

Extracting a cellular hierarchy from high-dimensional single-cell data

Peng Qiu

Department of Bioinformatics and Computational Biology
University of Texas MD Anderson Cancer Center

Flow / mass cytometry data

	Marker 1	Marker 2	Marker 3	...	Marker k (<40)
Cell 1	0.3	3.8	4.3		6.4
Cell 2	5.6	5.8	3.1	...	9.6
Cell 3	8.8	2.5	1.6		2.4
Cell 4	6.7	2.9	1.8	...	6.8
Cell 5	1.9	6.2	4.2		2.9
Cell 6	3.7	2.7	0.9	...	6.7
Cell 7	4.6	8.2	6		7
Cell 8	9.8	9.8	4.7	...	0.7
Cell 9	1.6	7.3	7		2.5
Cell 10	8.6	3.4	7	...	2.2
Cell 11	6.4	5.8	6.4		6.7
Cell 12	3.8	1.1	0.3	...	8.4
...
...
Cell N ($>100,000$)	2.3	0.2	8.2	...	3.9

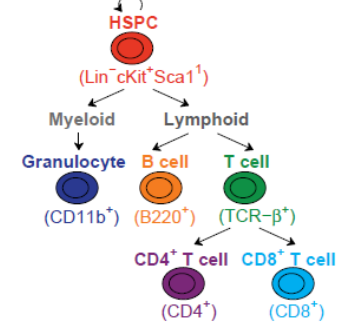
Biology questions

- How many cell types are there?
- How are different cell types related to each other?
- Does the cellular composition of a sample correlate with its overall phenotype?

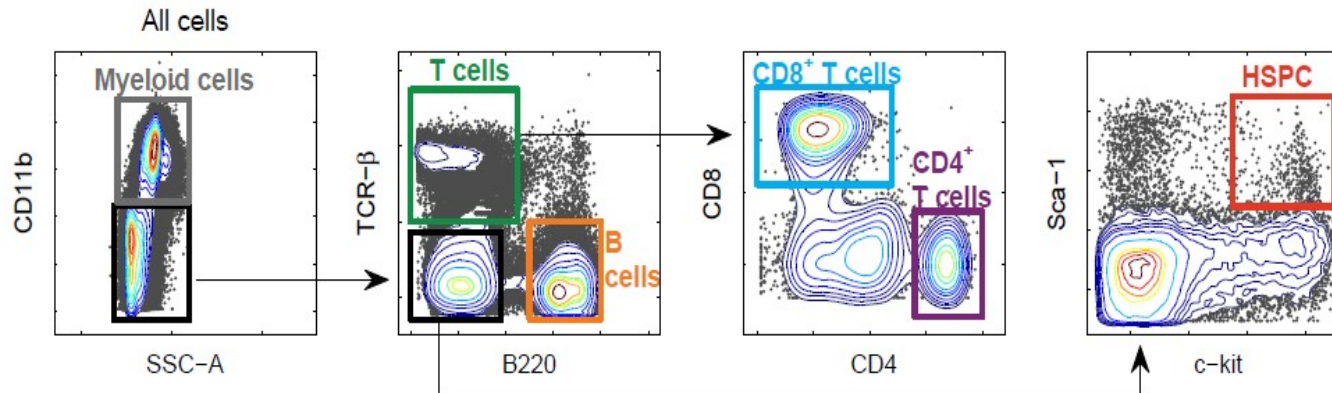
Introduction - gating

- Example data
 - 8-parameter flow cytometry
 - Mouse bone marrow
 - Parameters: c-kit, Sca-1, CD11b, B220, TCR-b, CD4, CD8

