Meet-EU 2021/2022
Genomic Compartment Detection

Oktawia Scibior
Maxime Gueudré
Mikal Daou

Table of contents:

I. Introduction

II. Materials and Method

III. Results

- **i. Detection of chromatin compartments**
- **ii. Detection of sub-compartments**

IV. Discussion

Detection of chromatin compartments

1. Introduction

A fundamental aspect of DNA nucleus organization is the fact that chromatin is made of loops, bringing close DNA sequences far from each other. These loops are thought to enable distal gene regulation. A number of methods have been developed to study these loops locus by locus but here we use results from a study by Rao et al[1] where they use HI-C, a method for genome-wide mapping of chromatin contacts. Rao et al showed that the genome contains 2 compartments, one enriched in open chromatin and one enriched in closed chromatin and 5 subcompartments each defined by intrinsic epigenetic marks. These results come along with other works attempting to understand 3D genome organization and its implication in gene regulation. Here, we focus on the compartmented aspect of the genome by studying gene expression levels and epigenetic characteristics of each compartment. More precisely, we first detect compartments using gene expression correlation matrices based upon the mapping of chromatin contacts, then we try to refine this result and detect sub-compartments based on epigenetic marks.

2. Materials and methods

Pre-processing of contact matrices

Contact map pre-processing included binning, filtering, iterative, Sequential Component Normalization (SCN), observed/expected normalization and Pearson correlation test. Binning procedure allowed to organize chromosome contacts into 100kb and 25 kb bins. The normalization was done by dividing each value of the contact matrix by the mean of the values located on the diagonal. The Pearson correlation test was applied to each pair of bins (correlation between column and raw). We thus obtained a correlation matrix that was used in the next steps.

Detection of the 2 main compartments

To detect 2 main chromosome compartments we calculated the first eigenvector of the correlation matrix. The sign of the eigenvector values indicates whether the corresponding location in the genome belongs to compartment A (open chromatin) or compartment B (closed chromatin). For the filtered correlation matrix, the obtained eigenvector values are positive for compartment A and negative for compartment B.

As the python linalg tool which we used for a calculation sometimes gives the results with flipped sign, we performed an additional gene density correlation test to assure ourselves that we found the correct configuration of compartments. (We calculated the correlation coefficient between gene density and obtained eigenvectors). The positive value suggested that the eigenvector isn't swapped.

Gene density analysis

We defined gene density as the number of genes per bin. The gene coverage per base pair was simply calculated by dividing gene density by resolution.

We analyzed the distribution of gene density to see if there were associations with either the distribution of compartment A or the distribution of compartment B. We calculated confusion tables between vector of compartments locations detected via "eigenvector" system and the compartments suggested by gene density. Obviously because all values are bigger or equal to 0, we cannot find compartment A and B based on the value sign. We used as a criterium the mean of gene density computed over all bins. Every bin above the mean corresponds to compartment A, those below the mean correspond to compartment B.

Detection of higher number of compartments.

The problem of efficient detection of more than two compartments is still open. The basic questions which can be asked in this context are:

- what is the optimal number of those potential compartments
- what is the hierarchy between them ? (Are they clearly located in the border on two main compartments A and B or they are at the equal level of organization hierarchy)

In [1,2] the idea of subcompartments is clearly stated. We propose several approaches to try to answer the questions above. Some of the methods needed prior assumption that chromatin subcompartments lie within the separate compartments A and B. We also made use of information about epigenetic marks intensity at gene locations. Below we describe our approaches:

1. Visual approach

- 3D visualization of chromosomes with 2 main compartments marked as well as selected epigenetics changes marked, using PyMol (spectrum/b-factor color setting).
- Epigenetic mark density visualization in bar plot form.

2. Hidden Markov Model

We have fitted hidden markov models on the correlation matrix , where the most possible sequence of hidden states signified susceptible compartments sequence . To be able to fit HMM, the initial information about the number of hidden states was needed. Therefore, we ran a few experiments with the number of states ranging from 3 to 10. To define the optimal number of compartments , we used the Akaike information criterion (AIC)(1) and the Bayes information criterion (BIC)(2) as well as silhouette metrics which is widely used in clustering optimisation[3].

$$AIC=2k - 2\ln(L) \quad (1)$$

$$BIC= k \ln(n) -2\ln(L) \quad (2)$$

where k is the number of estimated parameters, n is the nr of observations and L is the maximum value of the likelihood function of the model.

$$\text{Silhouette Score}=(b-a)/\max(a,b) \quad (3)$$

where a _averange intra-cluster distance i.e the average distance between each point within a, b-average inter-cluster distance i.e the average distance between all clusters.

