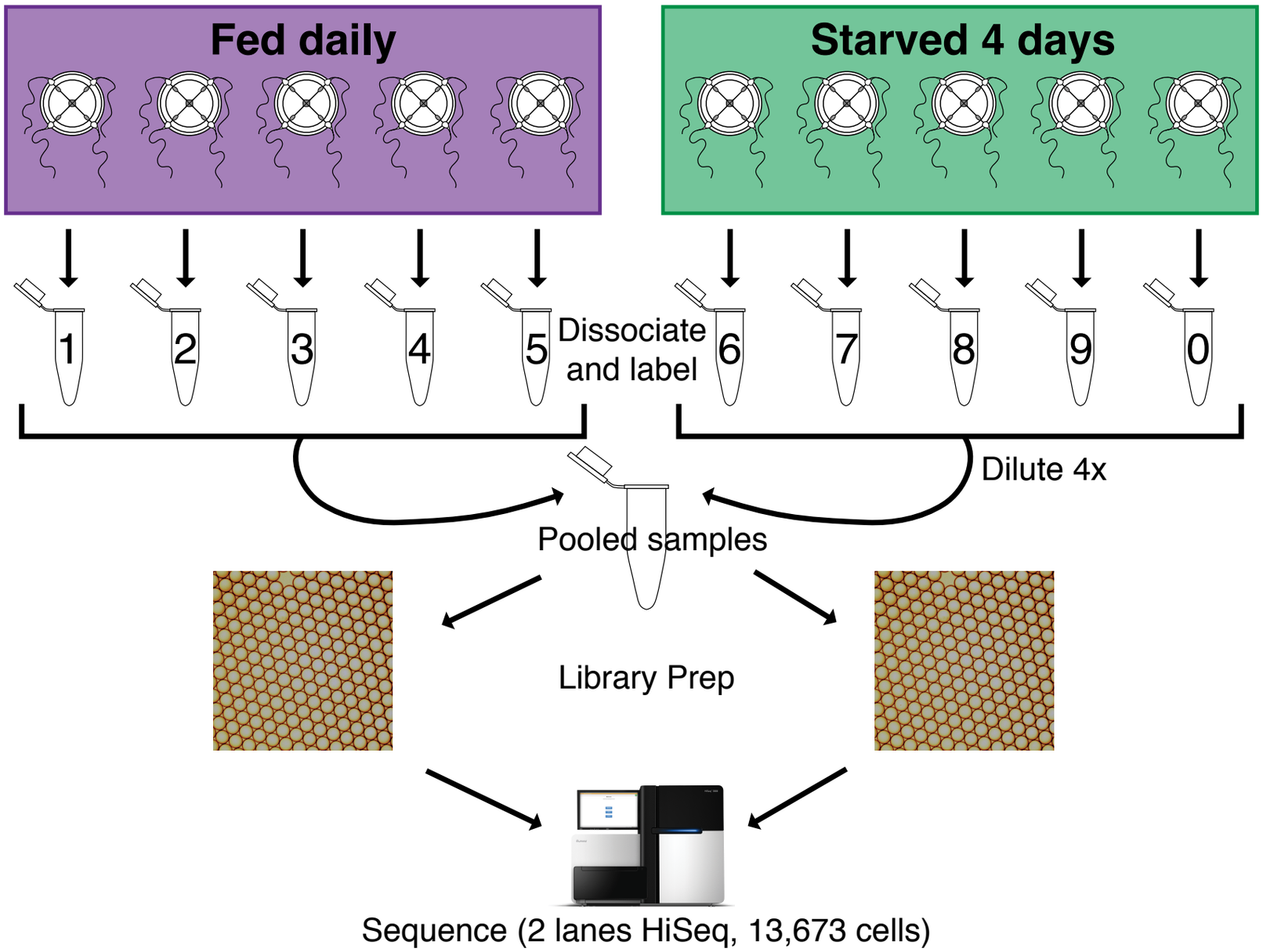
Clytia hemisphaerica



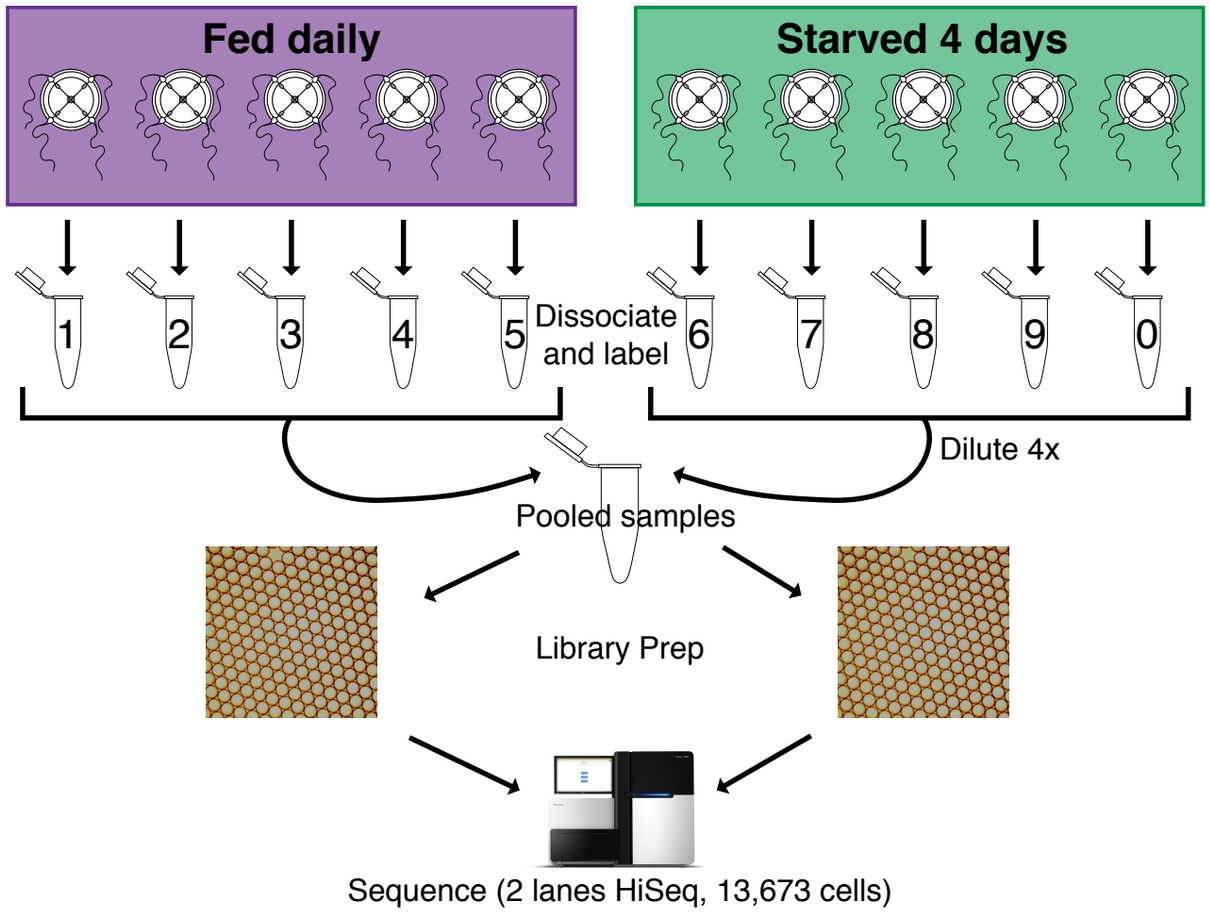
***Clytia* anatomy**

Adult *Clytia* medusa

