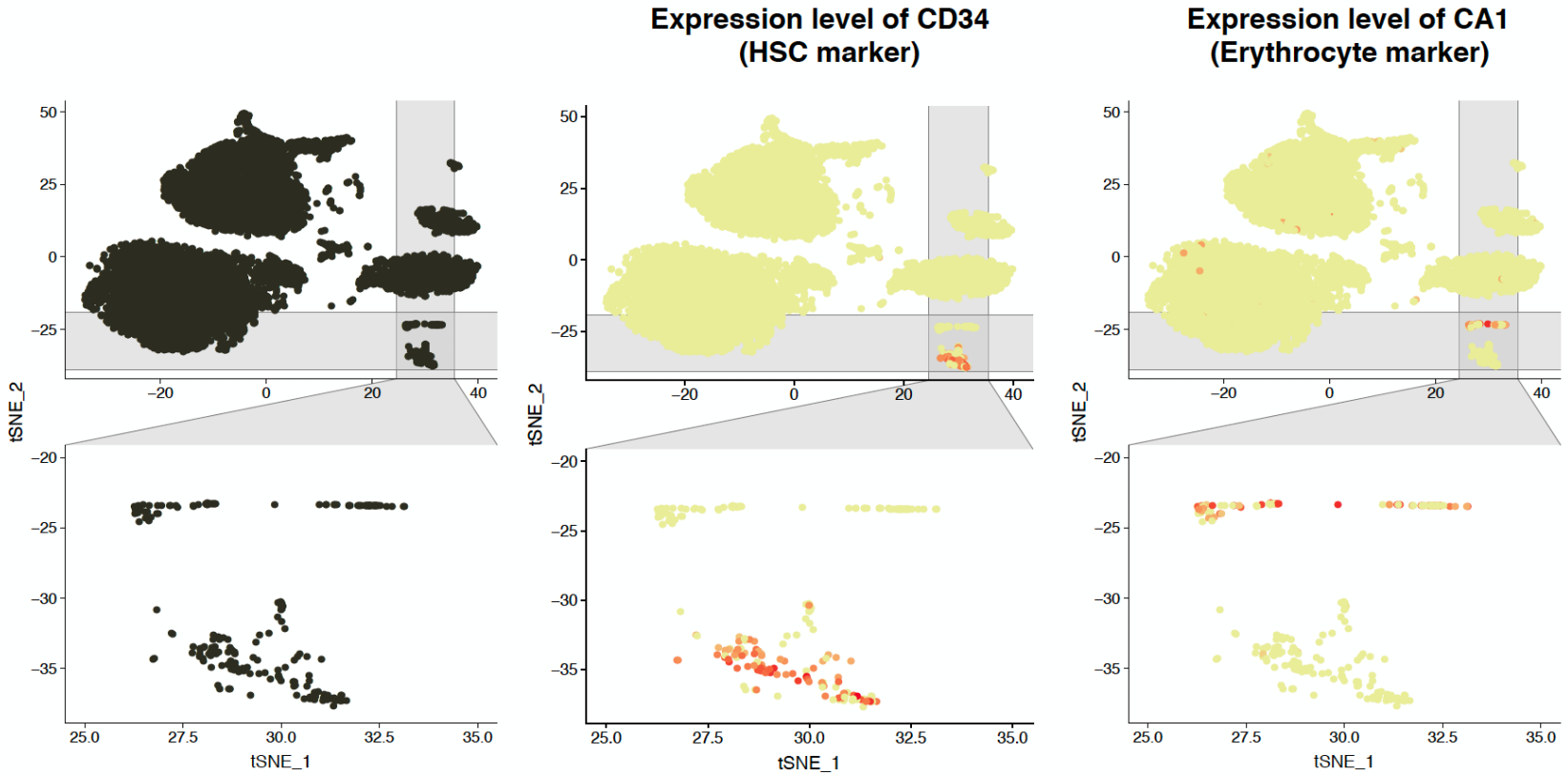
Investigating tumor heterogeneity in breast cancer single cells: a topic modeling approach

Loredana Martignetti / Gabriele Malagoli

Single cell club meeting – 20th September 2022

institutCurie

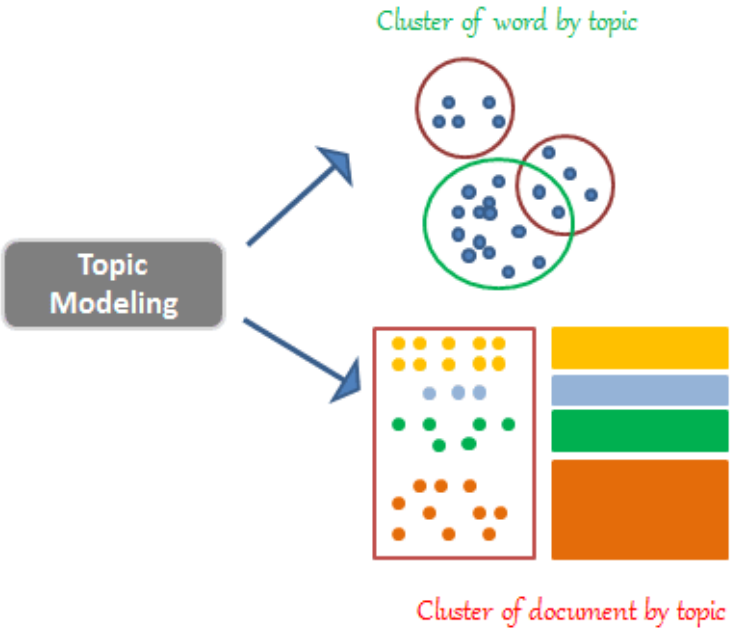
The problem of cell type identification from scRNA-seq data



Main issues:

