TÍTULO

CLASSIFICATION OF OPEN CHROMATIN REGIONS BASED ON ATAC-SEQ SIGNAL TOPOGRAPHY

AUTOR

Alberto Ríos Muñoz

Universidad Internacional de Andalucía, 2023

Máster Universitario en Análisis de Datos Ómicos
y Biología de Sistemas

Universidad de Sevilla
Universidad Internacional de Andalucía

Classification of open chromatin regions based on ATAC-seq signal topography

Alberto Ríos Muñoz[1], José Carlos Reyes[2,*], José Antonio Guerrero-Martínez[2,*]

[1]Faculty of Biology, University of Seville, C/Profesor García González 1, 41012 Seville, Spain

[2]Centro Andaluz de Biología Molecular y Medicina Regenerativa-CABIMER, Consejo Superior de Investigaciones Científicas-Universidad de Sevilla-Universidad Pablo de Olavide (CSIC-USE-UPO), Avenida Americo Vespucio 24, 41092, Seville, Spain

***Correspondence:**

José Antonio Guerrero-Martínez
jose.guerrero@cabimer.es

Jose Carlos Reyes
jose.reyes@cabimer.es

Keywords: ATAC-seq analysis, Peak shape clustering, Predictive model.

Abstract

Chromatin accessibility is key for the regulation of DNA expression and studies about it help to map the different transcriptional landscapes the cell can have under determined circumstances. The ATAC-seq assay employs the transposase Tn5 to identify regions of accessible chromatin, however, the proper software or pipelines to analyze the ATAC-seq data are still very scarce. Here we show that peak-shape based clustering and analysis developed for ChIP-seq data is also valid for ATAC-seq datasets. Our study provided information about how clusters showed different distribution of promoter and enhancer regions as well as distinctive signatures of histone marks and transcription factors associated to motifs. We also developed a prediction model to specify how peak shape can be useful for determining DNA elements' nature. These results show how peak shape provides useful information about the chromatinic state of the genes and reveal interesting biological insights about transcription regulation and up-regulated biological processes. This study can be the starting point for more ATAC-seq analysis studied in different cell lines, phases of the cell or pathologic circumstances in order to provide a general overview of the accessible chromatin regions, the transcriptional state of the cell and the epigenetic marks of DNA.

1 Introduction

The different levels of DNA compaction play a key role in the organization of DNA in the nucleus and allow fine regulation of gene expression. Depending on the degree of compaction, two types of chromatin are distinguished: heterochromatin, regions with transcriptionally inactive genes; and euchromatin, regions of less compact chromatin where gene expression does take place.

Of the levels of DNA organization, the nucleosome is the basic fundamental unit of compaction, consisting of the DNA strand coiled around a histone octamer made up of two copies of each of the four histone types (H2A, H2B, H3 and H4) (Albert et al., 2002). Regions of open chromatin show a lower nucleosome density or even nucleosome-free regions. In addition, active chromatin is marked by a specific combination of posttranslational modifications of core histone proteins (H3K27ac, H3K4me1 and H3K4me3) and by the presence of histone variants like H2A.Z and H3.3. In contrast, transcriptionally repressed genes are often organized within 'closed' chromatin domains, marked by different histone modifications (e.g. H3K27me3 and H3K9me3) (Siggens et al,. 2014). The degree of variability in DNA and histone modifications observed between cell types is different for different genomic elements, such as promoters, enhancers, or insulators. Therefore, different cell types, as well as the signalling cascades that produce differentiated cells, are characterized by a unique chromatin signature. The distribution of these regions varies according to cell stage and cell type, giving the cell a specific phenotype that can be used as a fingerprint to characterize a cell at a given time, as well as the effects of extracellular signals and to compare the response to stress, pathological and physiological states.

