IN SUPREMÆ DIGNITATIS 1343
UNIVERSITÀ DI PISA

SAILab
Siena Artificial Intelligence Lab

IIT
Consiglio Nazionale delle Ricerche

Network analysis for the integration of histone modification data to explain haematopoiesis

Federica Baccini

Dipartimento di Informatica, Università degli Studi di Pisa

Institute of Informatics and Telematics of CNR, Pisa

federica.baccini@phd.unipi.it

Pisa, March 23, 2020

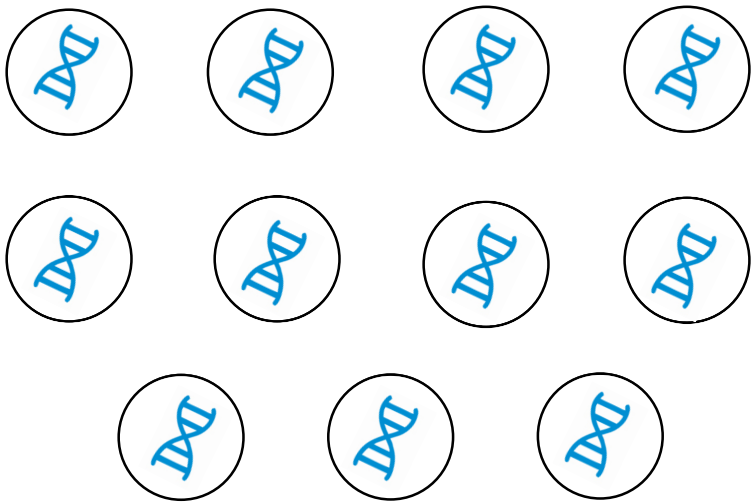
Outline

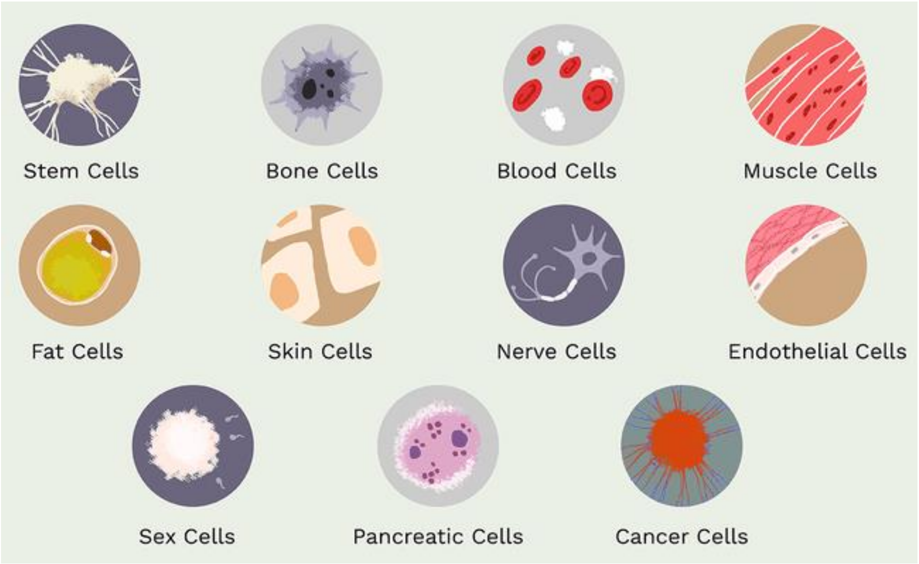
- Introduction to epigenetics and haematopoiesis

- Experimental analysis and methods:
    - Data description and processing
    - Hypothesis testing model

- Results
- Conclusions and further work

What is epigenetics?





