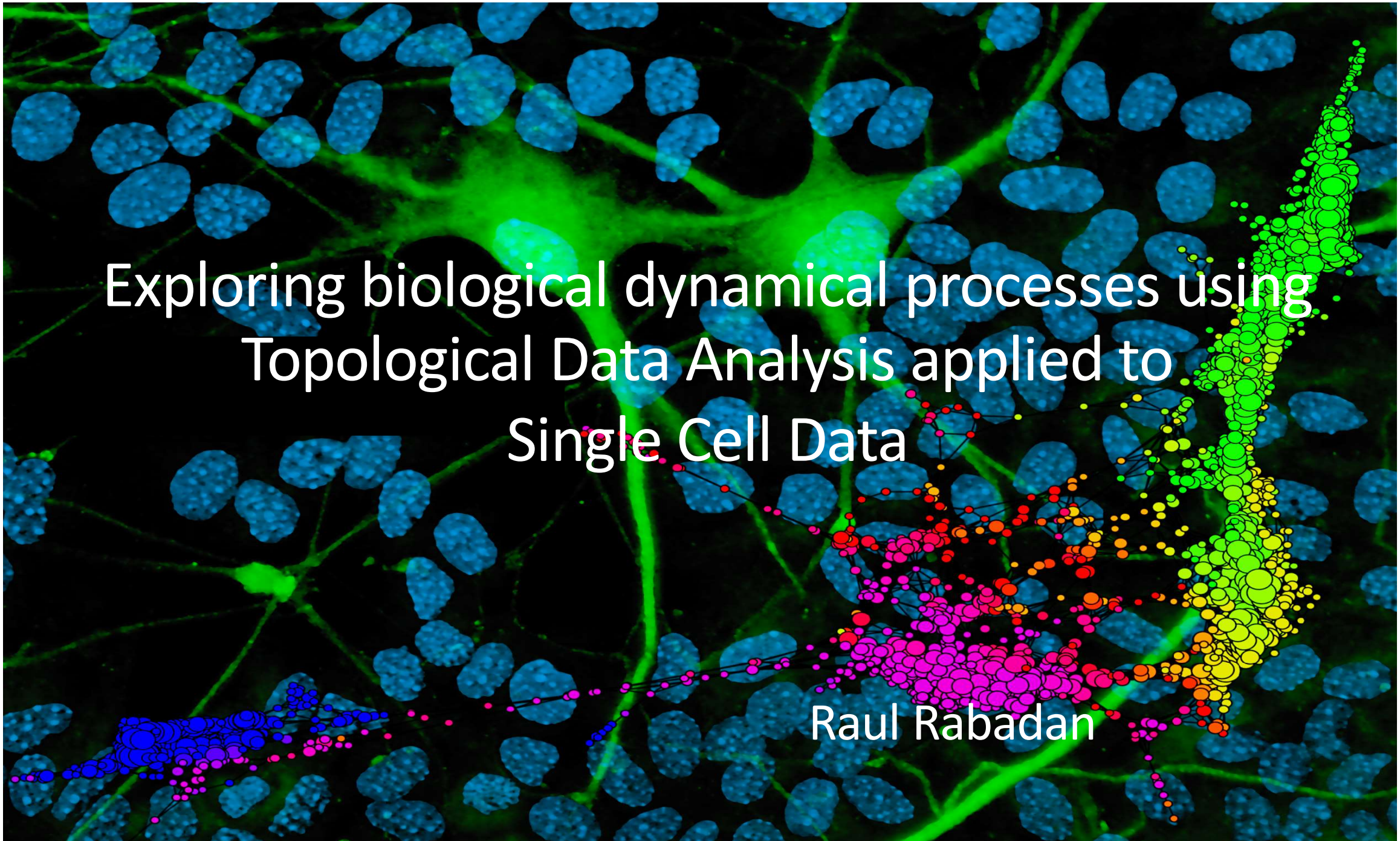


Exploring biological dynamical processes using Topological Data Analysis applied to Single Cell Data

Raul Rabadan

Outline

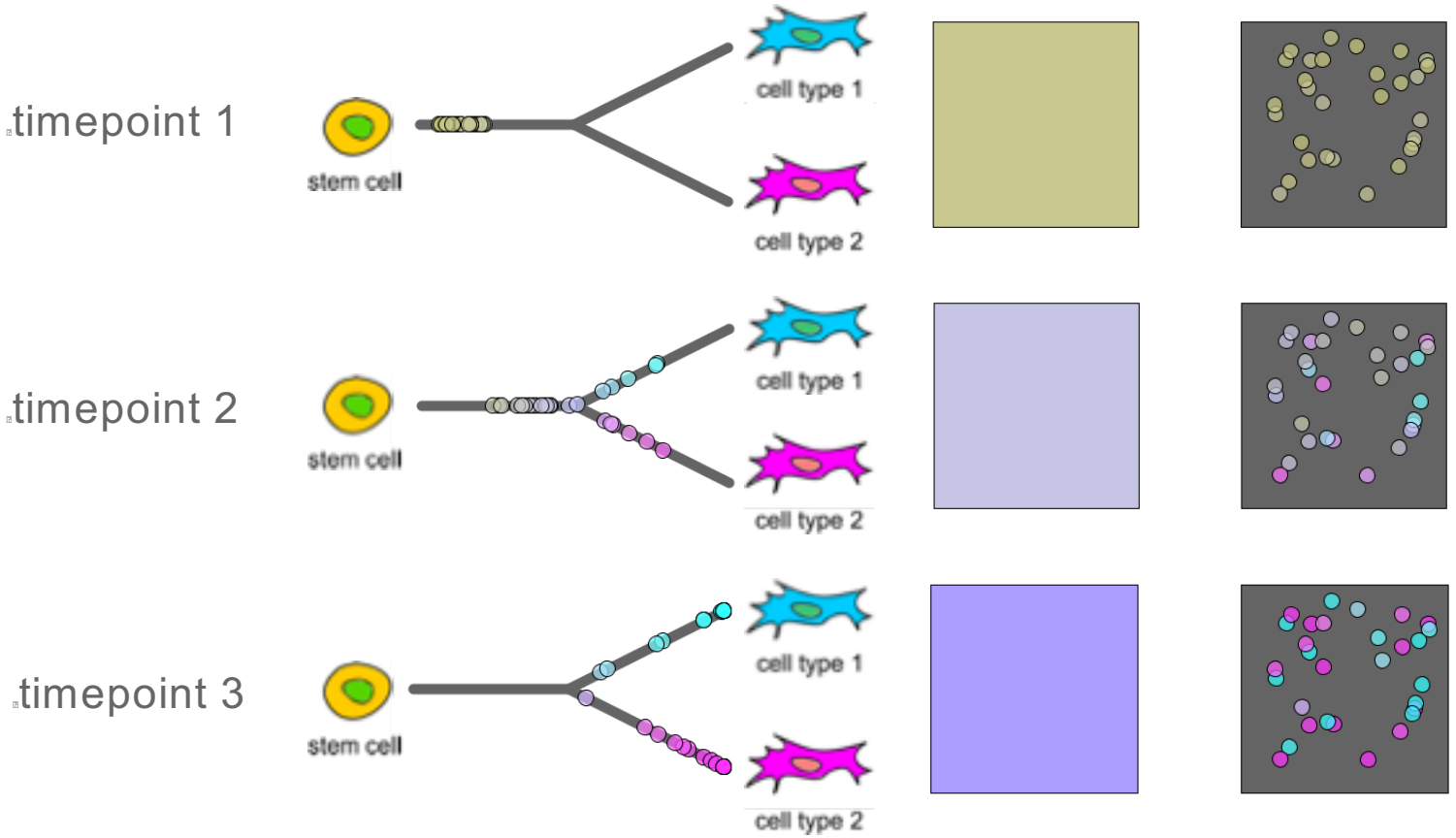
- Introduction to the abstract biological problem.
- Study single cell expression data using topological data analysis.
  - First example: development of motor neurons.
  - Second example: studying heterogeneity and evolution in cancer.
- Study HiC single cell expression data using topological data analysis.

Dynamic Problems in biology

- Cell Cycle.
- Development.
- Tumor heterogeneity and evolution.
- Spread of infectious diseases.
- Speciation.
- Many others…

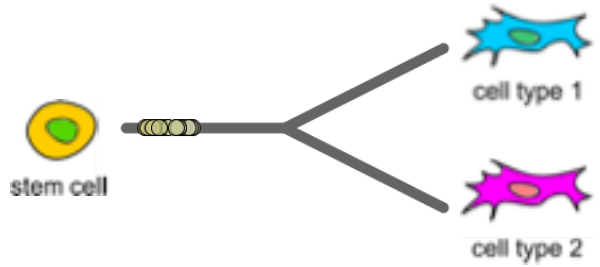
Single cell expression

- Single cell expression can capture snapshots of these processes that are missed in bulk analysis.



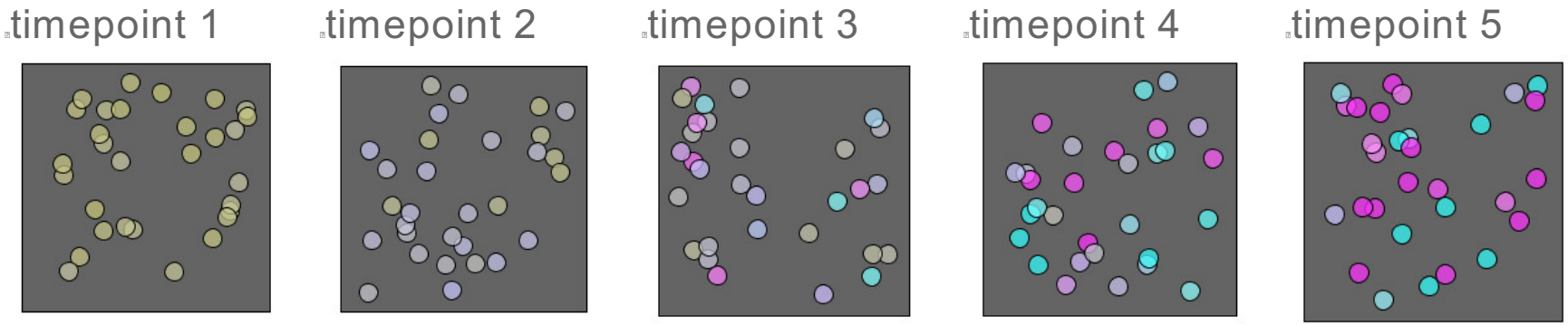
Characteristics of time processes

- **Heterogeneous processes**: different cells types are present simultaneously.
- **Continuous processes**: different cells present different transcriptional programs keeping memory of ancestral states.



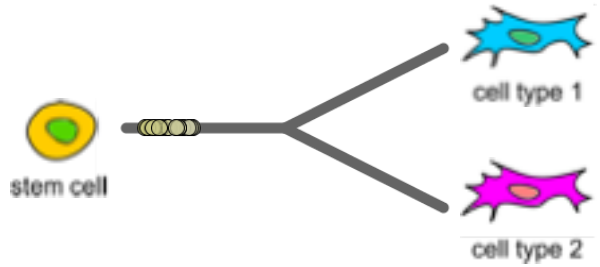
Characteristics of time processes

- **Heterogeneous processes**: different cells types are present simultaneously.
- **Continuous processes**: different cells present different transcriptional programs keeping memory of ancestral states.
- **Asynchronous processes**: different cells present different transcriptional programs evolving at different pace.



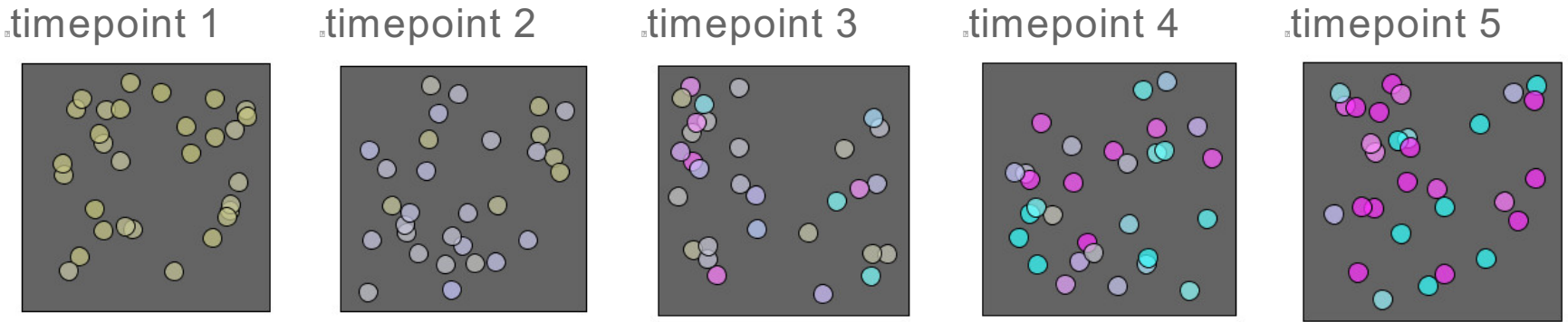
Characteristics of time processes

- **Heterogeneous processes**: different cells types are present simultaneously.
- **Continuous processes**: different cells present different transcriptional programs keeping memory of ancestral states.
- **Asynchronous processes**: different cells present different transcriptional programs evolving at different pace.
- **Branched or cycling processes**: structure from undifferentiated to diverse differentiated stages can be represented by tree like or cycling structures.



The problem

- Reconstruct biological relevant structures from snapshots of data. In particular considering:
  - **Continuity**: not capture by traditional clustering techniques.
  - **Asynchronicity**.
  - **Capturing low dimensional structures** (trees and cycles).



Further complications

- The problem is further complicated as the number of genes (dimensions where data is embedded) is of order 10,000, a number larger or similar to number of cells studied with most common single cell techniques.

Goals

- To develop a method that can
  1. capture continuous structures,
  2. capture asynchronous processes, and takes time into account,
  3. capture low dimensional features of the data (trees and cycles).
  4. perform statistics to identify transcriptional programs associated to different substructures (subpopulations or stages of differentiation).

Outline

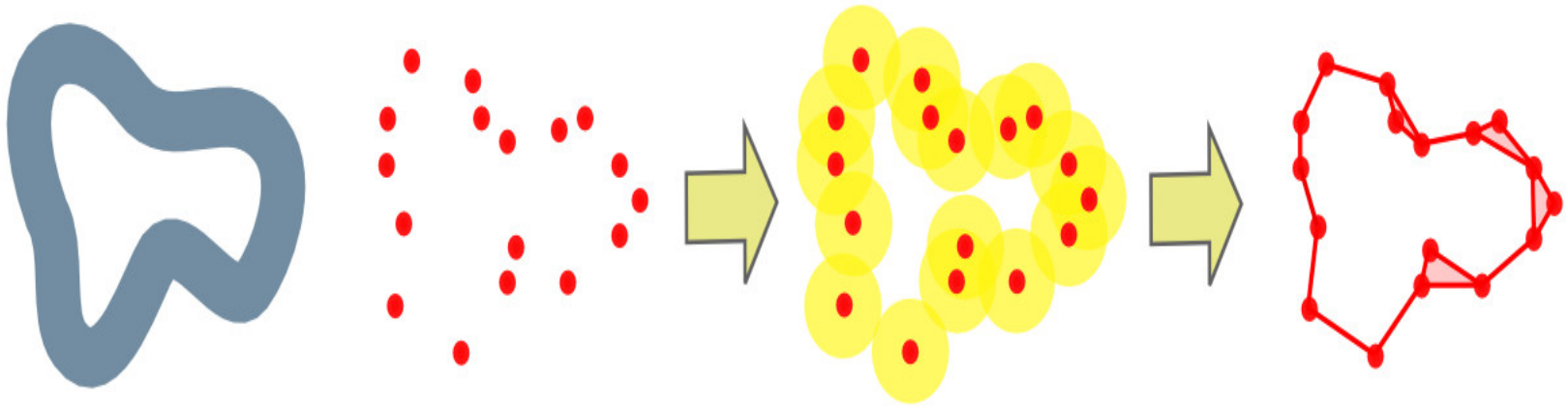
- Introduction to the abstract biological problem.
- **Study single cell expression data using topological data analysis.**
  - First example: development of motor neurons.
  - Second example: studying heterogeneity and evolution in cancer.
- Study HiC single cell expression data using topological data analysis.