Transposase-accessible chromatin assay by sequencing (ATAC-seq) is an alternative or complementary technique to MNase-seq, DNase-seq and FAIRE-seq for assaying chromatin accessibility. The results obtained from ATAC-seq are similar to those of DNase-seq and FAIRE-seq, however, ATAC-seq is gaining popularity because it does not require cross-linking, has a higher signal-to-noise ratio, requires a much smaller amount of biological material and is faster and easier to perform compared to other techniques (Yan et al., 2020). This technique involves the Tn5 hyper-reactive transposase to cut and integrate the adapters in the regions of accessible chromatin, while the chromatin in closed conformation will present steric hindrance making the insertion less probable. Therefore, amplifiable DNA fragments suitable for high-throughput sequencing are preferentially generated at locations of open chromatin (Buenrostro et al., 2013), such as promoters or enhancers. After processing ATAC-seq data, aligned reads return a pattern of peaks that represent the active regulatory regions in the genome. Though, ATAC-seq data have been widely used to identify regulatory elements, the analysis of the morphology of these peaks was poorly studied. ATAC-seq peaks in the same region of the genome, have been shown to vary in certain characteristics such as width, intensity, or number of peaks in different cell types or under different conditions. This suggests that analysis of the shape of the peaks can provide relevant information about the cellular regulome.

There are only two proper software that have been previously used for peak characterization based on the shape of ChIP-seq data. Those are SIC-ChIP (Shape Index Clustering for ChIP-seq peaks) (Cremona et al., 2015) and Fun-Chip. Given the scarce literature for the characterization of DNA regions from ATAC-seq studies based on peak morphology, in this study we have carried out the classification, analysis and annotation of the peaks generated by ATAC-seq performed in normal murine mammary gland epithelial cells (NMuMG) treated with TGFβ (Guerrero-Martinez et al., 2020), using the SIC-ChIP approach.

Topography is the science that studies the land shapes and forms of the surface, involves the recording of relief, the identification of specific landforms. This is also known as geomorphometry. By analogy to discipline, we can say that, in this study we have carried out a topographic analysis of the ATAC-seq peaks. This study revealed that morphological differences in peaks when classified according to five indices (height, area, width at half maximum height, number of local peaks and M-index) result in relevant biological information concerning chromatin signature and transcription factor binding motifs. Moreover, we developed a novel logistic regression model able to discriminate between regulatory elements based on ATAC-seq morphology indices.

2 Materials and Methods

2.1 Computational methods and statistical analysis.

Most of analyses were performed using R (v4.2.1), RStudio (v2021.09.0+351) and Bioconductor (v3.15). Data preparation and other specific analysis were performed using proper software within Ubuntu (20.04). Ggplot2 package was used for graphical representation.

2.2 Data acquisition

ATAC-seq paired-end data of NMuMG cells after 2 hours of TGFβ treatment was obtained from the ENA (European Nucleotide Archive) database with accession number SRR10485876 (Guerrero-Martinez et al., 2020).

ChIP-seq data for H3K4me1, H3K4me3 and H3K27ac histones marks were obtained from GEO under accession number: GSM4174040, GSM4174046, GSM4174034, respectively. The data correspond to NMuMG cells after TGFβ treatment from the same study.

JASPAR2022 vertebrate database was used to obtain transcription factor binding motifs information.

The list of transcription start sites was obtained from UCSC KnownGene annotation for mm9 mouse reference genome

2.3 Preprocessing of reads

ATAC-seq ENCODE pipeline (https://github.com/kundajelab/atac_dnase_pipelines) was used for ATAC-seq data alignment and peak calling, including pre and postanalysis steps like denoising and trimming of primers as well as quality control and statistical methods such as IDR for well conserved peaks selection between replicates. Alignment was fitted using mm9 mouse reference assembly. This analysis results in 39432 peaks.

GenomicAlignments package from Bioconductor was employed to generate signal files for ATAC-seq data. Three different signal files were obtained from the same ATAC-seq alignment after file, filtering sequencing fragments according to its length: ATAC (include all sequencing fragments), Open (include fragments with less than 100 bp, which correspond with fragments associated with Nucleosome Depleted fragments) and MonoNuc (include fragments with sizes in the range [180, 240 bp], which correspond with mononucleosomal fragments).

2.4 Clustering of regions

The peaks obtained were classified according to five indices corresponding to morphological characteristics of the peak: height, area, maximum width, number of local peaks and the M index. The clustering based on these indices was done using the SIC-ChIP pipeline (https://github.com/marziacremona/SIC-ChIP). Parameters related to the peak's indices were the same as in the original study.

To classify those peaks according to the five indices, k-means algorithm was employed obtaining ten different clusters for each different input: ATAC, Open, MonoNuc and a clustering using the combined indices for Nucleosome Depleted and Mononucleosomal fragments (Open-MonoNuc).

2.5 Clustering characterisation

ComplexHeatmaps R package used for generating heatmaps of shared regions, the heatmaps were scaled to rows and columns for each type of data input.

