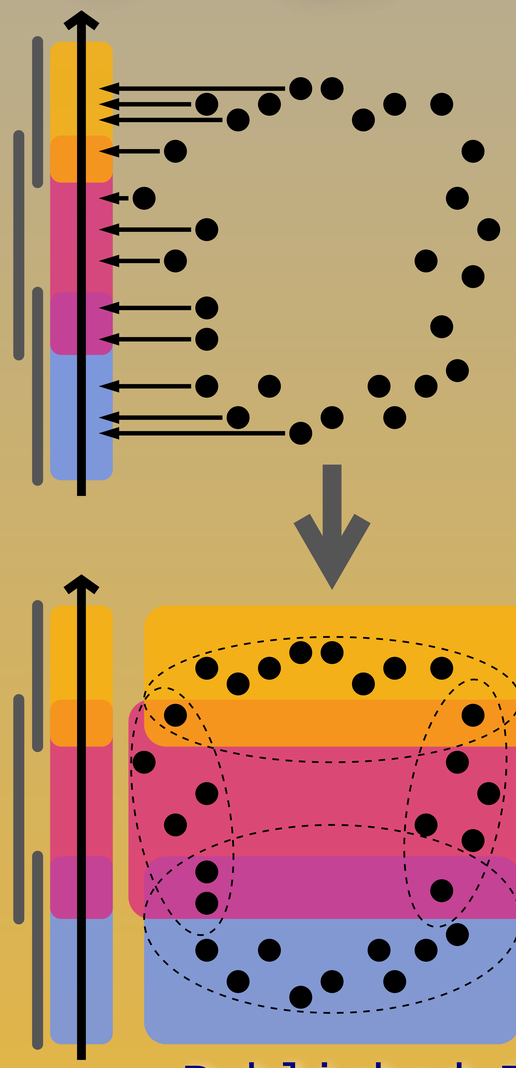
Detecting **novel** lung cancer **groups** with topology and Mapper



1. Filtering
2. Binning
3. Overlapping
4. Clustering
5. Linking

Published Paper DOI
10.1371/journal.pone.0284820

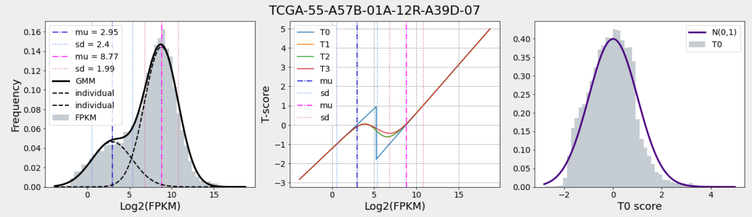
Genomics data analysis via spectral shape and topology

Erik Amézquita[1,2]
eah4d@missouri
Farzana Nasrin[3] **Kathleen Storey**[4] **Masato Yoshizawa**[5]

1. *Division of Plant Sciences & Technology, University of Missouri, Columbia, MO*
2. *Department of Mathematics, University of Missouri, Columbia, MO*
3. *Department of Mathematics, University of Hawaii, Manoa, HI*
4. *Department of Mathematics, Lafayette College, Easton, PA*
5. *School of Life Sciences, University of Hawaii, Manoa, HI*

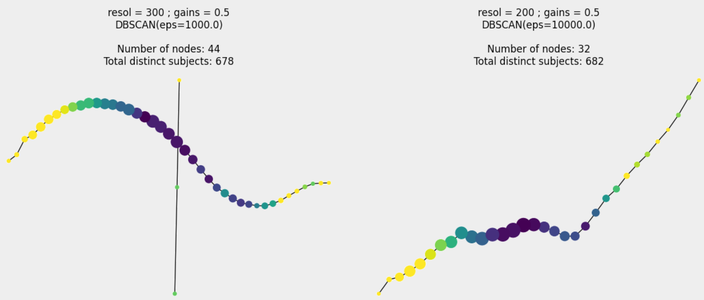
Materials and methods

	TCGA-62-A472-01A-11R-A24H-07	TCGA-93-A4JQ-01A-11R-A24X-07	GTEX-X4LF-0526-SM-3NMB6	GTEX-YF7O-0626-SM-4W21R	TCGA-44-6775-01A-11R-A278-07	GTEX-OXRO-0326-SM-33HBM	TCGA-78-7145-01A-11R-2039-07	TCGA-62-A465-01A-11R-A24H-07	GTEX-P4QS-0526-SM-2I3ET	TCGA-55-7907-01A-11R-2170-07
HIST3H2A	0.00	103.69	100.83	15.11	96.01	77.25	172.65	435.55	18.43	0.00
LIN7B	190.34	115.97	180.02	122.64	109.66	180.02	105.89	140.04	76.71	251.48
LXN	563.18	563.18	1233.75	1250.98	896.64	1369.04	389.72	409.15	698.41	961.07
CNKSR2	2.81	5.41	77.79	35.00	14.35	51.71	6.46	4.82	13.62	11.30
SCML1	265.87	166.73	195.72	128.79	130.60	293.07	119.26	102.25	82.87	60.82
...	...	...	...	...	...	...	...	...	...	...