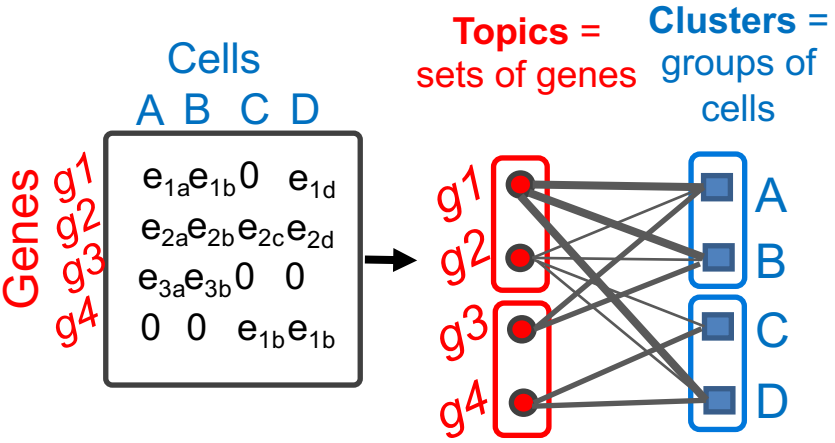
- Different options for clustering can give very different results
- Individual gene expression is noisy and not enough informative
- Error prone, time consuming approach

What is topic modeling



- It was originally developed as a text-mining tool
- A statistical model for discovering abstract ‘topics’ that occur in a collection of document based on word frequencies
- A form of unsupervised learning
- Simultaneous clustering of documents and associated words (keywords)

A topic modeling approach for clustering cells

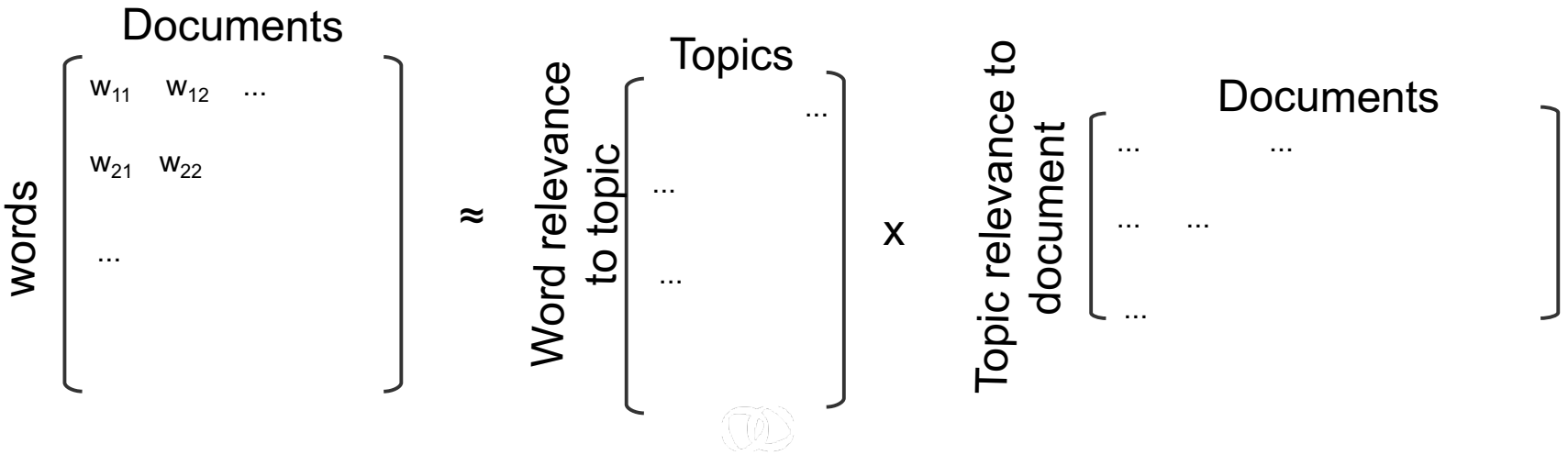


Application in bioinformatics:

- Clustering based on molecular profiles (scRNA-seq, scATAC-seq)
- Efficient for dealing with sparse and semi-sparse data
- Interesting for the interpretability of clusters (identify keywords = specific genes, markers)

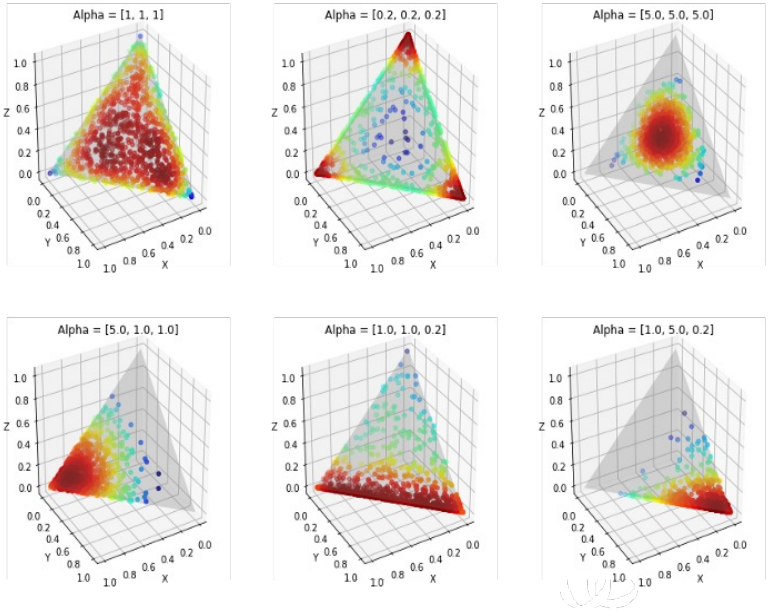
A brief overview of topic modeling algorithms