Histone modifications profiles were generated from intensity matrices obtained using ComputeMatrix from deeptools Suite (https://deeptools.readthedocs.io/en/develop/content/tools/computeMatrix.html). The intensity matrix was instructed to start in the centre point and extended 3 kb upstream and downstream using a bin size of 10 bp.

2.6 Enrichment of motives.

Motif search and enrichment for each of the clusters was carried out using the MEME suite tool, FIMO (https://meme-suite.org/meme/doc/fimo.html). For the enrichment analyses, scrambleFasta.pl from HOMER (http://homer.ucsd.edu/homer/motif/fasta.html) was used for obtaining 5x background sequences for each category.

Enrichment in each cluster was calculated as shown in the formula down below. Subsequently, a Fisher's test was performed to assign an enrichment *p*-value to each motif. In order to take the most significant motifs from each cluster, those with an enrichment greater than 1.5 and a *p*-value less than 0.05 were filtered out.

$$Enrichment = \frac{\frac{Motifs\ found\ in\ cluster\ i}{Motifs\ found\ in\ total}}{\frac{Motifs_{BG}{}^{*} found\ in\ cluster\ i}{Motifs_{BG}\, found\ in\ total}}$$

*Motifs$_{BG}$: Significative motives that appeared in the background sequences

2.7 Ontological analysis of motives

Clusters' genes associated to regions were classified according to their involvement in different cellular processes. This motives annotation was performed with the GREAT tool (https://github.com/jokergoo/rGREAT). The mode used was the basal plus. extension Further parameters for the analysis were adv_upstream = 50 kb, adv_downstream = 50 kb, adv_span = 1000 kb.

2.8 Prediction model based on binary logistic regression

A binary logistic prediction model to discriminate between regulatory elements was developed using a generalized linear model. For the construction of this model, the step validation and collinearity analysis were performed determining the 5 standardized indices were appropriate to use.

The model was subsequently validated both by the confusion matrix method and by ROC curves and AUC values, studying also the specificity and sensitivity.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives}$$

3 Results and Discussions

3.1 General workflow

The goal we tried to achieve is to characterize the peak regions based on their morphological indices. To achieve that, we followed the workflow illustrated in Figure 1, where from the raw data we got the .bed files filtered and cleaned which were used for clustering of the peaks using SIC-ChIP software. Once we got the clusters, to identify the DNA elements and draw more relevant biological information from them we first made sure the indices and clusters were correct, then we annotated the regions of each cluster to identify the significant DNA elements that were part of them, studying histone marks, most relevant motifs and biological processes involved. Finally, to achieve a better classification of promoter and enhancer regions we developed a prediction model based on the five indices.

3.2 Correlation between indices and identification of peaks

Previously, SIC-ChIP software had already proven to be valid for reliable clustering of ChIP-seq peaks according to their morphology (Cremona et al., 2015). In this work we studied if this software can be used to classify ATAC-seq peaks. Of the five indices computed in SIC-ChIP, two of them are related to peak morphological features and the other three to peak complexity, details of the indices are given below.

The height (h) and area (A) of the peak are related to its signal strength, i.e. the number of reads the region has. The maximum peak width ($w_{h/2}$) is a measure also related to the signal strength; this parameter consists of the width of the peak at half its maximum height.

The number of local peaks (p_{local}) is a parameter that must be smoothed to avoid an oversaturated signal of peaks, so it was estimated that to count the peak must be 50 nucleotides apart and that its difference with respect to the two contiguous local minima must be at least 20%.

The M-index (M/h) is a noise resistant and smoothed, measure of the complexity of the peak. The calculation of this index is based on counting the number of edges obtained by generating a tree of rooted nodes associated to a peak. This tree is built based on a depth function that each nucleotide X_i has with respect to its previous nucleotide (X_{i-1}). Three cases can occur, (1) when the function decreases, we move towards the root to the parent of the current node, (2) when the function increases, a new node is created and (3) when the function remains constant, nothing is done. To standardize this index was divided by the maximum peak height since the index M depends on this parameter.