Modest approach

- Instead of learning the whole space, which is theoretically unfeasible, let us take a very modest approach, learn its skeleton, just a low dimensional caricature of the space.
- Advantages:
  - Theoretically possible.
  - Biologically interpretable features are low dimensional.
- Disadvantage:
  - It misses most of information in dimension bigger than 1.

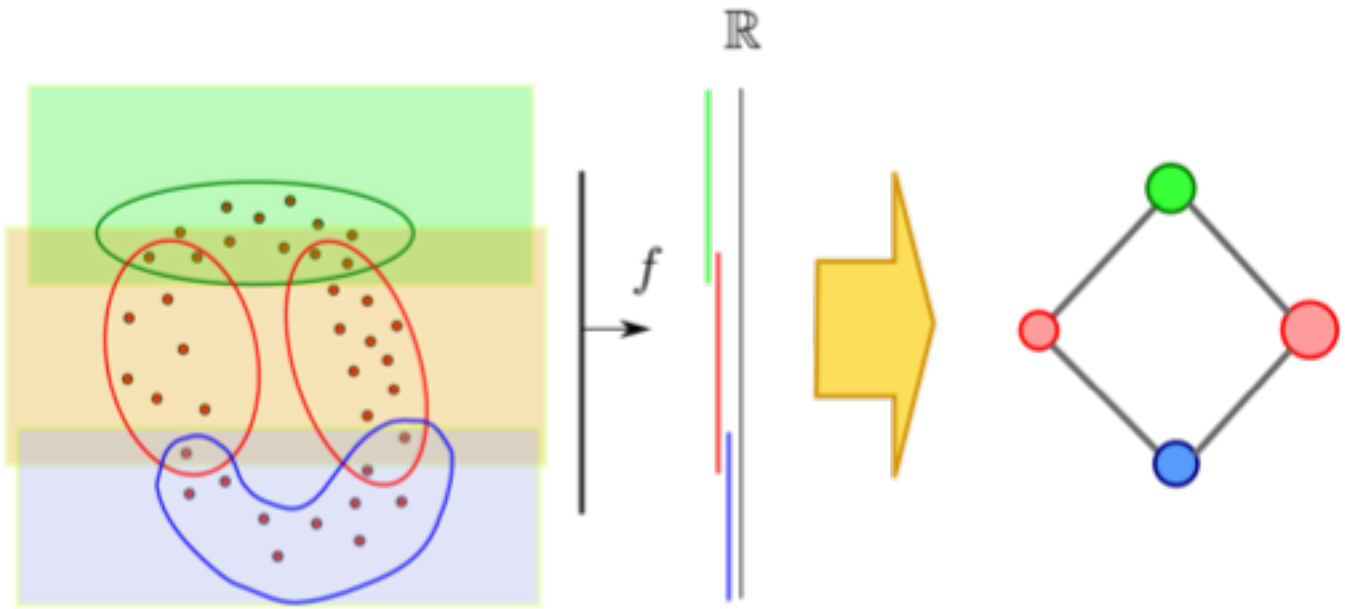
Topological Data Analysis (TDA)

- TDA: sets of techniques to learn continuous features of data.
- Usually by getting auxiliary objects as simplicial complexes.



Algorithm: Mapper

- Reconstruct the skeleton of a space from finite sampling of data. Captures and summarizes data.

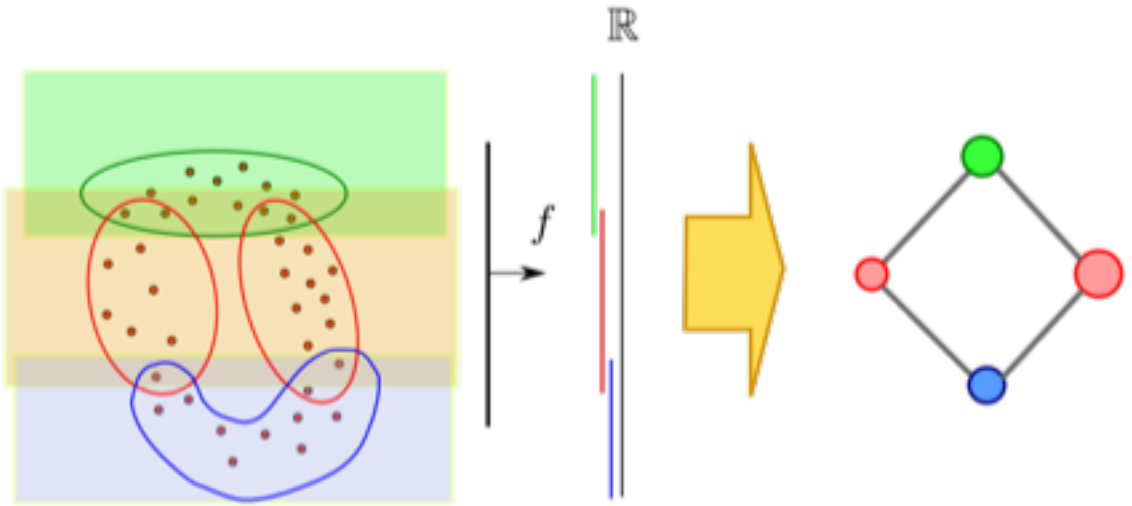


G. Singh , F. Memoli, G. Carlsson (2007)

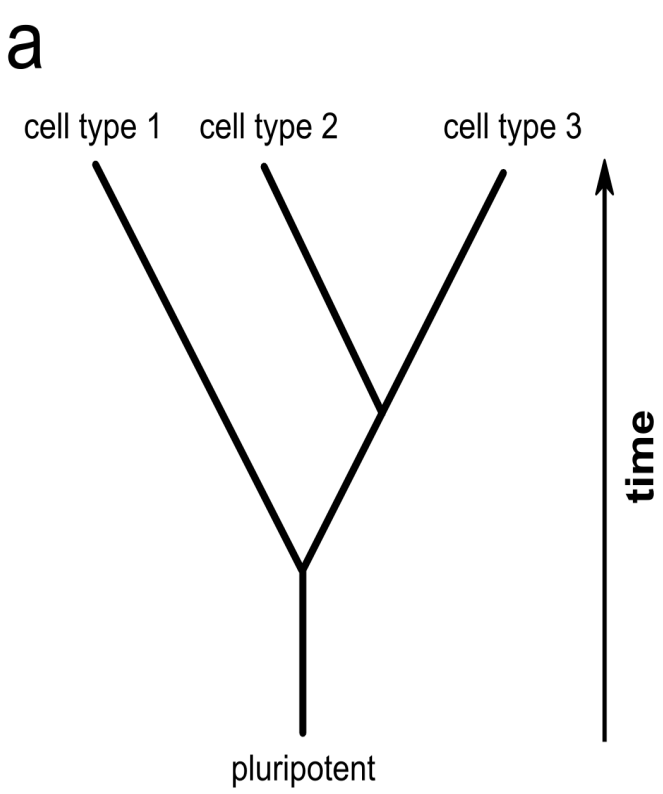
Algorithm: Mapper

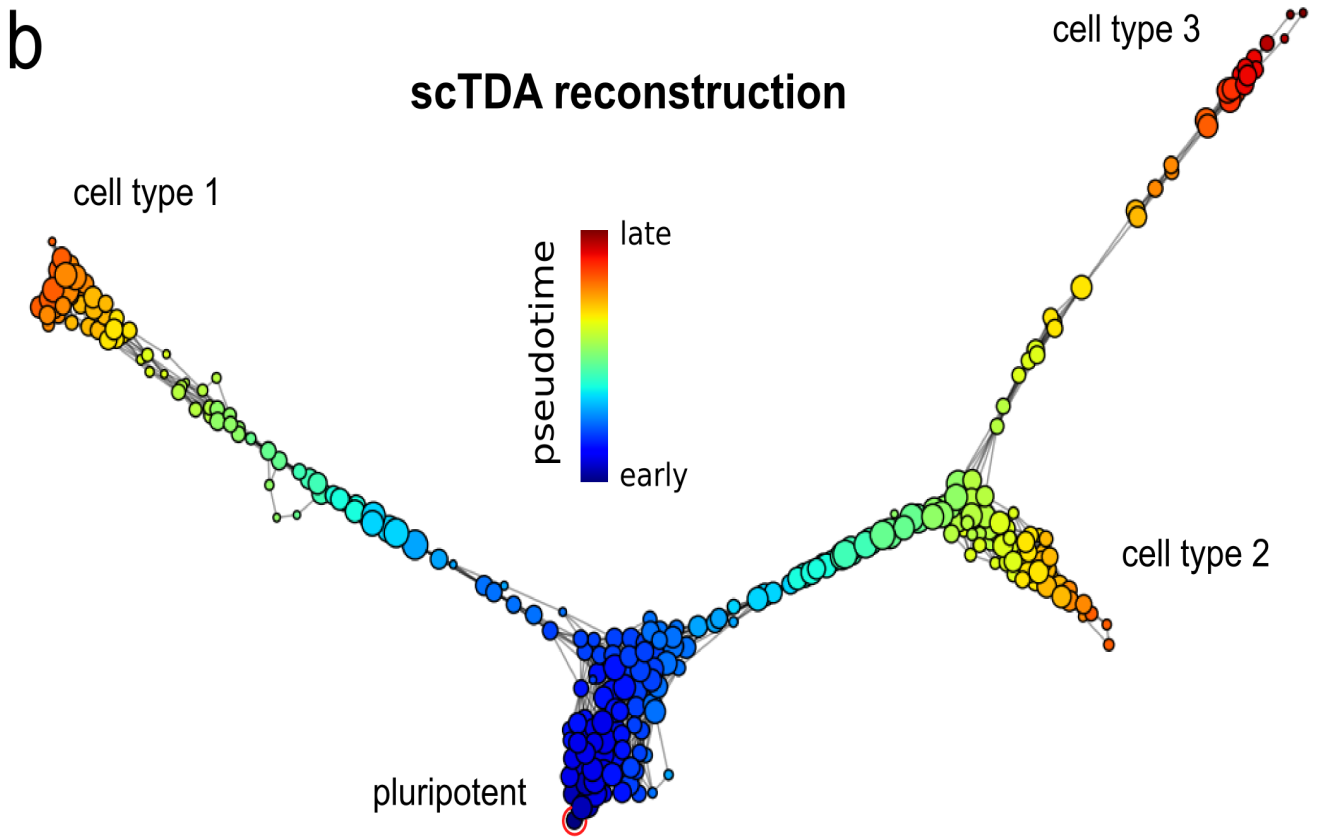
- A node represents a group of cells with similar transcription program (a "*cellular microstate*")

- An edge between two nodes represents a shared set of cells between the two nodes (a "*microstate transition*")



Synthetic data



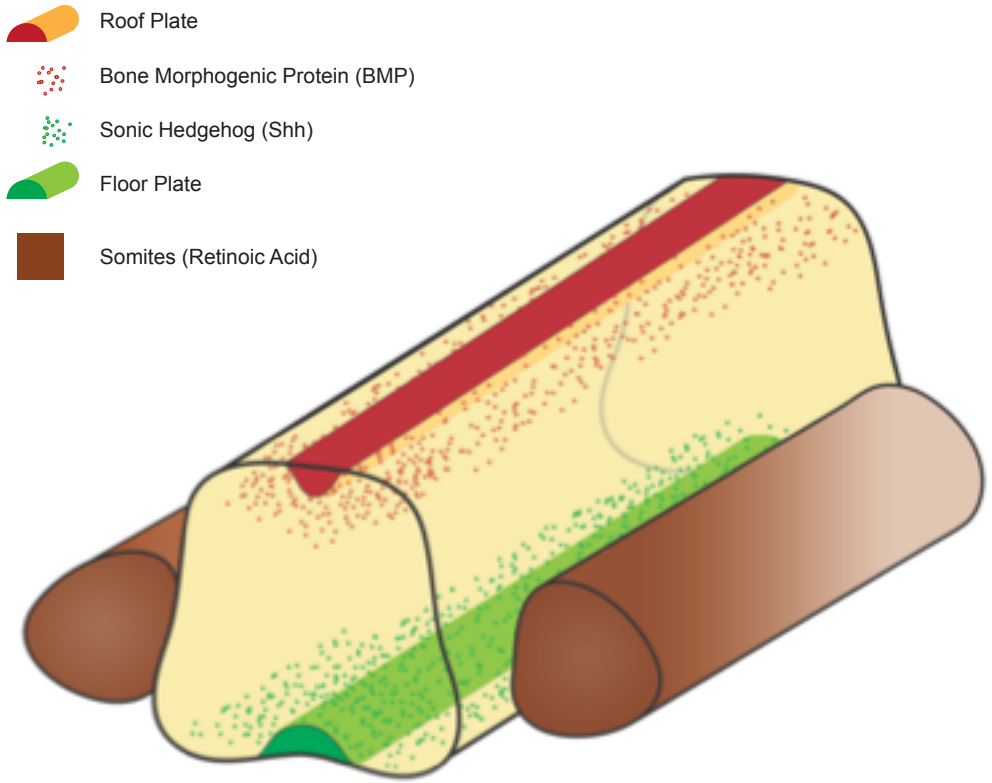


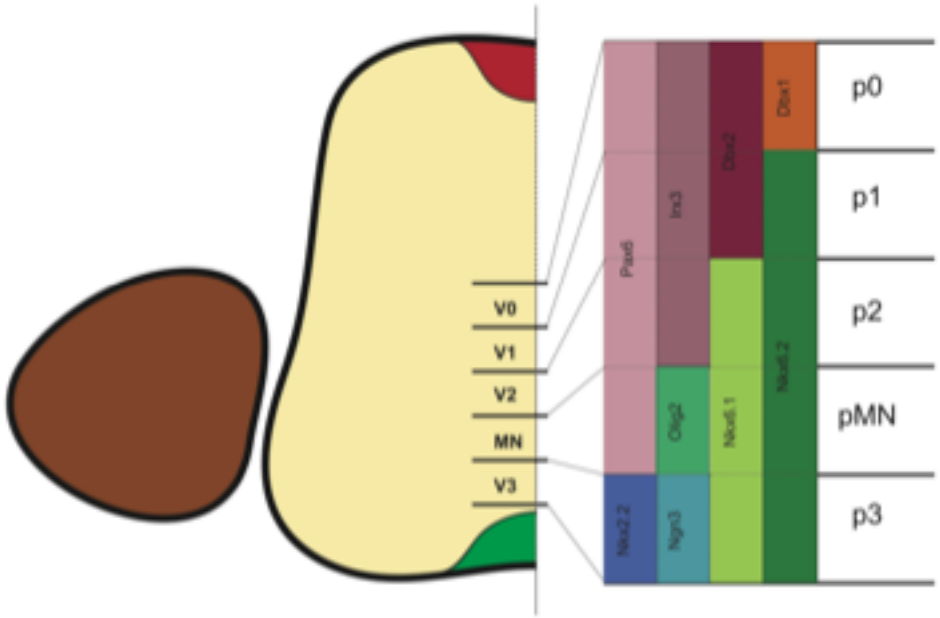
Outline

- Introduction to the abstract biological problem.
- Study single cell expression data using topological data analysis.
  - **First example: development of motor neurons**.
  - Second example: studying heterogeneity and evolution in cancer.
- Study HiC single cell expression data using topological data analysis.