An early topic model : LSA (latent semantic analysis)



- Reduce the dimension of the document corpus into a small number of topics
- Topics are hidden variables
- Deterministic algorithm

LSA → LDA : latent Dirichlet allocation



- Probabilistic model (David M Blei, Andrew Y Ng, Michael I Jordan (2003))
- Model the corpus of documents as a mixture of N-dimensional Dirichlet distributions
- Large α values (α >1) push the distribution to the middle of the triangle, whereas smaller α values push the distribution to the corners (α <1)
- It contains a large number of free parameters that can cause overfitting
- It requires to set the number of topics

LDA applied to bioinformatics

cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters & Stein Aerts

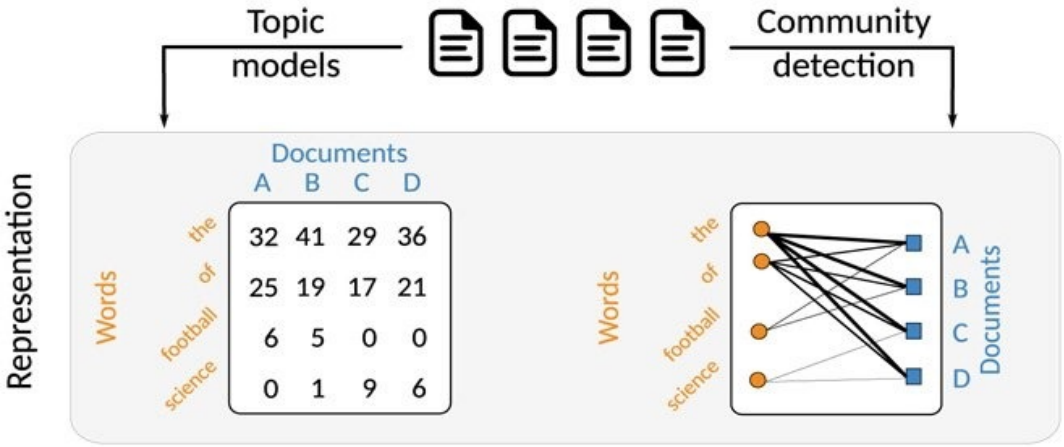
Nature Methods **16**, 397–400 (2019) | Cite this article

> Bioinformatics. 2020 Jul 1;36(Suppl_1):i474-i481. doi: 10.1093/bioinformatics/btaa403.

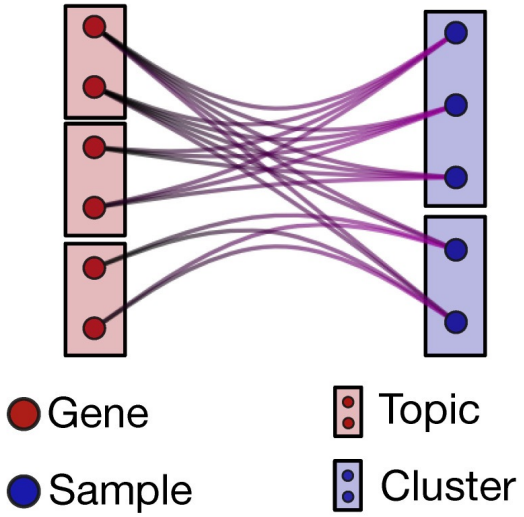
TopicNet: a framework for measuring transcriptional regulatory network change

Shaoke Lou [1], Tianxiao Li [2], Xiangmeng Kong [1], Jing Zhang [1], Jason Liu [1], Donghoon Lee [1], Mark Gerstein [1]

A network approach to topic models:
the problem of inferring topics becomes a problem of inferring communities



Gerlarch, Peixoto, et. al. 2018



Valle, Caselle, et. al. 2020

- Building a bipartite word-document network weighted on the word frequency
- Detect communities in the bipartite network
- hierarchical Stochastic Block Modeling (hSBM)

https://github.com/martingerlach/hSBM_Topicmodel

hSBM: some methodological details

- Non parametric algorithm
- Available in the **graph-tool** python library

Inference procedure based on Markov Chain Monte Carlo

0: Network partition initialization
1: **for** $j = 0$ **to** k **do**
2: Get a random move (i: $r \rightarrow s$) of node i from block r to block s
3: Calculate the improvement of the target function $\Delta F(i: r \rightarrow s)$
4: Accept the move (i: $r \rightarrow s$) with probability $p_A = \min(1, \exp(\beta \Delta F))$
5: **end for**

with F = network description length function ; k= num. of steps

https://github.com/martingerlach/hSBM_Topicmodel

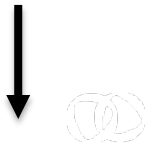
https://topsbm.github.io/TopSBM-tutorial.html

T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model. *Phys. Rev. E* **95**, 012317 (2017).