To classify peaks according to their morphology, three different type of signal data was used as input: (1) data produced from all ATAC-seq sequencing fragments (ATAC), (2) from sequencing fragments with length less than 100 bp, which correspond with Nucleosome Depleted Regions (Open) and (3) from sequencing fragments with lengths in range [180-240] bp, that correspond with fragments associated with mononucleosomes (MonoNuc). Finally, indices computed from Open and MonoNuc signal files were combined to classify peaks according both indices together as the fourth type of input. Results from ATAC signal is shown as main results, while the analyses from other signals was shown as complementary analysis and will be shown as supplementary figures.

The indices studied showed varying degrees of correlation, being the most highly correlated area and height, as would intuitively be expected, while the rest of parameters showed a poorer correlation between them (Figure 2a), correlation that was consistent for the other types of input as well (Supplementary Figure 1). We also plotted the distribution of indices values for each cluster (Figure 2b), which revealed that the clusters showed differences in the means of each of the indices studied. The largest differences are seen in the number of local peaks, being the height the index that shows the least differences between clusters. The distribution found in the vast majority of the clusters is around the means, with short outlier tails except in specific cases such as cluster 2 of the Open data (Supplementary Figure 2b), which has a wider distribution with larger tails of outliers, mainly due to the fact that cluster 2 has very few regions so that outliers have a strong impact on the distribution. As shown in the violin plots of Figure 2b from ATAC signal data the M-index and the local peaks were more unevenly distributed with less accuracy in the mean due to the higher variability even by using the standardized data, however, means still showed enough differences to be considered a distinctive signature of each cluster.

Peaks of each of cluster were inspected in IGV (Integrative Genomics Viewer) to assess correct clustering and some of the most representative ones of ATAC signal data are shown in Figure 3, while the rest of the inputs' clusters are shown on Supplementary Figure 3 showing that the shape of the peak was different depending on the type of input. All these data suggest that proposed indices are proved to be valid for a correct clustering of ATAC, Open, MonoNuc and combined Open-MonoNuc signal data.

3.3 Characterization of clusters

To study the correspondence between clusters identified with the 4 different kinds of clustering according to its input data, we plotted the percentage of overlapping of peaks between clusters (Figure 4a, Supplementary figure 4). Although a certain correspondence can be seen, heatmaps shown in Figure 4a revealed that there is no direct correspondence between the peaks of each of the clusters, showing that clusters identified in different clustering are distinct depending on which kind of input data to use. It is also noticeable that the composition of each cluster is heterogeneous in terms of their percentage of promoters and enhancers, finding clusters such as 1 and 5 (of the clustering performed with ATAC data) where almost 75% are promoters while others such as clusters 3 and 4 do not reach 10% (Figure 4b). This diversity in composition also occurred for the rest of the input data (Supplementary Figure 5), there was not a cluster in particular that remained consistently enriched in promoters or enhancers in all the 4 inputs of data.

On the heatmaps can be seen how there is a decent overlap of regions between certain clusters like cluster 1 of ATAC and cluster 5 of Open signal data, when attending to their composition of enhancers and promoters both clusters have a high percentage of promoter regions. This correlation can also be seen with the clusters 3 and 4 of ATAC and cluster 2 of Open signal data which have a high overlap of regions, in this case the three clusters are very poor on promoter regions.

Then, we wanted to study the chromatin signature of the different clusters. To achieve this, we studied the distribution of three different histone modifications (H3K4me1, H3K4me3 and H3K27ac) along with ATAC-seq reads distribution. In Figure 5a we plotted the shape of the peaks of the clusters. The appreciated topography of the clusters remained simple with one or two peaks varying in height and area. For example, the peaks of ATAC clusters' 1 and 10 are high and narrow with only one peak, while clusters 3, 4, 7 and 8 are generally shorter with very reduced area; peaks of clusters 2 and 9 are wider with two peaks. These different shapes are a good indication that the DNA regions sorted according to the proposed indices may be sufficient to achieve a good separation of the regions, providing critical information on the functions they play in cell regulation.

The H3K4me3 histone modification is generally restricted to narrow regions at the 5-terminus of the gene body (promoters), although a small subset of genes has a broad H3K4me3 domain that covers the majority of the coding region (Cao et al., 2017). Genes tagged with the broad epigenetic domain contain a number of epigenetic modifications complementary to trimethylation such as H3K27ac. These genes are thought to be involved in essential cellular identity and functions and have clinical potential as biomarkers for patient stratification (Beacon et al., 2021). In this study it can be seen that those clusters with high signal intensity for H3K4 trimethylation also correspond to high signal for lysine 27 acetylation and could therefore be these domain wide regions (Figure 5c-d). However further studies on other characteristic histone marks such as H4K12ac, H4K20me1, H2BK5me and H4R3me2a (Beacon et al., 2020) and ontology studies are needed to determine the cellular functions of these genes.