3. Unsupervised learning clustering algorithms.

Following [1] we tested K-Means and Hierarchical Clustering methods on correlation matrix to find the possible cluster/ compartments distribution . We also tested Spectral Clustering at the mathematical base of this method there is a calculation of relevant eigenvectors and clustering attribution consensus agreement. To find the optimal nr of clusters we used silhouette score.

4. p-value based approach

In this approach we made the assumption that subcompartments are strictly attributed to one of two main A, B compartments. We followed the procedure below:

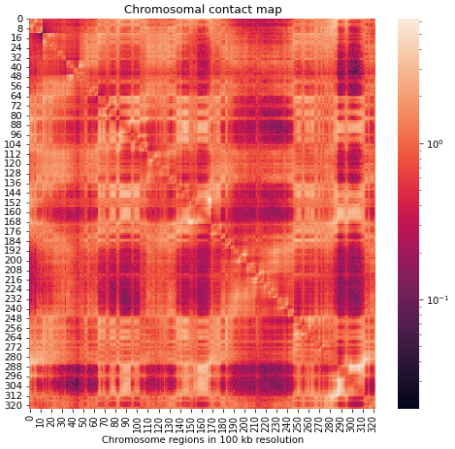
- ➔ Building two separate correlation matrices corresponding to active/inactive regions found in (see Materials and Methods: Detection of 2 compartments)
- ➔ HMM subcompartments determination for each separate correlation matrix.
- ➔ Welch test to determine **the significant difference (**with α**= 0.05)** in epigenetic marks density for each sub compartment .

We also calculated Pearson correlation coefficient between each epigenetic mark and each subcompartment (based on eigenvector values) without assumption of hierarchical organization of subcompartments.

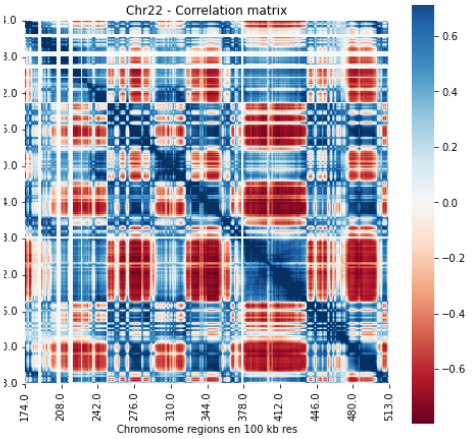
Results

HI-C uses DNA crosslinking, meaning that it links DNA sequences close to each other in the genome. The more often two loci are linked together, the closer they are to each other. The results of A HI-C protocol can then be used to build a contact matrix indicating, for each entry ij, the number of times locus i was found in contact with locus j. To identify genomic compartments, we followed the following method: normalization (see method) of the contact matrix in order to limit the impact of the distance between two loci on the data; the computation of a correlation matrix computed on the contact map from which we extracted the first eigenvalues. The correlation matrix tells us, for each pair of loci, if the variation in their gene expression levels are dependent, thus if the loci are in the same compartment or not. The sign (negative for closed chromatin and positive for open chromatin) of the eigenvectors indicates whether a locus was in an open chromatin enriched compartment or a closed chromatin enriched compartment. On the filtered (see materials and methods: pre-processing of the contact matrices) intra-chromosomal contact matrices (figure 1A), we can see regions of enhanced contacts on the diagonal (regions close to each other) and regions of few contacts. Correlation matrices (figure 1B and 1C) show clearly the relationship between proximity of genes and co-expression. Correlation matrices also show that compartments are not conserved across chromosomes.

A.



B.



C.

