Università degli Studi di Modena e Reggio Emilia

Dipartimento di Scienze della Vita
Corso di Laurea Triennale in Biotecnologie

Bioinformatics investigation of chromatin state discovery and genome annotation: analysis with ChromHMM

Laureando:	Relatore:
Tommaso Becchi	**Prof. Silvio Bicciato**
	Correlatore:
	Dott. Jimmy Caroli

Anno accademico 2018/2019

Abstract

Histones represent 80-90% of the entire chromatin proteins. Their role in DNA structural organization is of pivotal importance, as they allow the making of specific interactions with chromatin meant to compact and reduce the occupied space by DNA in the nucleus. This fundamental function is provided by their specific conformation: histones are in fact composed by a central core with a positive charge determined by the abundance of basic aminoacids as arginine and lysine. However, histones do not bear only this function: it was discovered that the N-terminal tails (which do not make any interaction with chromatin) can bear many different chemical modifications with key biological roles. These modifications can be of different types and on different histones: the most frequent ones are acetylation, methylation, phosphorylation and ubiquitination. It has been demonstrated that histone modifications are not random, but instead "mark" DNA regions with specific transcriptional activities. Given the great importance that these elements have in the analysis of DNA activity, different techniques have been developed for their study. One of the best approaches in terms of resolution and coverage is chromatin immunoprecipitation coupled with sequencing (ChIP-Seq), which returns the sequences linked to specific histones modifications. Once obtained these sequences, it is possible to divide the DNA in many regions, each of them characterized by the presence of certain combinations of modifications. These patterns and their associated transcriptional profiles provide insights into the identification of chromatin states. Chromatin states are biological elements with an important role in the determination of the DNA activity: each combination of histone modifications is associated to a specific transcriptional activity. Over the past few years many different bioinformatic tools aimed at the identification of these chromatin states have been developed, starting from the presence or absence of different histone modifications. Among these tools ChromHMM is one of the most used in literature, being able to create different models composed of variable numbers of chromatin states depending on the quantity and quality of input data. Given these premises, I used ChromHMM on publicly available ChIP-Seq data of histone modifications in different cell types. Briefly, I first optimized the parameters of the software using data and chromatin states from the Encode consortium. Then, I applied ChromHMM to classify the chromatin states of liver, lung and common myeloid progenitor cell lines, evaluating the consistency of the retrieved chromatin states through gene ontology functional characterization. Overall, ChromHMM proved to be reliable in identifying the chromatin states using a large number of histone modifications as input (Encode test data). At the same time, it has been able to generate good results also with a limited number of modifications, as in the case of the tested human tissues analyses.

Index

1. Introduction

DNA is located in the cell nucleus in a supercoiled conformation [1]. Given its enormous length (over two meters of DNA into each nucleus), several approaches to compact this molecule are required. One of the pivotal structures aimed at this function are histones, proteins characterized by a positive charge which enables interactions with the negative backbone of DNA. Histones are composed by a central core, defined by four different subunits in duplicate [2][3], around which the DNA is wrapped. Each of these subunits contains several N-terminal tails that are free, not involved in any direct interactions with the chromatin. These tails are able to interact with other elements in the nucleus, such as enzymes whose function is to modify some aminoacids which are highly conserved in the aminoacid sequence of the tails, performing modifications such as acetylation, methylation, phosphorylation and ubiquitination [4].

These are not random phenomena; instead these changes have an important role to mark DNA regions with different transcriptional activity [5][6]. For these reasons, many techniques have been developed to enforce the study of regions associated to these single modifications [7][8]. One of the most used techniques is *ChIP-Seq*, (Figure 1) in which the combined use of immunoprecipitation and sequencing allow to obtain a large number of data in few time with high resolution [9].

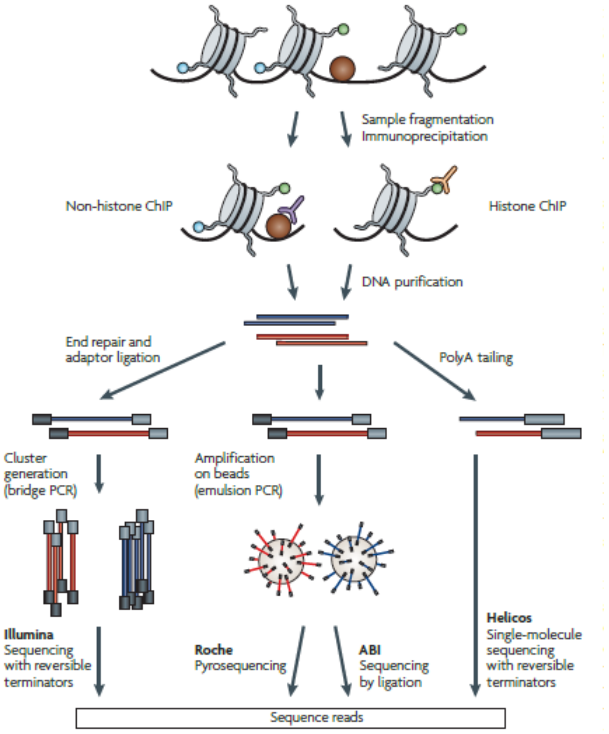

Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads

Figure 1: With ChIP-Seq it is possible to obtain the DNA fragments associated to a certain histone modification and retrieve their sequence. The process consists of three steps: fragmentation, immunoprecipitation and sequencing. Fragmentation can be achieved with sonication or with mNase digestion depending on the experiment goal. For the immunoprecipitation, specific antibodies are used: they can be against a modification or against a protein of interest. Finally, for the sequencing many different techniques can be adopted as Illumina, pyrosequencing, ligation and Helicos sequencing. (Adapted from Park, *et al*. [9])