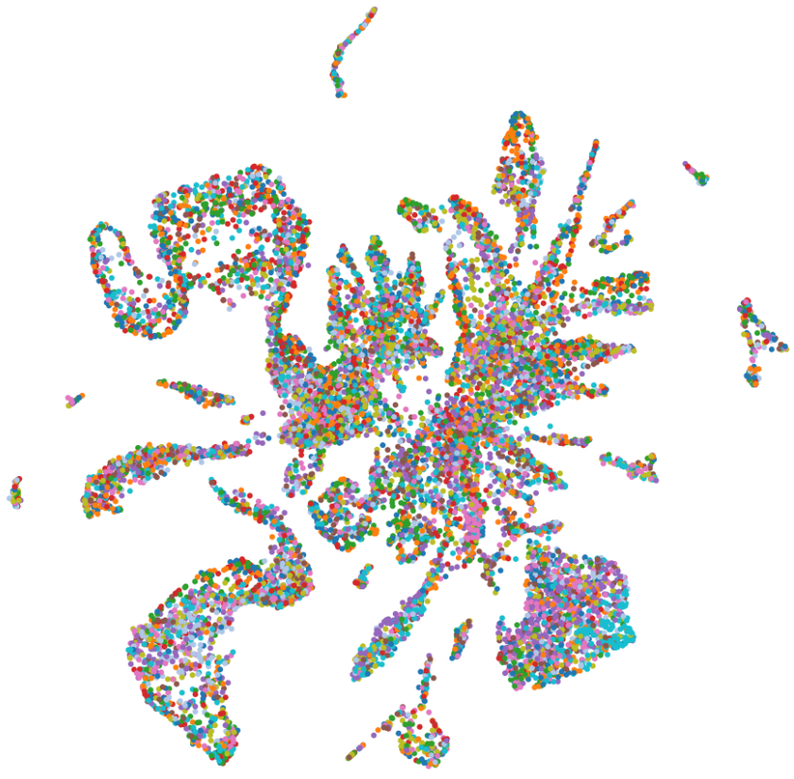
Experiment: single-cell RNA-seq



A view of the results

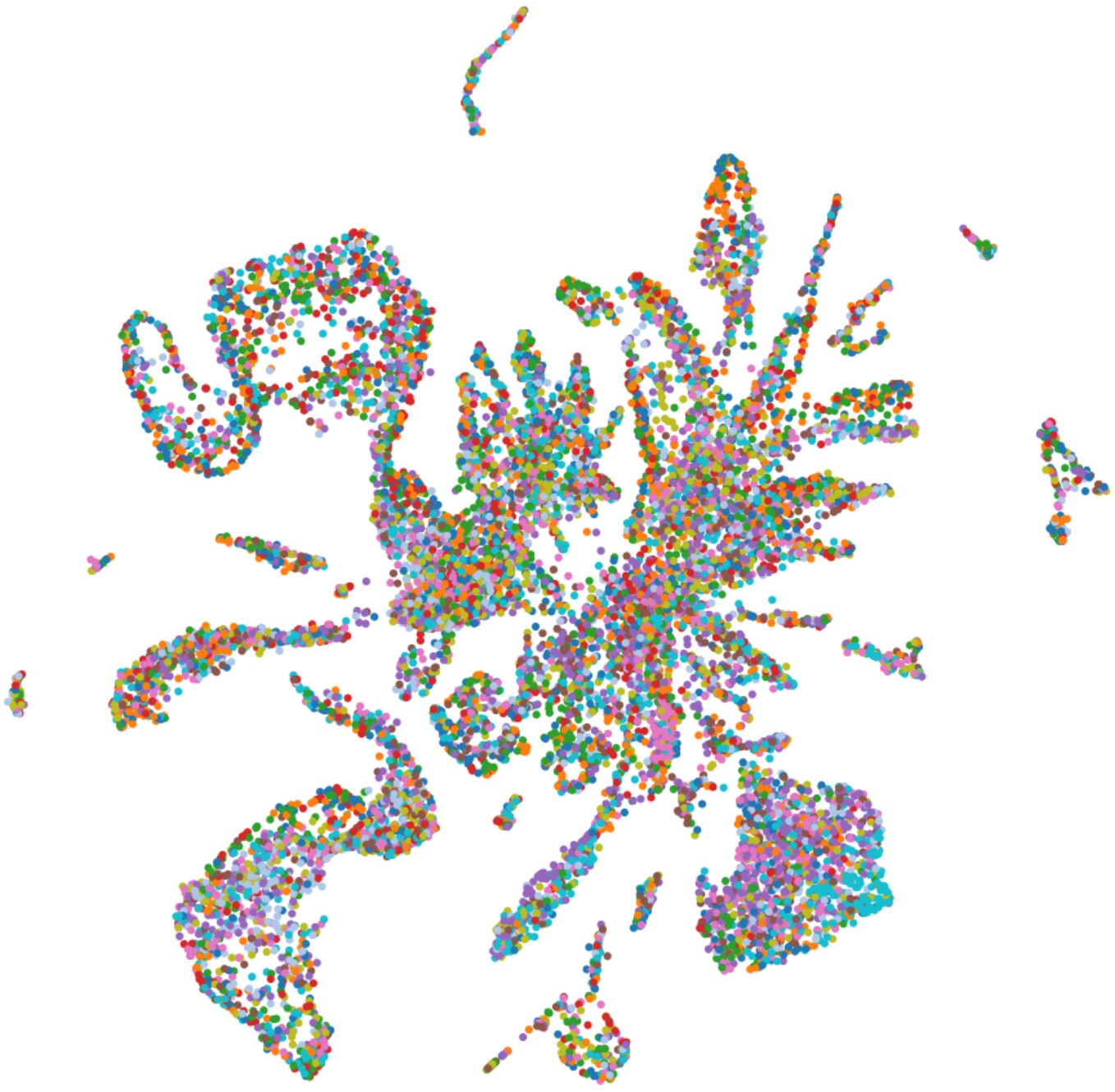