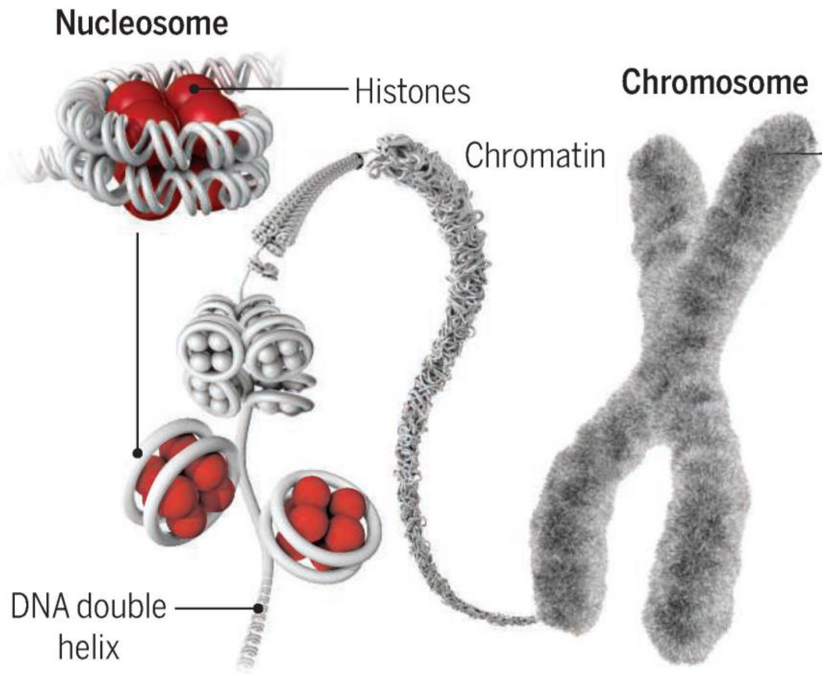
All the cells have same DNA…

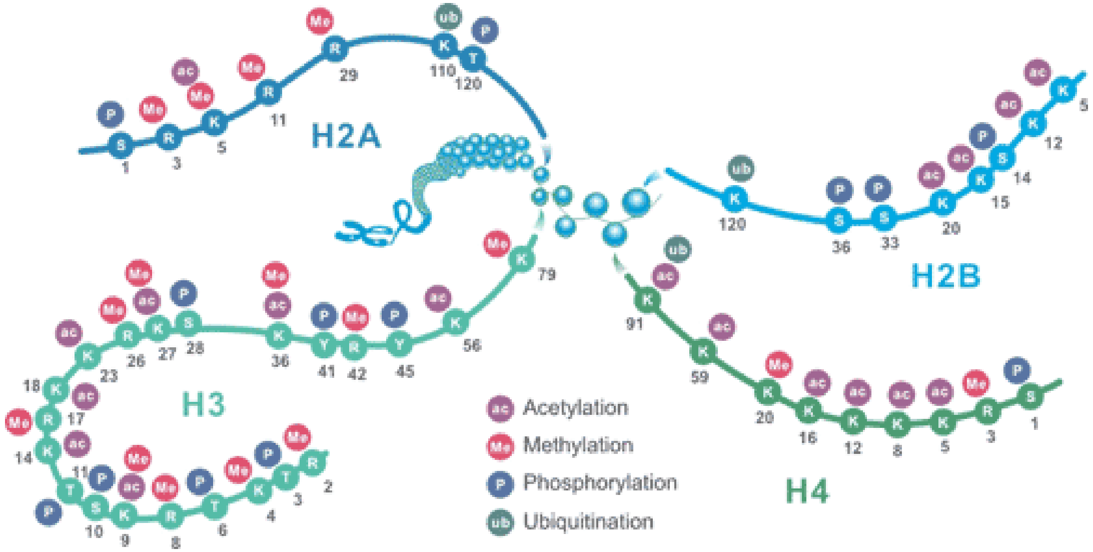
…but there are many types of different cells

REGULATION OF GENE EXPRESSION THROUGH MODIFICATIONS

EPIGENETICS

Histone modifications





Histones are protein complexes around which DNA binds. They allow DNA to assume a compact structure (chromatin), and to finally organize into chromosomes.

Histones and, predominantly, their N-tails, can be subject to chemical modifications that can act as promoters or inhibitors of gene expression.

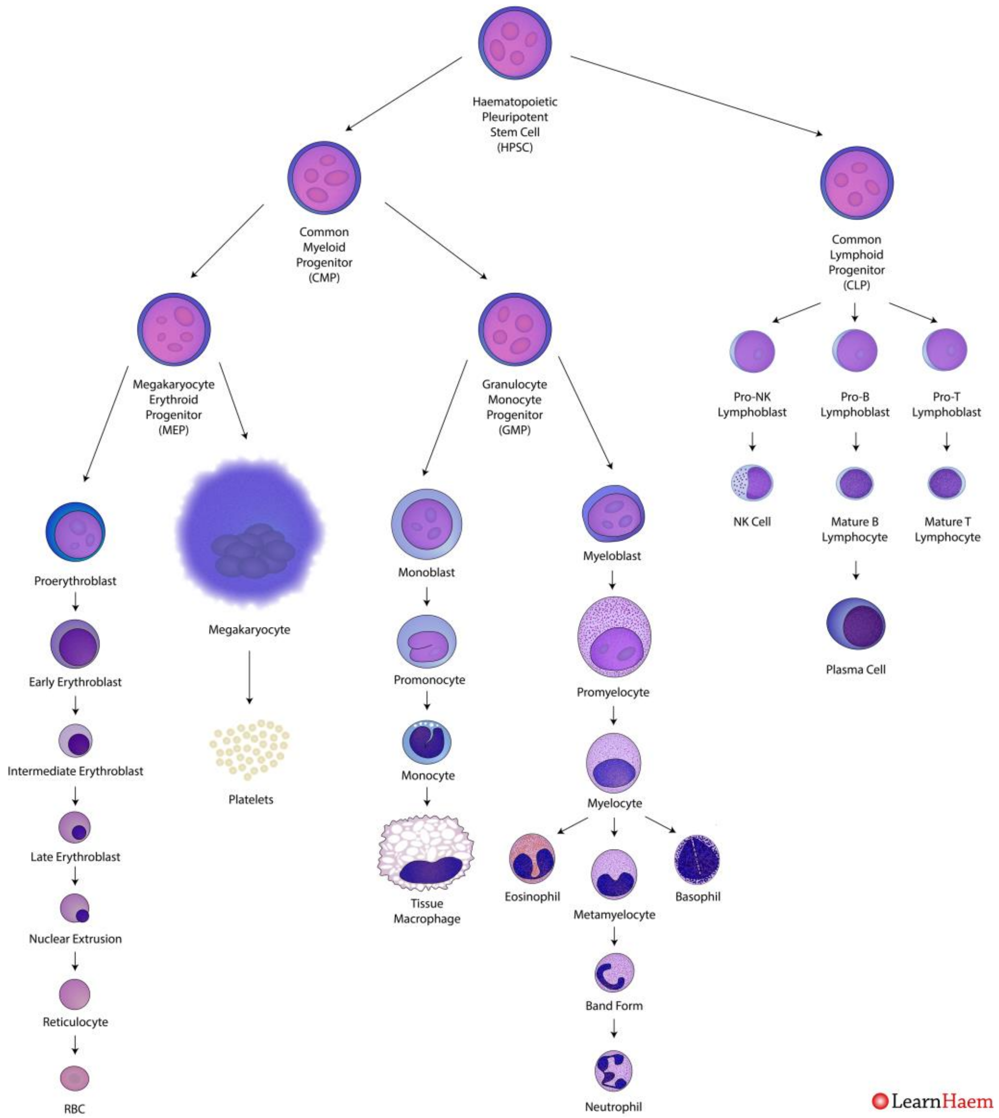
The process of haematopoiesis

Haematopoietic (multipotent) stem cell

Progenitors (oligopotent)

Precursors (MEP and GMP)