When ChIP-Seq is used to investigate DNA-binding proteins, the chromatin is sheared by sonication [10] into small fragments, which generally have a length of 200–600 bp. Differently, when ChIP-Seq is used to determine histone modifications, micrococcal nuclease (MNase) digestion without crosslinking is most often used to fragment the chromatin because it removes linker DNA more efficiently than sonication, therefore providing more precise mapping of each nucleosome [9]. After the fragmentation step, an antibody specific to the protein or to the histone modification of interest is used to immunoprecipitate the DNA–protein complex [4]: by performing these analyses, it is possible to extract only the DNA fragments which bind the histones bearing that specific modification. Finally, the released DNA is sequenced to determine the reads bound by the protein. Many different options are available for the sequencing step, as for instance Illumina, pyrosequencing, ligation and Helicos sequencing [11]. Each technique has advantages and disadvantages and it is preferred depending on many different factors [12]. Analysing the totality of the sample for a single cell line, histone modification reads frequencies are observed. The greater this frequency is, the greater the intensity of the peak of the signal relative to that modification in that region is registered. ChIP-Seq analysis show a peak distribution for the modifications on the DNA[13]: there are some regions in the center of the peak where the presence of the modification is strong, regions on the side of the peak where it decreases and regions far from the peak where its frequency is zero. By observing the distribution of different histone modifications in various ChIP-Seq experiments, it is possible to see that DNA fragments are rarely associated to a single modification, but rather to multiple ones (with different frequencies; Figure 2). Therefore, some DNA regions are associated with combinations of modifications which are not randomly present on DNA, they are characteristic of specific chromatin conditions as promoter, inhibitors and others. The association between modifications and transcriptional activity of correlated chromatin regions can be used to derive potential chromatin states [14]. These elements are of major importance when study the functional utilization of the genome, since they can be used to highlight potential regions of activity of chromatin sections not directly correlated to protein encoding activity [15]. For example, inside these regions many functional genome elements can be found, such as promoters, enhancers and repressors [16]. The particularity of these analyses consists in the identification of how the activity of these elements changes in the different cell lines where they can be active, weakly active and inactive.

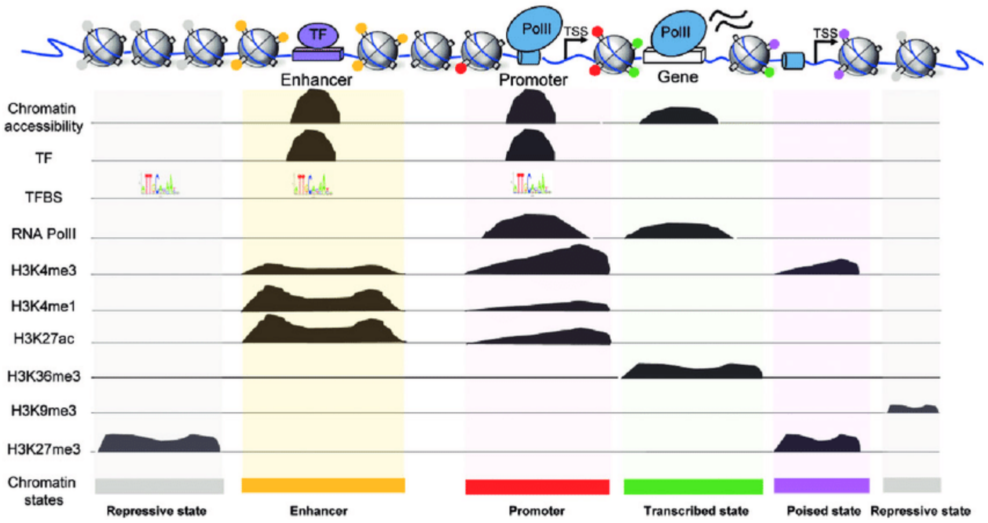


Figure 2: Specific DNA regions with a certain transcriptional function (promoters, inhibitors, ...) are associated with pattern of modifications peaks. These patterns are not casual: they are repeated for all the regions with the same function in all the cell type, allowing the correlation between them and the biological functions to identify chromatin states. (Adapted from Jiang, S. & Mortazavi, A. [13])

2. Aim of the thesis

The role of histones in the transcriptional activity of DNA is becoming increasingly important, fueling the development of techniques that study histone modifications and the DNA associated with them. One of the main approaches in this regard is ChIP-Seq, a method that combines immunoprecipitation and sequencing to determine in a precise manner which regions of the chromatin are associated with histones that carry certain modifications. ChIP-Seq was extensively applied to investigate histone modifications and the aim of this thesis is the application of a computational tool to inferred chromatin states from ChIP-Seq data of histone modifications. For this purpose, I used ChromHMM, a software designed to investigate the composition of chromatin states and the subdivision of the genome starting from the presence or absence of modifications in histones associated with a DNA region. Briefly, I first optimized the parameters of the software using data and chromatin states from the Encode consortium. Then, I applied ChromHMM to classify the chromatin states of liver, lung and common myeloid progenitor cell lines, evaluating the consistency of the retrieved chromatin states through gene ontology functional characterization.

3. Materials and methods

ChromHMM is a bioinformatics tool designed to investigate the combinations of histone modifications, eventually leading to the definition of specific chromatin states. To get this result it uses a multivariate hidden Markov model based on probabilistic modeling of both the combinatorial presence and absence of multiple marks [17] [18].