Hematopoiesis in mouse

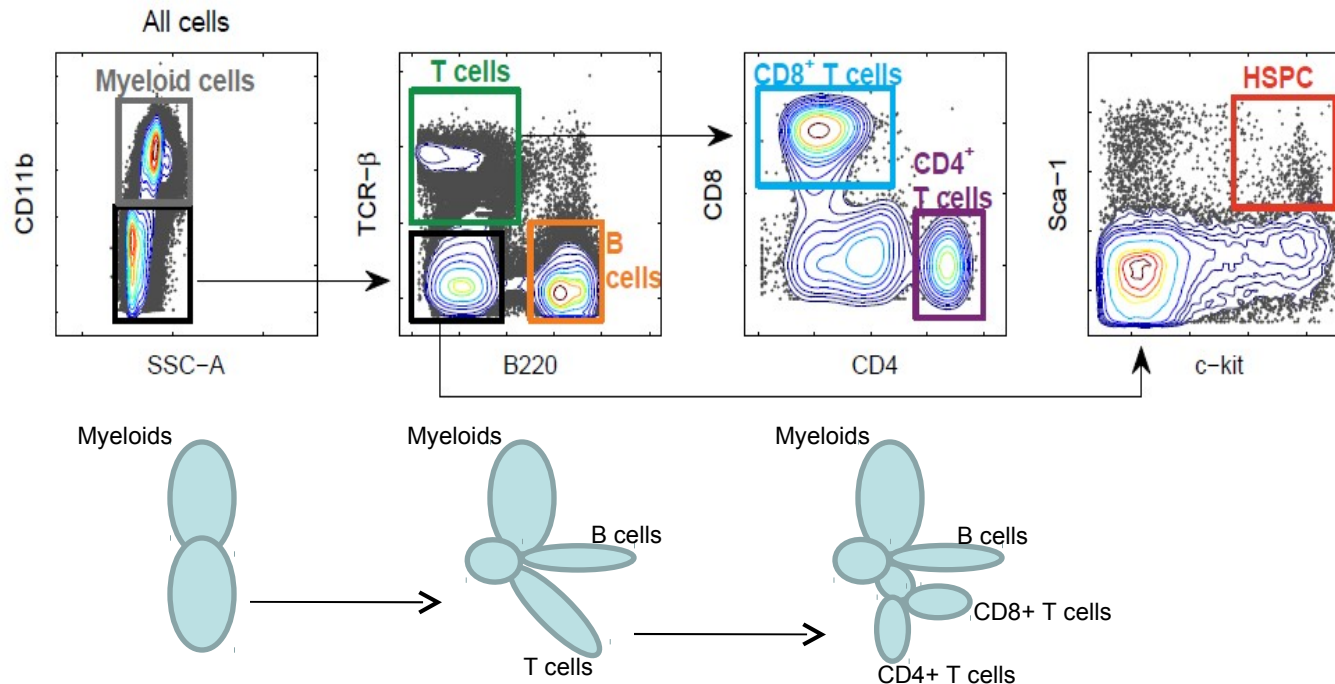


- Traditional analysis: Gating



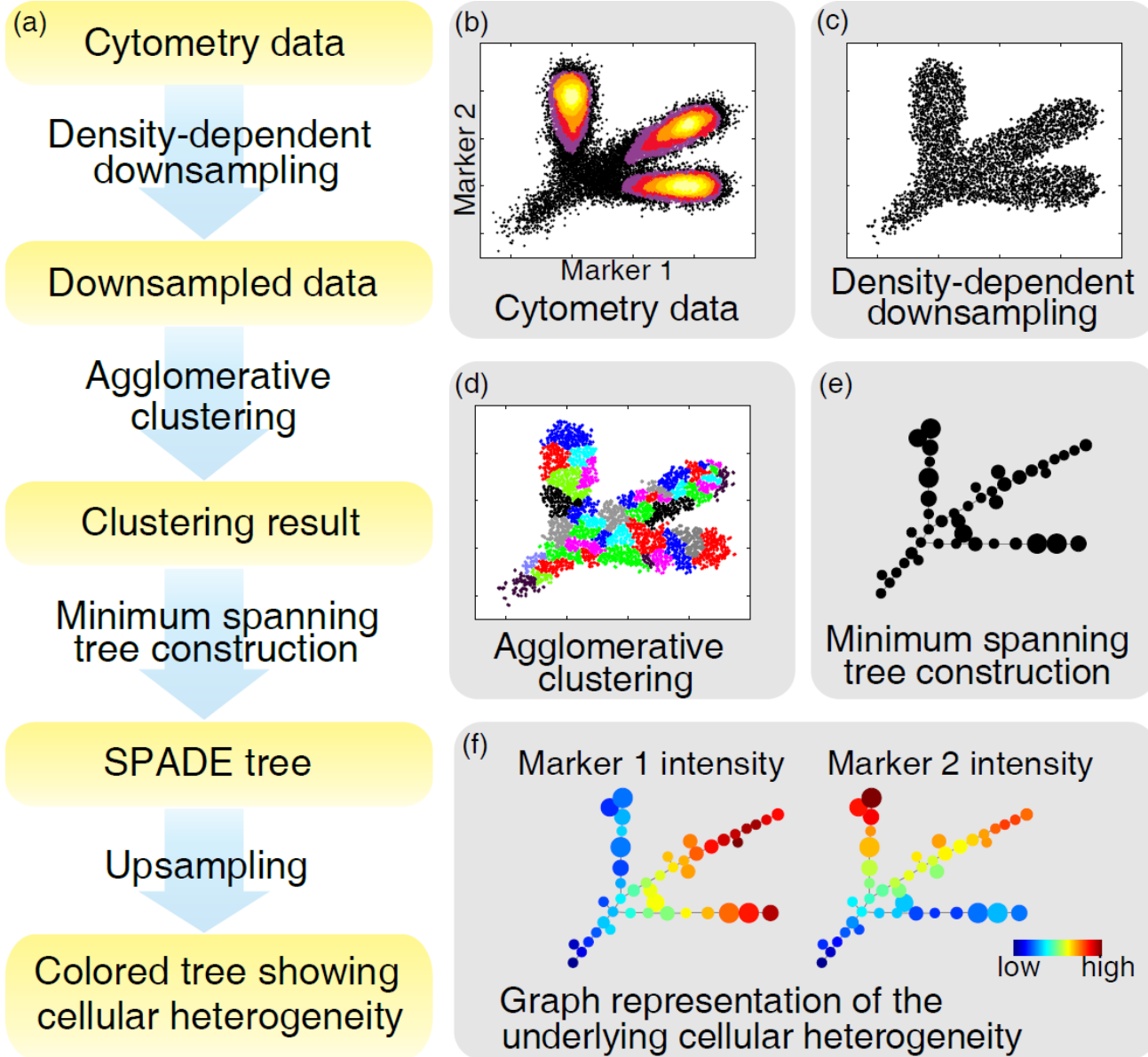
Basic idea

- Consider the data as a point cloud
- Extract the shape of the cloud

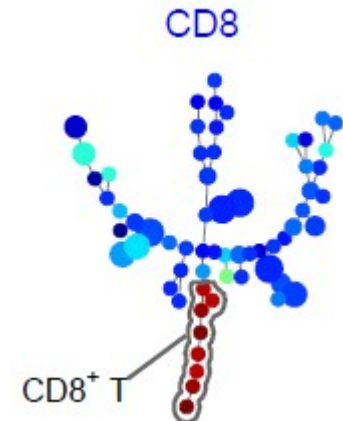
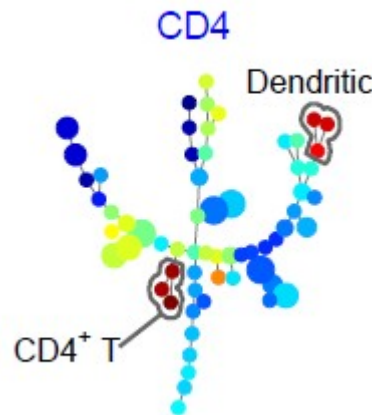
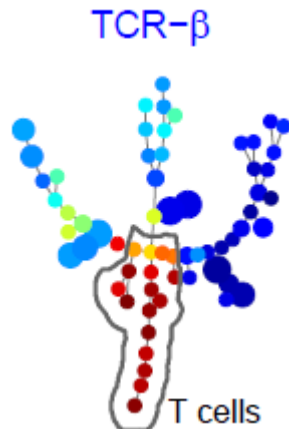
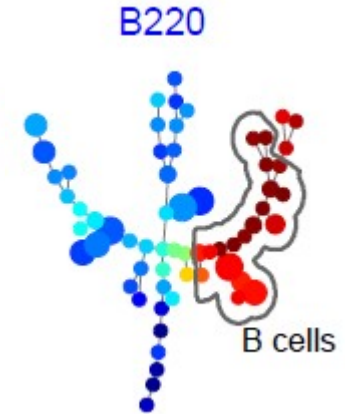
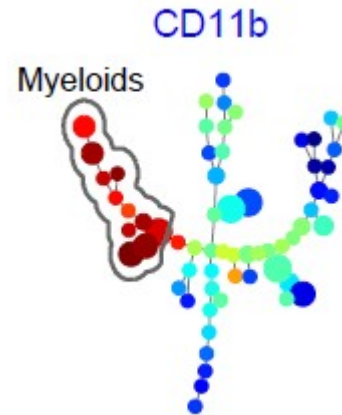
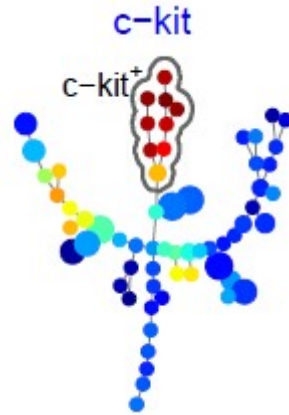
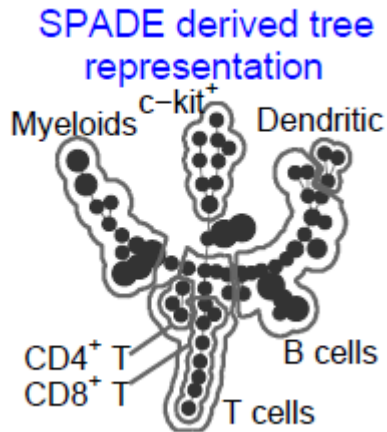
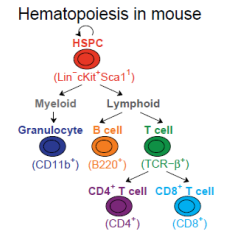