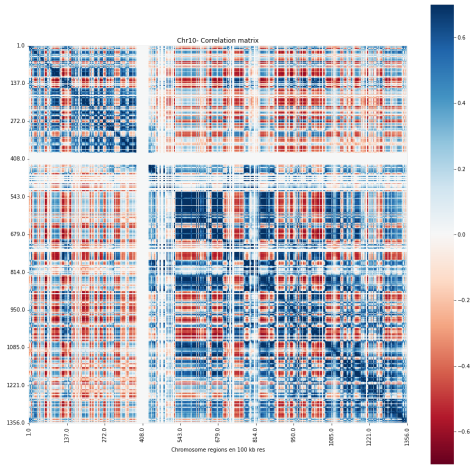


Figure 1. A. Chromosome 22 intra-chromosomal contact matrix resolution 100 kb. **B.** Chromosome 22 intra-chromosomal correlation matrix, resolution 100 kb. **C.** Chromosome 10 intra-chromosomal correlation matrix, resolution 100 kb with centromere region visible.

By calculating the eigenvectors (figure 2A) of the intra-chromosomal contact matrices, we were able to find 2 compartments. Our eigenvectors match the gold standard results. Positive eigenvector values indicate regions enriched in intra-chromosomal contacts and open chromatin (compartment A). Negative eigenvector values indicate regions with few intra-chromosomal contacts and are enriched in closed chromatin (compartment B). Furthermore, that the distribution of the number of genes sequenced for each bin (fig. 2B) seems to match that of the first eigenvector of each bin, suggesting that compartment A is richer in genes than compartment B. It is worth noting that gene annotation depends on the sequencing quality of each gene which in turn depends on the base pair covering of each gene. Our gene annotation results may contain bias due to that.

A. **B.**

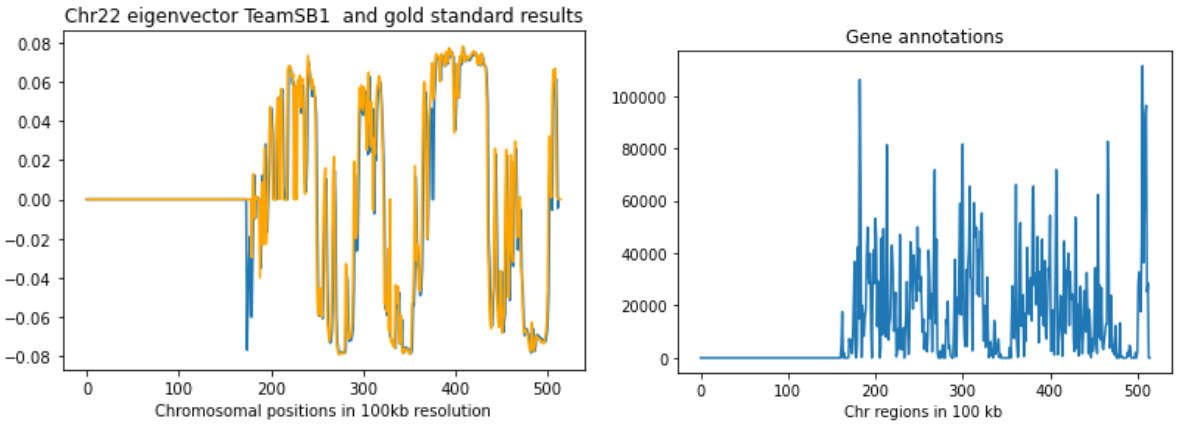


Figure 2. A. Comparison of our team's first eigenvectors for the correlation matrix of chromosome 22, 100kb resolution (blue), with the gold standard results (yellow). Positive values (red) indicate compartment 1 and negative values (blue), compartment B. **B.** Gene annotations for chromosome 22, 100 kb resolution.

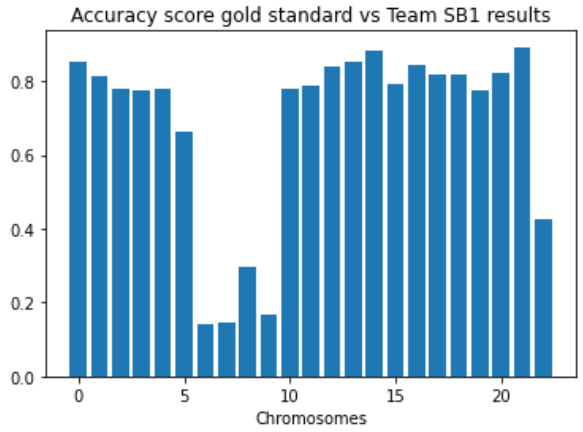


Fig. 3 Accuracy score between gold standard results and our team's results.