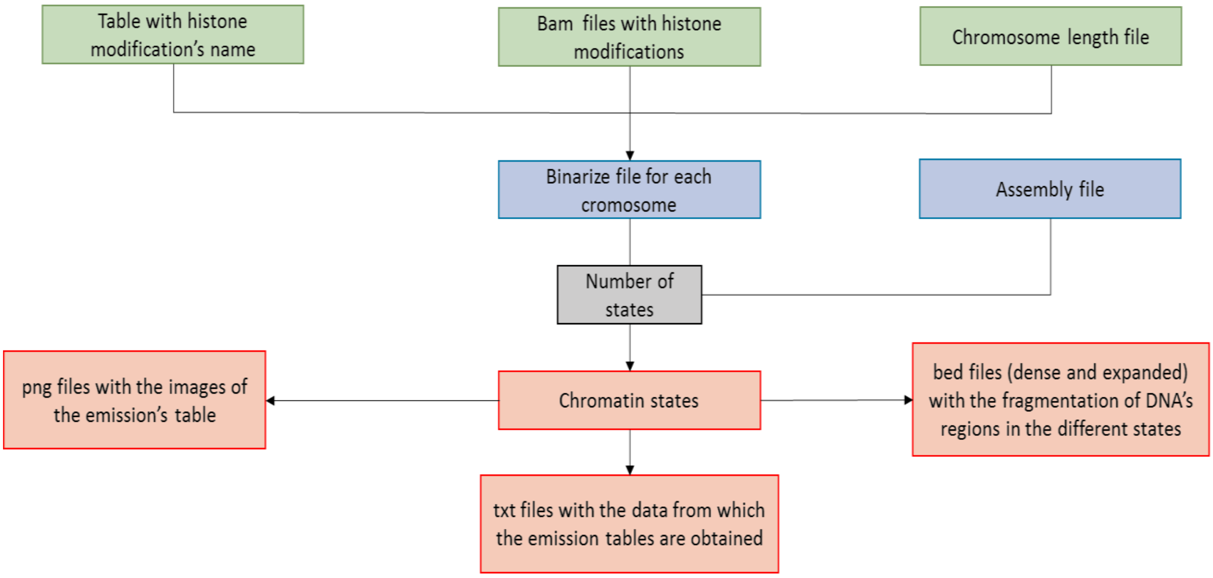


Table with histone modification's name
Bam files with histone modifications
Chromosome length file
Binarize file for each cromosome
Assembly file
Number of states
png files with the images of the emission's table
Chromatin states
bed files (dense and expanded) with the fragmentation of DNA's regions in the different states
txt files with the data from which the emission tables are obtained

Figure 3: The process by which ChromHMM generates the chromatin states combinations starting from the histone modifications files with ChIP-Seq results. To get the single chromosomes binary files you need to provide a chromosome length file (referred to the species of interest) and a table where the various input files are referred to. After that the program also requires a file with the genome assembly as input to obtain the states combinations in addition to the binary files.

ChIP-Seq allows to obtain data related to the association frequency between every chromatin fragment and a single histone modification. These results are the main input for ChromHMM: the starting information is composed by the complete genome, divided in bin and value of the association with the specific modification. The best formats to express data of this type are bed and bam format. BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track whereas BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments.

In both cases the first step is the binarization. To get it a "*cell mark file table*" is a fundamental input to provide: it is a tab delimited file in which each row contains the cell type or other identifier for a group of marks, the associated mark, and the name of the BED (or BAM) file.

```
cell1 mark1 cell1_mark1.bed
cell1 mark2 cell1_mark2.bed
cell2 mark1 cell2_mark1.bed
cell2 mark2 cell2_mark2.bed
```

Binarization returns files in which for each DNA segment the presence or absence of the histone modification is expressed with 0 (absence) or 1 (presence). ChromHMM generates a binary file for each chromosome using the information relative to every modification in input. Binary files are the input for the model generation. ChromHMM grants the possibility to define the number of chromatin states you want to get in output. This is very useful because there is not a correct default: in fact, according to the input modifications type and number even the best model can change. Therefore, before generating the models there is no way to know how many states are required to obtain a good genome partition. The number of desired states is a very important parameter because it could lead to results that may contain errors or show lacking of information: with too few states, few results are obtained, leading to an incomplete analysis, while a too high number forces the program to an excessive dispersion of the information throughout the fragmentation of the states. The difficulty of such study lies in determining for each analysis which model is the best. The execution time to generate the models changes according to the number of calculated states, ranging from minutes to hours. The composition in modification of states is not a fixed value that is found in an immediate time, more steps are needed to arrive at the desired result. Models creation is performed by an iterative process: many different models are continuously created and from time to time they are modified with some corrections. The program starts from the frequency of association with the modifications for each bin and it begins to create combinations of states, adapting them to the number of states imposed by the user. At first between two successive iterations there are large deviations while as the model improves the corrections are increasingly minimal. The process ends when ChromHMM is no longer able to improve the requested model. ChromHMM re-splits the DNA in new bins, different from those defined via ChIP-Seq, and assigns to each of them one of the created states. The assignment is based on the analysis of the modifications pattern on the bin. Their default length is 200 base pairs, but it is possible to increase the resolution entering a lower value. Nearby bins can be associated to the same state and then grouped together; in any case all the regions have a length that is a multiple of the minimum bin. When the model generation is completed ChromHMM returns many different outputs:

- *webpage_N.html:* a webpage which contains the commands used to learn the model and links to all the files generated as part of the output.
- *emissions_N.txt:* a tab delimited text file containing the emission parameters of the hidden Markov model
- *emissions_N.png, emissions_N.svg:* these files display the contents of emissions_N.txt as a heatmap in .png and .svg image formats respectively (Figure 4a)

- *transitions_N.txt:* a tab delimited text file containing the transition parameters of the model
- *transitions_N.png, transitions_N.svg:* these files display the contents of transitions_N.txt as a heatmap in .png and .svg image formats respectively. (Figure 4b)
- *model_N.txt:* this file contains the emission, transition, and initialization parameters of the hidden Markov model in a format for which ChromHMM can parse all the parameters
- *celltype_N.bed (segments, dense and expanded):* these files are a segmentation of the genome containing ChromHMM's genome annotation for a cell type in an easy to parse format.