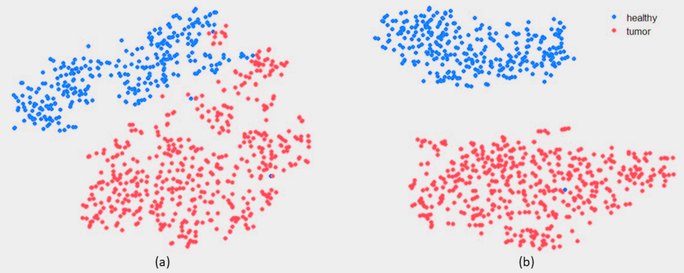
- FPKM counts of RNAseq data from human **lung** tissue:
  - ↪ 19,648 genes per sample.
  - ↪ 314 healthy samples (GTEx).
  - ↪ 500 cancerous samples (TCGA).
- Fit a Gaussian Mixture Model (GMM):
  - ↪ Accurate transformation to a unimodal Gaussian.
  - ↪ Consider ***Z*-scores** onward.
- Compute pairwise correlation across all samples
  - ↪ Filter data by **mean correlation** value.

Novel subgroups revealed with Mapper



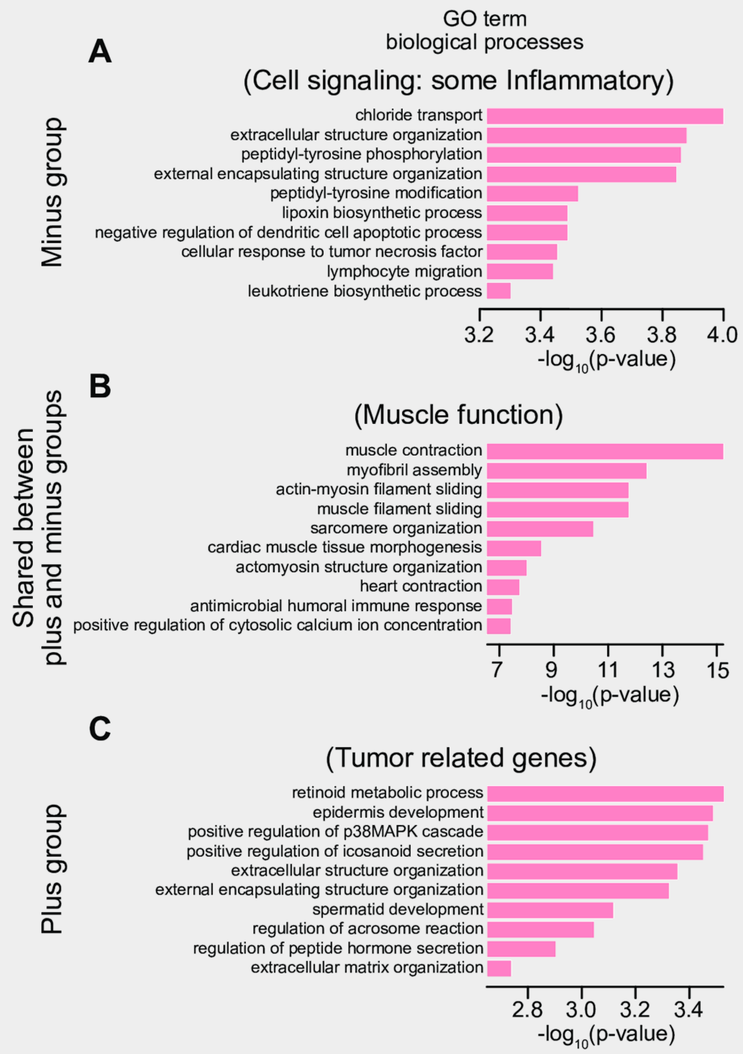
- Vary the number of **bins** $60 \leq b \leq 110$.
- Vary the **overlap** percentage $30 \leq p \leq 80$.
- Bright yellow: 100% cancerous samples.
- Deep purple: 100% healthy samples.
- **Regardless** of parameters:
  - ↪ **Healthy** samples tend to stay at the **center**.
  - ↪ **Cancerous** samples are **split** between both ends.

Split not captured by tSNE



- tSNE separates healthy from cancerous samples **but**,
- Fine structural **details** are **lost** with tSNE:
  (a) Using FPKM, (b) Using GMM Z-scores.

GO Enrichment Analysis



- **Two possible processes for forming lung cancer**.
- Strand **-1**: Mostly tumor cells
  - ↪ Primarily upregulating inflammatory reactions.
- Strand **+1**: Mixed bag but high risk
  - ↪ Environmental factors and tumor gene interactions.
- Similar conclusions when analyzing KEGG pathways.

Acknowledgements

This research was supported by National Institute of General Medical Sciences - Centers of Biomedical Research Excellence (COBRE) grant number P20GM125508.