The visible low accuracy area is very likely due to the partial inversion of the eigenvector calculated. We noticed that in some cases only the part of the eigenvector values had flipped sign. Those parts corresponded to longer or shorter branches of the chromosome. We could notice it only by comparison with the gold standard. We didn't use any automatic tool to correct it and we kept the gene density test we mentioned above as an indicator. Obviously the gene density test works well in case of full inversion of an eigenvector but not a partial one.

Detection of more than 2 compartments

As we were sure about the correctness of the 2 main compartments detection, we tried a few methods to detect more than 2 compartments based on chromosome 22 contact maps of human cell GM12878. Even though a visual approach is not an exhaustive method and rather gives an intuition, it's still useful for observing some obvious dependencies. Fig.4 shows we can clearly state that epigenetic marks 1, 2 and 5 manifest themselves in the non active regions. We confirmed this for epigenetic marks 1 and 2 by 3D visualization (Fig. 5), where higher intensity of these marks clearly falls into the inactive region. Pymol visualizations also show the variances of intensity of epigenetic marks giving some confirmation that subcompartments are attributed to particular compartments and aren't located in between (are on the same level of hierarchy). Furthermore, the heterochromatin region is more susceptible to present intricate spatial organization.

We next attempted to determine a strict number and the localization of subcompartments using experiments with machine learning methods. To do so, we first fitted a Gaussian Hidden Markov Model to the correlation matrix of GM12878 chr 22, taking as a parameter the number of hidden states from 2 to 10. We calculated AIC , BIC and the silhouette score. AIC and BIC criterions did not point towards an optimal number of clusters but indicated optimal values ranging from 5 to 8 (both included). The silhouette score showed that the optimal number of hidden states is 2. (Fig 6.b) The same silhouette score calculated for clustering methods (KMeans, Hierarchical , Spectral) also indicated 2 as the optimal number of clusters. These results are contradictory to those of AIC and BIC. However, silhouette score measures consistency inside the cluster and separation from the others clusters. The obtained result may prove again the hierarchical organization of compartments and subcompartments. Further visualization of cluster distributions obtained by the methods mentioned above (fig.8) suggests that "new appearing" clusters have a tendency to appear within the same already defined clusters. This is coherent with our previous analysis. However the question of the optimal number of clusters remains open.

Next, we calculated Welch-test p-values to determine if there is some significant difference between mean epigenetic marks intensities of each subcompartment. (fig.7a.) For compartments A and B, we ran independent subcompartment detection using HMM and clustering techniques (2 subcompartments for each region). We noticed some significant differences for regions 3,4,7 of compartment B. This proves that epigenetic changes are a great criterion to analyze compartments. We repeated the same experiment with 3 subcompartments (Fig.7b). Even though fig.7B shows more significant cases, the Bonferroni correction needs to be applied and it reduces the number of significant cases to 1. The next step would be to increase the number of subcompartments within each region and observe up to which value significant differences will remain.
A similar approach was used on calculation of the correlation coefficient between epigenetic marks and subcompartments (more precise eigenvector values corresponding to particular regions)(fig.9). We didn't make any assumptions about hierarchisation of compartments. However this method could be also used in the opposite case.

Discussion:

The protocol of detection of two main chromatin regions is already well defined. Despite technical difficulties, we obtained expected results. We found 2 nuclear compartments, one enriched in open chromatin (compartment A) and one enriched in poor chromatin (B). We also found that compartment A has more genes than compartment B, although that last result needs to be put in perspective with the gene density of each compartment. The comparison of our results showed a generally good accuracy despite some flaws in compartment detection in longer or shorter branches of chromosomes, even though we were able to correct them after detection.