Another important aspect to highlight is the H3K4me1/H3K4me3 ratio. It is known that high presence of H3K4me1 is a signal of enhancers, whereas a high presence of H3K4me3 is related to promoters (Soldi et al., 2017). According to the intensity plots, all clusters show a high monomethylation signal, however clusters 3, 4, 6, and 7 show the highest K4me1/K4me3 ratio since their level of trimethylation is the lowest. This corresponds to the low percentage of promoters in these clusters being 5%, 7%, 30%, and 11% respectively, meaning they are highly enriched in enhancers. Furthermore, H3K27ac is a typical mark of active enhancers. Since clusters 3, 4 and 7 have low level of H3K27ac, it is possible that these enhancers are in a not fully active configuration. Enhancers that present H3K4me1 but not H3K27ac are often called poised or primed (Crispatzu et al., 2021). Therefore, it is possible that these three clusters are enriched in poised enhancers. This relation was consistent for the other 3 types of inputs as it can be seen in the Supplementary figures 6-8.

In conclusion these results show that peaks clustering based on their morphology allow a general distinction of the regulator regions according to their nature, observing clusters with a high percentage of enhancers with high H3K4me1/H3K4me3 ratio, while those clusters enriched in promoters have a low H3K4me1/H3K4me3 ratio.

3.4 Search and enrichment of motifs

Then we asked if clusters with different morphology are enriched in motifs of specific transcription factors. Prior to performing the enrichment analysis, a series of random sequences of similar composition to our study sequences, known as background sequences, were generated to ensure that the motif classification was consistent and to avoid biases due to the A-T and C-G composition of the sequences.

To achieve this, we computed the enrichment of transcription factor binding motifs found in JASPAR vertebrate database for each cluster against the rest of the clusters. Most significant motifs for each cluster are shown in Figure 6.

It is remarkable the correlation between the TFs that appear on clusters and their composition of enhancers or promoters, for example in clusters mainly dominated by promoter regions (cluster 1 and 5) we find NF-YA and NF-YB two subunits of the NF-Y protein which is known for binding directly on the CCAAT-box of promoters (Mantovani, 1999). and E2F family proteins that binds to the TTTCCCGC site in the target promoter sequence and is highly involve in cell proliferation due to its role in the control of the transition from phase G1 to S (Gaubatz et al., 2000). On Figure 6 of ATAC data, another group of factors that had high significance in both enrichment and p-value were Sox2, Smad4 and TEAD3 of clusters 3 and 4. All involved in cell proliferation, differentiation, and maturation, Sox2 participating in the maintenance of the pluripotent form of embryonic cells, Smad4 being an important transcription factor in the TGFβ signalling pathway and TEAD3 being a key factor regulating epithelial cell maturation (Adachi et al., 2010; Zhao et al., 2018; Li et al., 2020). All these factors are typically enhancer binding proteins.

The motif enrichment analysis on the other inputs showed some relevant ones such as the case of the transcription factor CTCF which is highly represented in cluster 2 of the mononucleosomes data (Supplementary Figure 9b) and present in other clusters. This factor consists of 11 highly conserved zinc fingers with which it can bind to multiple regions of the genome. One of the unique functions of CTCF is its insulator function. Insulators are short nucleotide sequences that establish boundaries between nearby genomic domains. When CTCF binds to an insulator sequence, it prevents communication between an enhancer and a gene promoter by blocking gene transcription (Kim et al., 2015). It also plays a key role in the 3D organization of the genome and the maintenance of topologically associated DNA domains (TADs) by maintaining the structure of the loops as two of these proteins interact with each other to isolate segments of the genome, thus favouring the connections within the domain itself and allowing for a more fine-grained regulation of the domain (Ghirlando & Felsenfeld, 2016).

Analysis revealed cis-regulatory logic through known motifs (e.g., AP-1, ETV, ZNF and ELK sites) and less common ones (e.g., CTCF, Tead and NF1). Many DNA-binding transcription factors, which recognise these cis-motifs, are markedly up-regulated. However, the clustering did not allow a fine and clear separation of the motives from each cluster only providing information about the most the general cis or trans regulation of them.