- Method:
Spanning-tree Progression Analysis of Density-normalized Events (SPADE)

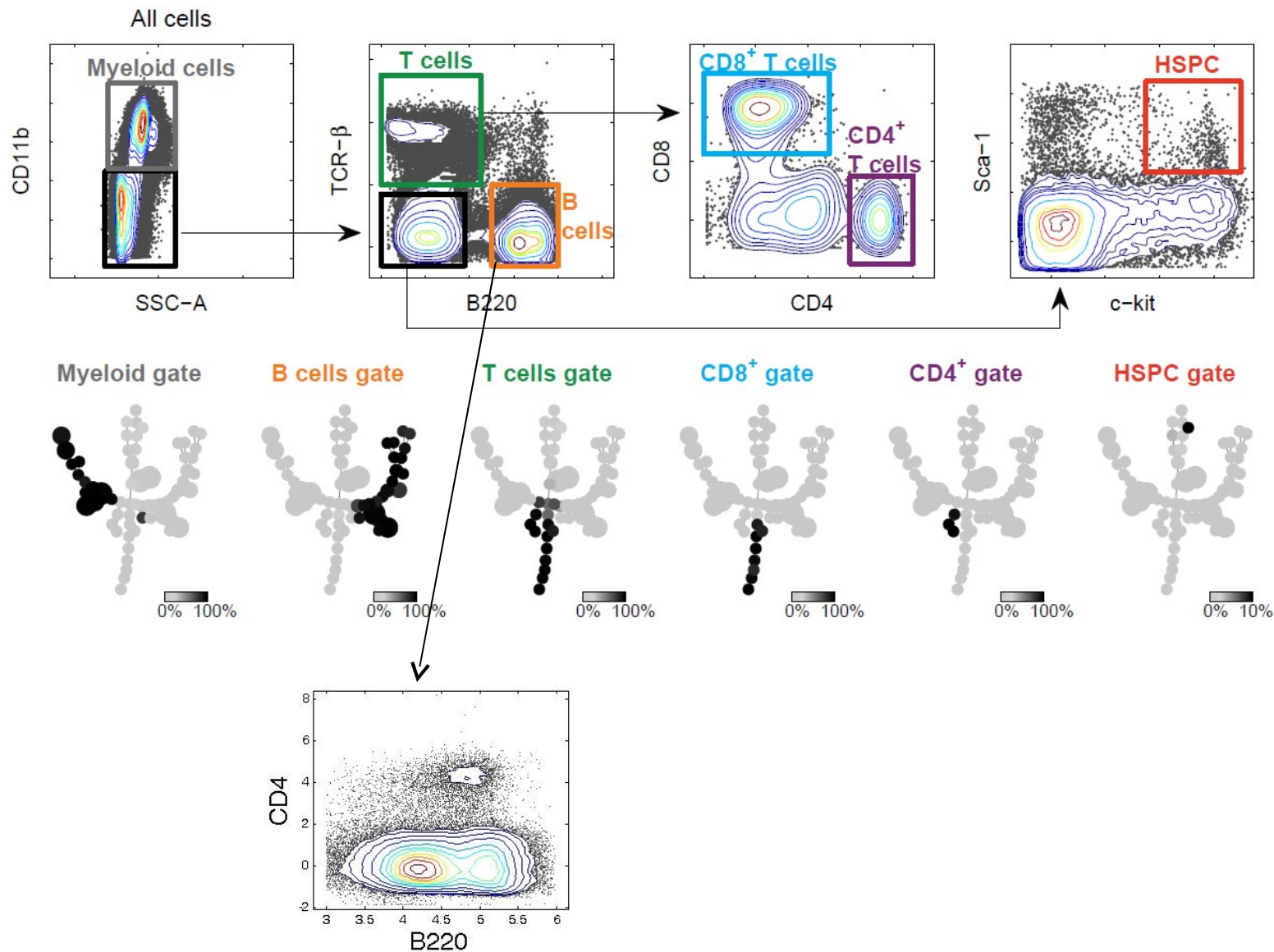
SPADE



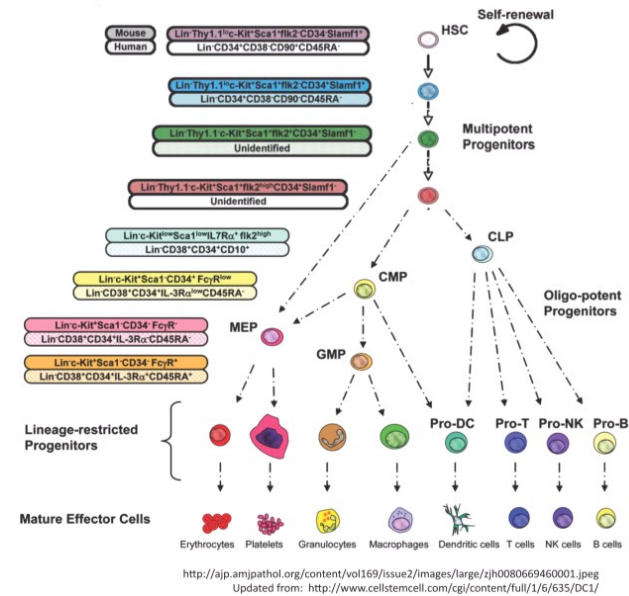
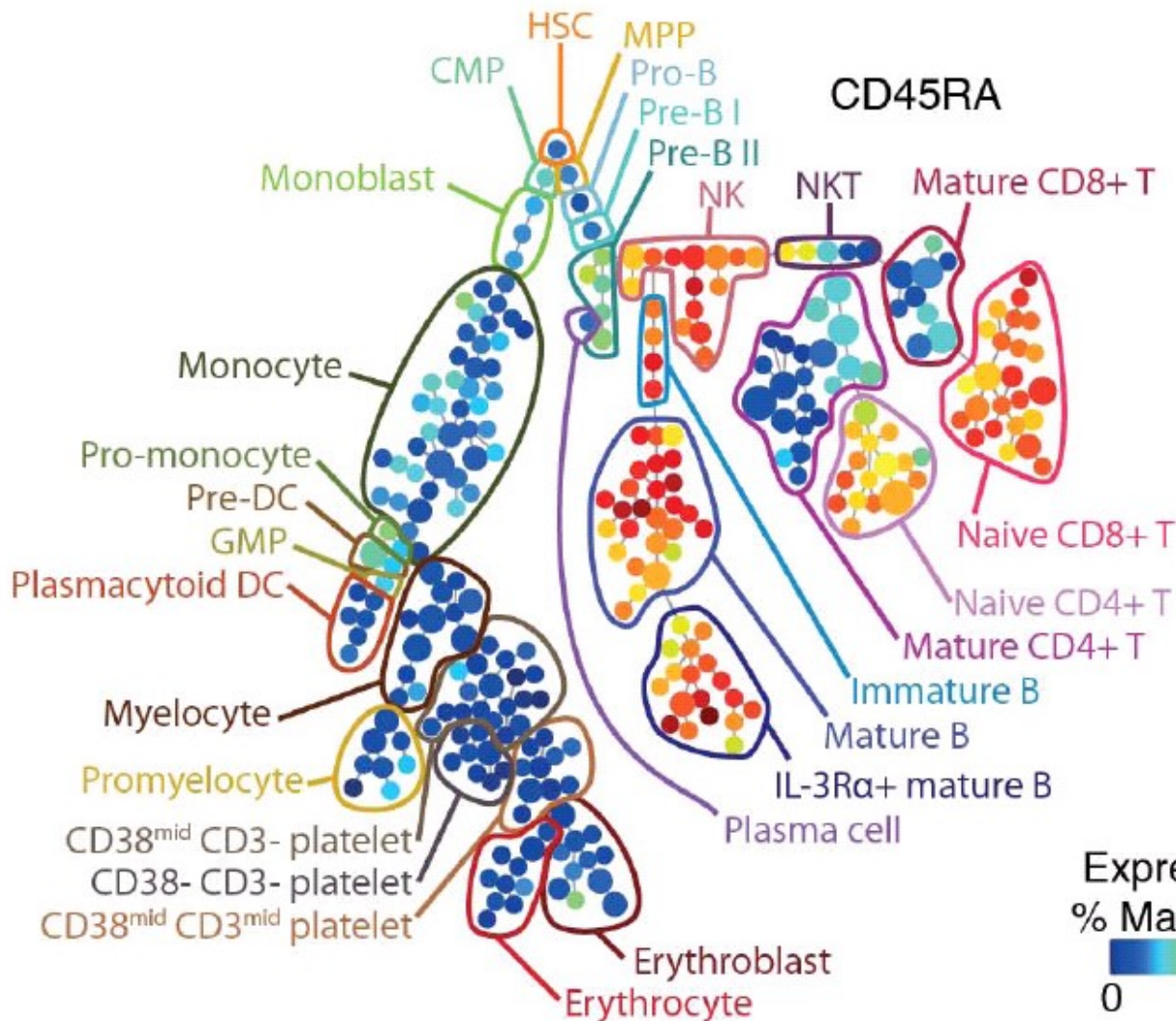
SPADE applied to mouse bone marrow data