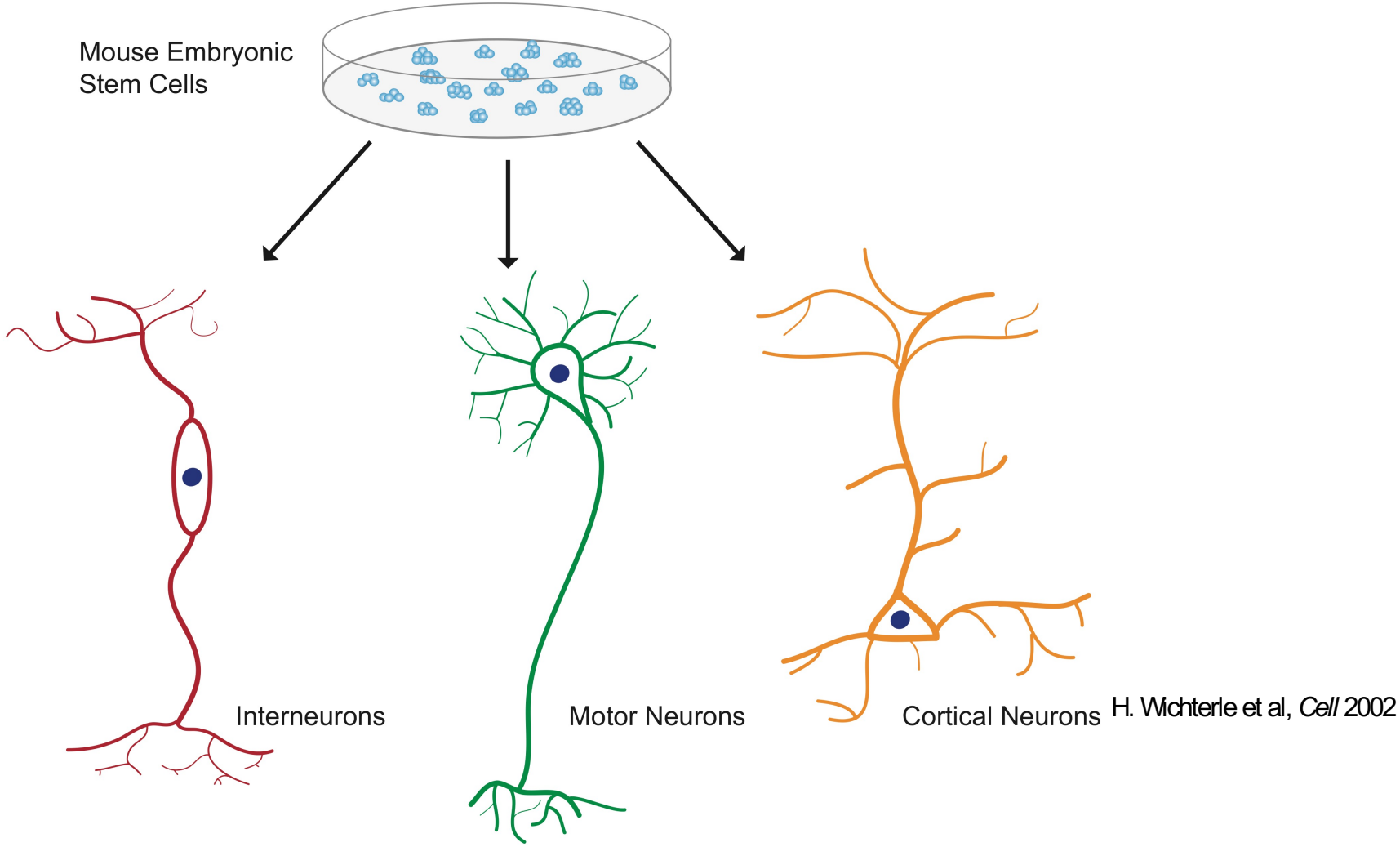
The developing mammalian spinal cord

- Neurons with cell body in the spinal cord and axons innervating muscles. Involved in various neurodegenerative disorders (e.g. ALS).





Stem cell based generation of neurons



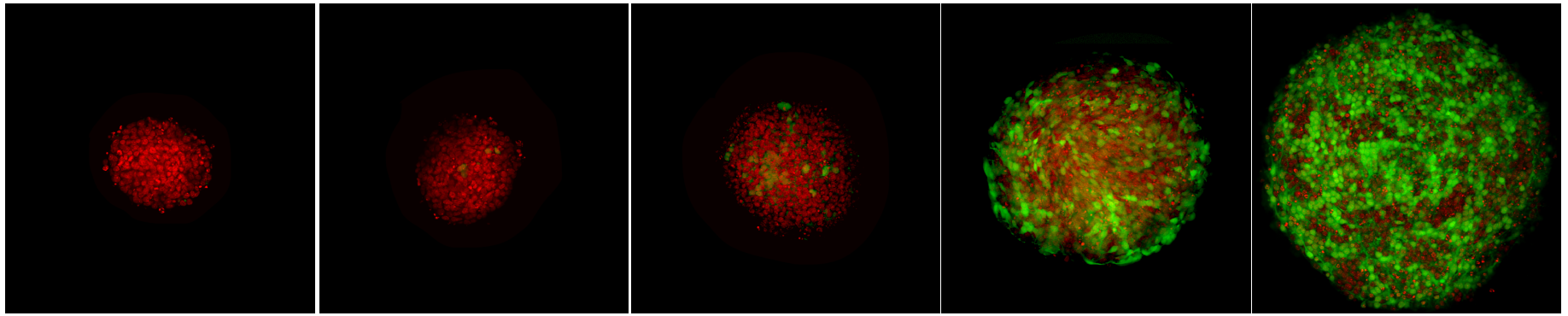
H. Wichterle et al, *Cell* 2002

Motor neuron development *in vitro*

During development MNs differentiate from the neural tube. MNs can be produced in vitro from mESCs by induction with sonic hedgehog and retinoic acid.

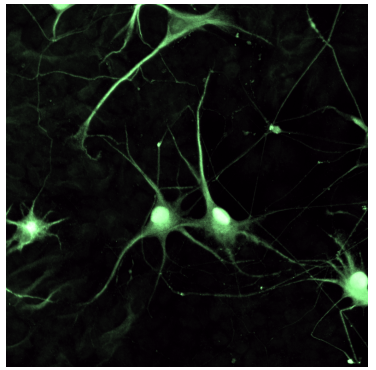
Day 2

Day 6



Induction
RA and Shh

Induction
RA and Shh



Hb9::eGFP
Nuclear Marker

RA = Retinoic Acid
Shh = Sonic Hedgehog

Abbas Rizvi, Maniatis Lab

Some questions about the process

Some key question about this process:

- What cell types and states arise throughout the differentiation?
- What are the transcriptional signatures (coding and non-coding)associated with different stages?
- What is the full cohort of transcriptional regulators controlling the differentiation process?

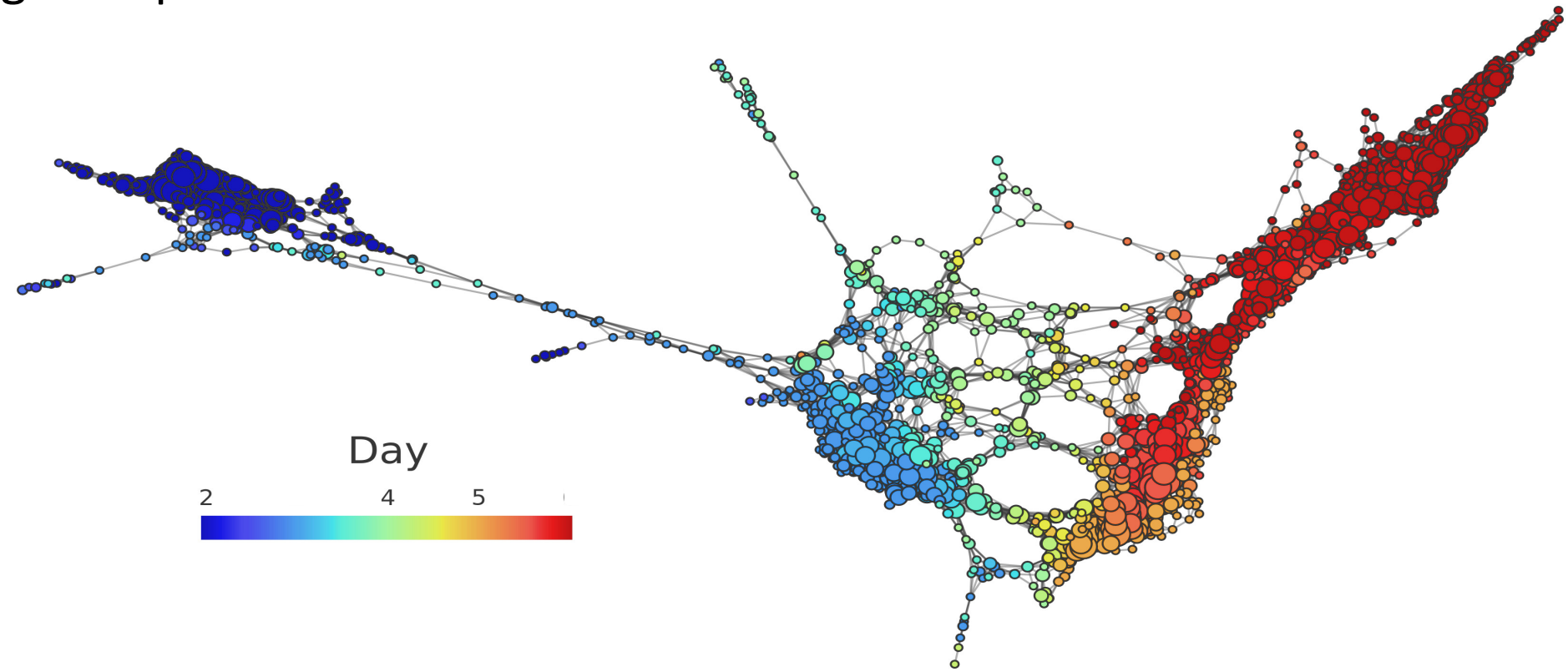
To answer these questions we performed single-cell RNA sequencing of 2,744 cells spanning days 2 to 6 of differentiation

Nature Biotechnology (2017).

Topological representation

We constructed a representation based on MDS and Pearson correlation similarity.

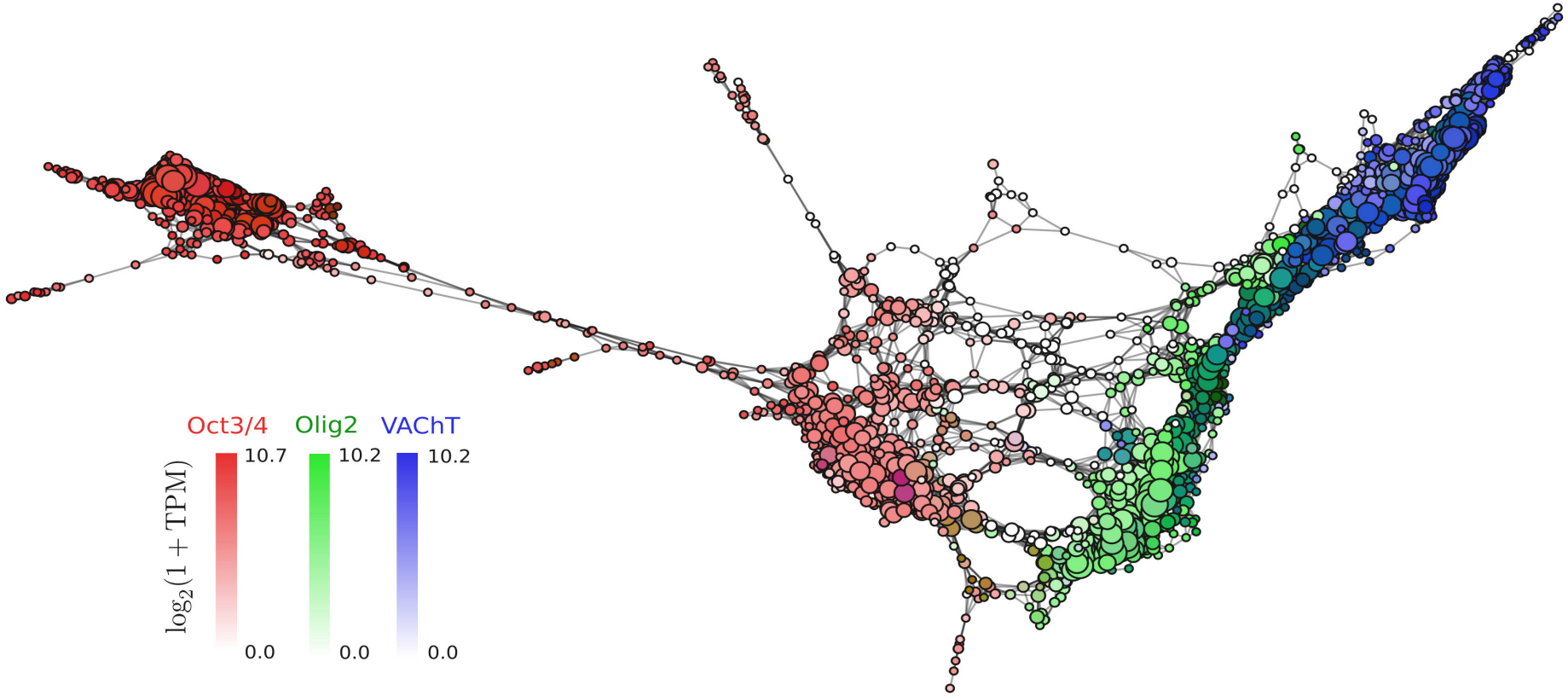
Topological representation follows differentiation time-line.



Nature Biotechnology (2017).