As for the detection of subcompartments, we tested several approaches. [1] presents results of subcompartement determination when the optimal /correct number of subcompartments is stated as 5 and 6 (for chr 19). This number could be an indication for us however it was found by analysis of intra and inter-chromosomal contacts while our analysis was based only on intra-chromosomal contacts. Additionally, the combination of both types of analysis provides complementary information. For instance, it would tell us if the compartments detected on the basis of intra and inter-chromosomal contacts are in the same location on a given chromosome, in which case, gene activity (compartment A) of a given chromosome could be dependent of the chromosome's interchromosomal regulations. Moreover, here we only analyzed results based on 100kb resolution. To pursue our analysis, considering different resolutions would be relevant to see how compartment delimitation varies with respect to resolution. A drawback of our report is that most of our results are based on measures coming from the same cell (GM12878), as our work mainly has an illustrative purpose.

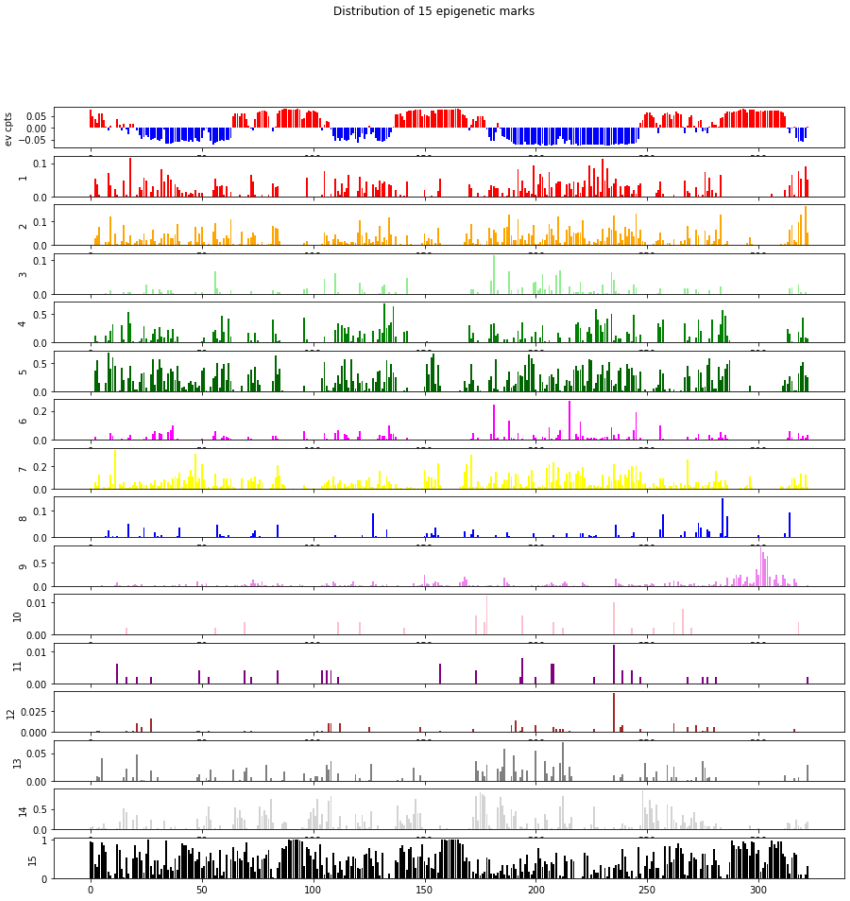


Fig. 4. Visualization of the distribution of 15 epigenetic marks (starting from the 2nd barplot) and the first eigenvector values of correlation matrix for GM12878 chr 22 res. 100kb (first barplot).

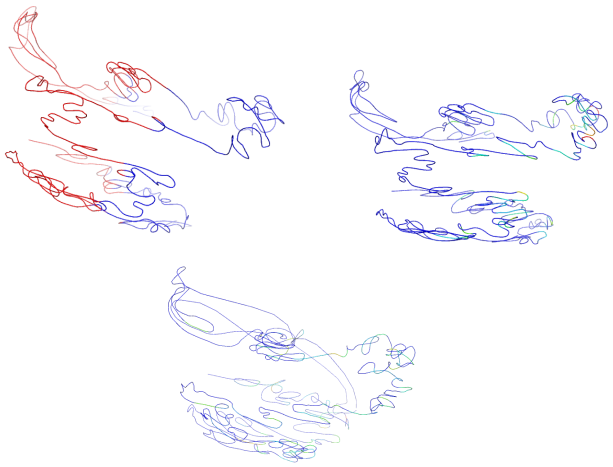


Fig. 5. (left) 3D visualization of chr 22 with two main compartments (red: active, blue: non-active) marked as well as epigenetic mark 1 (right) and epigenetic mark 2 (bottom).