Mature cells



LearnHaem

Differentiation capability and self-renewal

Proliferation capability

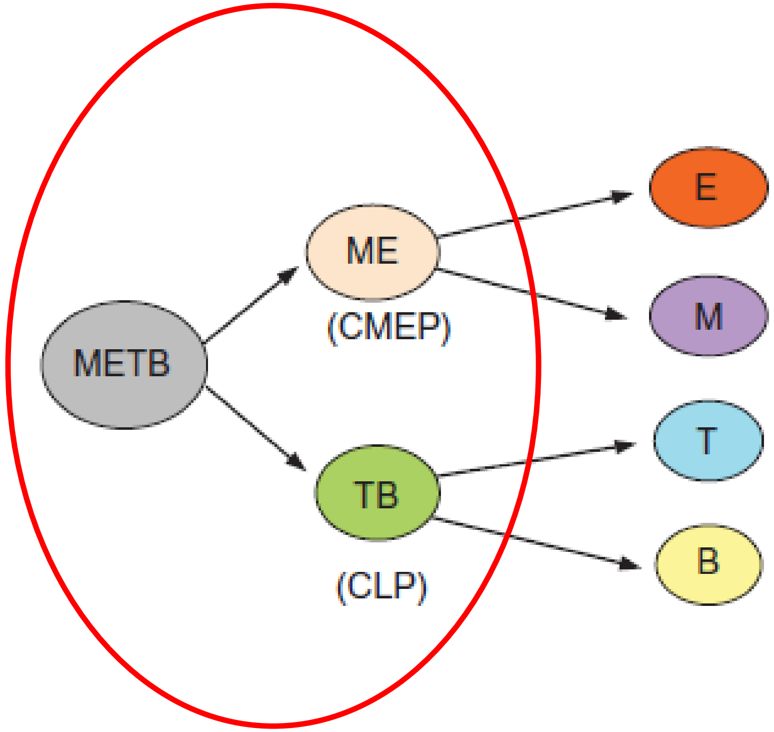
Challenges to the classical model

- Studies have highlighted that the myeloid potential is maintained in both the lymphoid and myeloid lineages.

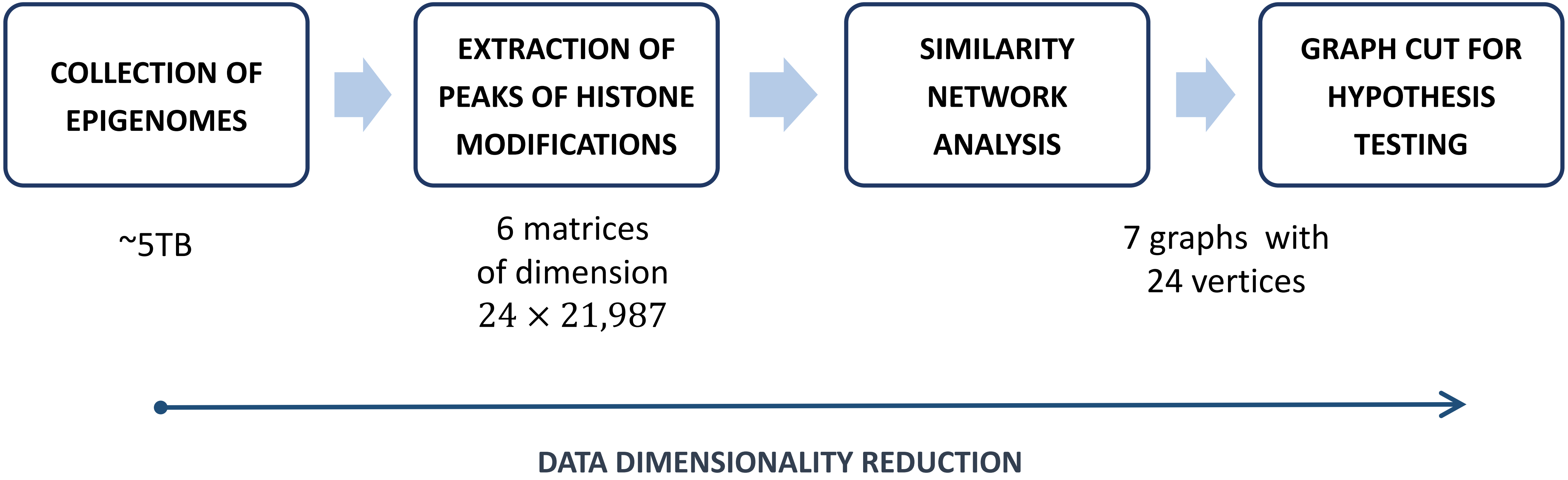
Questions:

- Does Epigenetics play a role in the process of haematopoiesis?

- Is it possible to build a model for testing the classical hypothesis on the first hierarchical subdivision?