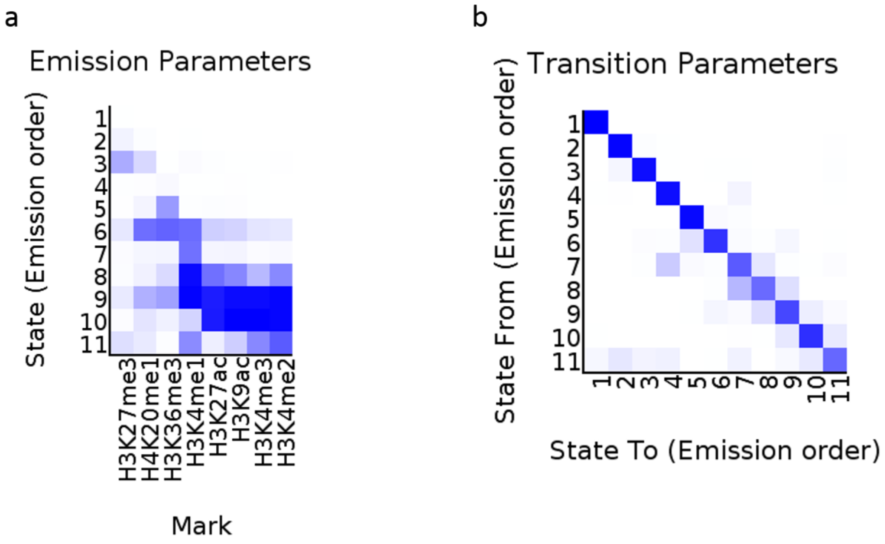
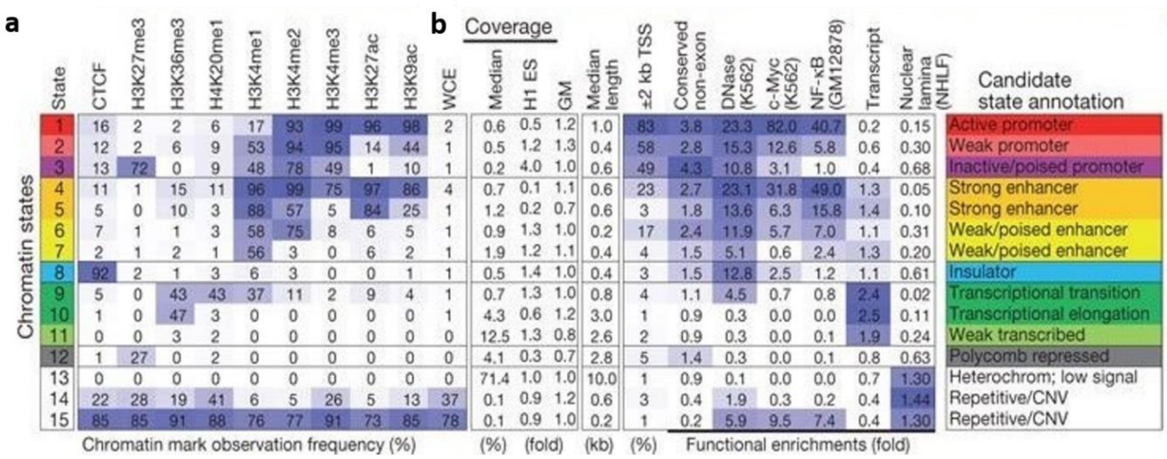


Figure 4: 4a is the table with the frequency of each modification in the different states for the 11-states combination. Various shades of blue represent the different modifications expression. **4b** is the table with the frequencies relative to the transition parameters for each state in the 11-states combination. The first row and first column are header rows and columns respectively. The values correspond to the probability under the model conditioned on being in the state of the row of transitioning to the state of the column

Furthermore, it is possible to generate chromatin states giving as input the modifications on more than one cellular line. This should improve the quality of the result because the tool has more data to use. As a result, a single model is obtained for the desired number of states and this model is used on both cell lines. In this case in the input table it is necessary to specify to which cell line the different BAM or BED files refer so that the tool can create the division of the genome for both lines.