943 M reads | 1,367 ± 257 cells/animal | 705 genes/cell

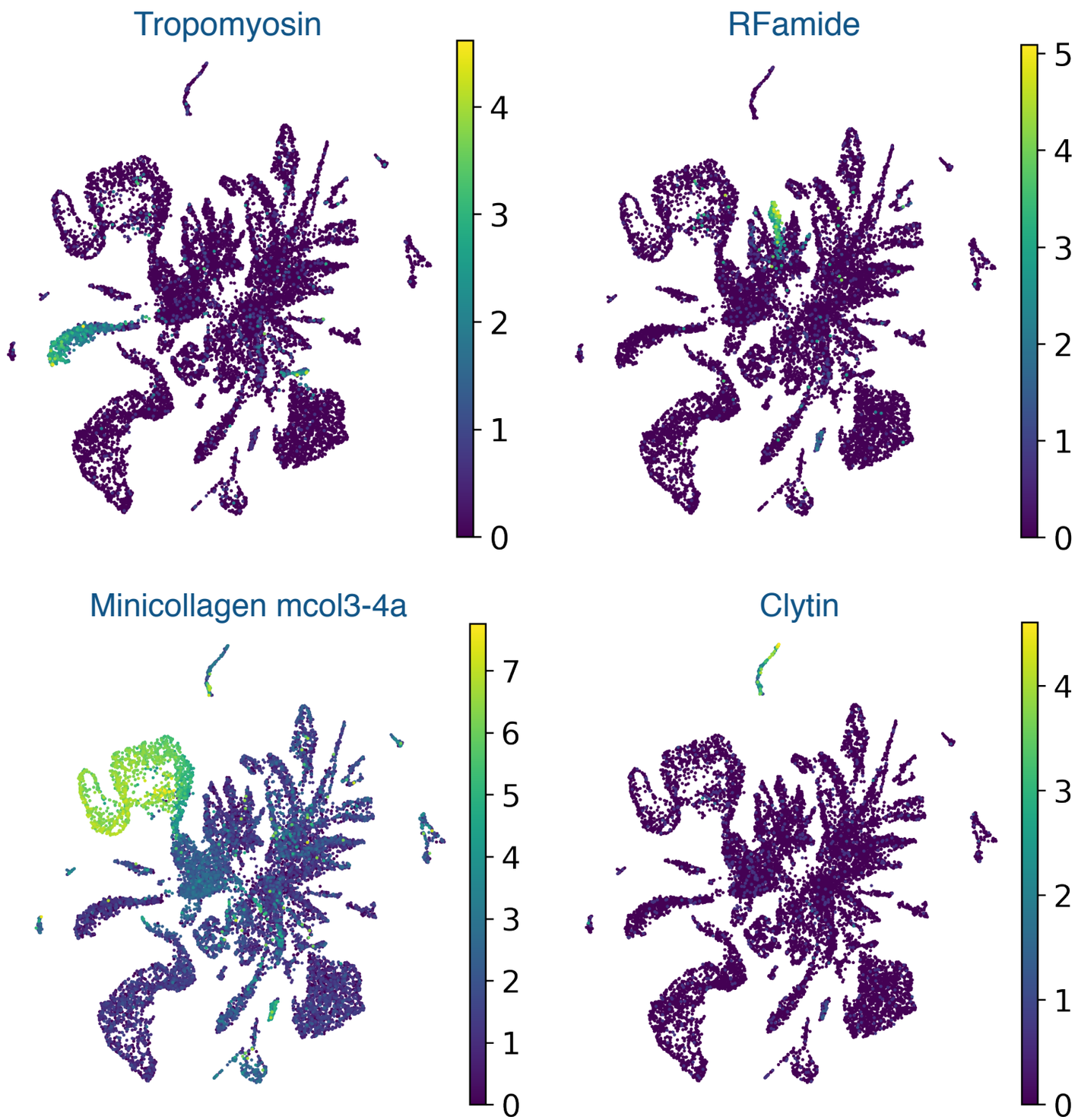


Organism	Cell yield
1 - Fed	1,235
2 - Fed	1,305
3 - Fed	1,274
4 - Fed	990
5 - Fed	1,807
6 - Starved	1,064
7 - Starved	1,467
8 - Starved	1,470
9 - Starved	1,785
10 - Starved	1,276