3.5 Ontology analysis

Since clusters have shown to have distinctive characteristics and different motives associated to genes, we carried out an ontological analysis in hope to find if the genes associated to the regions of each cluster were involved in similar biological processes. To achieve that we performed the annotation using the rGREAT tool with the basal plus extension mode in which each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

Subsequently, also with the GREAT tool, the genes associated to the peaks were annotated and classified according to their annotation in Gene Ontology (GO) of biological process, obtaining different processes among the clusters, but at the same time the processes found within clusters were more related to each other. As shown in Supplementary Table 1, clusters 1, 4 or 5 of the ATAC-seq data are involved in different biological processes, while cluster 1 is very significantly enriched in regions associated with genes related to the regulation of RNA metabolism, DNA repair and chromatin accessibility, cluster 4 has regions mostly related to cell migration and angiogenesis, and cluster 5 seems to contain genes involved in protein maturation and modification in the endoplasmic reticulum and Golgi (See Supplementary Tables 1-4).

This ontology analysis revealed interesting biological insights about the morphological based clustered regions. Showing how the shape of the peak can be related with the biological functions of the genes associated to the DNA regions.

3.6 Logistic regression model

An important part of the peaks analysis is to determine the nature of the regions in order to find more about their role in the regulation of DNA expression. That is why we developed a binary logistic regression model with the standardized data of the five morphological indices to determine on the basis of the morphological indices of the peak whether the region is considered a promoter or enhancer. The variables were first evaluated by a Step test and a collinearity analysis to determine if they would provide enough significant information to the model.

The five indices were analyzed to see how promoters and enhancers were distributed (Figure 7a), with the former having a higher mean for all indices as well as a higher range and standard deviation, although this is again due to the fact that fewer outliers have a greater impact on the distribution. Once again, this distribution was consistent in the four types of inputs (Supplementary figures 10-12a).

The model was evaluated, and it proved to be very robust according to the tests performed. A high accuracy rate was obtained for promoters (>70%) and even better for enhancers (>85%). This better classification of the enhancers is mainly due to their higher abundance. The model was also validated via Roc Curves method, achieving as well good results of confidence since the area under the curve was above 0.80 in all the four types of data (Figure 7b, c; Supplementary figures 10-12b).

While clustering itself did not allow a separation of the promoter and enhancer regions of the peaks, this model involving the morphological indices did help to clarify the differences between the two gene elements with a high accuracy rate, making it interesting for mapping promoter enriched or enhancer enriched domains in the genome. On Figure 7d we show the comparison of percentages of promoters' regions of each cluster versus the percentage we got from the predicted model seeing that the predictions are very close to the true values. It is also interesting evaluate some of the peaks that are not predicted correctly. For example, some enhancers act also as promoters of lncRNA (Lam et al., 2014). Therefore, it is possible that some peaks annotated as enhancers but that our model predicts as promoters have a hybrid function.

4 Conclusions

The classification and characterization of genome regions based on the characteristics of the peaks generated in ChIP-seq studies has been validated on many occasions and it has been shown that the characteristic morphology of each peak can provide relevant biologically meaningful information. In this analysis we have shown how this classification method is also applicable to ATAC-seq data. The clusters obtained all presented a distinctive signature in terms of their distribution and the overlap was minimal. Furthermore, thanks to the study of histone marks, it was possible to characterize how each cluster had a characteristic distribution with a biological meaning in relation to the percentage of promoters and enhancers, which was later contrasted and corroborated by both the literature and the analysis of the motifs. However, the separation by clusters does not allow a reliable separation of the function of the genes under distinctive ontological terms, nor does it allow a direct determination of which regions are promoters and which are enhancers. For this reason, the binary logistic regression model was developed to provide more information about the composition of each cluster. Further analysis can be performed to enrich this information such as enhancer RNA analysis or including more parameters into the logistic model to refine it and enhance its predictive power making it able to identify other elements such as insulators or poised enhancers. Overall, the results we got are a very interesting starting point, revealing general information of the nature of the DNA elements studied, for more analysis to be performed using this clustering method for ATAC-seq data.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author Contributions

JCRR and JAGM provided the data and planned out the experiments, JAGM contributed with the code and general idea of analysis, JAGM and ARM carried out the computational analysis, JCRR, JAGM and ARM evaluated and checked the results, ARM wrote the final version of the manuscript, JCRR, JAGM supervised the manuscript.