4. ChromHMM validation on known data

ChromHMM has been designed to study the division of DNA into chromatinic states, thus enabling better cell characterization. However, the process that leads to the generation of states is not immediate and requires several steps, which should be tested and fully optimized. On this purpose, the first part of my thesis was aimed at the validation of the analysis protocol with ChromHMM, using already generated and publicly available results obtained with this software. Many of these results are present on Encode[19], an international collaboration of research groups which has the goal to build a comprehensive list of functional elements in the human genome and regulatory elements that control cells and circumstances in which a gene is active. Thus as a reference I used data downloaded from this consortium where the chromatin states analyses concerning nine different cell lines are uploaded (https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeBroadHmm&db=hg19). ChIP-Seq results were used to generate these tracks and the chromatin states were learned from these binarized data using a multivariate Hidden Markov Model (HMM) that explicitly models the combinatorial patterns of observed modifications. Encode analyses have been obtained starting from 8 modifications data and 1 transcription factor data: H3K27me3, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K20me1 and CTCF. This input data quantity provides a complete panel of information for the generation and identification of the main chromatin states. In total, fifteen states were used to segment the genome, later grouped and colored to highlight predicted functional elements. For each of the nine cell types (GM12878, H1-hESC, K562, HepG2, HUVEC, HMEC, HSMM, NHEK, NHLF) each 200 base pair interval (bin) was then assigned to its most likely state under the model. This genome partition is summed up in dedicated BED files, which will be used as the comparison element for my analyses, providing this way not only a visual but also a numerical approach to rank results. ChromHMM generates the chromatin states characterizing them based on the different presence of the modifications but without assigning them a specific name. The composition in histone modifications of the different known states is constant in the various cell types and it is known, therefore it is possible to assign a name to each ChromHMM state by comparing the results with data in literature (Figure 5) [14].



a											b Coverage												
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Median	H1 ES	GM	Median length	±2 kb TSS	Conserved non-exon	DNase (K562)	c-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV
	Chromatin mark observation frequency (%)										(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						

Figure 5: Comparison matrix used to assign to each state generated by ChromHMM a known state comparing the percentage presence of the modifications (a). Data relating to statistical and enrichment analyses are also present, allowing the different states to be identified and distinguished better (b). In this study the chosen result is a 15 states model and it is an optimal comparison model because 12 different types of states are present to allow a complete validation. (Modified from Jason Ernst *et al* [14])

4.1 K562

On Encode there are nine studies resulting from ChromHMM analysis which can be used as a validation comparison. One of these concerns K562 cells, immortalized myelogenous leukemia cells [20] extracted from the blood characterized by the presence of BCR-ABL fusion gene. Encode analyses were performed with an input of 9 different chromatin marks (8 histone modifications plus 1 transcription factor). However, to generate chromatin states in these cells I selected 8 histone modifications, thus providing a similar but not equal input to that used in the comparison results. Files related to these modifications were also downloaded from Encode where it is possible select files related to ChIP-Seq experiments for many different modifications in different cell lines (https://genome.ucsc.edu/encode/dataMatrix/encodeChipMatrixHuman.html). The results are uploaded in different format as BAM, BED, bigWig, broadPeak, narrowPeak and fastq. Each of these formats sums up in a different way the information resulting from ChIP-Seq, the choice of the type of input therefore greatly influences the subsequent steps. The most suitable for ChromHMM are BAM and BED for which there are specific commands while the other formats are recognized with more difficulty. I selected BAM files of H3K27me3, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K20me1. Compared to Encode reference the only difference is the lack of CTCF data, so I have to verify if with a very similar input the tool still manages to produce correct results.

4.1.1 ChromHMM results

The first step in the generation of chromatin states is the signal binarization: starting from the 8 chosen modifications using the *BinarizeBam* command I generated the binarized files, one for each chromosome. The comparison analysis was performed using as reference assembly hg19, thus this version of the genome was selected for the analyses. Once the binary files were obtained, I could generate the chromatin states models with *LearnModel* command. In this step *numstates* is the fundamental parameter, defining the numerosity of the states created by ChromHMM. This value allows me to run many different analyses, so I have the possibility to choose the best model among them. The best model is the one showing a good abundance of different and clearly distinguishable chromatin states. Given these premises, I generated the models ranging from 7 states to 20, even though the Encode reference featured only 15 states, this way scanning the software power. For each state many different outputs are produced (as previously stated) and those of greatest interest are the files relating to the expression frequency of each modifications in the various states. These results are presented both as an image (Figure 6) and as a text file: text files with data of modifications frequency in the states are the most useful results, allowing to perform numeric comparison with Encode ones. By editing the text files of the various emission frequencies with Excel I obtained tables similar to those I used for the analysis comparisons (Figure 5). ChromHMM generates unknown states numbered from 1 to the number of states requested, thus assigning a label to each state is required. This assignment is made on the basis of the frequency matrices looking for states that have a high correlation with those featured in the reference table (Figure 5). In some cases, the numerical correspondences are elevated, so it is easy to assign the name of the state while for other states this does not occur. Nevertheless, every generated state in each of the 7 to 20 state combinations has been investigated and renamed according to histone modifications composition. Some of the results of this first part of the analysis are shown in Table 1. During this step, I observed that in the combinations with a low number of states (Table 1a) this association is often immediate even if the states obtained are few, providing a limited set of information. Increasing the number of states, the result increases in importance and new states appear while the states previously identified remain often unchanged (Table 1b). By further increasing the number of states to be generated, a subsequent fragmentation is obtained: the same state is separated into several similar states to reach the required number, creating information redundancy.