Outline and dimensionality reduction



COLLECTION OF EPIGENOMES

~5TB

EXTRACTION OF PEAKS OF HISTONE MODIFICATIONS

6 matrices of dimension $24 \times 21{,}987$

SIMILARITY NETWORK ANALYSIS

GRAPH CUT FOR HYPOTHESIS TESTING

7 graphs with 24 vertices

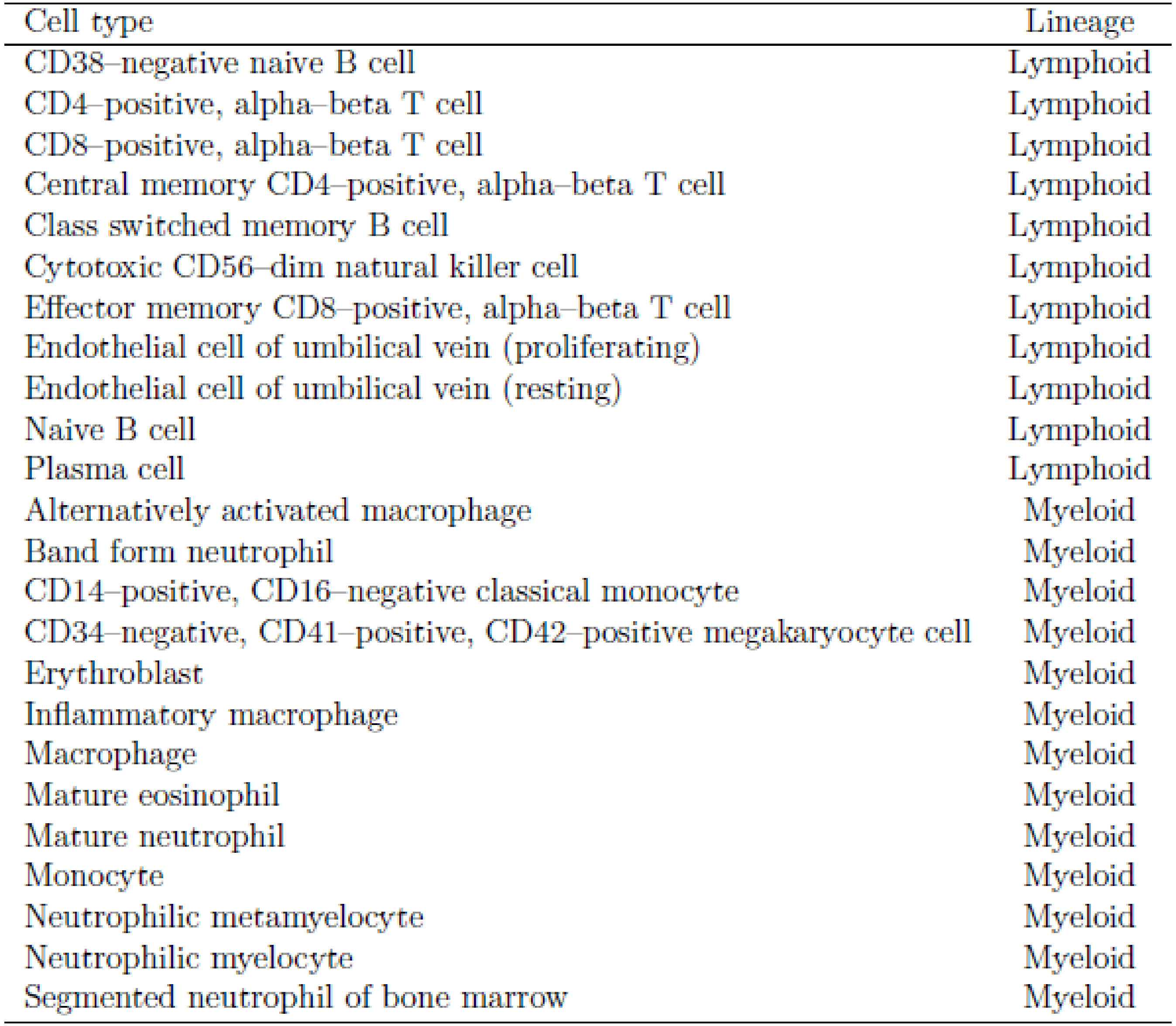
DATA DIMENSIONALITY REDUCTION

Data collection and organization-1

# of cellular types : 24

# lymphoid: 11

# myeloid: 13

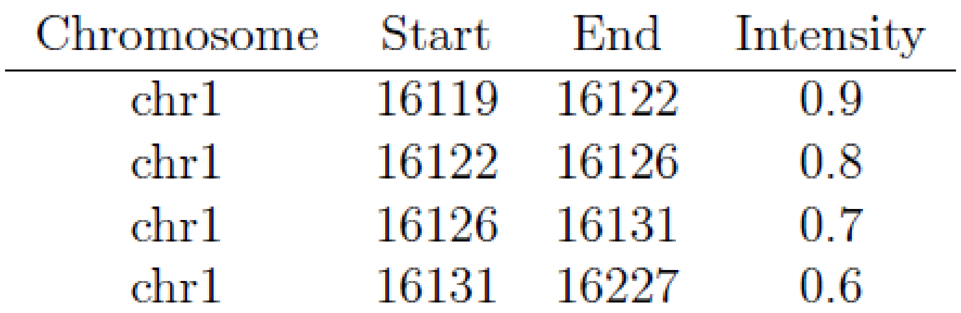


Cell type	Lineage
CD38–negative naive B cell	Lymphoid
CD4–positive, alpha–beta T cell	Lymphoid
CD8–positive, alpha–beta T cell	Lymphoid
Central memory CD4–positive, alpha–beta T cell	Lymphoid
Class switched memory B cell	Lymphoid
Cytotoxic CD56–dim natural killer cell	Lymphoid
Effector memory CD8–positive, alpha–beta T cell	Lymphoid
Endothelial cell of umbilical vein (proliferating)	Lymphoid
Endothelial cell of umbilical vein (resting)	Lymphoid
Naive B cell	Lymphoid
Plasma cell	Lymphoid
Alternatively activated macrophage	Myeloid
Band form neutrophil	Myeloid
CD14–positive, CD16–negative classical monocyte	Myeloid
CD34–negative, CD41–positive, CD42–positive megakaryocyte cell	Myeloid
Erythroblast	Myeloid
Inflammatory macrophage	Myeloid
Macrophage	Myeloid
Mature eosinophil	Myeloid
Mature neutrophil	Myeloid
Monocyte	Myeloid
Neutrophilic metamyelocyte	Myeloid
Neutrophilic myelocyte	Myeloid
Segmented neutrophil of bone marrow	Myeloid

[1] Source of the data: https://epigenomesportal.ca/ihec/

Data collection and organization-2

- Epigenomes record the intensity of 6 histone modifications:
    - H3K27ac
    - H3K27me3
    - H3K36me3
    - H3K4me1
    - H3K4me3
    - H3K9me3



Chromosome	Start	End	Intensity
chr1	16119	16122	0.9
chr1	16122	16126	0.8
chr1	16126	16131	0.7
chr1	16131	16227	0.6

- Samples from diseased donors were filtered out.

Counting peaks per gene

- **Computation of peaks** of each histone modification in every epigenome*.*

- **Count of the number of peaks per gene**[2] in each sample (# genes considered: $21{,}987$), for each modification.

- Construction of **6 matrices** (one for each histone modification), where for a generic matrix $\boldsymbol{M}$, $\boldsymbol{M_{ij}} = \mathbf{number\ of\ peaks\ of\ sample}\ \boldsymbol{i}\ \mathbf{in\ gene}\ \boldsymbol{j}$.