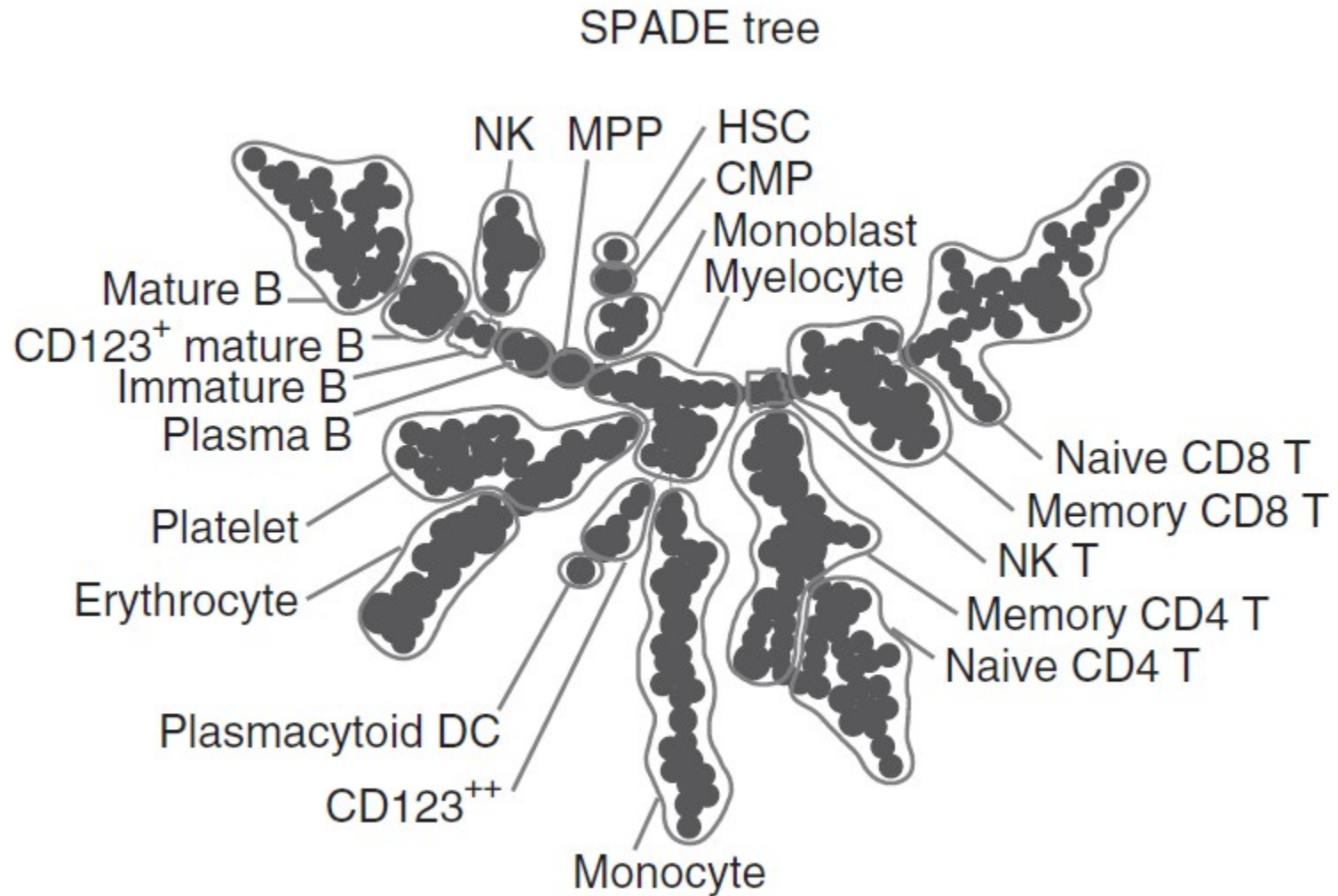
SPADE vs. gating

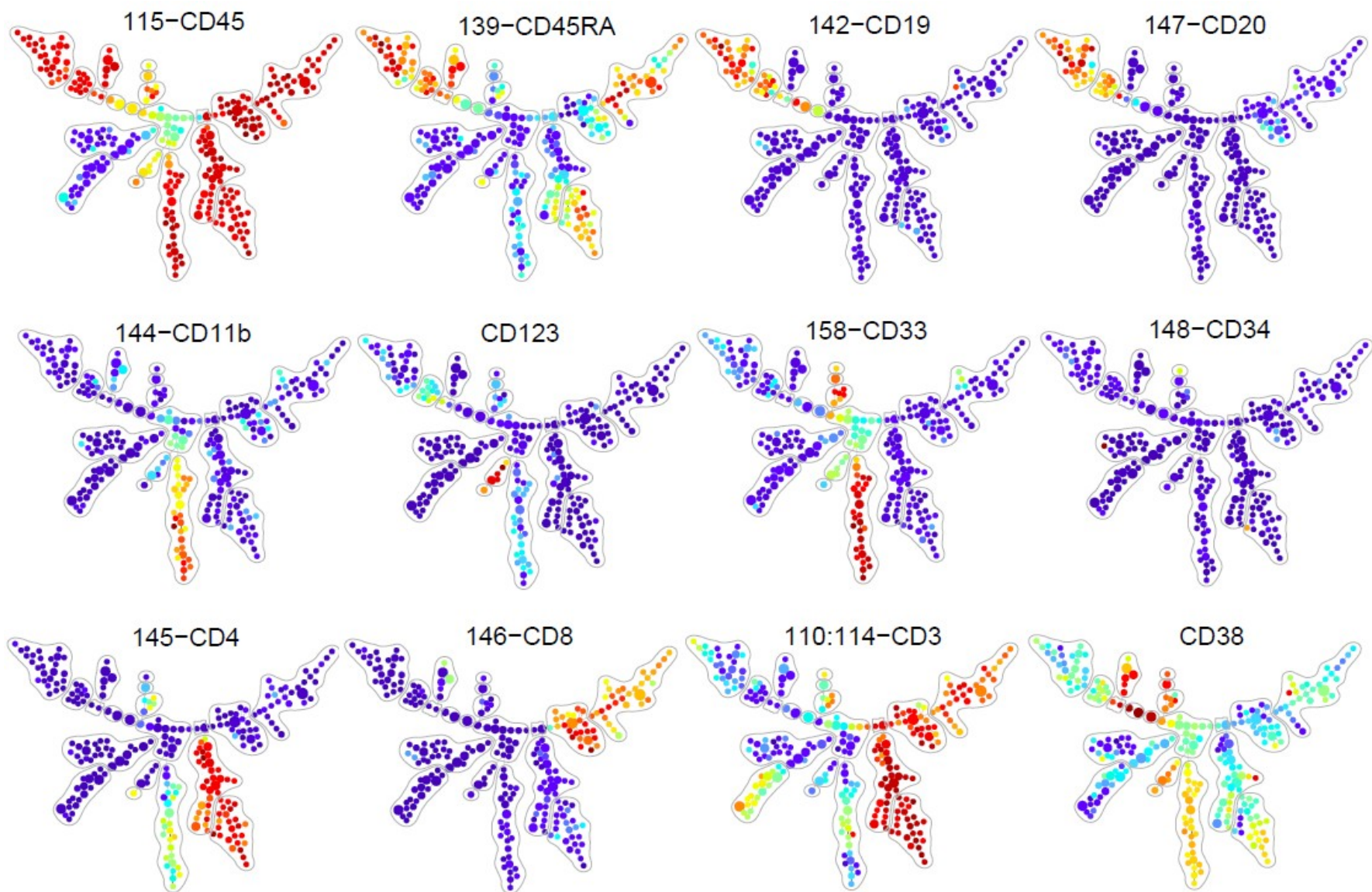


SPADE applied to human bone marrow data



SPADE applied to CyTOF data of human BM





Challenge 2: Normal vs AML

- 359 subjects
 - 316 normal subjects
 - 43 AML samples
- 8 Tubes per subject
- Channels per tube: FSC+SSC+5 colors

Challenge 2: Normal vs AML

Since the overlap among the 8 different staining panels/tubes is minimal, we consider them separately.