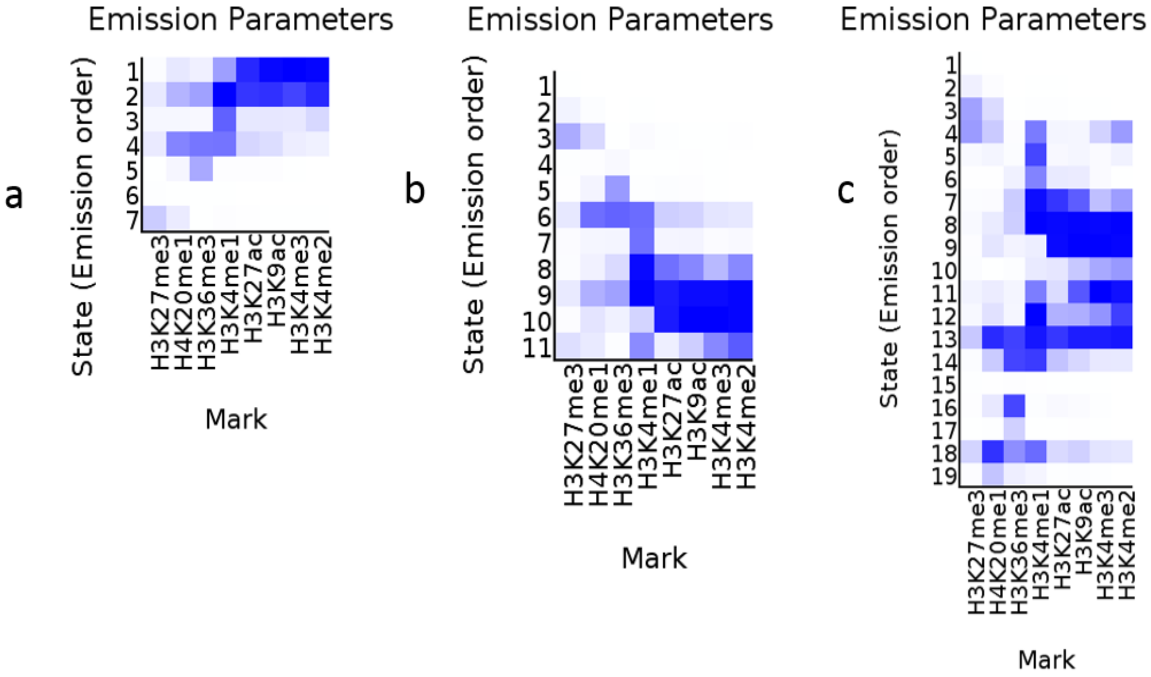
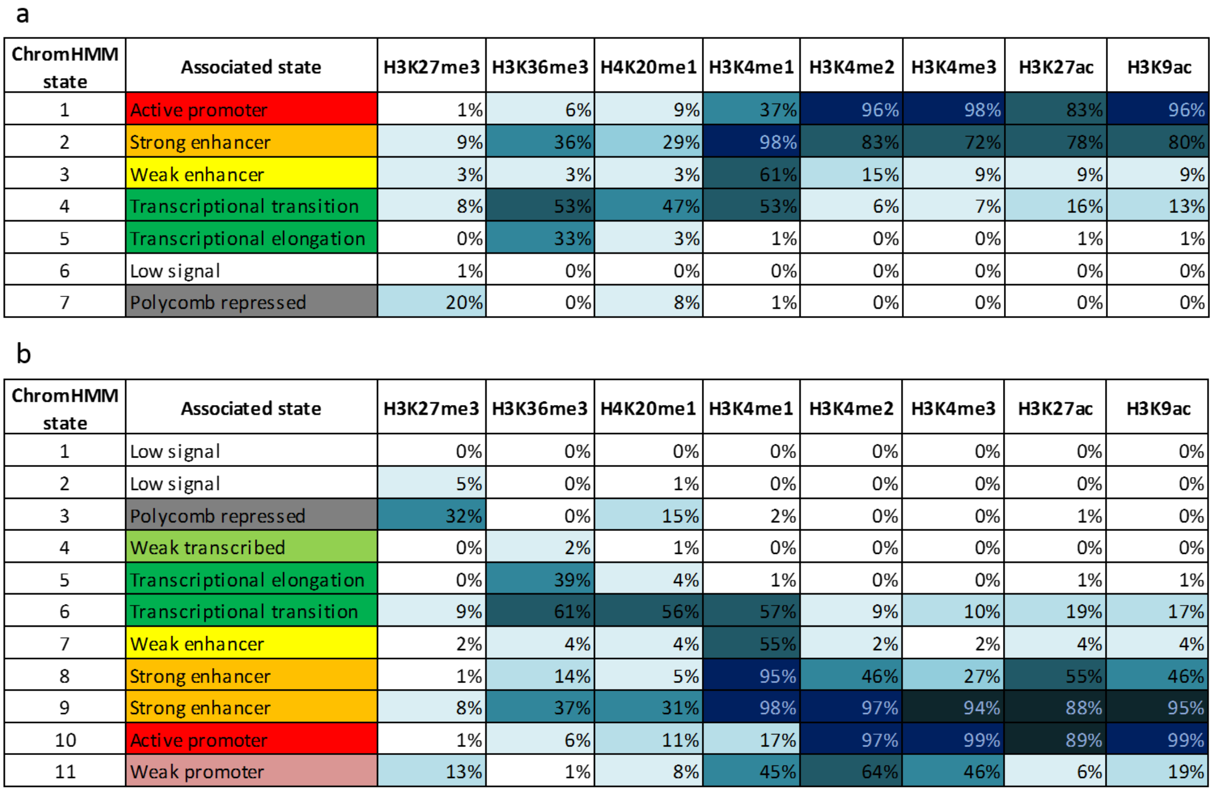


Figure 6: The table with the frequency of each modification in the different states for three different models. The different expression is represented by the various shades of blue. ChromHMM uses the reference table in input to assign the correct modification name to each column. It is observed that the combinations present in models with low number of states (a) are often present also in the other models (b) while in combinations with a high number of states the same state is repeated (c).