Output of a topic model

- P(word|topic) = probability of associating a gene with a topic

- P(topic|document) = probability of associating a topic with a document

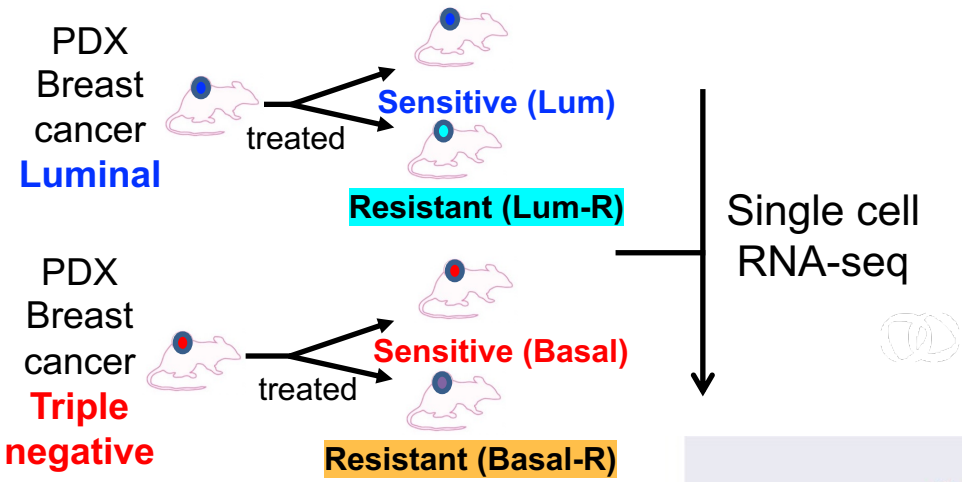


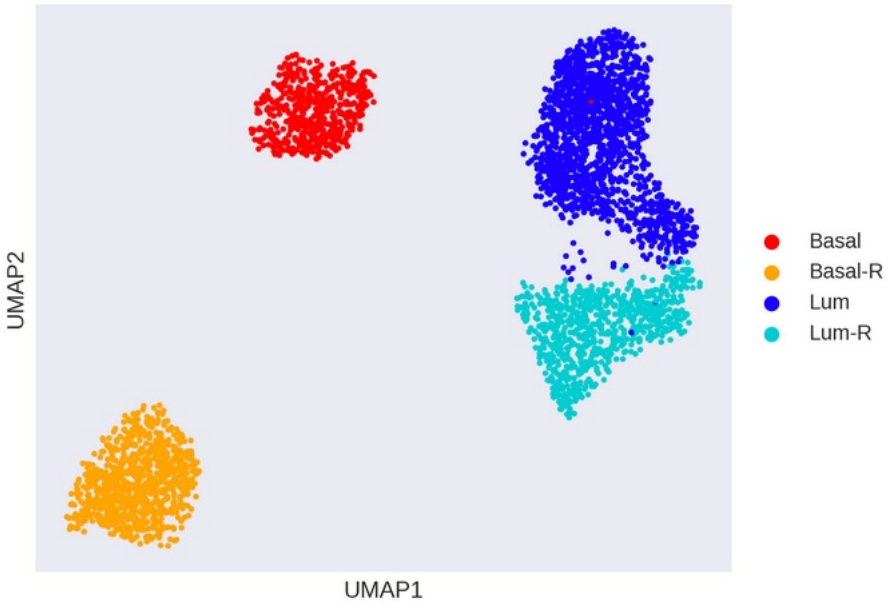
  - Classify documents into topics
  - Identify keywords that characterize our topics

Application to breast cancer scRNA-seq

PDX models of breast cancer with acquired resistance to treatment

[Grosselin K et. al, Nat Genet 2019] – Models and data from Institut Curie and ESPCI



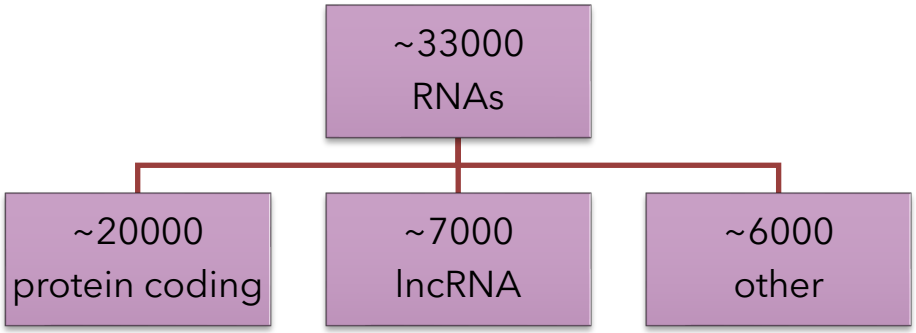


Aim of the analysis

- Clustering cells of the four PDX models

- Identify signatures associated to drug-resistant cells

Pre-processing



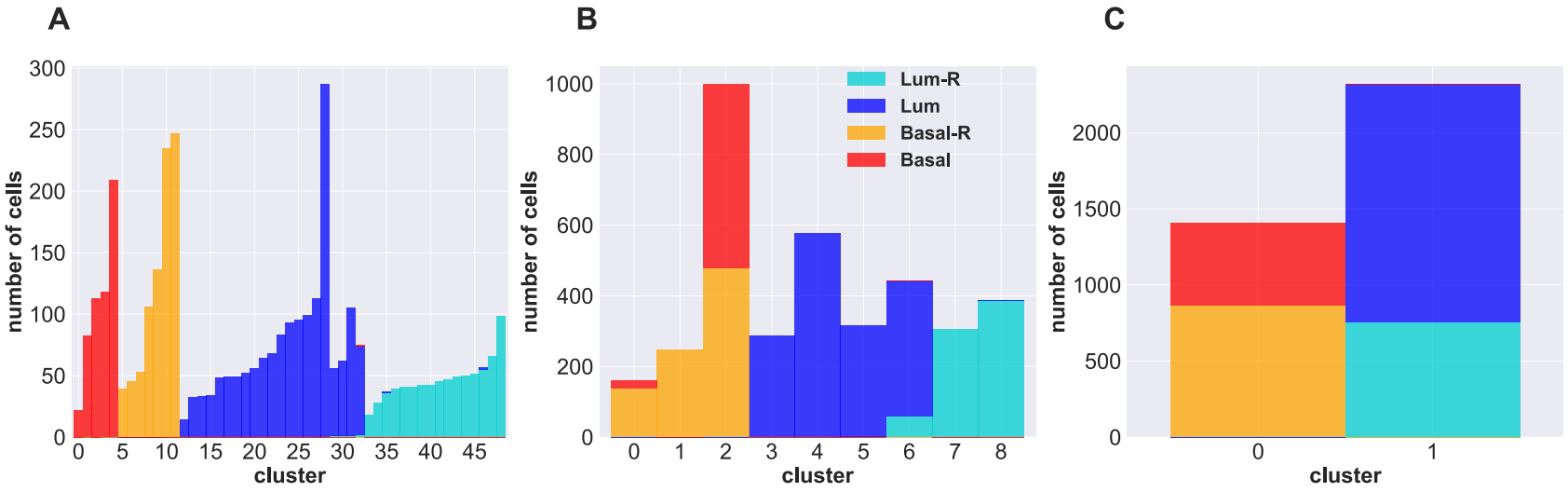
Gene selection	Normalization	Filter	Result
Select the class of RNAs you are interested in (lincRNAs, mRNAs,...)	Normalize total counts by library size	Select highly variable genes	2111 protein coding HVGs