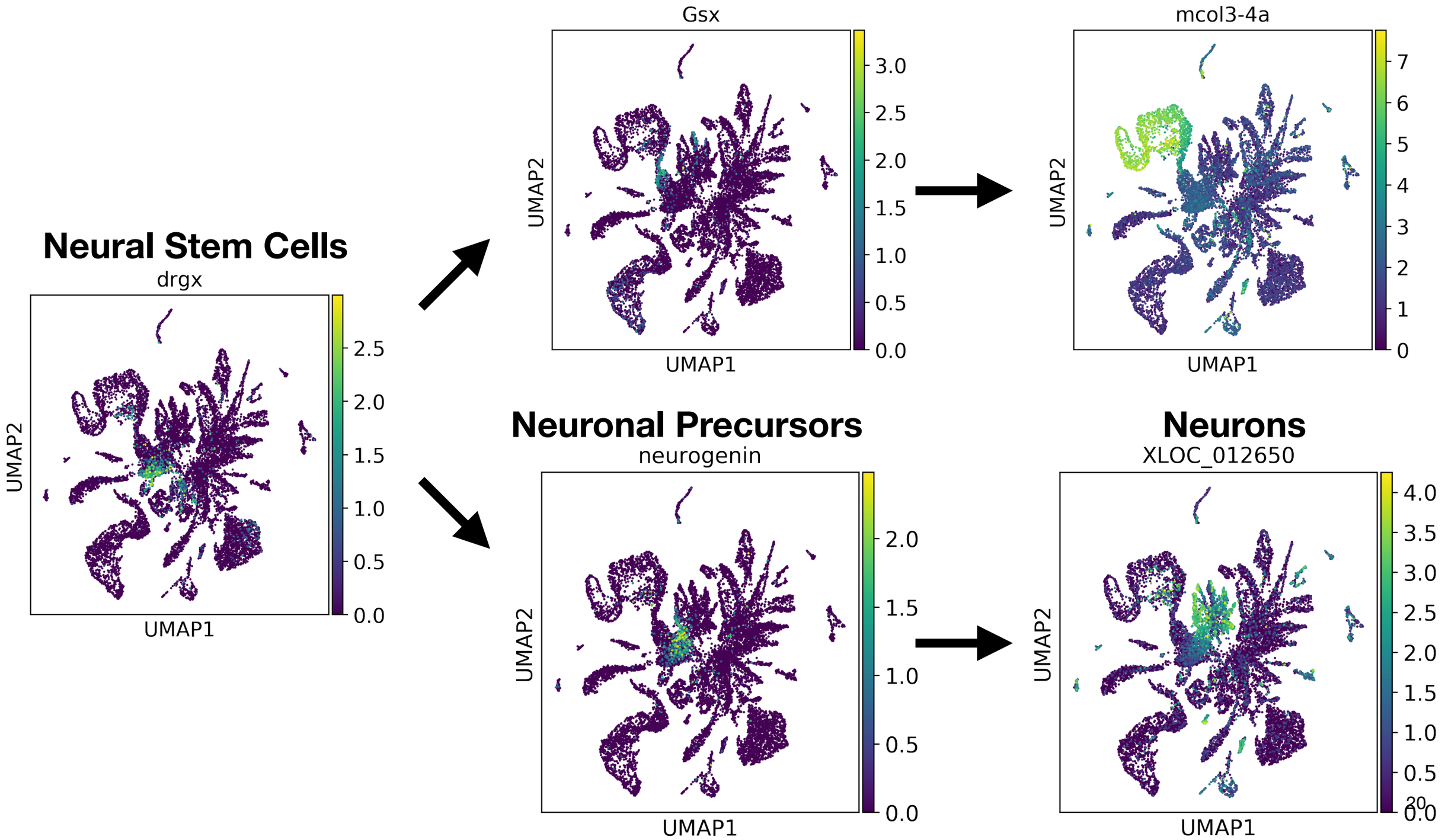
The t-SNE



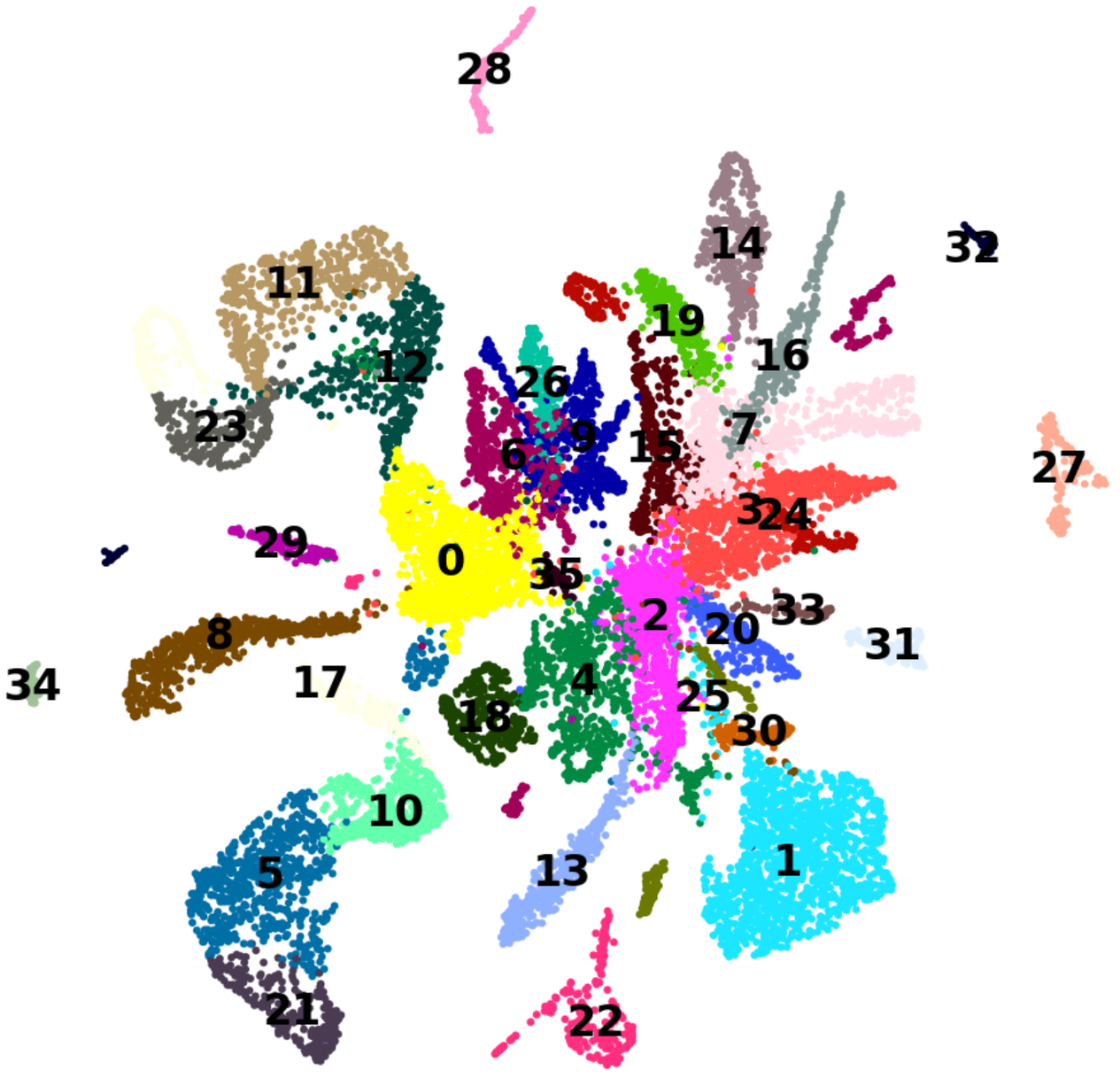
Biological interpretation of mathematically meaningful clusters