[2] http://ftp.ensembl.org/pub/release-76/gtf/homo_sapiens/

Data cleaning and construction of cell type matrices

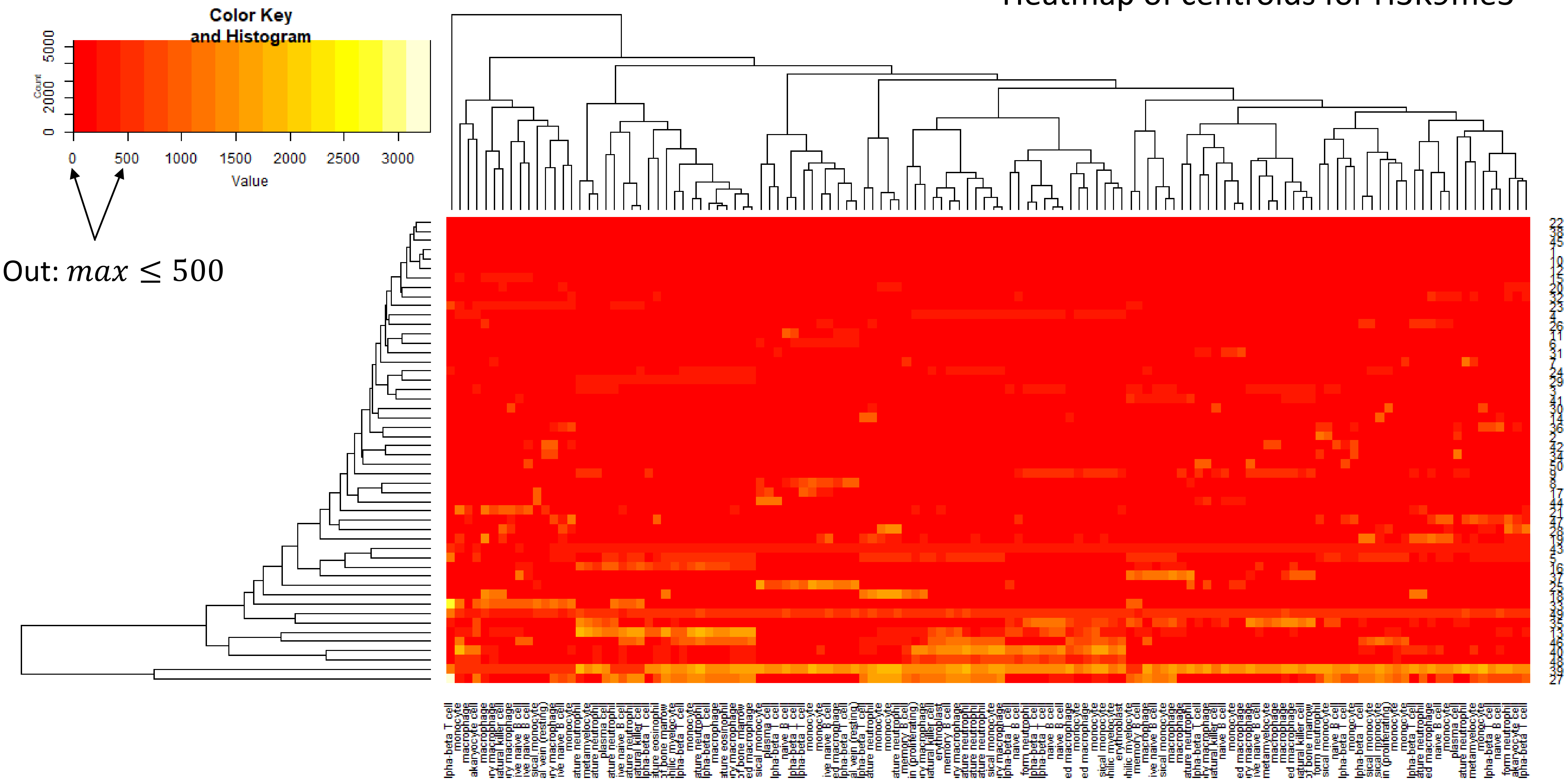
$n = \#samples$
$m = \#genes$

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \xrightarrow{\text{average of samples from the same cell type}} \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{24,1} & \cdots & x_{24,m} \end{bmatrix}$$

Elimination of «flat» genes using k-means clustering on genes profiles

Construction of **6 matrices**, by averaging the profiles of samples of the same cell type (dimension $24 \times m$)

Data cleaning: an example



Similarity network analysis

- **Similarity Network Fusion**[1] is a tool that has the aim of aggregating multiple types of information collected on the same set of experimental units.

$$\begin{tikzcd}
M_1 = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m_1} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m_1} \end{bmatrix} \arrow[dr] & M_2 = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m_2} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m_2} \end{bmatrix} \arrow[d] & \dots & M_l = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m_l} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m_l} \end{bmatrix} \arrow[dll] \\
 & \text{SNF} \arrow[d] & & \\
 & \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,n} \end{bmatrix} & &
\end{tikzcd}$$

[1] Wang, Bo & Mezlini, Aziz & Demir, Feyyaz & Fiume, Marc & Tu, Z. & Brudno, Michael & Haibe-Kains, Benjamin & Goldenberg, Anna. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*. 11. 10.1038/nmeth.2810.

SNF

- For each count matrix, a **similarity matrix**, based on a *scaled exponential similarity kernel*, is constructed.

- The six matrices are fused through a **Cross Diffusion Process (CrDP)**.