7 Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

8 Acknowledgments

Special thanks to both José Carlos Reyes and José Antonio Guerrero-Martínez for their guidance and patience.

9 Data and Code Availability Statement

Data can be found as stated at Data Acquisition segment publicly accessible. ATAC-seq reads can be found in ENA's database (https://www.ebi.ac.uk/ena/browser/view/SRR104858769). Data from the histone marks can be found in geo's repository (https://www.ncbi.nlm.nih.gov/geo/) under the correspondent accession numbers clarified before. Motif list was taken from JASPAR's database (https://jaspar.genereg.net/downloads/).

The ENCODE pipeline to process ATAC-seq data can be found in (https://github.com/kundajelab/atac_dnase_pipelines). The software to perform the clustering of regions was done following SIC-ChIP pipeline (https://github.com/marziacremona/SIC-ChIP). The rest of the code to perform the analysis can be found publicly posted on GitHub (https://github.com/AlbertoRiMu/TFM_ARM_ATAC-seq-Analysis.git).

10 References

Adachi, K., Suemori, H., Yasuda, S., Nakatsuji, N., & Kawase, E. (2010). Role of SOX2 in maintaining pluripotency of human embryonic stem cells. Genes To Cells. doi: 10.1111/j.1365-2443.2010.01400.x

Alberts B, Johnson A, Lewis J, et al. (2002) Molecular Biology of the Cell. 4th edition. New York: Garland Science; p. 207. ISBN-10: 0-8153-4072-9

Beacon, T.H., Delcuve, G.P., López, C. et al. (2021) The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. Clin Epigenet 13, 138. doi: 10.1186/s13148-021-01126-1

Beacon, T., Xu, W., & Davie, J. (2020). Genomic landscape of transcriptionally active histone arginine methylation marks, H3R2me2s and H4R3me2a, relative to nucleosome depleted regions. Gene, 742, 144593. doi: 10.1016/j.gene.2020.144593.

Buenrostro, J., Giresi, P., Zaba, L. et al. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218. doi: 10.1038/nmeth.2688.

Cao, F. et al. (2017) "Super-enhancers and broad H3K4me3 domains form complex gene regulatory circuits involving chromatin interactions," Scientific Reports, 7(1). Available at: https://doi.org/10.1038/s41598-017-02257-3.

Cremona, M.A., Sangalli, L.M., Vantini, S. et al. (2015). Peak shape clustering reveals biological insights. BMC Bioinformatics 16, 349. doi: 10.1186/s12859-015-0787-6.

Crispatzu, G., Rehimi, R., Pachano, T. et al. (2021). The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. Nat Commun 12, 4344 https://doi.org/10.1038/s41467-021-24641-4

Kim, S., Yu, N., & Kaang, B. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. Experimental &Amp; Molecular Medicine, 47(6), e166-e166. doi: 10.1038/emm.2015.33.

Ghirlando, R., & Felsenfeld, G. (2016). CTCF: making the right connections. Genes &Amp; Development, 30(8), 881-891. doi: 10.1101/gad.277863.116.

Guerrero-Martínez, J.A., Ceballos-Chávez, M., Koehler, F. et al. (2020). TGFβ promotes widespread enhancer chromatin opening and operates on genomic regulatory domains. Nat Commun 11, 6196. doi: doi.org/10.1038/s41467-020-19877-5.

Gaubatz, S., Lindeman, G. J., Ishida, S., Jakoi, L., Nevins, J. R., Livingston, D. M., &amp; Rempel, R. E. (2000). E2F4 and E2F5 play an essential role in pocket protein–mediated G1 control. Molecular Cell, 6(3), 729-735. doi:10.1016/s1097-2765(00)00071-x

Hanahan, D., & Weinberg, R. (2000). The Hallmarks of Cancer. Cell, 100(1), 57-70. doi: 10.1016/s0092-8674(00)81683-9

Lam, M.T.Y. et al. (2014) "Enhancer RNAS and regulated transcriptional programs," Trends in Biochemical Sciences, 39(4), pp. 170–182. doi: 10.1016/j.tibs.2014.02.007.