Topological representation

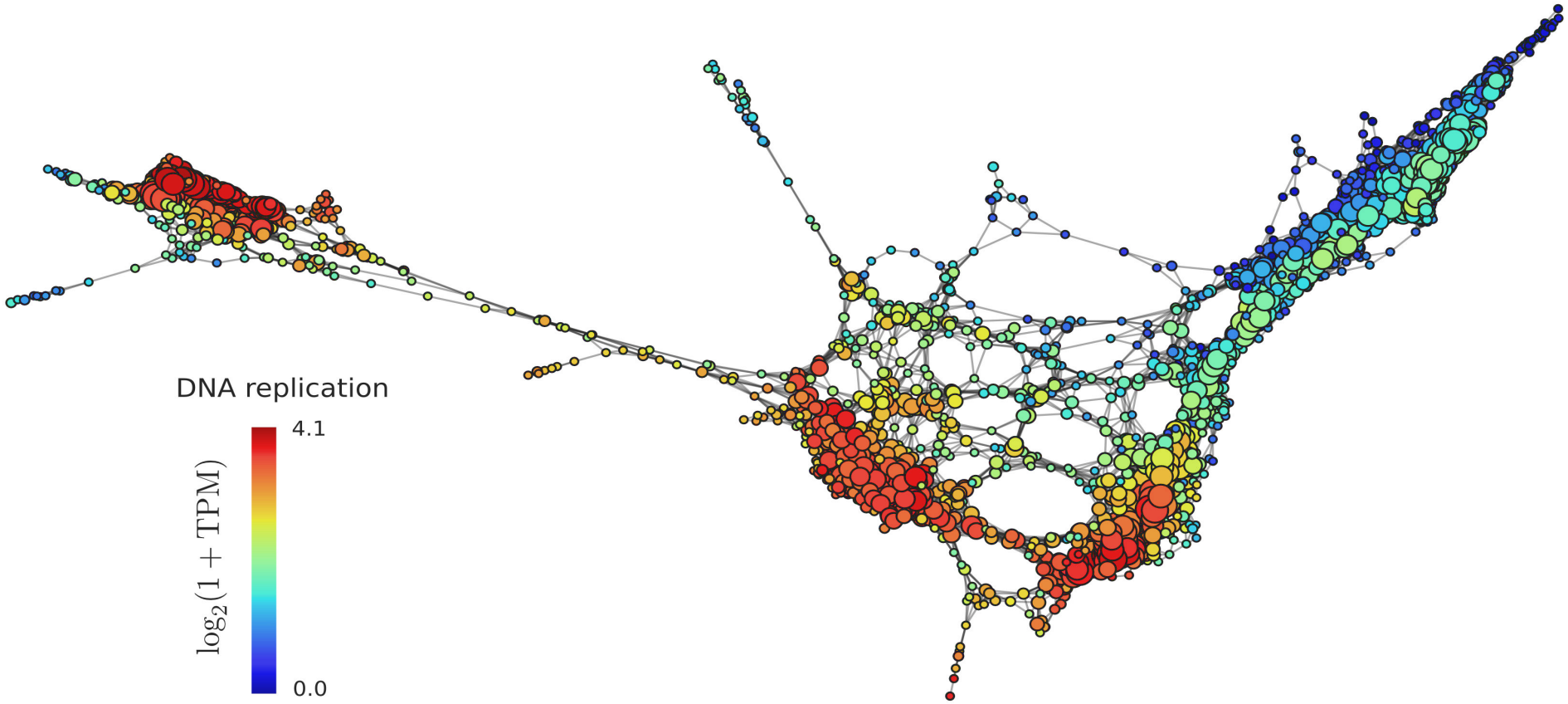
The topological representation captures known molecular markers of developmental progression of cellular states.



Nature Biotechnology (2017).

Topological representation

The topological representation captures cell cycle progression and post-mitotic arrest.



Nature Biotechnology (2017).

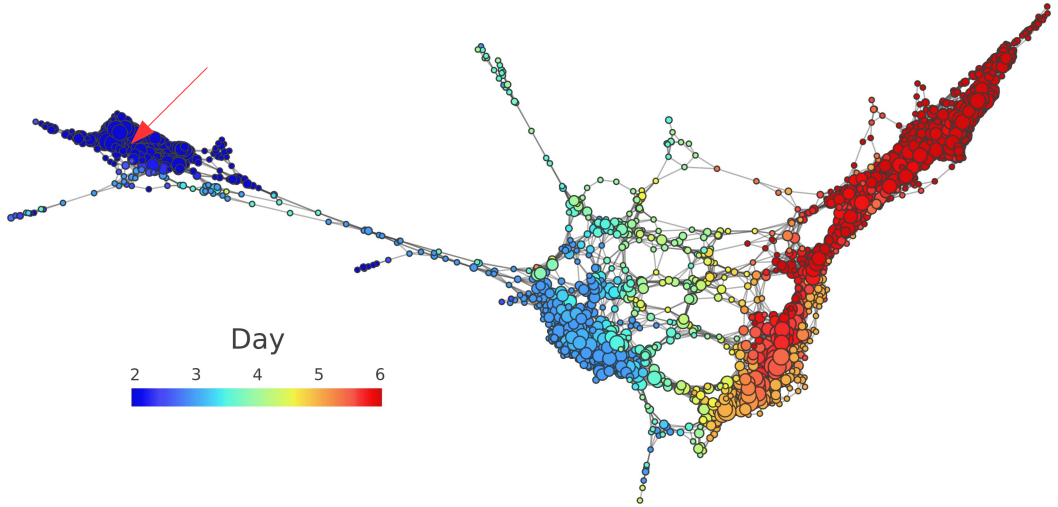
scTDA

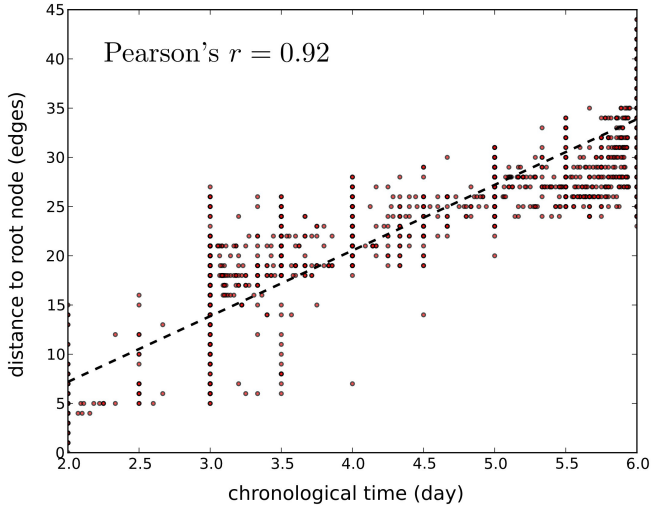
- We build upon the TDA representation to
  - Incorporate chronological information to identify distinct states.
  - Perform statistics on genes associated to different microstates to quantify local cell states that share a common transcriptional program.

- For lack of better name, we call it single-cell Topological Data Analysis (scTDA)

scTDA

Using time to infer progression: we define root node as the microstate that maximizes correlation between graph distance and chronological sampling time. It corresponds to the less differentiated cell microstate.





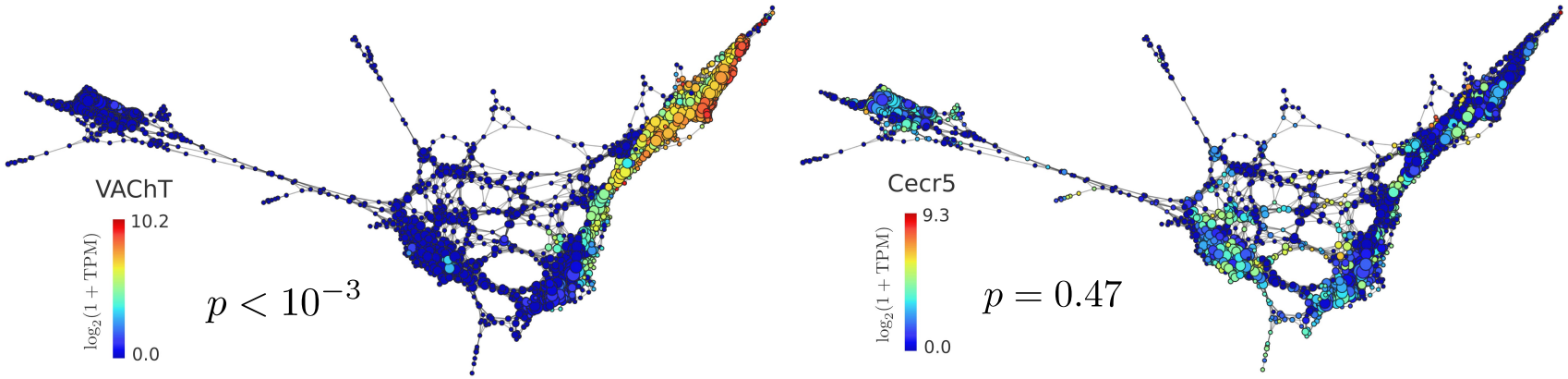
$$d_{\text{root}} \simeq a_0 + a_1 t$$

scTDA

Gene connectivity quantifies whether nearby cells in the topological representation express a given gene more than random.

$$S(g) = \frac{N}{N-1} \sum_{\alpha,\beta \in \Gamma} \frac{e_\alpha(g) A_{\alpha\beta} e_\beta(g)}{\left(\sum_{\gamma \in \Gamma} e_\gamma(g) \right)^2}$$

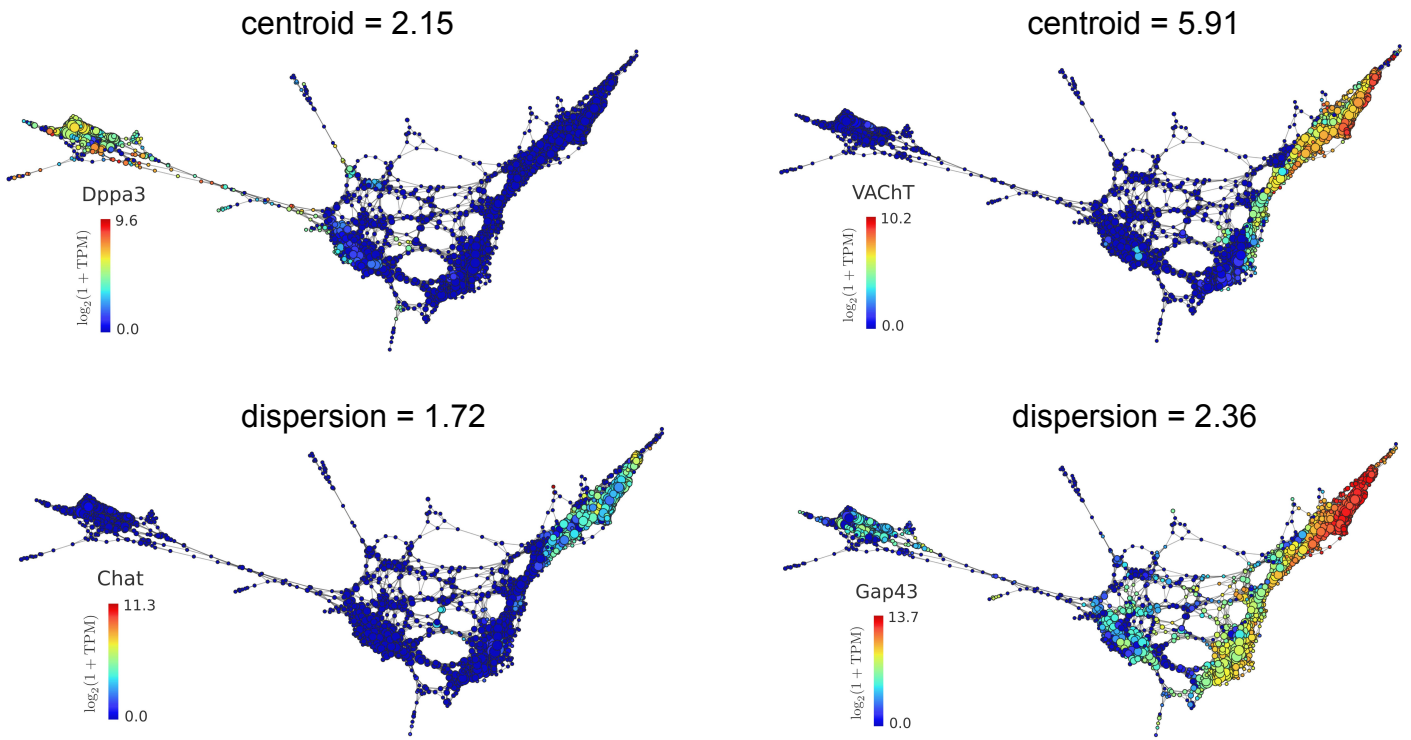
A null distribution is built for each gene using a permutation test.



scTDA

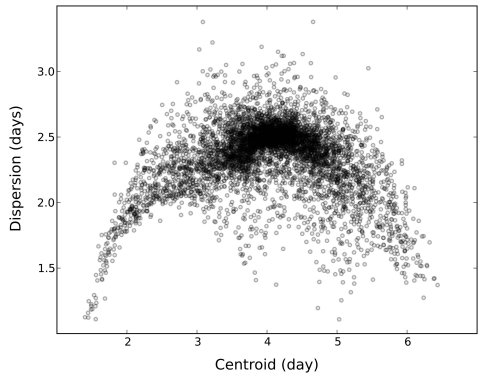
Centroid and dispersion locate where the gene is expressed and its variance.

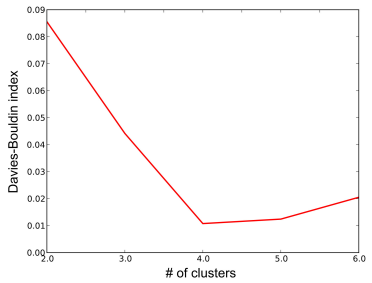
$$c(g) = \frac{1}{a_1}\left(\frac{\sum_{\alpha\in\Gamma} d_{\mathrm{root},\alpha} e_\alpha(g)}{\sum_{\beta\in\Gamma} e_\beta(g)} - a_0\right) \qquad d(g) = \frac{1}{a_1}\left(\sqrt{\frac{\sum_{\alpha\in\Gamma}(d_{\mathrm{root},\alpha} - c(g)a_1 + a_0)^2 e_\alpha(g)}{\sum_{\beta\in\Gamma} e_\beta(g)}} - a_0\right)$$

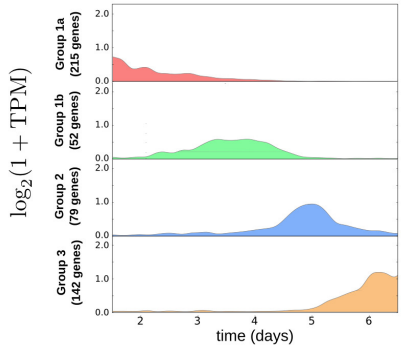


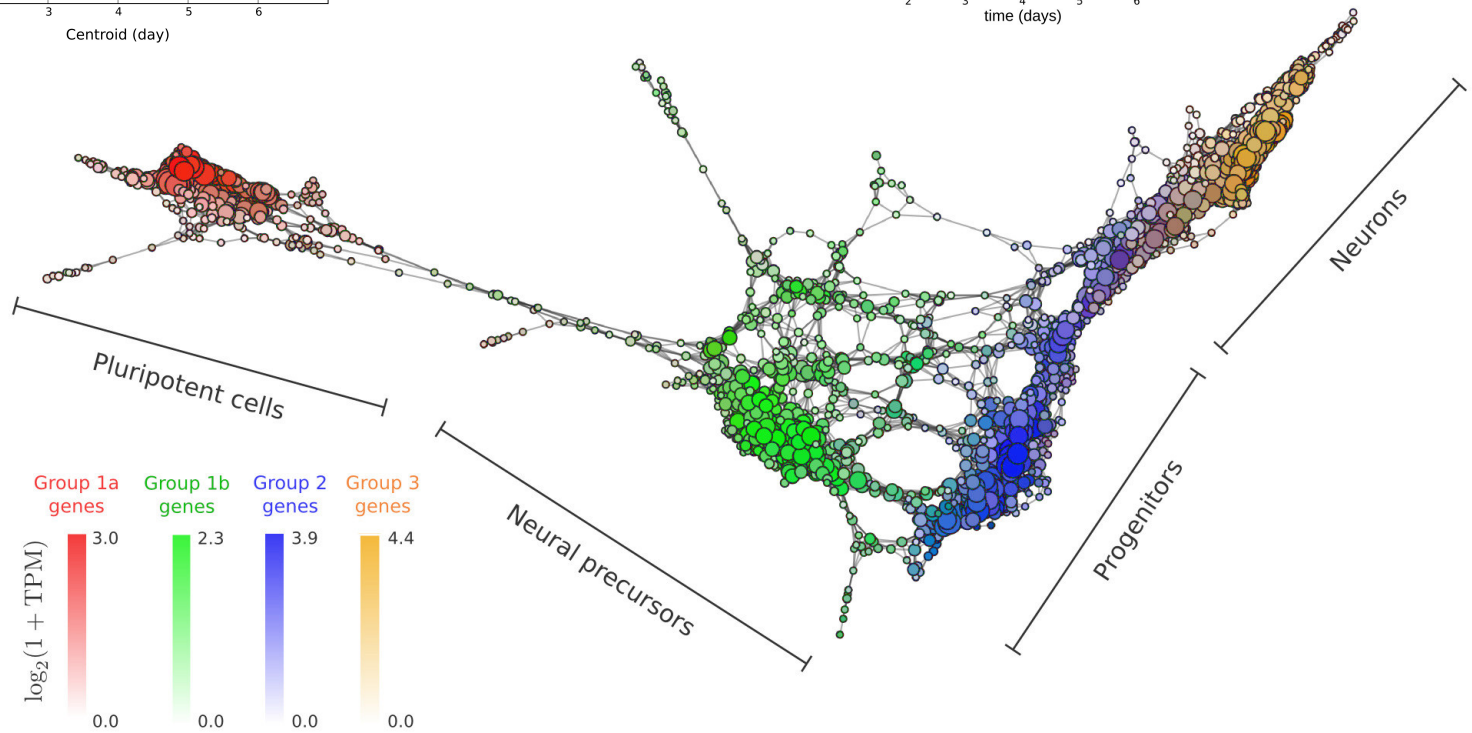
Transient cell types

Cell types/states are characterized by clusters of low dispersion genes.