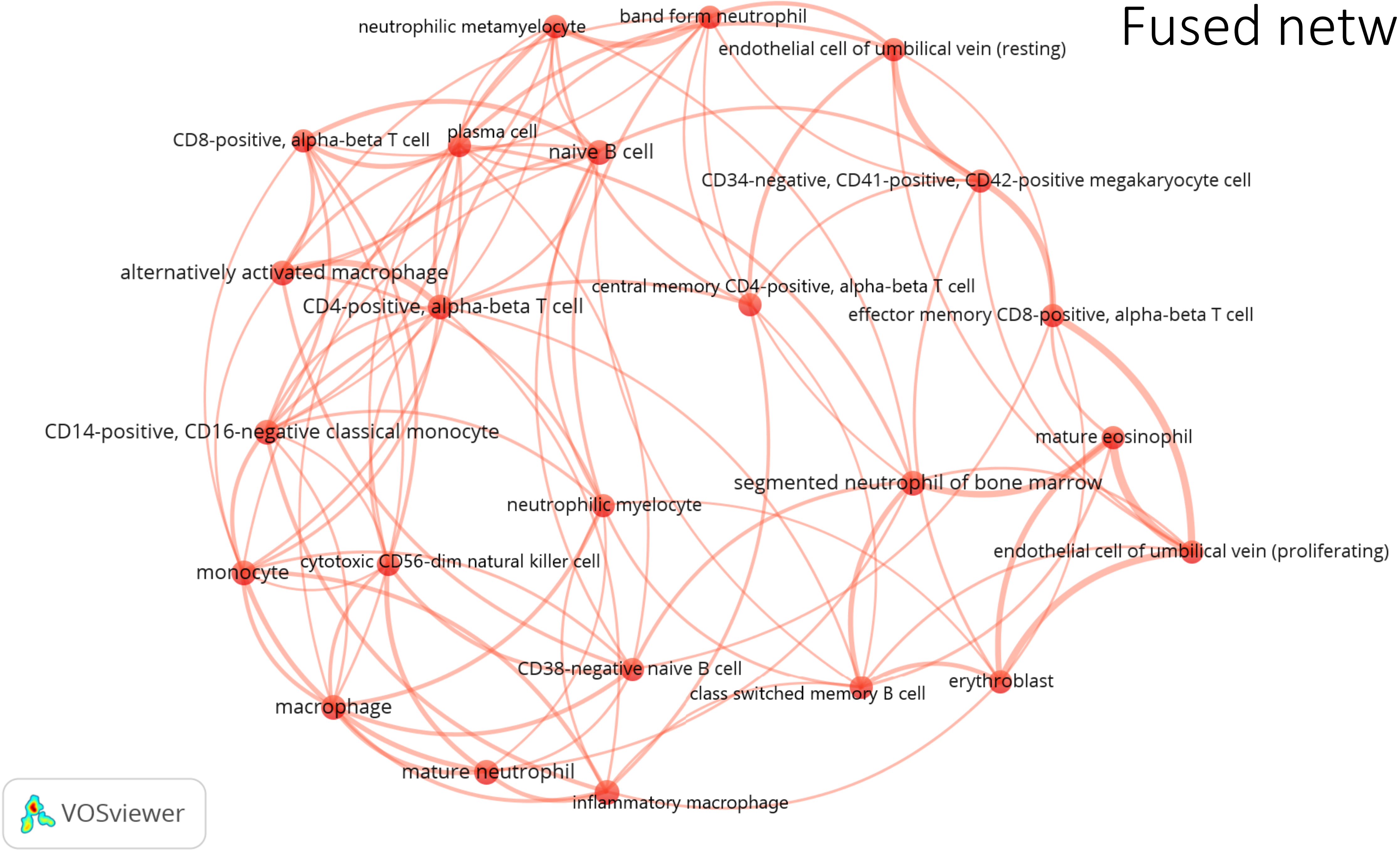
General updating rule for the fusion of m networks:

$$P_{t+1}^{(\nu)} = S^{(\nu)} \times \left(\frac{\sum_{k \neq \nu} {P_t}^{(k)}}{m-1} \right) \times \left(S^{(\nu)}\right)^T$$