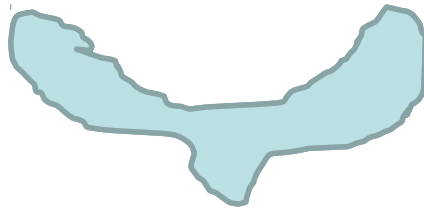
Therefore, we have 359 fcs files to compare.

Challenge 2: Normal vs AML

Since the overlap among the 8 different staining panels/tubes is minimal, we consider them separately.

Therefore, we have 359 fcs files to compare.

Tube2 Sample1



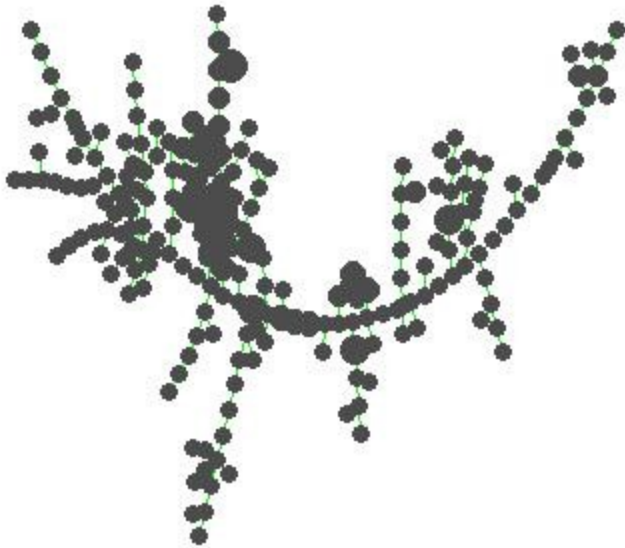
Tube2 Sample2



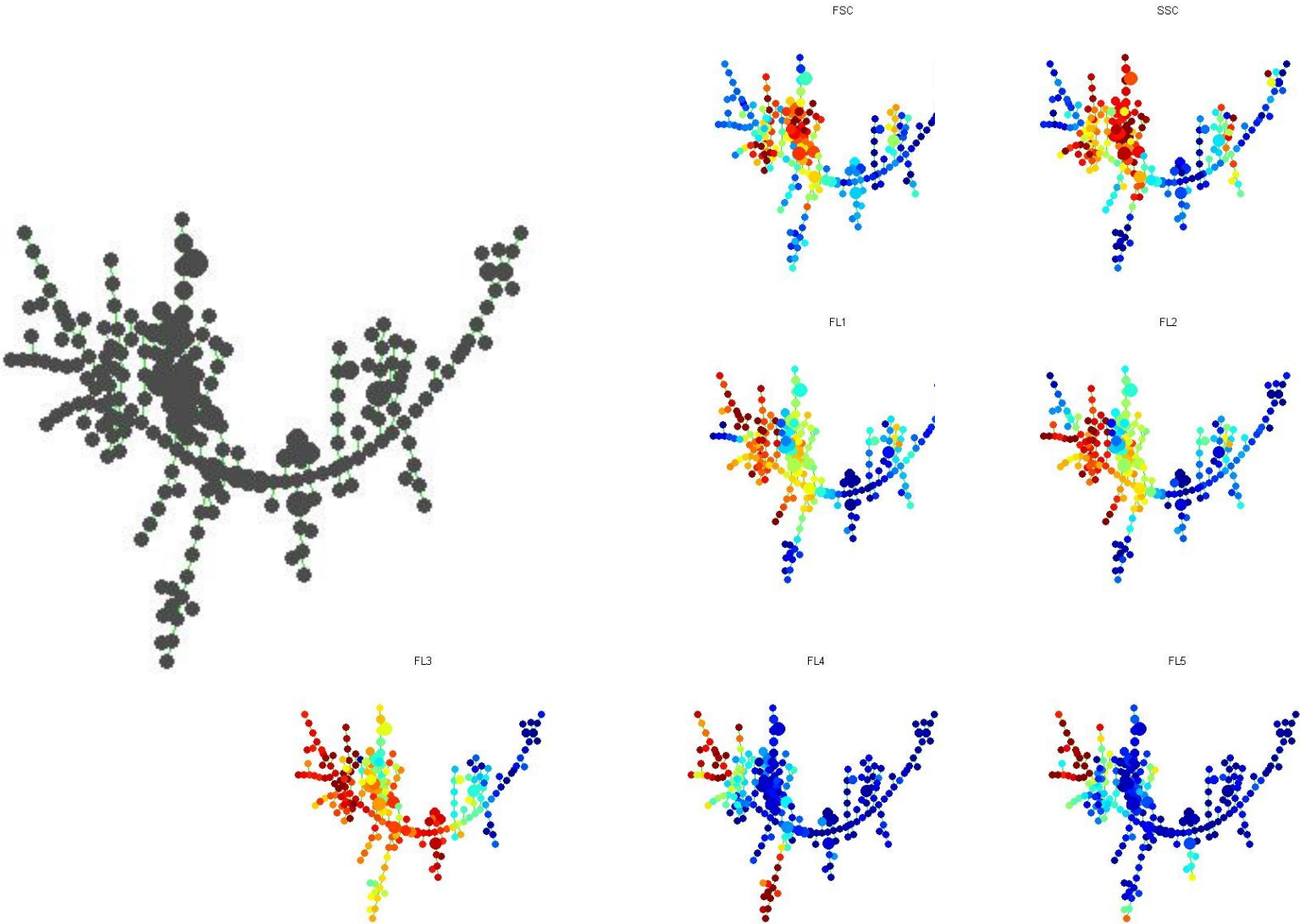
...