Cells clustering obtained with hSBM applied to the mRNA expression dataset

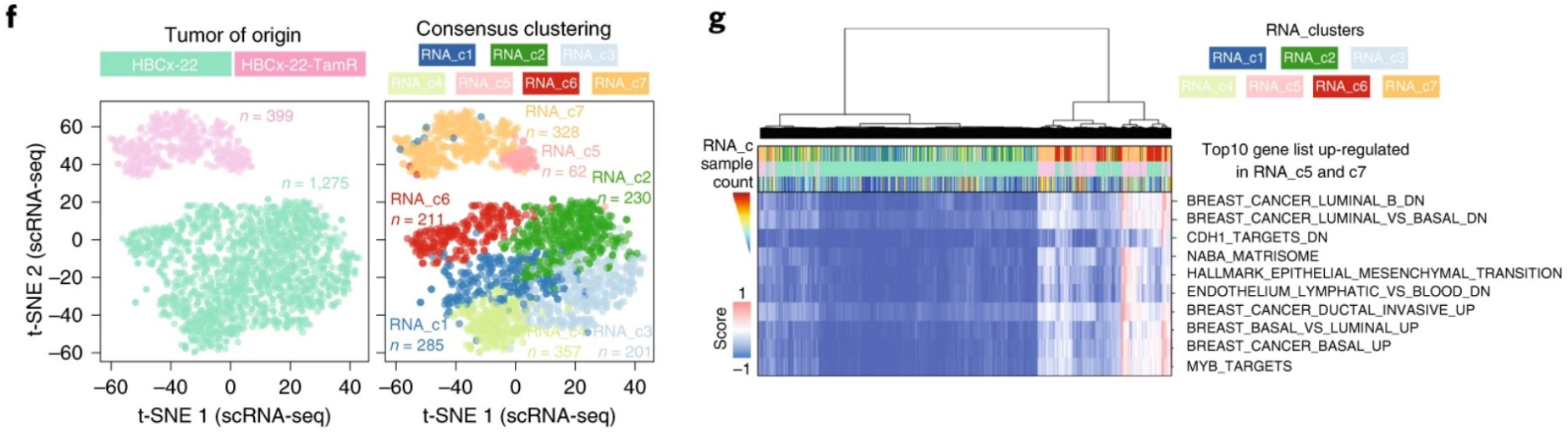
Level 1 > 40 clusters

Level 2 = 9 clusters

Level 3 = 2 clusters



Clustering of scRNA-seq obtained in Grosselin K et al



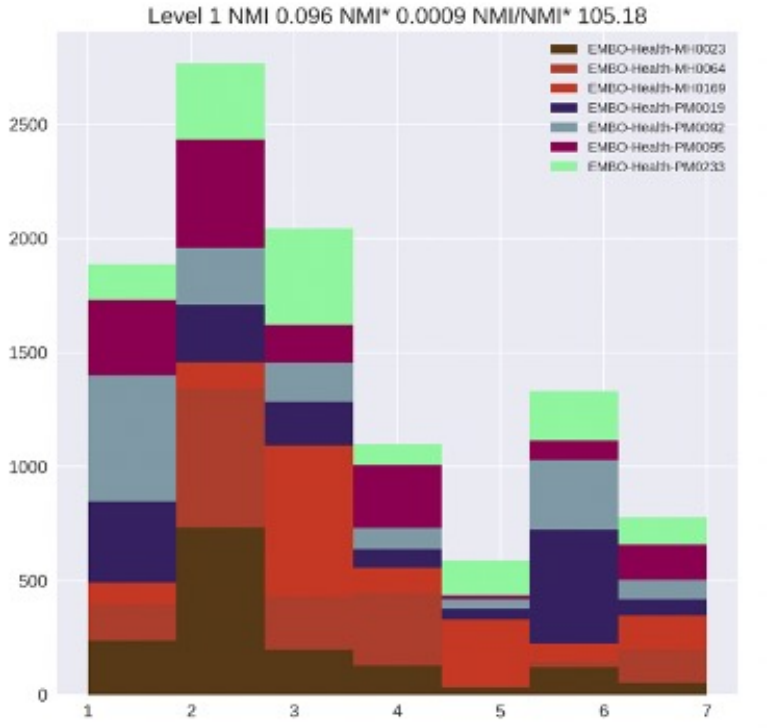
Testing for batch effect by using a dataset of healthy cells

EMBO J. 2021 Jun 1;40(11):e107333. doi: 10.15252/embj.2020107333. Epub 2021 May 5.