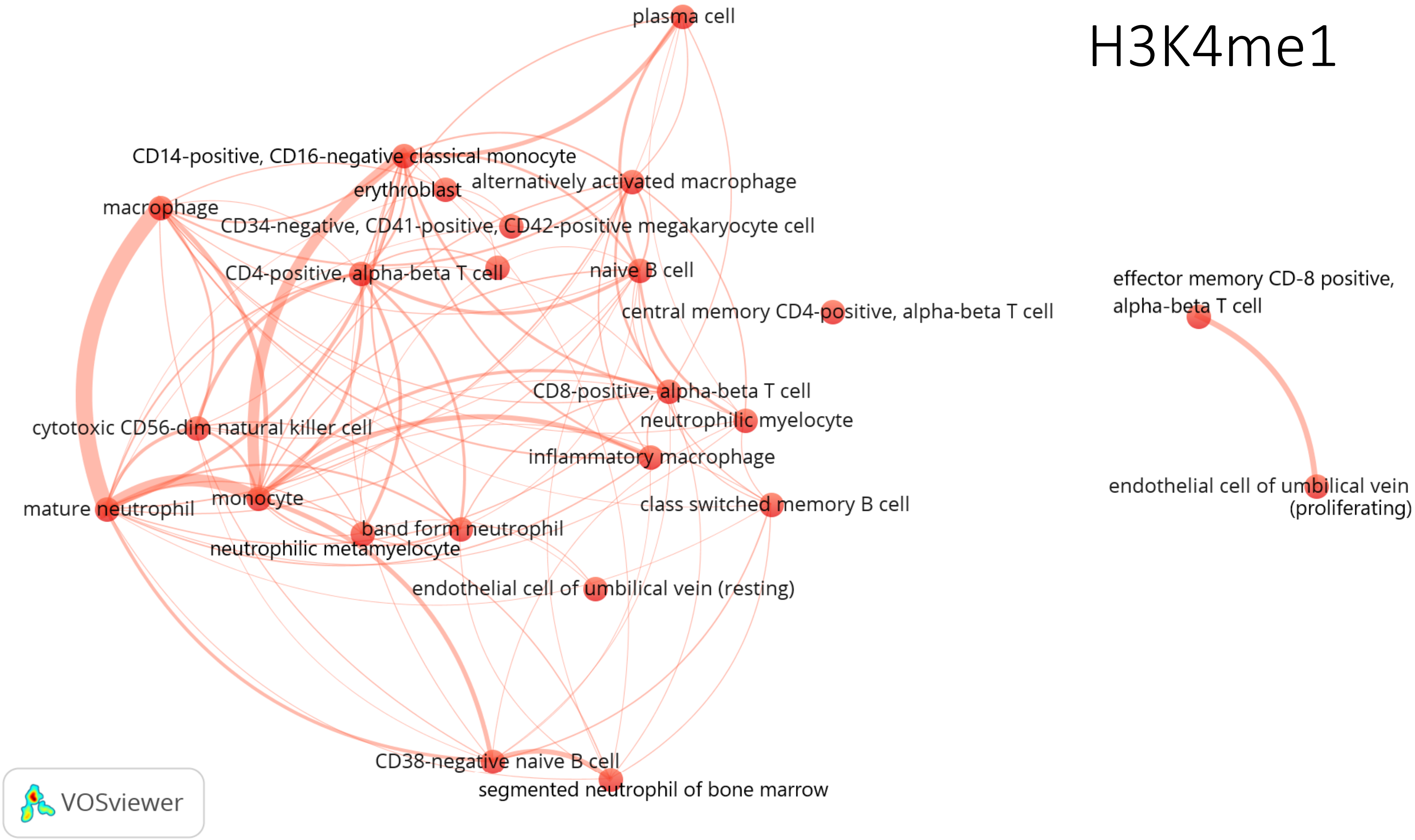
S → *local affinity matrix*

P → *status matrix*

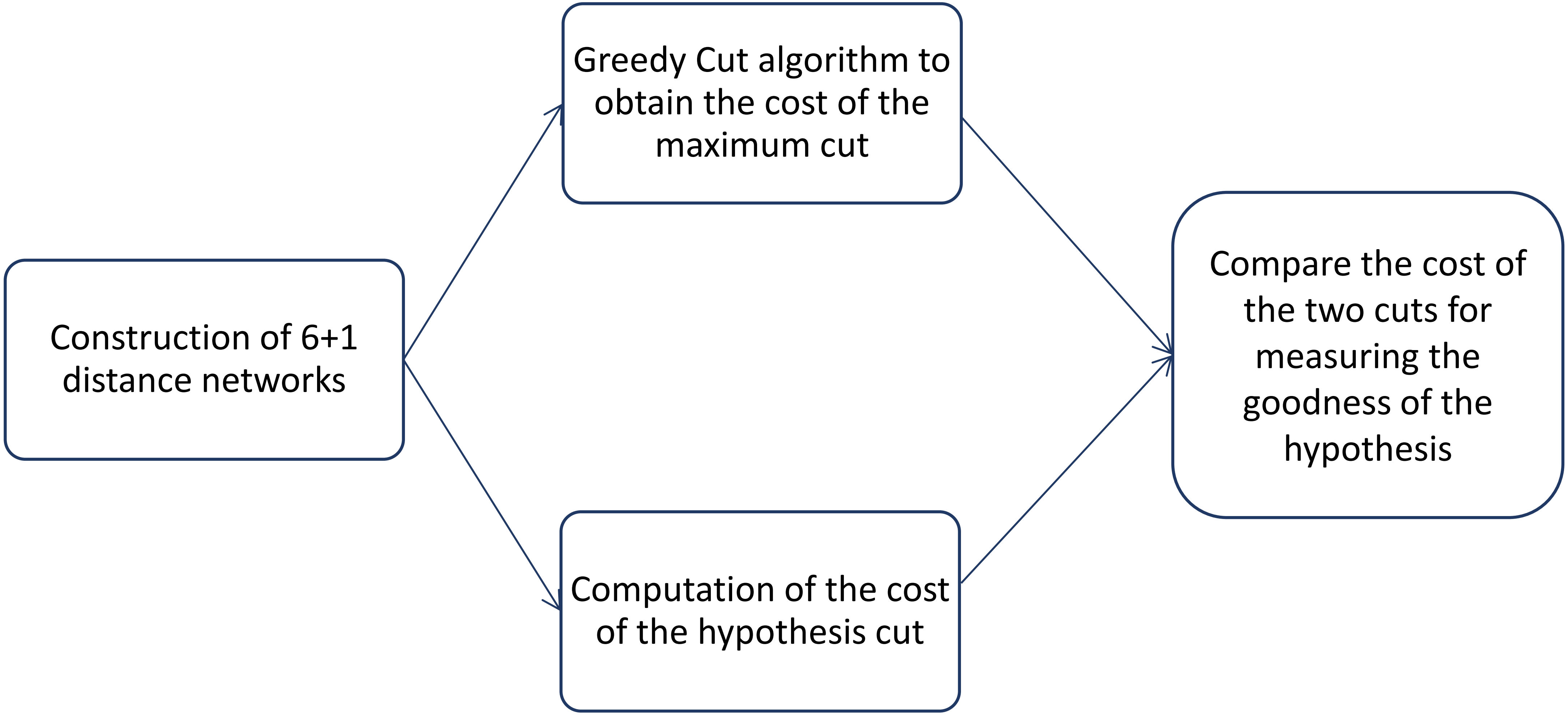
Fused network


neutrophilic metamyelocyte
band form neutrophil
endothelial cell of umbilical vein (resting)
CD8-positive, alpha-beta T cell
plasma cell
naive B cell
CD34-negative, CD41-positive, CD42-positive megakaryocyte cell
alternatively activated macrophage
central memory CD4-positive, alpha-beta T cell
CD4-positive, alpha-beta T cell
effector memory CD8-positive, alpha-beta T cell
CD14-positive, CD16-negative classical monocyte
mature eosinophil
segmented neutrophil of bone marrow
neutrophilic myelocyte
endothelial cell of umbilical vein (proliferating)
cytotoxic CD56-dim natural killer cell
monocyte
CD38-negative naive B cell
erythroblast
class switched memory B cell
macrophage
mature neutrophil
inflammatory macrophage
VOSviewer

H3K4me1


plasma cell
CD14-positive, CD16-negative classical monocyte
erythroblast
alternatively activated macrophage
macrophage
CD34-negative, CD41-positive, CD42-positive megakaryocyte cell
CD4-positive, alpha-beta T cell
naive B cell
effector memory CD-8 positive, alpha-beta T cell
central memory CD4-positive, alpha-beta T cell
CD8-positive, alpha-beta T cell
neutrophilic myelocyte
cytotoxic CD56-dim natural killer cell
inflammatory macrophage
endothelial cell of umbilical vein (proliferating)
mature neutrophil
monocyte
class switched memory B cell
band form neutrophil
neutrophilic metamyelocyte
endothelial cell of umbilical vein (resting)
CD38-negative naive B cell
segmented neutrophil of bone marrow
VOSviewer

Hypothesis testing: outline



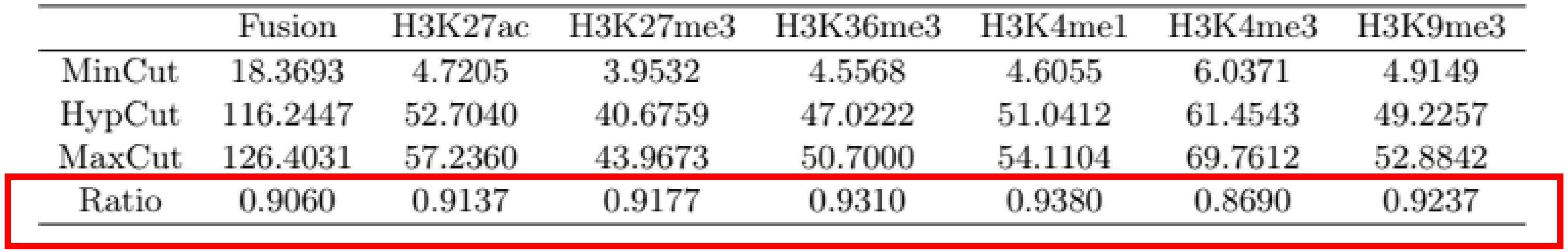
Construction of 6+1 distance networks

Greedy Cut algorithm to obtain the cost of the maximum cut

Computation of the cost of the hypothesis cut

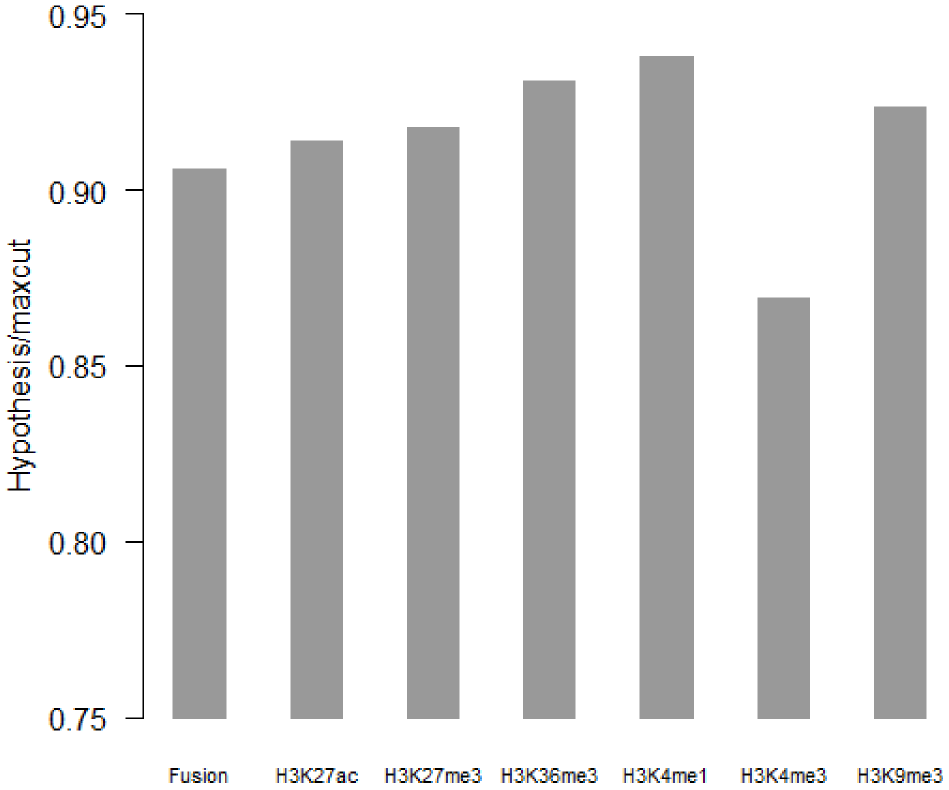
Compare the cost of the two cuts for measuring the goodness of the hypothesis

Results



	Fusion	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3
MinCut	18.3693	4.7205	3.9532	4.5568	4.6055	6.0371	4.9149
HypCut	116.2447	52.7040	40.6759	47.0222	51.0412	61.4543	49.2257
MaxCut	126.4031	57.2360	43.9673	50.7000	54.1104	69.7612	52.8842
Ratio	0.9060	0.9137	0.9177	0.9310	0.9380	0.8690	0.9237

$$ratio = \frac{cost\ of\ the\ hypothesis - mincut}{cost\ of\ the\ max\ cut\ - mincut}$$



Conclusions

- Histone modifications may have a role in the haematopoietic cell differentiation process.
- **SNF + hypothesis testing** strongly supports the hypothesis of differentiation into the myeloid and lymphoid lineages…
- …but the similarity analysis suggests that a hybrid model could be more appropriate at higher differentiation level.

Further work

- Testing different hypotheses on haematopoiesis.
- Application of the model to network of diseased cells, and possible individuation of anomalies related to pathologies.

References

Wang, Bo & Mezlini, Aziz & Demir, Feyyaz & Fiume, Marc & Tu, Z. & Brudno, Michael & Haibe-Kains, Benjamin & Goldenberg, Anna. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*. 11. 10.1038/nmeth.2810.

Bo Wang, Jiayan Jiang, Wei Wang, Zhi-Hua Zhou, and Z Tu. Unsupervised metric fusion by cross diffusion. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2997–3004, 06 2012.

Vikas Bansal and Vineet Bafna. Hapcut: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* (Oxford, England), 24:i153–9, 09 2008.

Palshikar, Girish. Simple algorithms for peak detection in time-series. (2009). *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*. Vol. 122.

Xhemalce, B., Dawson, M. A., & Bannister, A. J. (2006). Histone modifications. *Reviews in Cell Biology and Molecular Medicine*.