Apply SPADE to the union
of the two clouds

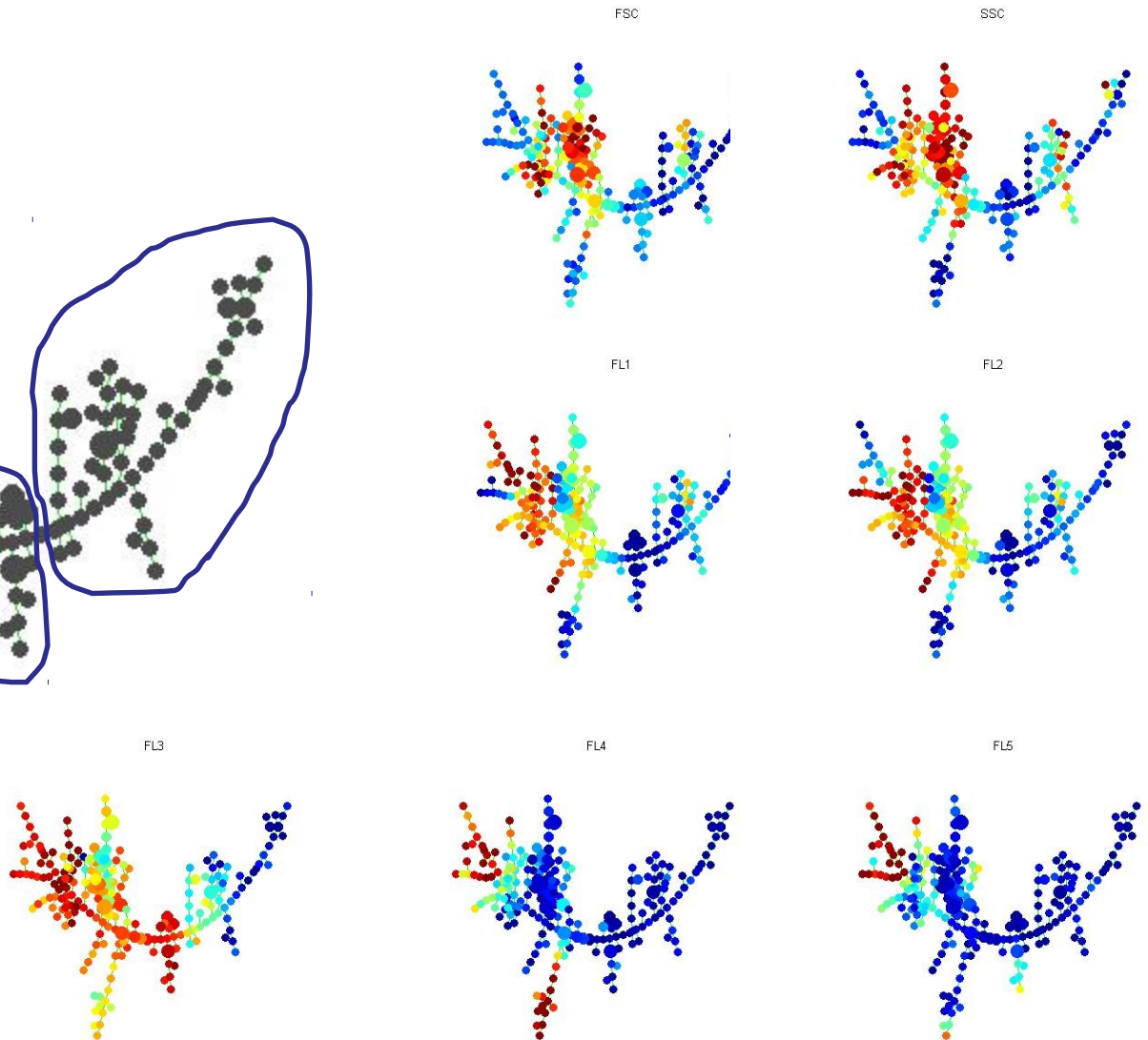
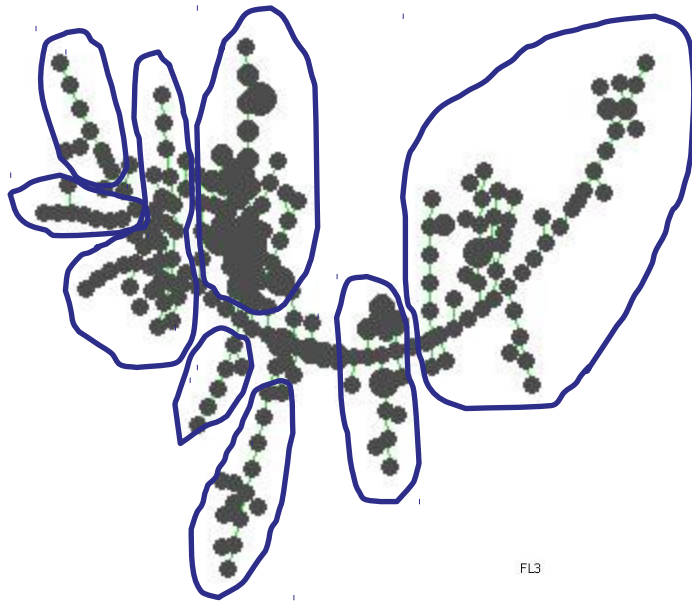
SPADE tree for Tube 2