A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast

- We took healthy breast cells from 7 different donors
- We run hSBM on the dataset of healthy cells

Testing for batch effect by using a dataset of healthy cells



Clusters are composed by cells coming from different individuals

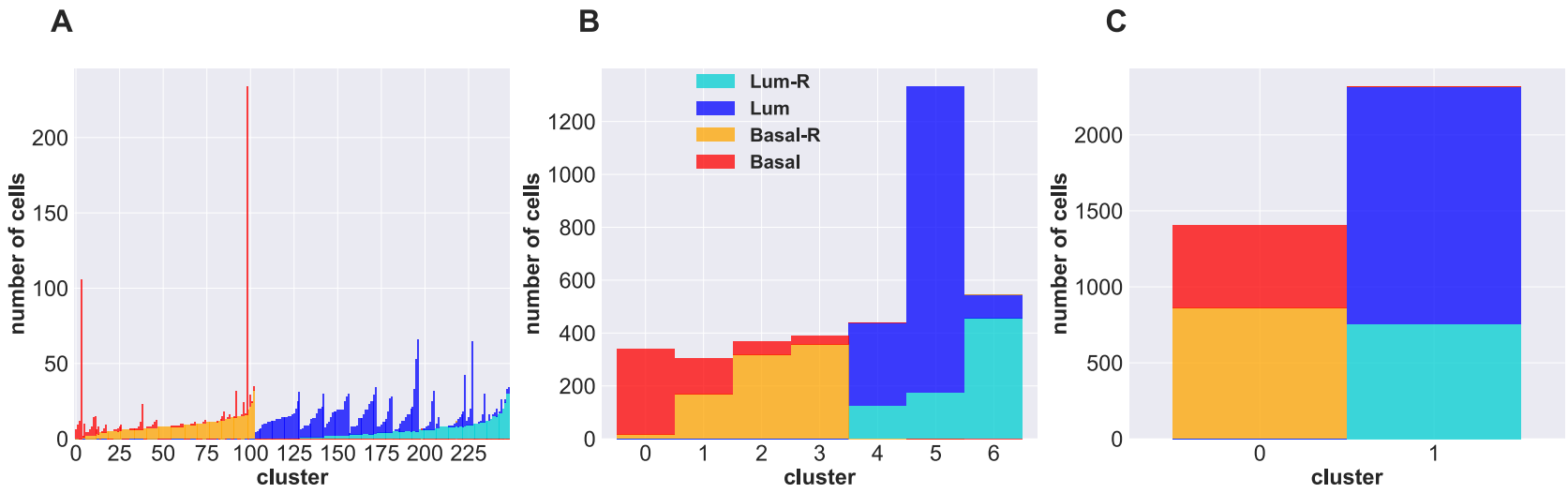
hSBM solution is not affected by batch effect

Cells clustering obtained with hSBM applied to the lncRNA expression dataset

Level 1 > 200 clusters

Level 2 = 7 clusters

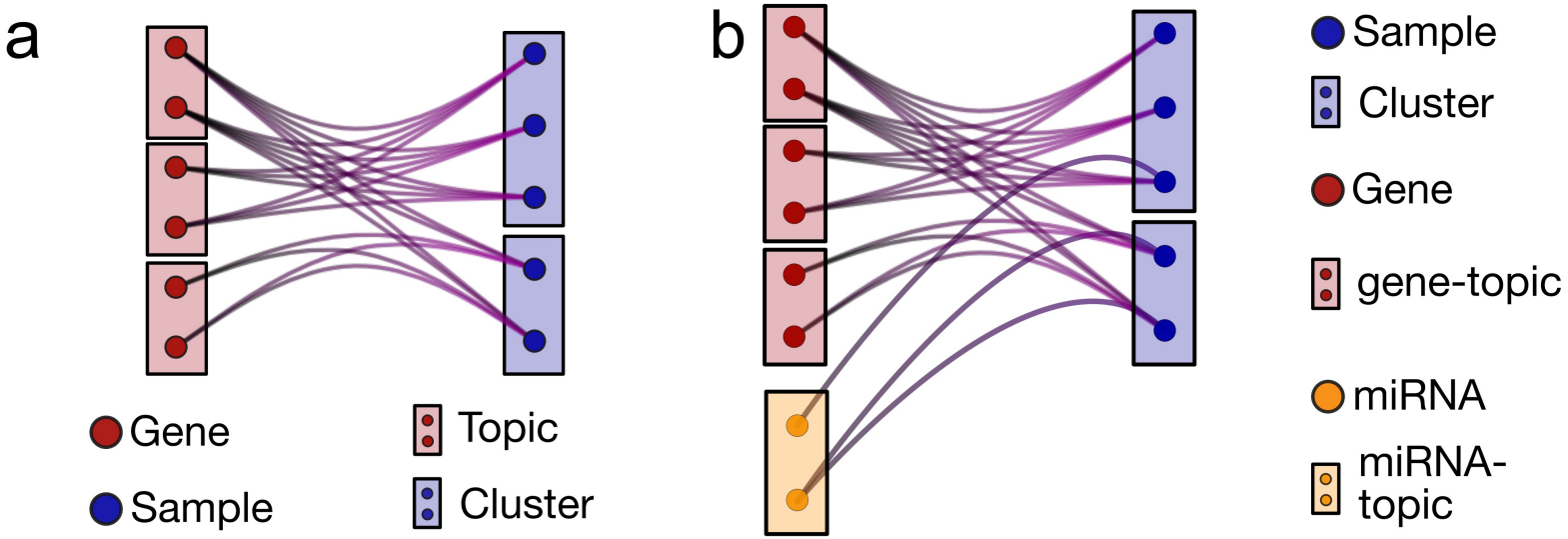
Level 3 = 2 clusters



Integrative clustering of mRNAs and lncRNAs with **Multiomics Topic Modeling**

Hyland, C.C.; Tao, Y.; Azizi, L.; Gerlach, M.; Peixoto, T.P.; Altmann, E.G. Multilayer networks for text analysis with multiple data types. EPJ Data Sci. **2021**