Looking at all the combinations obtained, the one that expresses the best results starting from the modifications given as input is the 11-state one (Table 1b) which will be the model selected for further discussion. The difference between the number of states chosen (11) and that of the reference analysis (15) is due to the presence in the latter of the CTCF data so it is normal that a higher number of states are found. Focusing on the 11-state model we can observe that all the states of the reference analysis are identified except for the "Insulator" state (which is characterized by the presence of CTCF that is missing in my analysis), "Inactive/poised promoter" and "Repetitive/CNV" states. Furthermore, only the known "Low Signal" and "Strong Enhancer" states are associated with more than one unknown state and this increases the possibility of making a clear comparison without ambiguous results



a

ChromHMM state	Associated state	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac
1	Active promoter	1%	6%	9%	37%	96%	98%	83%	96%
2	Strong enhancer	9%	36%	29%	98%	83%	72%	78%	80%
3	Weak enhancer	3%	3%	3%	61%	15%	9%	9%	9%
4	Transcriptional transition	8%	53%	47%	53%	6%	7%	16%	13%
5	Transcriptional elongation	0%	33%	3%	1%	0%	0%	1%	1%
6	Low signal	1%	0%	0%	0%	0%	0%	0%	0%
7	Polycomb repressed	20%	0%	8%	1%	0%	0%	0%	0%

b

ChromHMM state	Associated state	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac
1	Low signal	0%	0%	0%	0%	0%	0%	0%	0%
2	Low signal	5%	0%	1%	0%	0%	0%	0%	0%
3	Polycomb repressed	32%	0%	15%	2%	0%	0%	1%	0%
4	Weak transcribed	0%	2%	1%	0%	0%	0%	0%	0%
5	Transcriptional elongation	0%	39%	4%	1%	0%	0%	1%	1%
6	Transcriptional transition	9%	61%	56%	57%	9%	10%	19%	17%
7	Weak enhancer	2%	4%	4%	55%	2%	2%	4%	4%
8	Strong enhancer	1%	14%	5%	95%	46%	27%	55%	46%
9	Strong enhancer	8%	37%	31%	98%	97%	94%	88%	95%
10	Active promoter	1%	6%	11%	17%	97%	99%	89%	99%
11	Weak promoter	13%	1%	8%	45%	64%	46%	6%	19%

Table 1: 2 different emission tables with the association of the known states. It is observed that in the combinations with low number of states (**a**) the association between state of ChromHMM and the known state is often immediate but the result obtained has little utility given the low number of states found. Combinations with high numbers of states instead identify many different known states but to reach the number required the tool must fragment them a lot and some repeat themselves. The best results are obtained with intermediate number combinations of states (**b**) where the ChromHMM state association is known to be almost unambiguous and a high number of different states are identified.

As a further comparison I analyzed the data of Coverage, Median Length and distance of ± 2 kb from a Transcription Start Site (TSS). The Coverage of each state indicates the percentage of the entire genome occupied by regions associated with that state. To calculate it, I added up the length of all segments associated with a state and then divided this value by the total length in kb of the reference genome. Median Length represents the average length in kb of all the regions associated with a single state. To obtain this value I calculated the sum of the length of the regions associated with a state and then I divided the resulting value by the number of regions associated with that state. The value ± 2 kb TSS indicates the percentage of the regions associated to a state that are at a maximum distance of 2 kb from a TSS in both directions of reading the DNA. Therefore, this value indicates how much the regions of a state are close to genes effectively transcribed. To do this, I used information in the BED file relating start and end point of each regions with a file containing the location of all TSS in the genome. The summary of this analysis is shown in Table 2. Looking at the results, I observed that state 10 labeled as "Active Promoter" shows the highest percentage (69%) in the proximity of TSS which is in line with literature [21], since the promoters are regions located near active genes.

This comparison holds true also for the state labeled as “Weak promoter”. To confirm the correctness of these results we checked on the reference table, where the highest value in the ± 2 kb TSS column is that of the “Active Promoter” state. Another state with an high value in this column is state 1 of ChromHMM which is associated with “Low signal”: in this state all the chromatin regions that have not been associated with any state starting from the type of input used are enclosed, even though they may have a transcriptional activity. This explains the high percentage of proximity to TSS.

ChromHMM state	Associated state	Median Length (kb)	Median Coverage (%)	± 2 kb TSS (%)
1	Low signal	185,0	49,5	56
2	Low signal	15,0	0,7	31
3	Polycomb repressed	4,5	0,6	22
4	Weak transcribed	4,4	19,2	20
5	Transcriptional elongation	7,1	4,5	19
6	Transcriptional transition	1,1	13,9	15
7	Weak enhancer	0,6	6,3	15
8	Strong enhancer	0,5	1,0	19
9	Strong enhancer	0,8	2,5	39
10	Active promoter	1,1	0,9	69
11	Weak promoter	0,5	0,9	39

Table 2: Summary table of the results of the genome analysis for each state. The Low Signal state has high values because it contains all the DNA that ChromHMM has not associated with any state with these modifications therefore it presents within it a great variety of elements. The values in the ± 2 kb TSS column are consistent with what is expected from the biological function of the states

After a first analysis based on the visual comparison with the reference matrix it was necessary to see if there was also a relationship in the numerical data at the base of the reference frequency matrix. Thus, I compared results obtained with ChromHMM with data provided by the Encode analysis. This way, l wanted to check how much the regions identified by ChromHMM that I associated to a known state were correlated with those identified by Encode. For this comparison I used *bedtools intersect* function which allows to screen for overlaps between two sets of genomic features. In this step I related each of the 11 states obtained from my data with each of the 15 states present in the Encode analysis. The goal was to observe the greatest correspondences in the states that I associated only based on the modifications pattern and the same state in Encode. These matches are expressed as a percentage indicating which fraction of all the regions associated with a state in my analysis are associated with a specific Encode state. Table 3 shows the overlapping percentages of the corresponding states. Looking at the table with the summary of this analysis, it is clear that the assignment of the states is consistent since all the overlaps between corresponding states have high percentages, except for state 11.