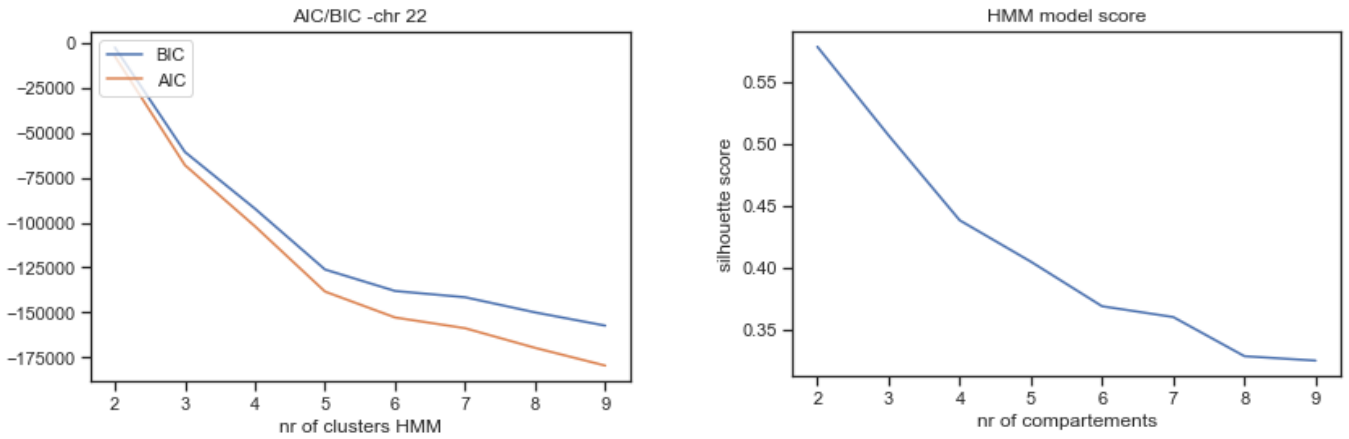


Fig. 6. A. (left) AIC/BIC curves as a function of the number of states for chr 22 _res. 100kb, **B.** Silhouette score as a function of the number of compartments (right).