Li, J., Tiwari, M., Xu, X., Chen, Y., Tamayo, P., & Sen, G. (2020). TEAD1 and TEAD3 Play Redundant Roles in the Regulation of Human Epidermal Proliferation. Journal Of Investigative Dermatology, 140(10), 2081-2084.e4. doi: 10.1016/j.jid.2020.01.029

Li, M., & Flavell, R. (2008). TGF-β: A Master of All T Cell Trades. Cell, 134(3), 392-404. doi: 10.1016/j.cell.2008.07.025

Mantovani, R. (1999). The molecular biology of the CCAAT-binding factor NF-Y, Gene, 239(1) 15-27, ISSN 0378-1119, doi: 10.1016/S0378-1119(99)00368-6.

Siggens, L., & Ekwall, K. (2014). Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. Journal Of Internal Medicine, 276(3), 201-214. doi: 10.1111/joim.12231.

Soldi, M., Mari, T., Nicosia, L., Musiani, D., Sigismondo, G., & Cuomo, A. et al. (2017). Chromatin proteomics reveals novel combinatorial histone modification signatures that mark distinct subpopulations of macrophage enhancers. Nucleic Acids Research, 45(21), 12195-12213. doi: 10.1093/nar/gkx821.

Wilson, S., Billings, P., D'Eustachio, P., Fournier, R., Geissler, E., & Lalley, P. et al. (1990). Clustering of cytokine genes on mouse chromosome 11. Journal Of Experimental Medicine, 171(4), 1301-1314. doi: 10.1084/jem.171.4.1301

Yan, F., Powell, D., & Curtis, D. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biology, 21 22. doi: 10.1186/s13059-020-1929-3.

Zhao, M., Mishra, L., & Deng, C. (2018). The role of TGF-β/SMAD4 signaling in cancer. International Journal Of Biological Sciences, 14(2), 111-123. doi: 10.7150/ijbs.23230.

Figure Legends.

Figure 1. Diagram of the workflow followed for the analysis of the ATAC-seq peaks clusters' indices, annotation of the clusters' peaks and elaboration of the predictive model. Packages and pipelines specified on the graph next to each step

Figure 2. a) Scatter plot representing the Pearson correlation between each of the 5 indices for the ATAC standardized data. Indices names: h, peak height; A, peak area; $w_{h/2,}$ maximum peak width; p_{local}, number of local peaks; M/h, M-index normalized to h. b) Violin plot representing the distributions of each of the ATAC clusters in each of the indices.

Figure 3 Regions of significant peaks in each of the clusters of the ATAC data taken from IGV. Peaks were taken with a window size of 48 kb, each of the clusters were aligned with the complete genome ATAC-seq data.

Figure 4. a) Heatmaps summarizing the number of overlapping regions of each of the ATAC signal data clusters with those of open chromatin (Open), mononucleosome regions (MonoNuc) and the combination of both (Open MonoNuc). The heatmaps are scaled respect to the number of elements of each of ATAC clusters. The summatory of each row represents the size of each ATAC cluster and the summatory of each column represents the size of each cluster for the other three input data. b) Percentage of the clusters that are identified as promoter regions for the ATAC data based on the list of TSS regions file.

Figure 5. Plots representing the distribution of the corresponding marks of the ATAC data. The intensity of each bin is plotted at a central point of the cluster regions and spread 3kb in each direction. a) Distribution of the ATAC signal data reads over the different clusters. b) Distribution of the H3K4me1 histone mark, c) Distribution of the H3K4me3 histone mark, d) Distribution of the H3K27ac histone mark.

Figure 6. Profile of the transcription factors associated to the most significant motifs found in the analysis for each cluster of the ATAC data. The transcription factors shown are filtered by enrichment and p-value (Enr > 1.5, p-value < 0.05).

Figure 7 Relevant information and graphics about the prediction model of ATAC signal data. a) Distribution of the five morphological indices depending on the nature of the classified region promoter

or enhancer. b) Representation of the area under the curve evaluation method. ROC Curve representing sensibility over. c) Confusion matrix for model validation predictions being on the x axis and observations on y axis. d) Percentaje comparing the promoter regions predicted by the model for each cluster versus the true percentaje of true promoter regions of each cluster.

Table headers

Supplementary table 1-4. Enrichment tables summarizing the most relevant Gene Ontology biological processes of the genes associated to the regions of each of the clusters for the four types of input data, in order being 1) ATAC, 2) Open, 3) MonoNuc, 4) Open MonoNuc. Tables were obtained using the rGREAT annotation analysis, sorted by binom and HypeRank q-value in excel and taken the first 20 most significant processes.