A hint at dynamics: neuronal differentiation



Towards a Clytia hemisphaerica cell atlas



- 0 Neural stem cells
- 1 Blank
- 2 Blank
- 3 Blank
- 4 Blank
- 5 Conodopine-2
- 6 SynTag5 neurons
- 7 Canal endoderm
- 8 Muscle
- 9 GLWamide neurons
- 10 Conodopine-1
- 11 Mid Cnidocytes
- 12 Early cnidocytes
- 13 Bell rim endoderm
- 14 Blank
- 15 Blank
- 16 Blank
- 17 Late cnidocytes 1
- 18 Blank
- 19 Blank
- 20 Blank
- 21 Conodopine-3
- 22 Neuronal subtype
- 23 Late cnidocytes 2
- 24 Blank
- 25 Neuronal subtype
- 26 RFamide neurons
- 27 Neuronal subtype
- 28 Bioluminescent neurons
- 29 Gastrovascular endoderm
- 30 Blank
- 31 pp9 neurons
- 32 Blank
- 33 Blank
- 34 Blank
- 35 Oocytes

Questions relevant to extracting information

- What is the “right” approach to **dimension reduction**?
  - PCA, t-SNE, column subset selection, … ?
- What is the “best” **clustering** algorithm?
  - Louvain, k-means, k-medoids, …
- How should cell **imputation** be performed?
  - NNMF, affinity propagation, …
- How should **differential expression** be applied to identifying markers?
  - Linear regression, negative binomial modeling …

Thank you

Jase Gehring
molecular biology and biochemistry

Vasilis Ntranos
information theory

Valentine Svensson
mathematics and statistics

Lynn Yi
computer science, statistics and medicine

Sina Booeshaghi
mechanical engineering

Eduardo Beltrame
molecular biology

Taleen Dilanyan
chemistry

Kristjan Eldjarn
computer science

Gennady Gorin
chemical engineering

Lambda
computational biology

Nadia Volovich
biophysics

Fan Gao (Bioinformatics Core)
bioinformatics

Kerkchoff Laboratories of Biology, Caltech