SPADE tree for Tube 2

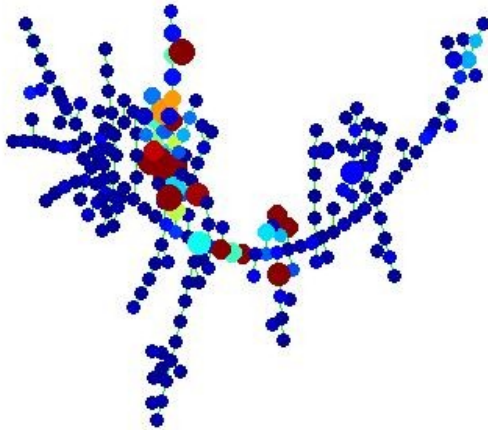


SPADE tree for Tube 2

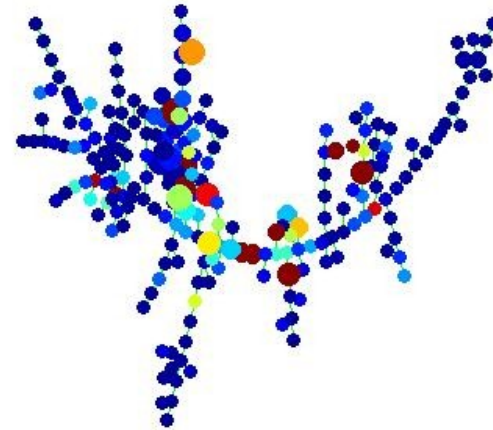


SPADE tree for Tube 2

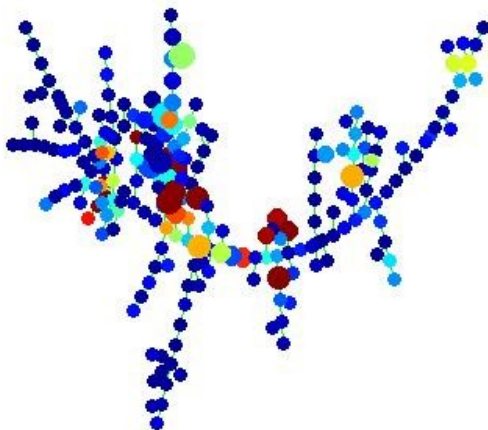
CellFreq Subject 001



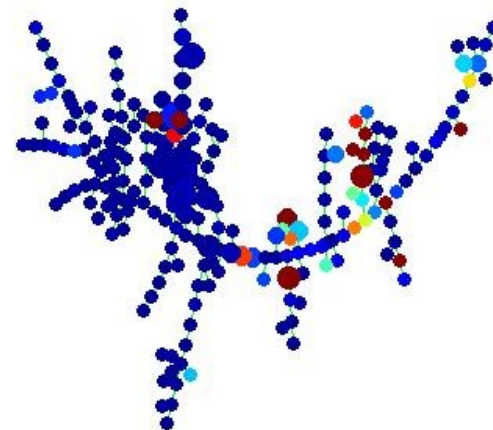
CellFreq Subject 005



CellFreq Subject 007



CellFreq Subject 009



RELIEF classifier & Earth Mover's Distance

Earth Mover's Distance:

a metric to compare two probability distributions over a structured domain.

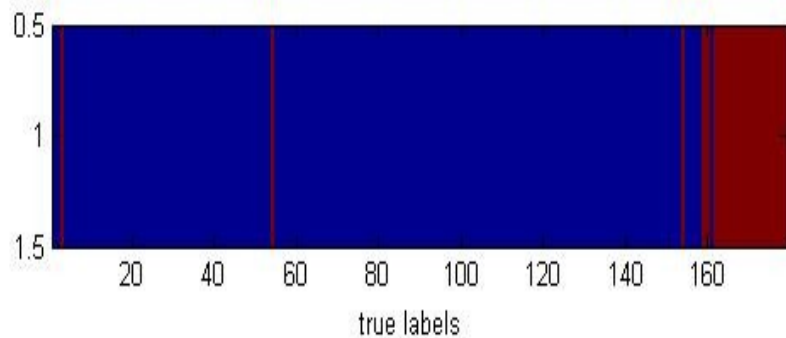
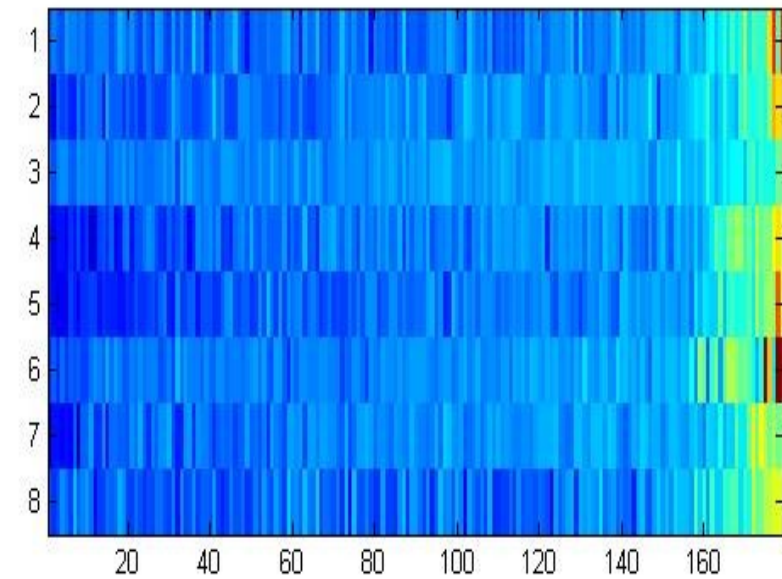
RELIEF classifier

for each testing sample, find its nears normal (N_N) and its nearest AML(N_{AML})

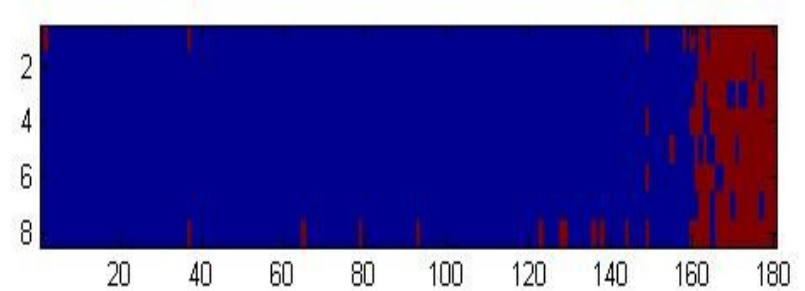
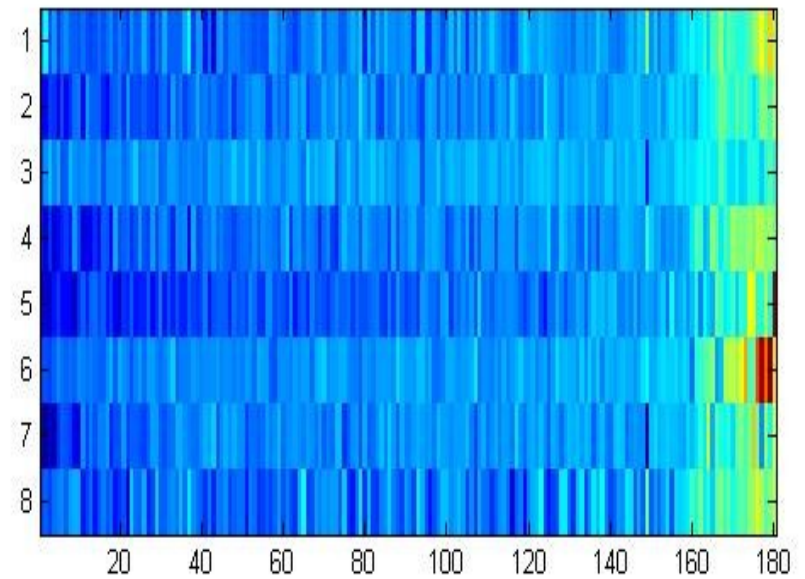
compute the following score: $\text{dist-to-}N_N - \text{dist-to-}N_{AML}$