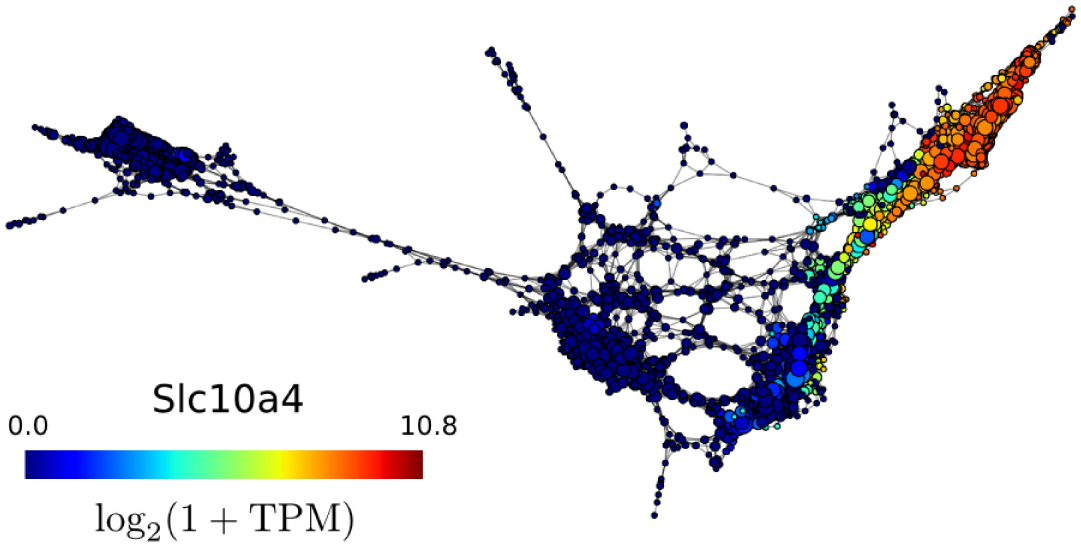






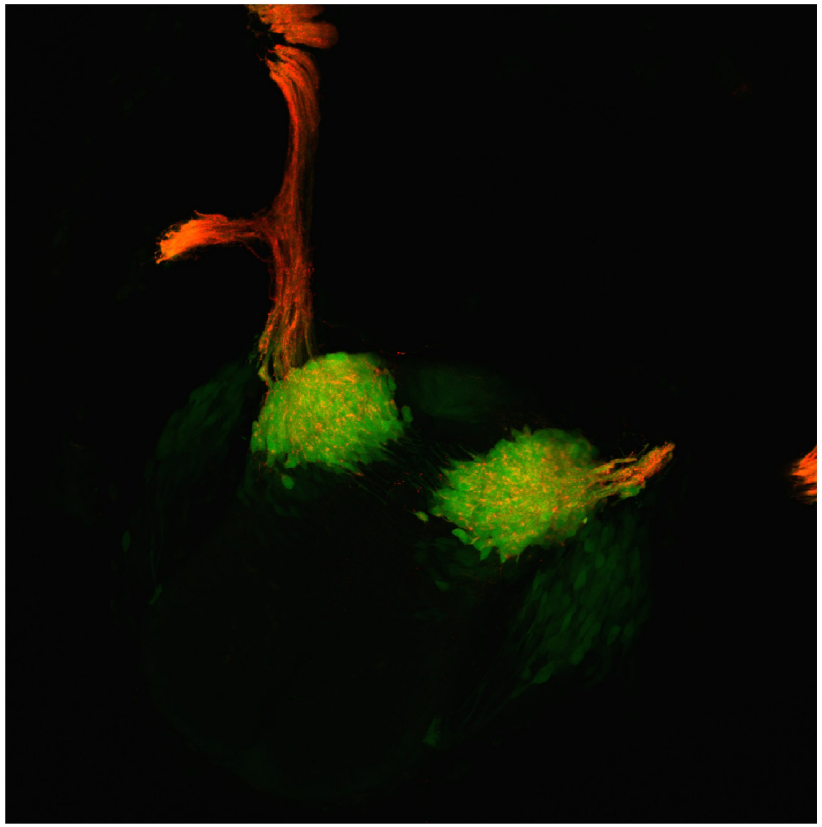
Novel surface markers

For instance, we can predict novel cell surface markers for specific cellpopulations:



E9.5

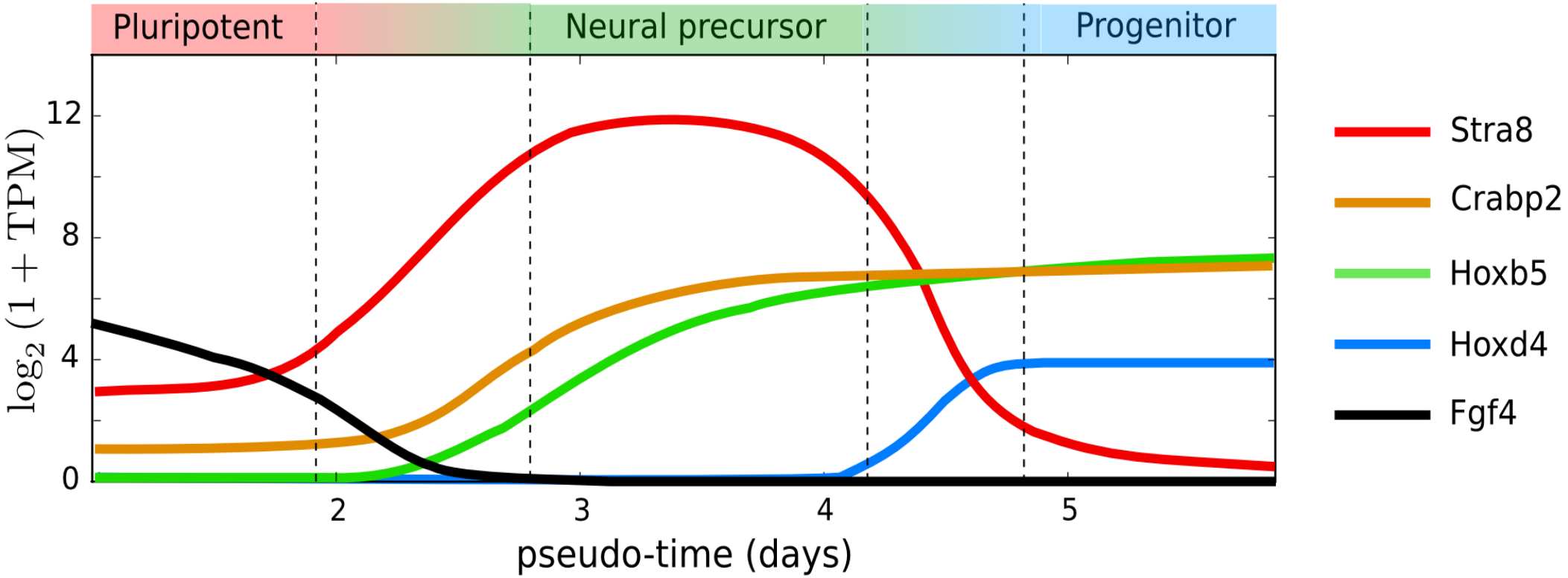
Slc10a4, Mnx1::Egfp



Abbas Rizvi, Maniatis Lab

Characterizing transitions

Or identify the molecular transitions that govern the differentiation process and their timing:



Outline

- Introduction to the abstract biological problem.
- Study single cell expression data using topological data analysis.
  - First example: development of motor neurons.
  - **Second example: studying heterogeneity and evolution in cancer.**
- Study HiC single cell expression data using topological data analysis.

Intratumor heterogeneity

Characterizing the cellular composition of a tumor and its microenvironment is essential for designing therapeutic strategies that can foresee the effect of clonal selection and infiltrating cells

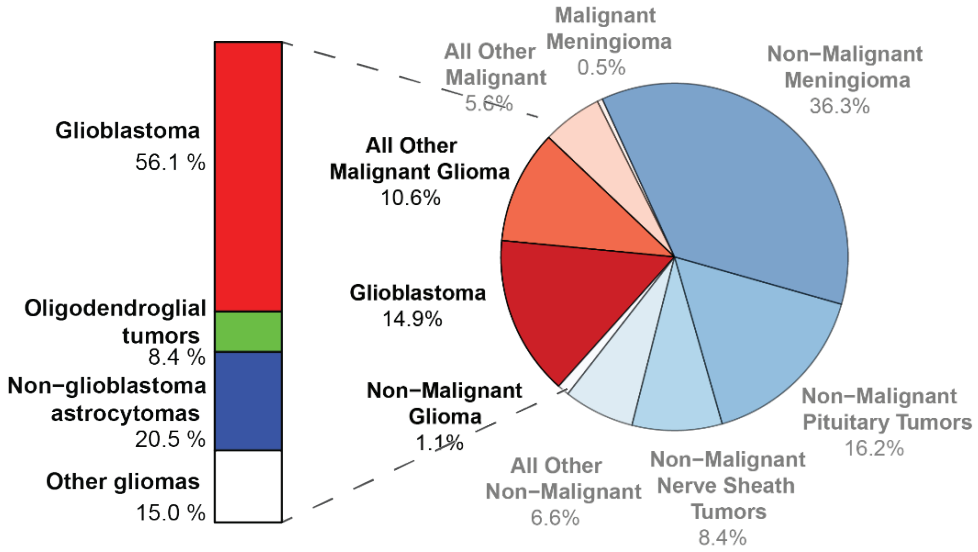
Can we characterize the clonal structure of a tumor at the transcriptional level?

How does transcriptional heterogeneity relates to genomic heterogeneity?

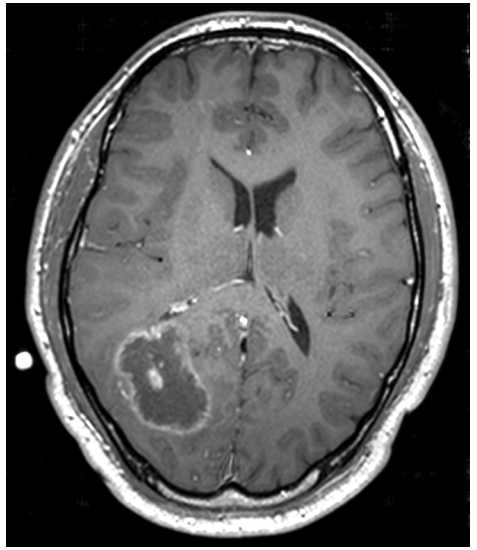
How does intra-tumor heterogeneity and micro-environment composition relate to each other and affect the response to therapy?

Glioblastoma multiforme

- Glioblastoma multiforme (GBM), or Glioblastoma, or Grade IV Astrocytoma is the most common and most aggressive malignant primary brain tumor.
- GBM is usually involving glial cells, and the incidence is about 2~3 cases per 100,000 person life-years.
- 5-year survival 3.3%.

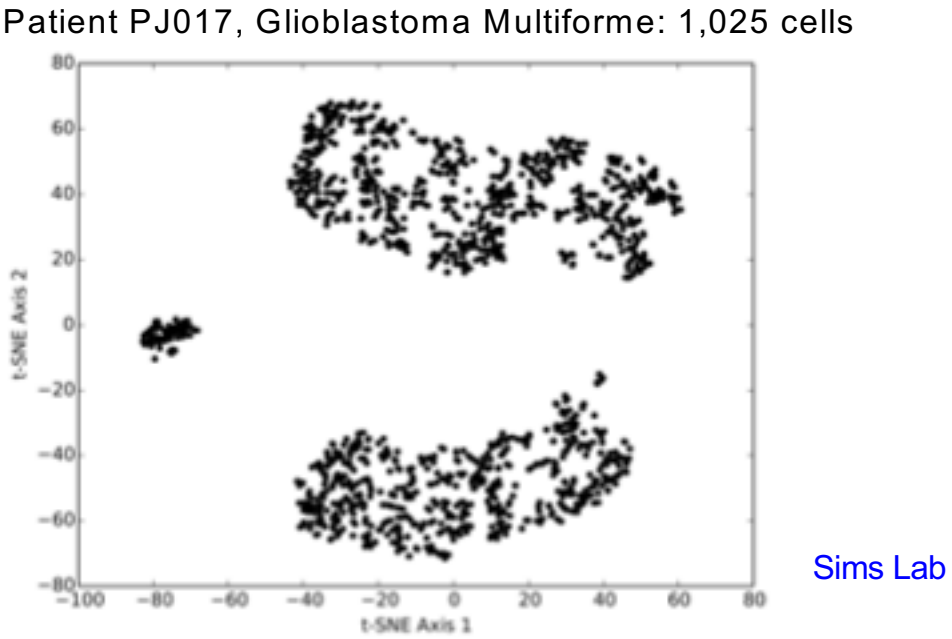


CBTRUS, 2017



Single cell data from tumors

Current unsupervised methods for the analysis of single-cell expression data from tumors are based on dimensional reduction, followed by clustering and differential expression analysis:

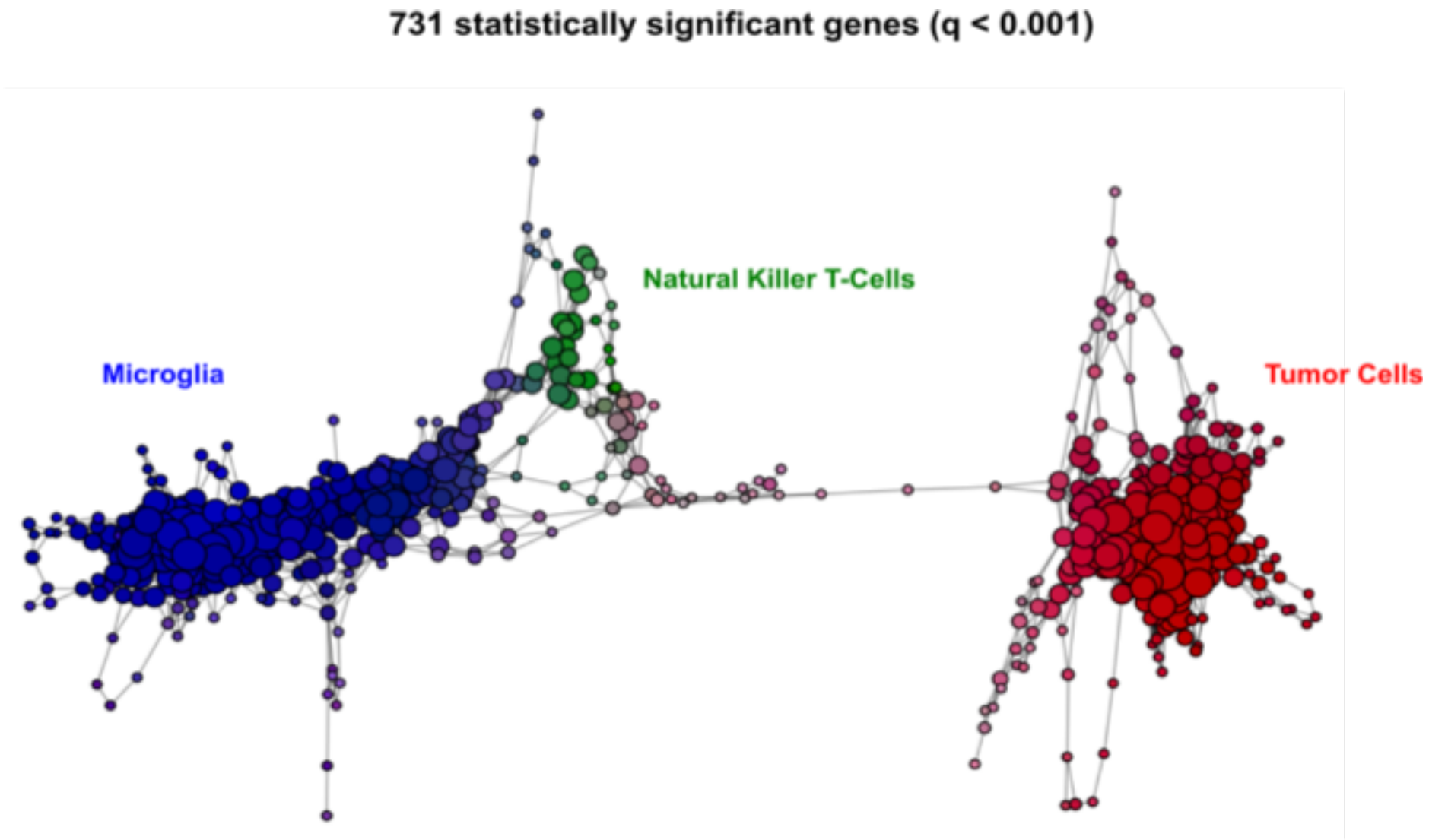


Sims Lab

Limited power to capture structure within clusters: dynamic epigenetic states, cellular subtypes, differentiation axes, etc.