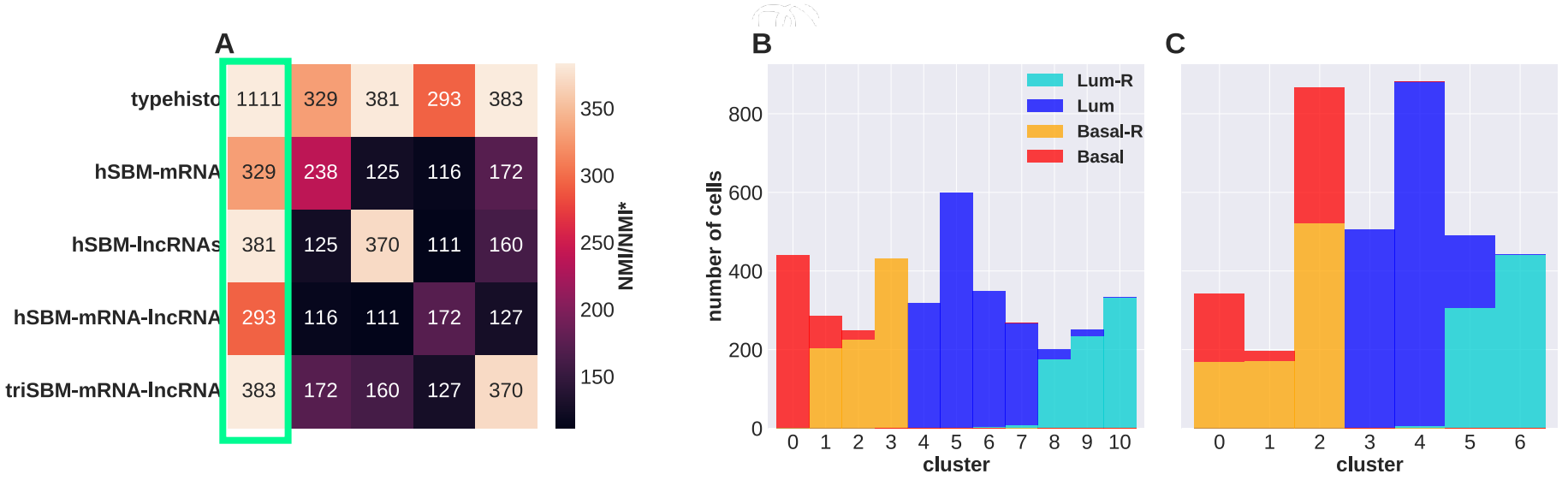
Valle, F. nSBM: Multi Branch Topic Modeling. Zenodo 2021. Available online: https://zenodo.org/record/6120683



Integrative clustering of mRNAs and lncRNAs with **Multiomics Topic Modeling**

Strategy 1: concatenated mRNAs and lncRNAs expression matrices

Strategy 2: multi-omics topic modeling by tri-partite hSBM



Functional enrichment analysis of topics

JOURNAL ARTICLE

Molecular signatures database (MSigDB) 3.0 FREE

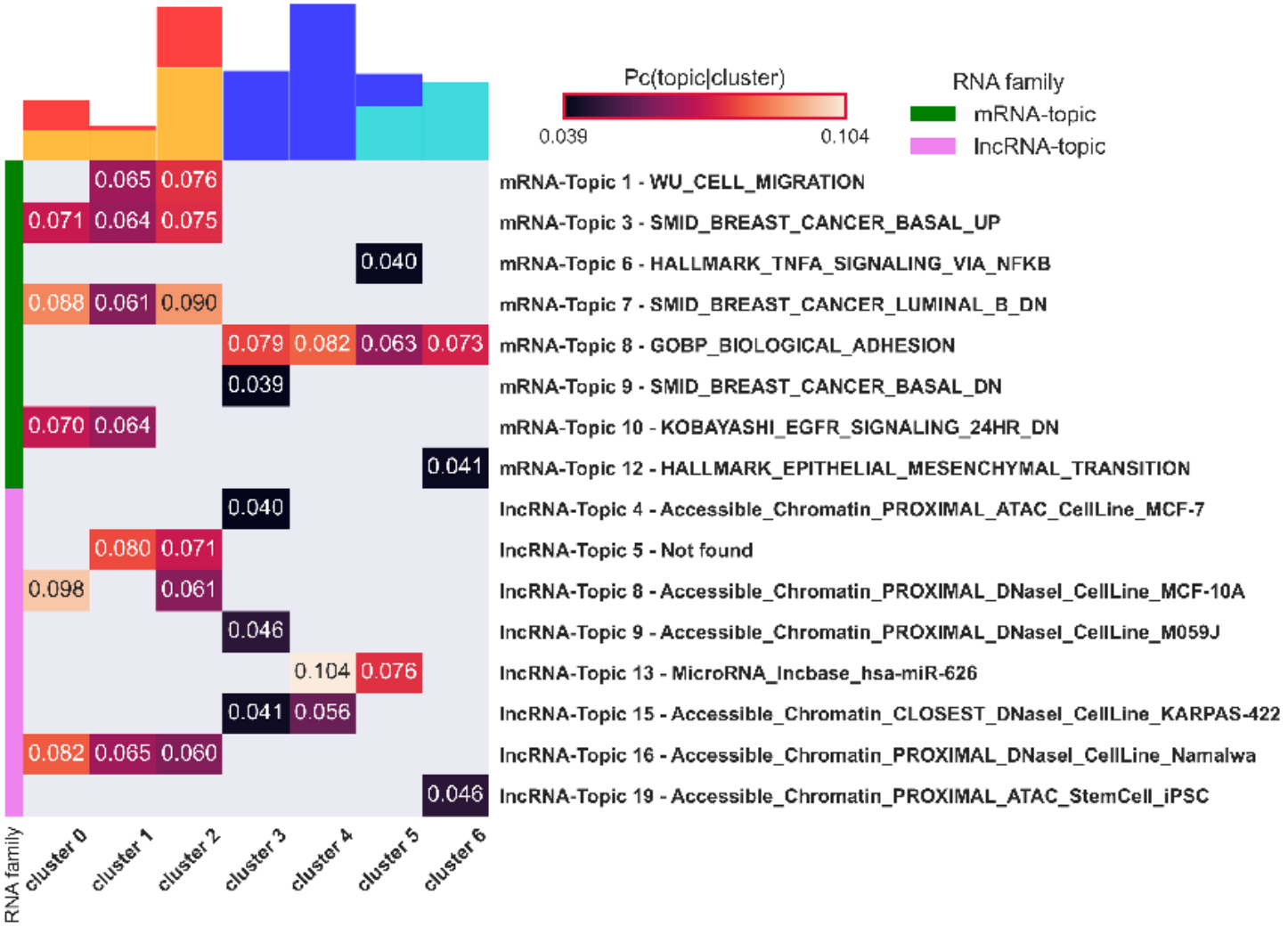
Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, Jill P. Mesirov ✉ Author Notes