A. **B.**

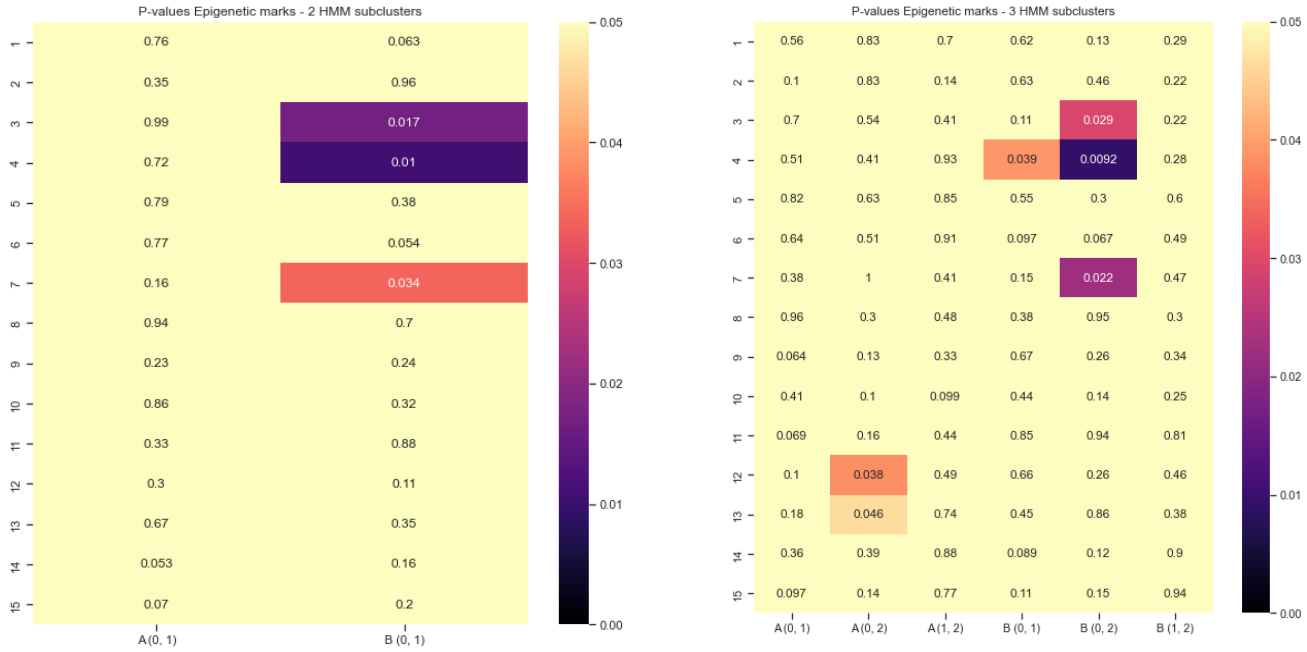


Fig. 7. A, B Table of significance, null hypothesis - epigenetic marks means in two subcompartments included in compartement A and B ,chr 22_res. 100kb , for 2 and 3 subcompartments detection.

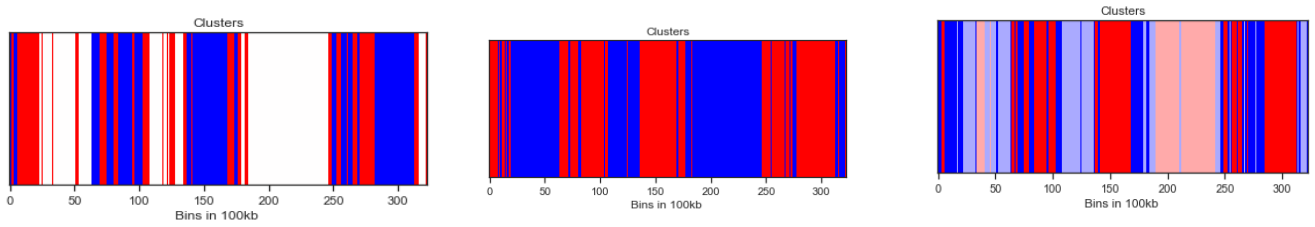


Fig. 8. HMM Subcompartments location for states 2 (top left), 3 (right) and 4 (bottom left).

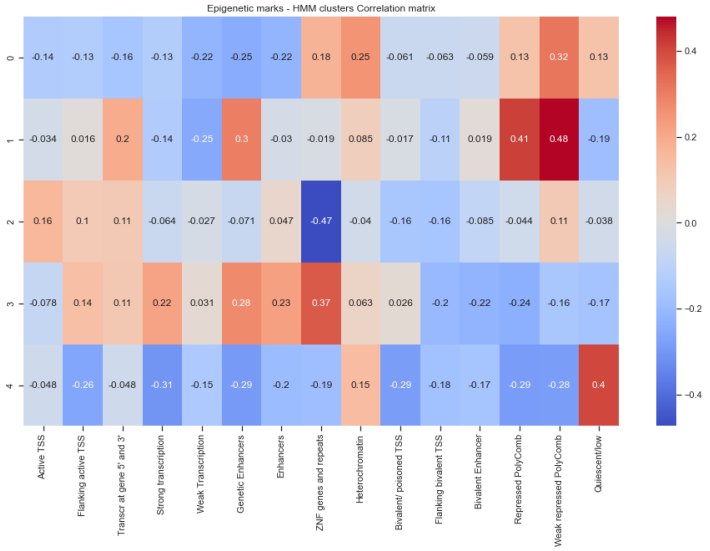


Fig. 9. Table of correlation coefficient between each epigenetic mark and each of the 5 subcompartments detected in spectral clustring.

Bibliography:

[1] *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin looping.* Rao et al.
Published:December 112014DOI:https://doi.org/10.1016/j.cell.2014.11.021

[2] *Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes*
Y.Liu, published 10 May 2021 , Nature Communications

[3]https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/]

[4]https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced