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
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

7 graphs with 24 vertices

GRAPH CUT FOR HYPOTHESIS TESTING

DATA DIMENSIONALITY REDUCTION
Data collection and organization-1

# of cellular types : 24
# lymphoid: 11
# myeloid: 13

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

1Source 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

  - 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 \( M \), \( M_{ij} = \text{number of peaks of sample } i \text{ in gene } j \).

Data cleaning and construction of cell type matrices

\[ n = \#\text{samples} \]
\[ m = \#\text{genes} \]

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

\[
\begin{bmatrix}
  x_{1,1} & \cdots & x_{1,m} \\
  \vdots & \ddots & \vdots \\
  x_{n,1} & \cdots & x_{n,m}
\end{bmatrix}
\]

average of samples from the same cell type

\[
\begin{bmatrix}
  x_{1,1} & \cdots & x_{1,m} \\
  \vdots & \ddots & \vdots \\
  x_{24,1} & \cdots & x_{24,m}
\end{bmatrix}
\]

Construction of 6 matrices, by averaging the profiles of samples of the same cell type (dimension \(24 \times m\))
Data cleaning: an example

Out: $\max \leq 500$

Heatmap of centroids for H3K9me3
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.

\[
M_1 = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,m_1} \\
\vdots & \ddots & \vdots \\
x_{n,1} & \cdots & x_{n,m_1}
\end{bmatrix} \quad M_2 = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,m_2} \\
\vdots & \ddots & \vdots \\
x_{n,1} & \cdots & x_{n,m_2}
\end{bmatrix} \quad \ldots \quad M_l = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,m_l} \\
\vdots & \ddots & \vdots \\
x_{n,1} & \cdots & x_{n,m_l}
\end{bmatrix}
\]

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:

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

$S \rightarrow$ local affinity matrix

$P \rightarrow$ status matrix
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

<table>
<thead>
<tr>
<th></th>
<th>Fusion</th>
<th>H3K27ac</th>
<th>H3K27me3</th>
<th>H3K36me3</th>
<th>H3K4me1</th>
<th>H3K4me3</th>
<th>H3K9me3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HypCut</td>
<td>116.2447</td>
<td>52.7040</td>
<td>40.6759</td>
<td>47.0222</td>
<td>51.0412</td>
<td>61.4543</td>
<td>49.2257</td>
</tr>
<tr>
<td>MaxCut</td>
<td>126.4031</td>
<td>57.2360</td>
<td>43.9673</td>
<td>50.7000</td>
<td>54.1104</td>
<td>69.7612</td>
<td>52.8842</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.9060</td>
<td>0.9137</td>
<td>0.9177</td>
<td>0.9310</td>
<td>0.9380</td>
<td>0.8690</td>
<td>0.9237</td>
</tr>
</tbody>
</table>

\[
\text{ratio} = \frac{\text{cost of the hypothesis} - \text{mincut}}{\text{cost of the max cut} - \text{mincut}}
\]
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