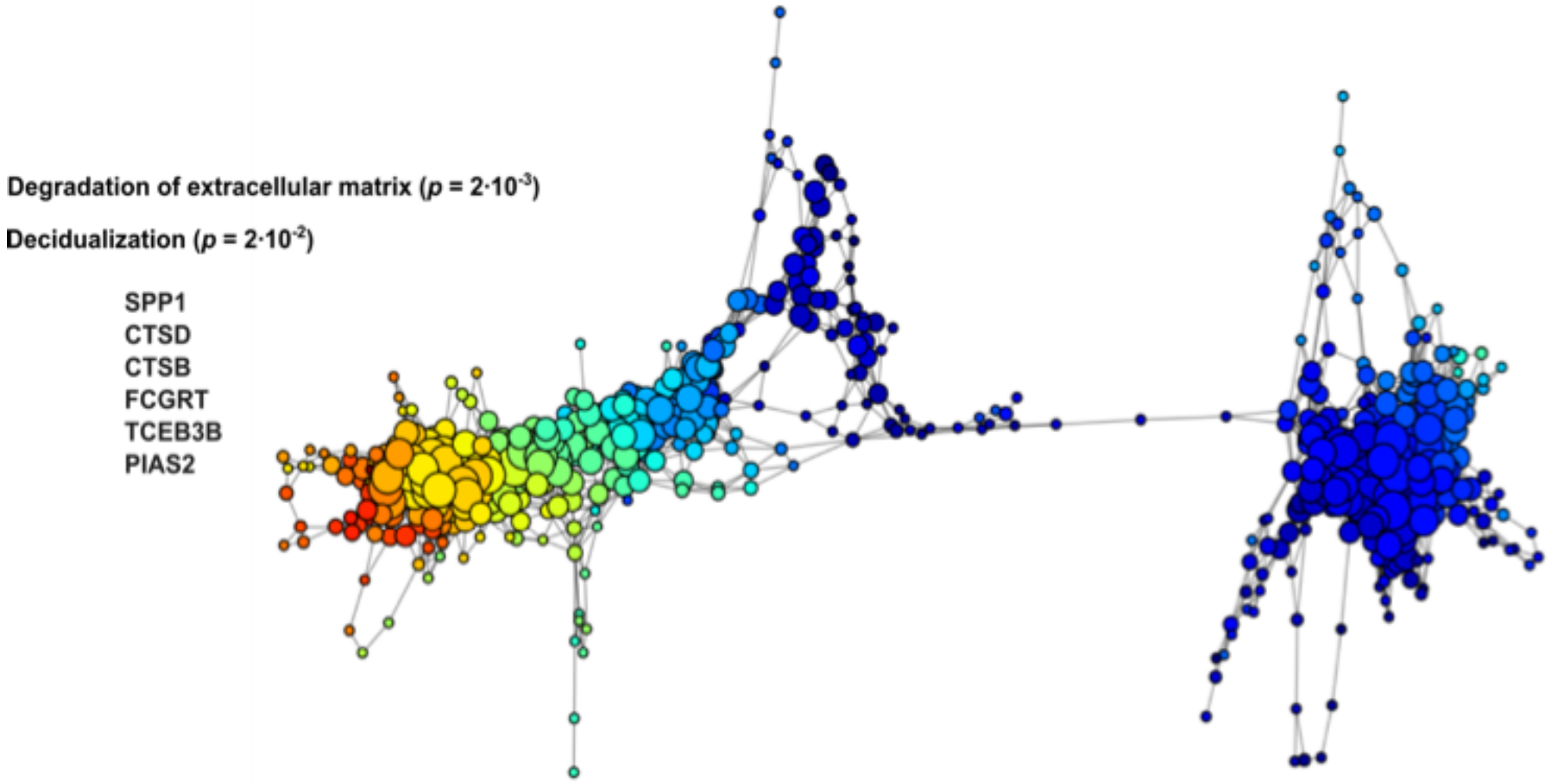
Limited power to capture relationships between clusters.