LncSEA: a platform for long non-coding RNA related sets and enrichment analysis

Jiaxin Chen, Jian Zhang, Yu Gao, Yanyu Li, Chenchen Feng, Chao Song, Ziyu Ning, Xinyuan Zhou, Jianmei Zhao, Minghong Feng ... Show more
Author Notes

Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D969–D980,

Functional enrichment analysis of topics



Computational time to run hSBM on different types of datasets

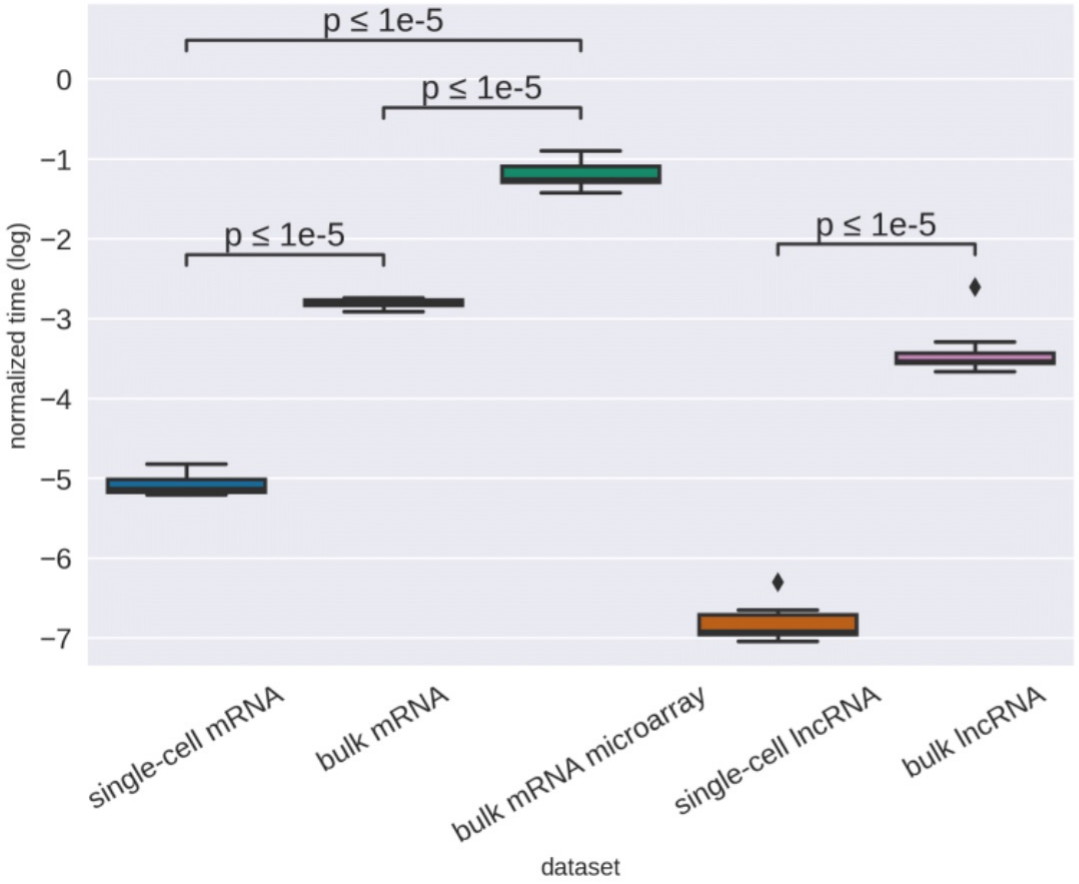


Figure 16: Box plots showing the time (normalised) that hSBM took to run with each data set. The p-values are the result of the t-test between two sets of normalised times.

Work in progress

- Exploring the signatures (keywords) of the Lum-LumR cluster
- Implementing a measure of the stability of the hSBM solution
- Analysis of the sc-Chipseq dataset

For more details:

Comprehensive analysis of long non-coding RNAs in breast cancer using topic modeling.

Gabriele Malagoli, Filippo Valle, Emmanuel Barillot, Michele Caselle, Loredana Martignetti

doi: https://doi.org/10.1101/2022.09.13.507779

Conclusions

- Topic modeling is an interesting parameter-free clustering method
- Useful for clustering of sparse and semi-sparse datasets (scRNA-seq, lncRNA analysis)
- In progress: useful for identifying relevant topics (= sets of genes)

Acknowledgments

- U900 Unit
Computational Systems Biology of Cancer team
Bioinformatics platform

- University of Torino, BioPhys team

- Grosselin Kevin et al (for data availability)

- Martin Gerlach, Wikimedia foundation

- graph-tool development team

Inserm
Institut national
de la santé et de la recherche médicale

aviesan
ITMO Cancer

Thank you for your attention