RELIEF classifier & Earth Mover's Distance

Training samples



Testing samples



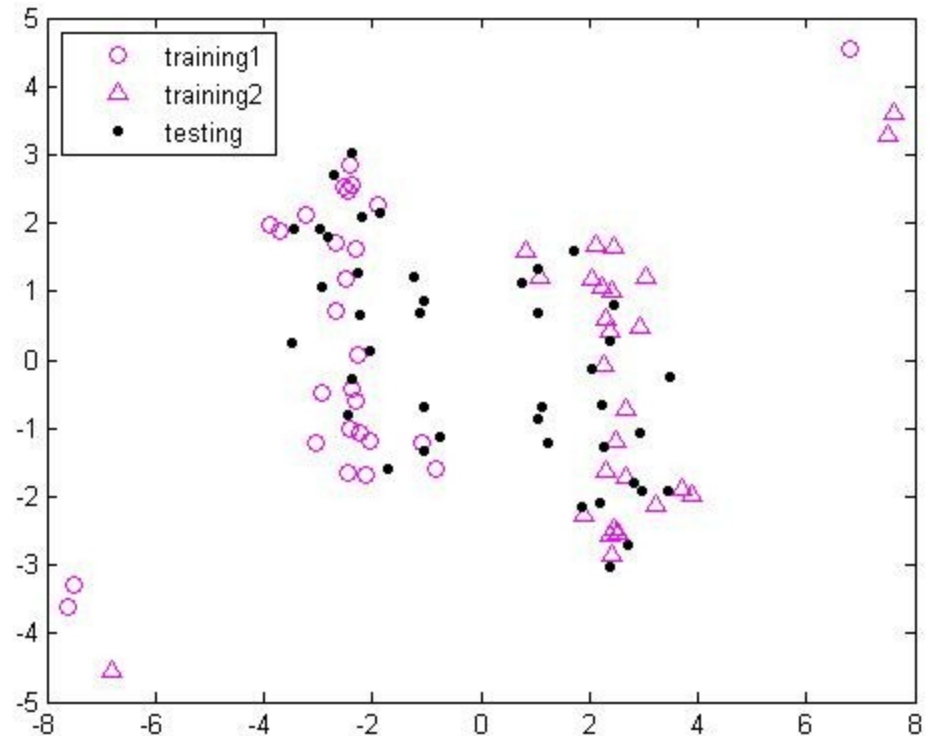
Challenge 3A

Use 48*2 samples to derive a SPADE tree

Compute cell freq distribution for each sample

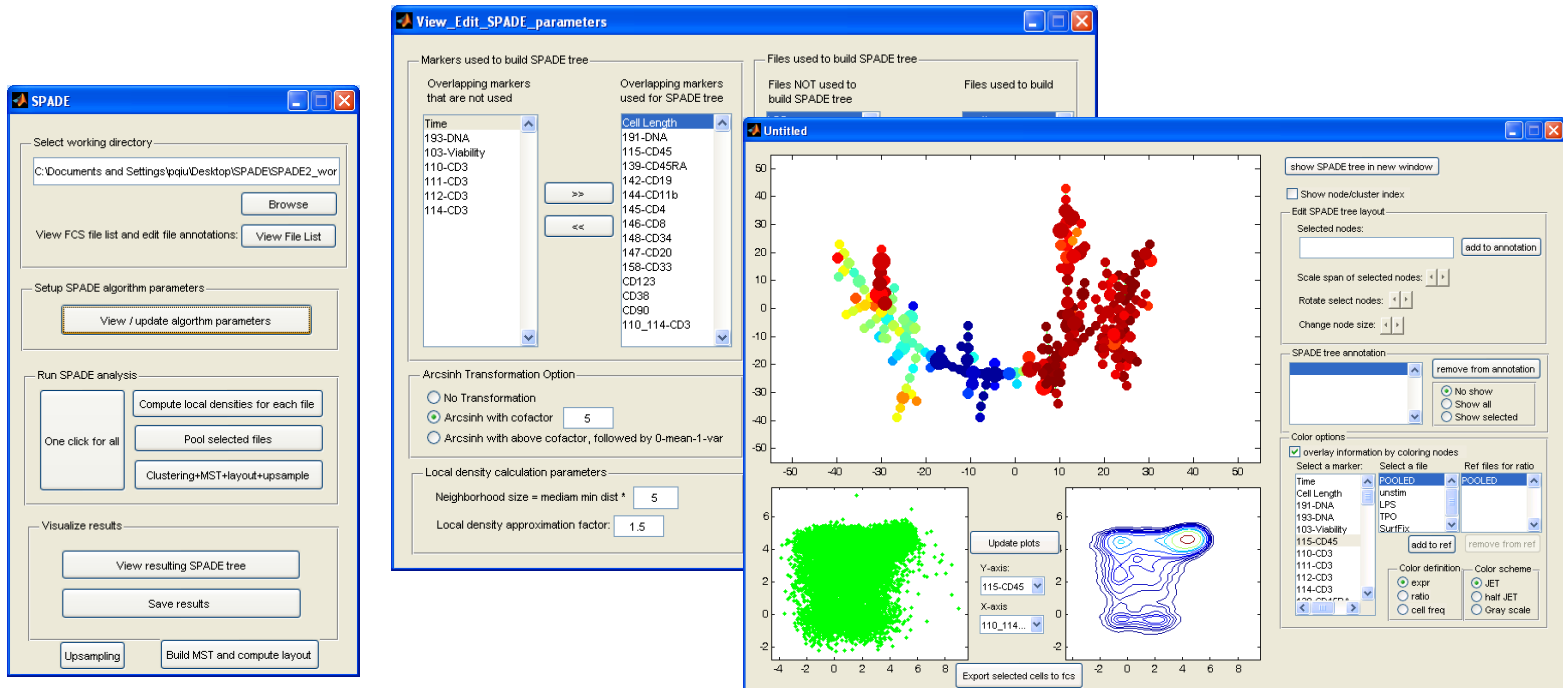
For each sample, compute its distribution – the distribution of its paired sample.

PCA



Summary

- Using SPADE, we can:
 - Identify cell types
 - Compare multiple samples



Acknowledgement

- Sylvia Plevritis
- Garry Nolan
 - Erin Simonds, Sean Bendall,
 - Kenny Gibbs
 - Karen Sachs, Michael Linderman, Rob Bruggner
 - Matt Clutter, Tiffany Chen