Patient PJ017



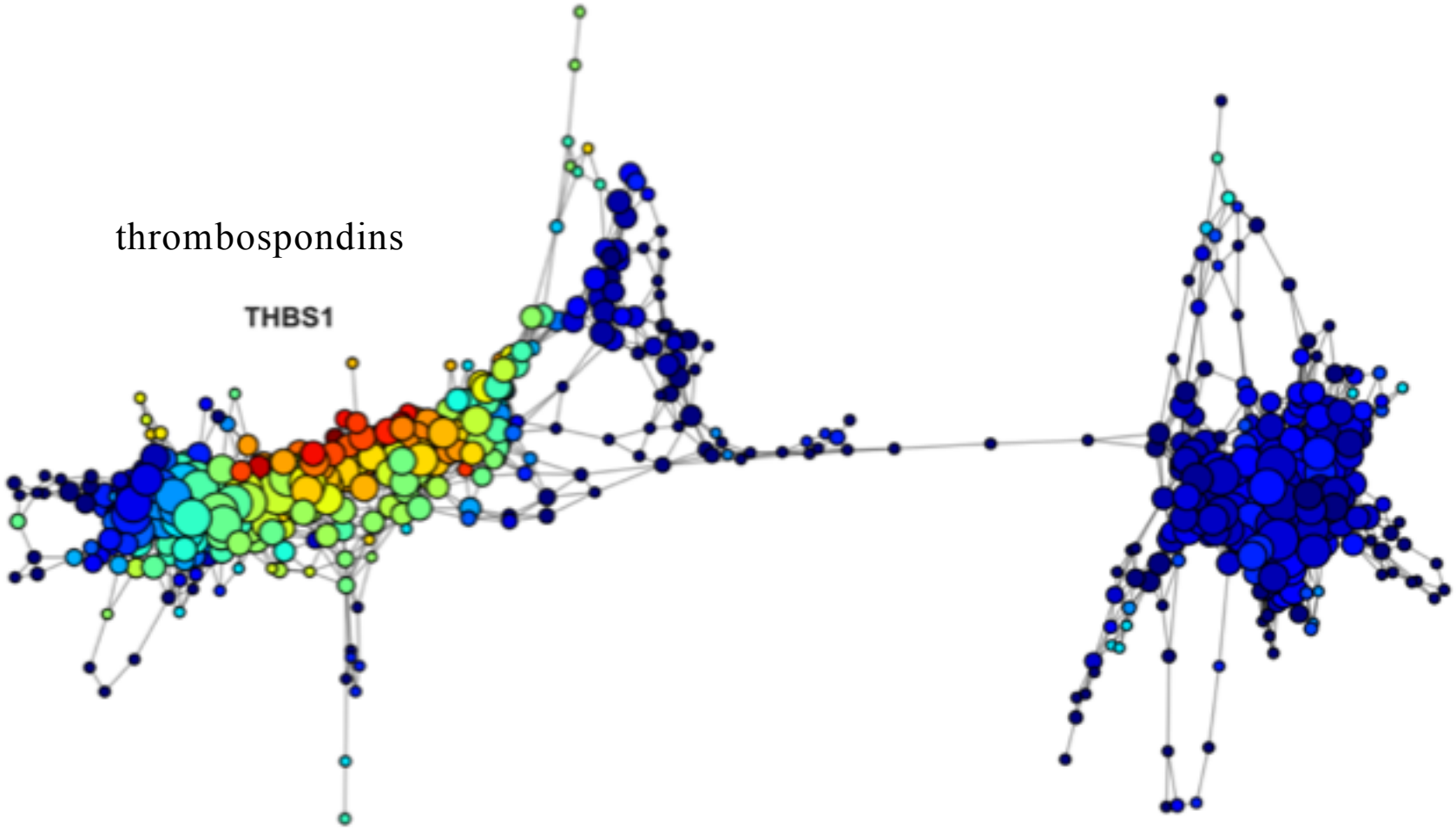
Data from Peter Sims Lab

Patient PJ017: microglia



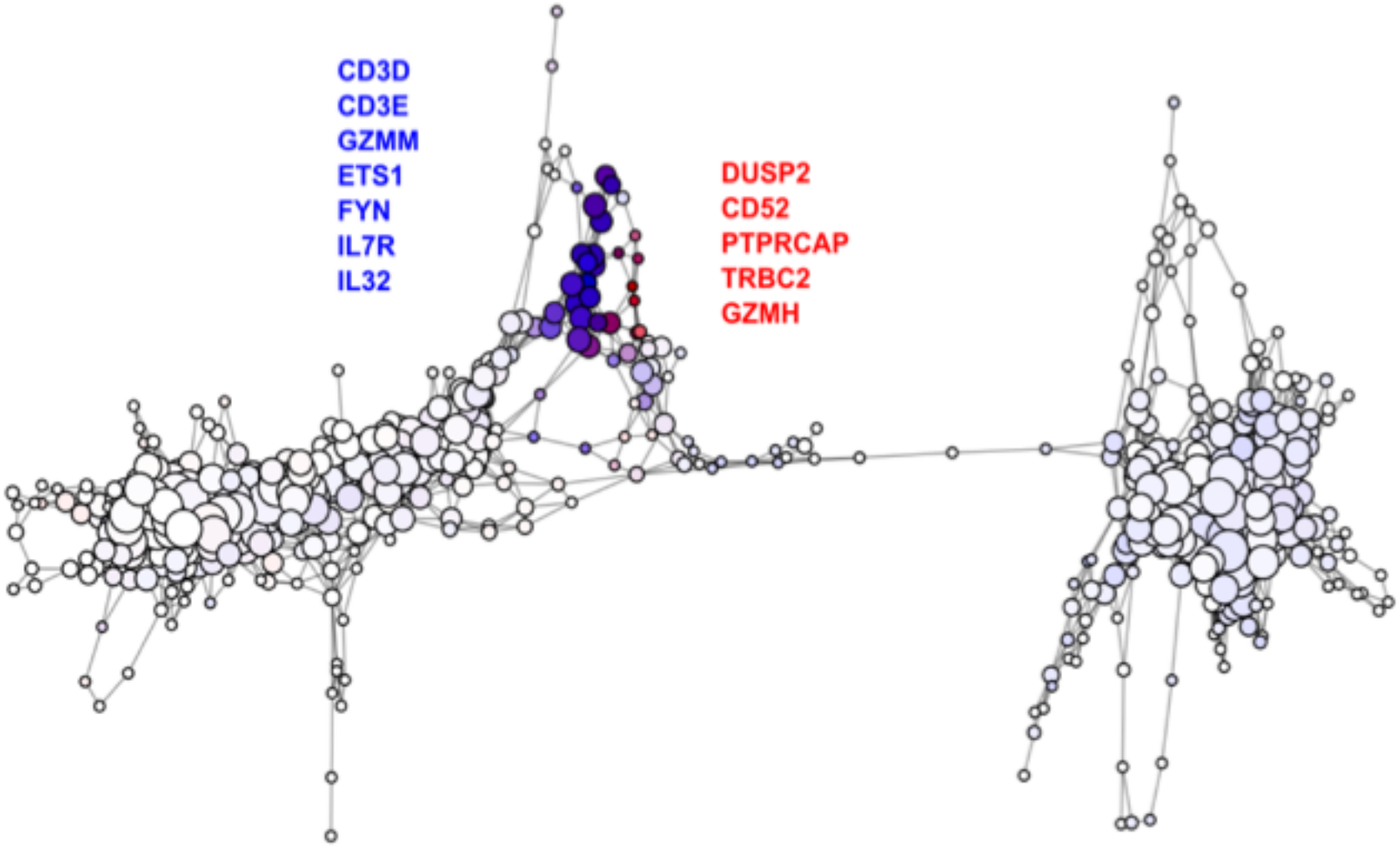
Data from Peter Sims Lab

Patient PJ017: microglia



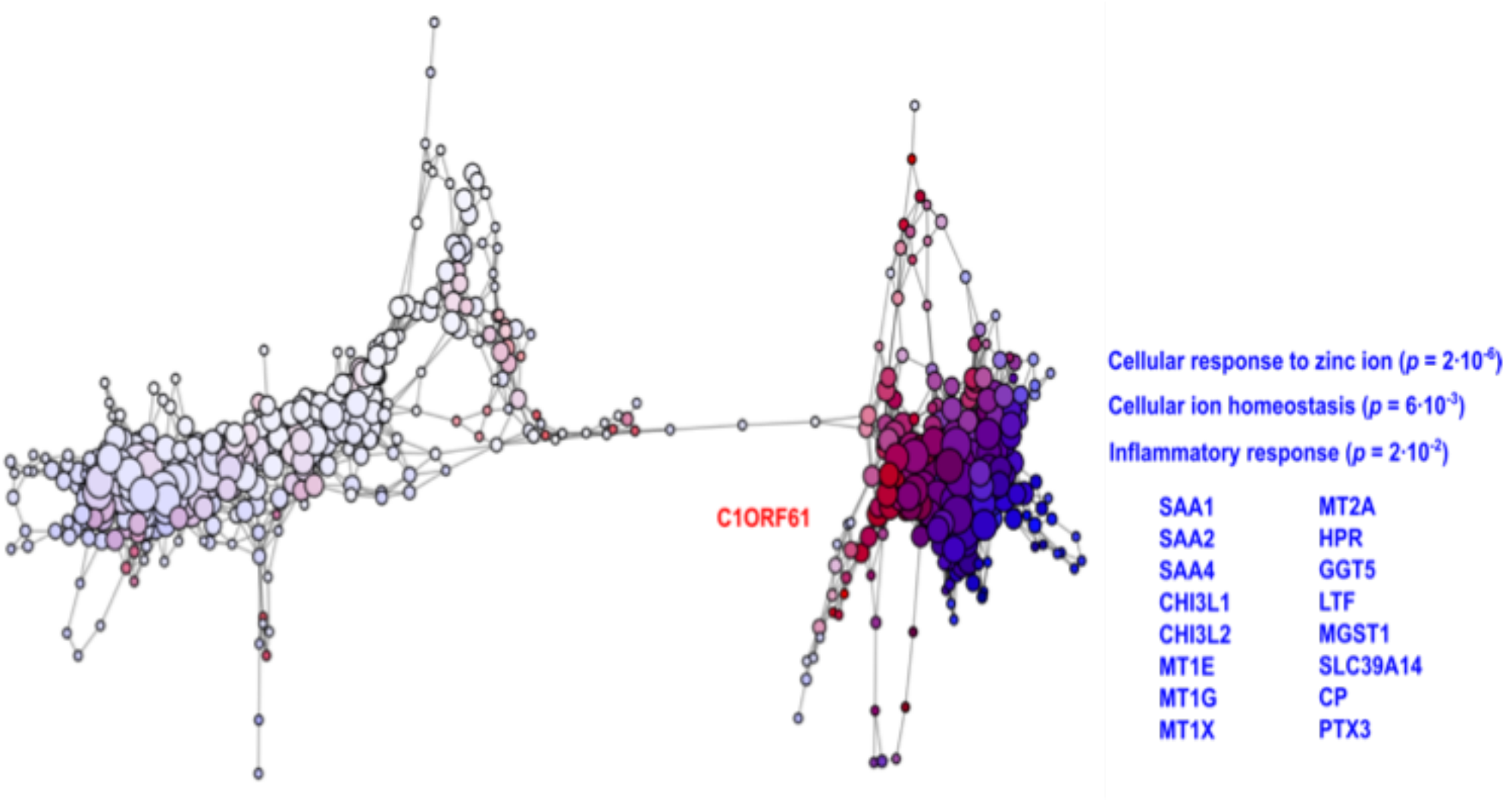
Data from Peter Sims Lab

Patient PJ017: NK T-cells



Data from Peter Sims Lab

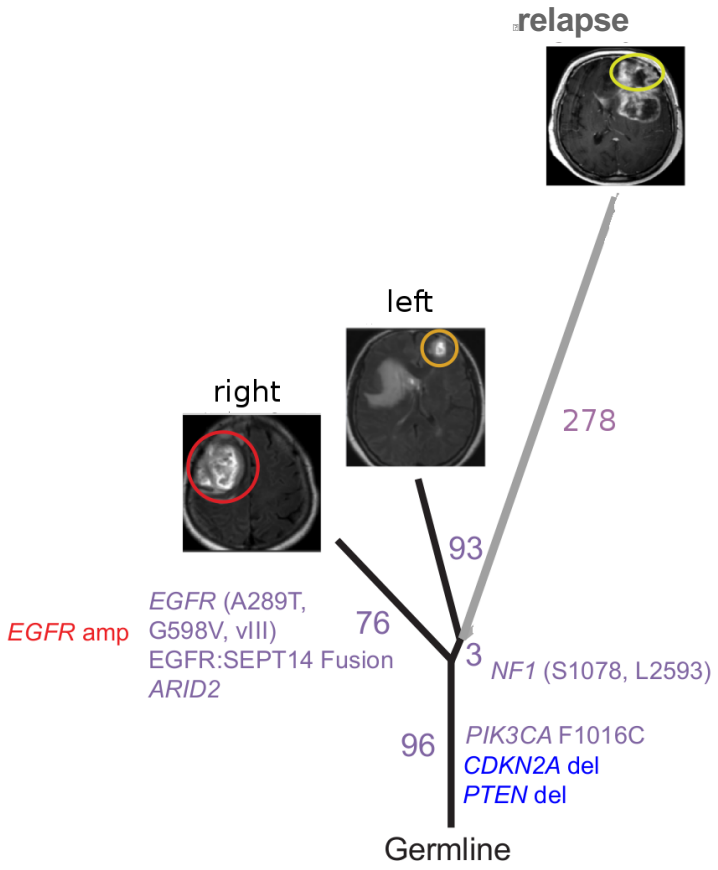
Patient PJ017: tumor cells

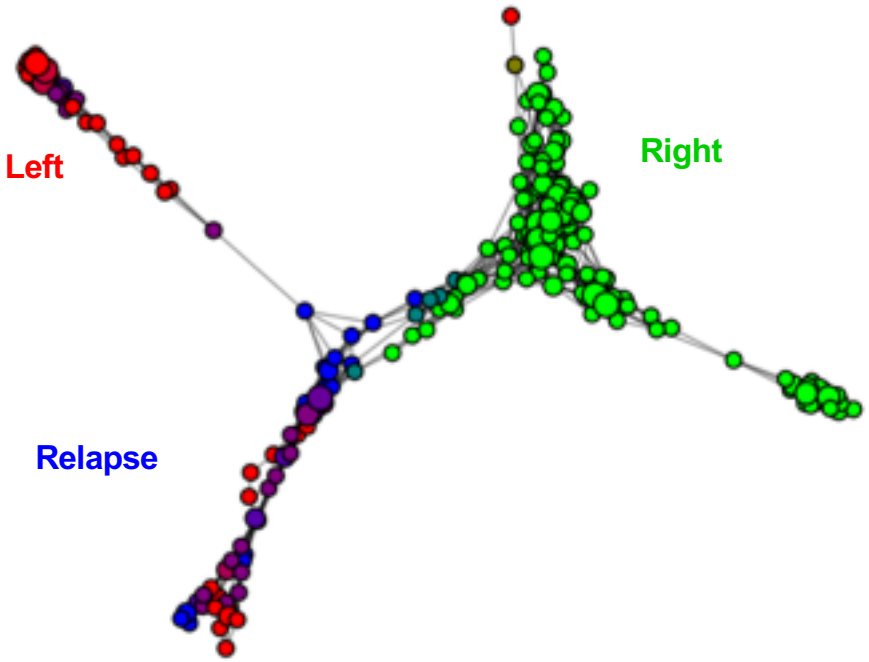


Data from Peter Sims Lab

Patient GBM09

94 cells from 3 samples (2 primary tumors, 1 relapse)



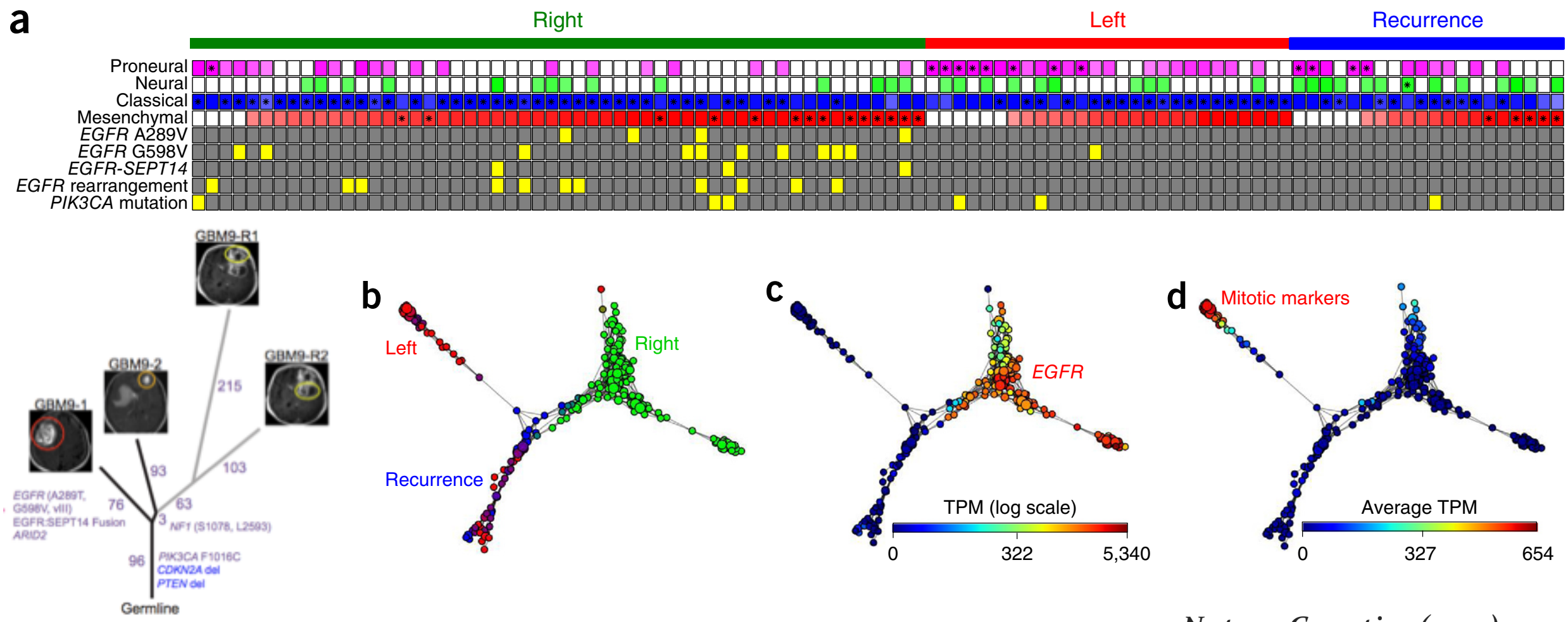


Nature Genetics (2016).

Nature Genetics (2017).

Topological Data Analysis on single cell transcriptomic data

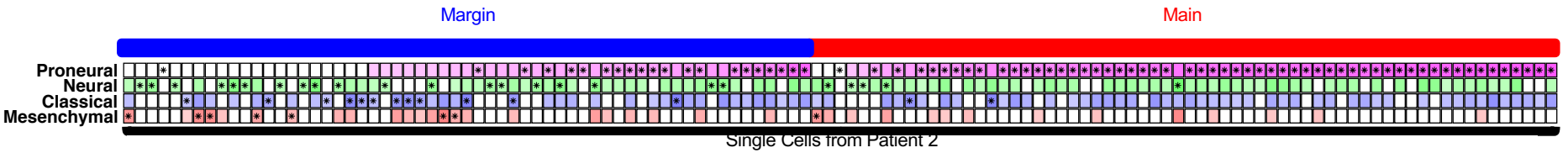
Different subtypes coexist in tumor
Similar coexpression patterns in original site and recurrence

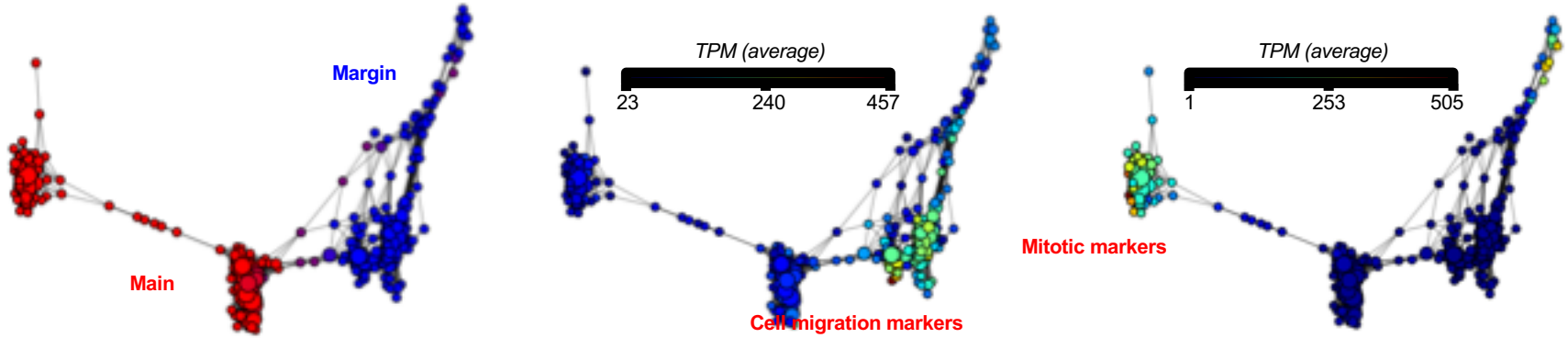


Nature Genetics (2017).
Nature Biotechnology (2017).

Topological Data Analysis on single cell transcriptomic data

Active replicating cells locate in main tumor
Migrating cells in margin





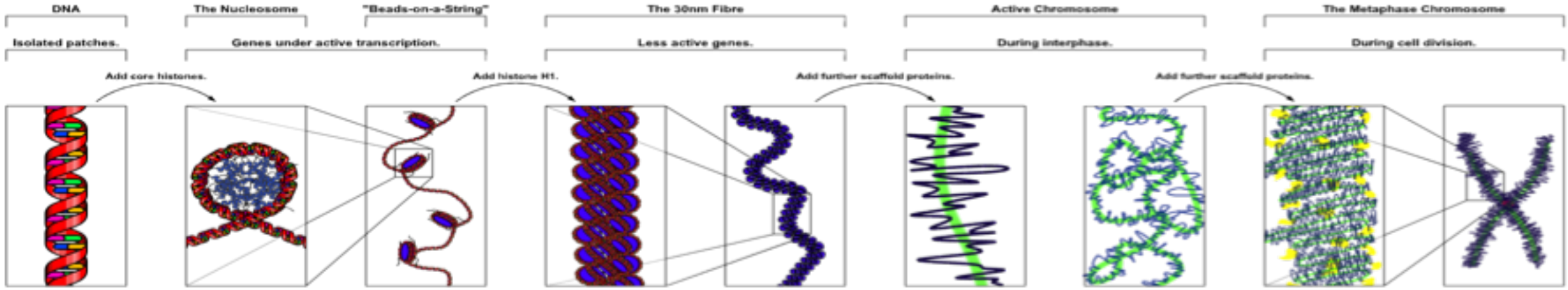
Nature Genetics (2017).
Nature Biotechnology (2017).

Outline

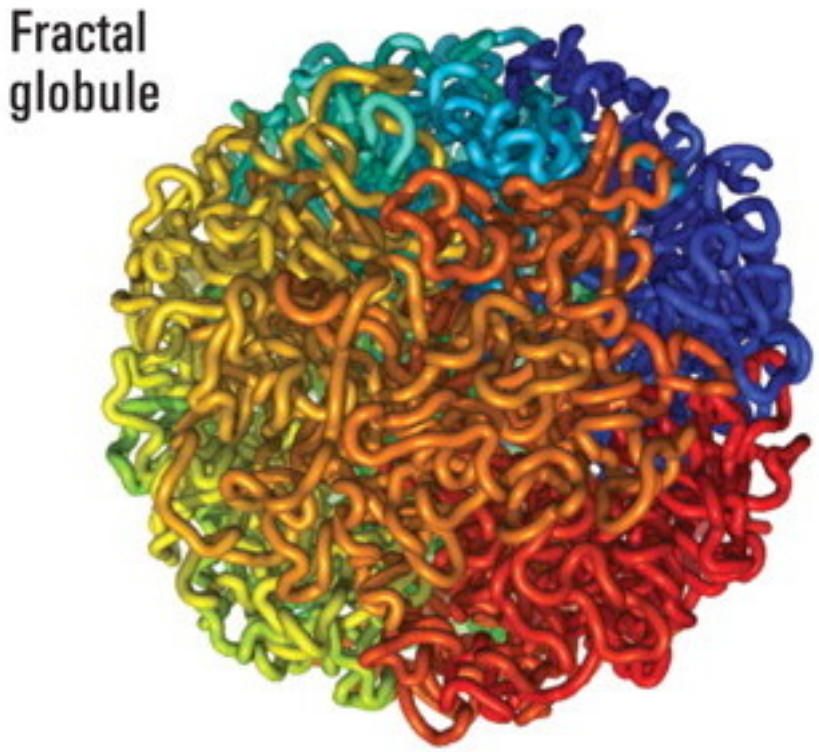
- Introduction to the abstract biological problem.
- Study single cell expression data using topological data analysis.
  - First example: development of motor neurons.
  - Second example: studying heterogeneity and evolution in cancer.
- **Study HiC single cell expression data using topological data analysis**.

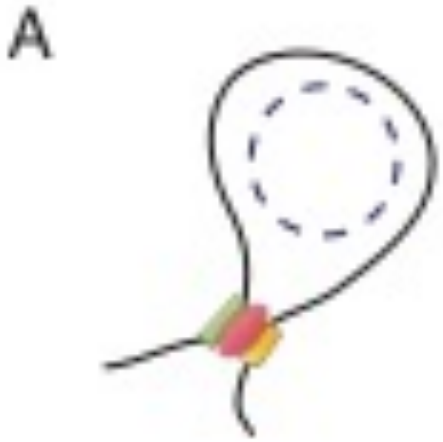
Scales

- A million fold problem:
  - Human genome 2 x $3.3.10^9$ bases ~ 2 meters in length.
  - Cell nucleus is ~ few microns. Fold a million times.
- The fold has functional structure at different scales:
  - Nucleosomes ~200 bases.
  - Long distance promoters ~ 1Mb.
  - Transcription localizes in active sites (transcription factories: active gene transcription unit).
  - Topologically associating domains: ~10 Mb.
- Organized structure at different scales.



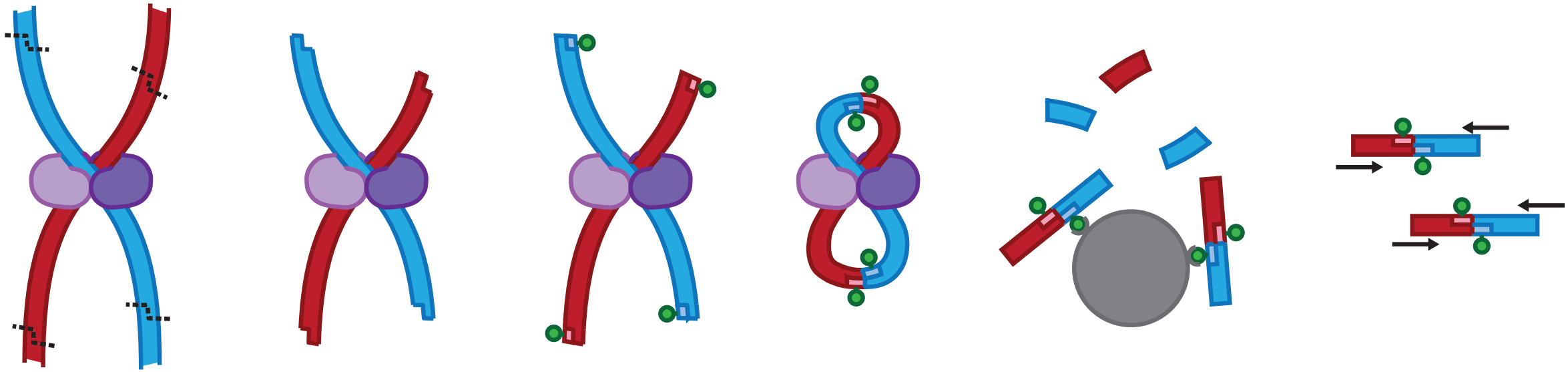
Functional loops in chromatin structure at all scales





Chromosome Conformation Capture

- Genomic sequencing allows unprecedented high throughput genome wide information.
- Several variants: 3C, 4C, 5C, Hi-C.



Crosslink DNA

Cut with restriction enzyme

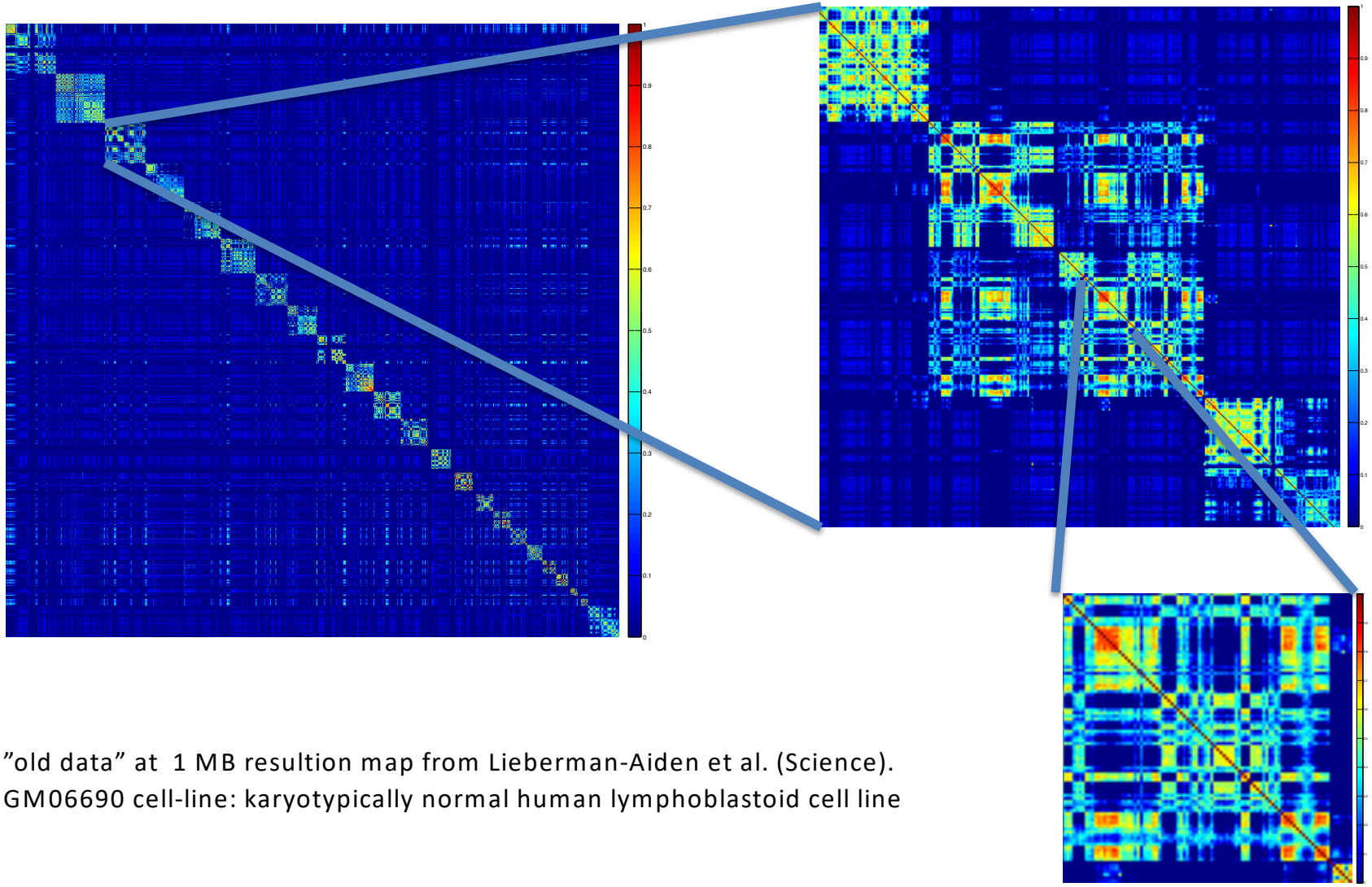
Fill ends and mark with biotin

Ligate

Purify and shear DNA pull down biotin

Sequence using paired-ends

Data

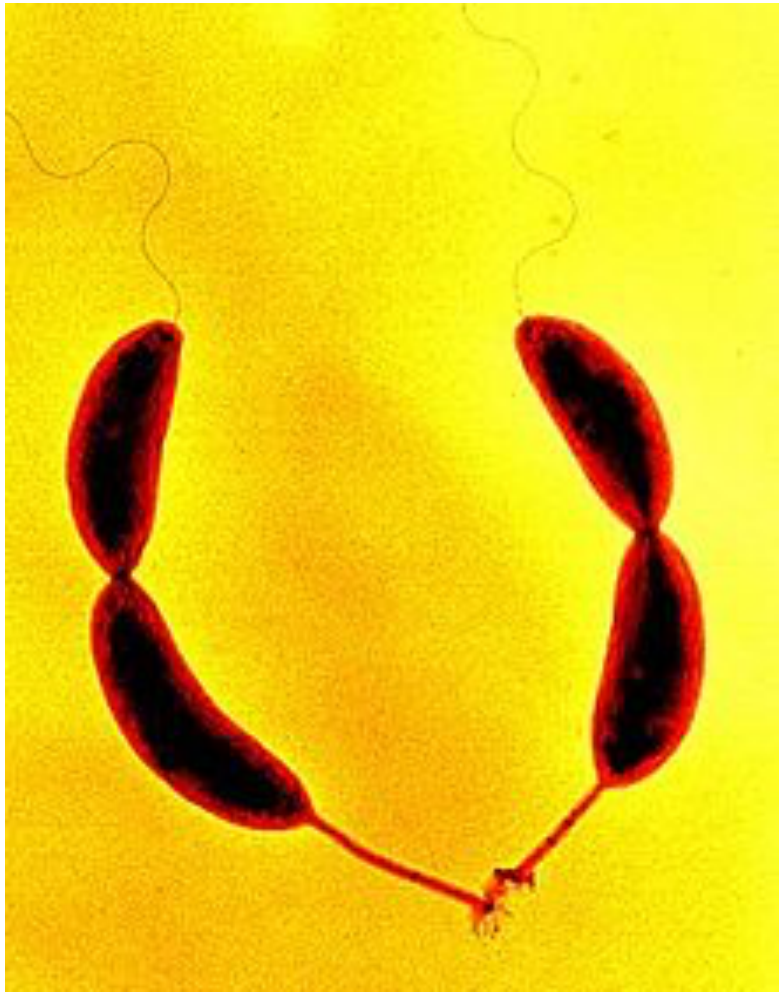


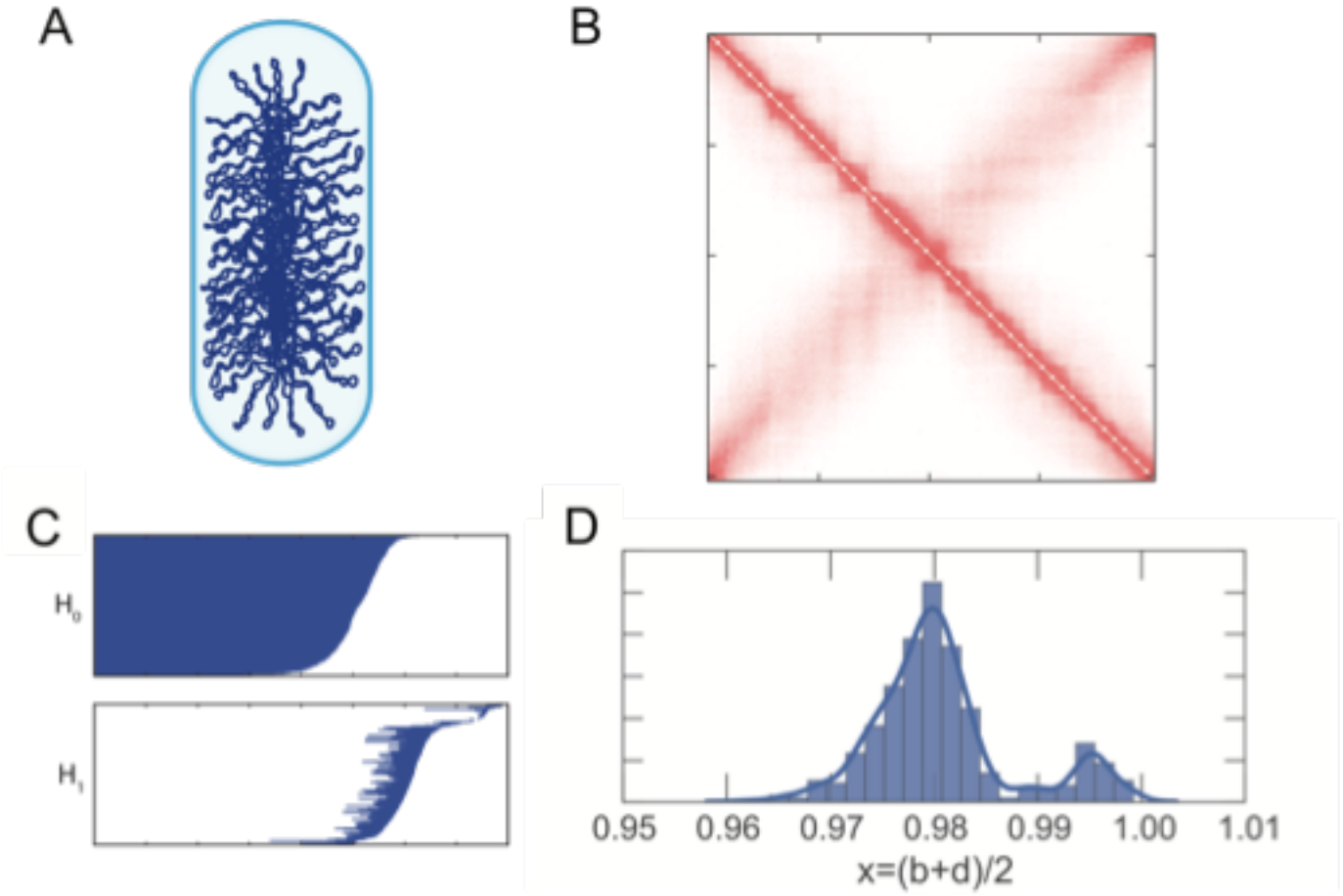
”old data” at 1 MB resultion map from Lieberman-Aiden et al. (Science).
GM06690 cell-line: karyotypically normal human lymphoblastoid cell line

Persistent Homology view

- In humans: Embedding of 23 pairs of 1 dimensional intervals into a compact space with well-defined structures. Two types of distance between any two points in the genome: 1D and 3D.
- Hi methods sample interactions in ensemble of cells. The contact probability is inversely related with 3D distance.

Caulobacter crescendus





Conclusions

- Many biological problems can be study through single cell genomic technologies (independent but related by underlying biology):
  - Transcription
  - DNA structure.

These processes are continuous, asynchronous, heterogeneous, with interesting low dimensional features (clusters, trees, cycles, etc).

- There is a need to
  - characterize structures at different scales,
  - Uncover biological processes associated,
  - how these processes change is space and time.

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Center for Topology of
Cancer Evolution and Heterogeneity

A member of the National Cancer Institute's Physical Sciences in Oncology Network

- Columbia University:
  - Departments of Systems Biology and Biomedical Informatics:

    Andrew Chen, Anthea Monod, Chioma Madubata, Daniel Rosenbloom, Erik Ladewig, Francesco Brundu, Ioan Filip, Jiguang Wang (HKUST), Junfei Zhao, **Luis Aparicio**, Luis Perez, **Mykola Bordyuh**, Oliver Elliott, **Pablo Camara** (Upenn), Tim Chu, Wesley Tansey, Zhaoqi Liu.
  - Department of Systems Biology:

    Peter Sims.
  - Department of Biochemistry:

    Abbas Rizvi, Elena Kandor, Tom Roberts, Tom Maniatis.
  - Institute for Cancer Genetics and Irving Cancer Center:

    A. Iavarone, A. Lasorella.

- Samsung Medical Center (Seoul, Korea): **Do-Hyun Nam**.