This means that most of the DNA regions are associated to the same state both in the Encode analysis and in my analysis. Although the input data of my study lacked information related to the CTCF, ChromHMM was able to segment the genome into extremely similar states to those highlighted by Encode. To visualize in an even more evident way the high correspondence between my analysis and that of Encode, alluvial plots have been designed for each of the investigated chromatin states. These graphs were implemented with *Alluvial* function which can be used to visualize any type of change in group composition between states. Looking at the graphs obtained it is evident how for most states generated by ChromHMM the highest correspondence occurs with the same state of the Encode reference. The highest correspondence is highlighted with the use of a specific color to clearly visualize how each state relates to those of comparison. Some of these graphs are shown in Figure 7.

ChromHMM state	Associated state	Overlap (%)	Encode Reference
1	Low signal	98,8	Low signal
2	Low signal	77,4	Low signal
3	Polycomb repressed	81,3	Polycomb repressed
4	Weak transcribed	56,8	Weak transcribed
5	Transcriptional elongation	62,5	Transcriptional elongation
6	Transcriptional transition	69,5	Transcriptional transition
7	Weak enhancer	67,6	Weak enhancer
8	Strong enhancer	48,1	Strong enhancer
9	Strong enhancer	70	Strong enhancer
10	Active promoter	79,9	Active promoter
11	Weak promoter	27,2	Weak promoter

Table 3: The table shows the relationship obtained using *bedtools intersect* function between the states generated by ChromHMM and the states present in the analysis on Encode. In the first column there are the numbers that ChromHMM assigns to each unknown state generated. The second column shows the names of the known states that I first assign based on which modifications are present in the combinations. The last column has the same known states used in the analysis performed with ChromHMM on Encode. The third column shows the percentages with which the regions of the "Associated State" column coincide with the regions associated with the corresponding state in the "Encode Reference"

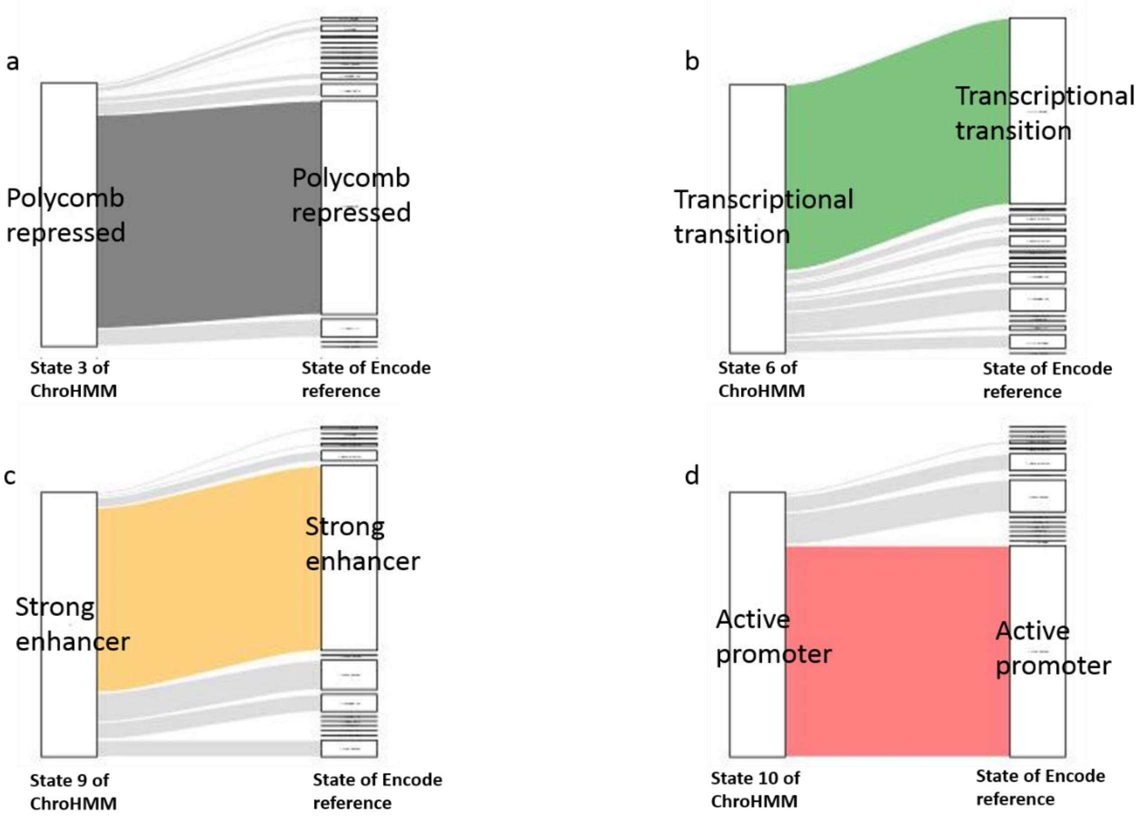


Figure 7: Four Alluvial graphs which show the correlation between corresponding states in my analysis and the Encode one. In each graph the major overlap is colored to highlight the states that are more similar. From these results the good quality of the performed analysis is observed: the states that match most are those that have the same biological function. This relationship is particularly evident for the "Polycomb repressed" state (**a**) and "Active promoter" state (**d**). Some states as "Transcriptional transition" (**b**) and "Strong enhancer" (**c**) have greater fragmentation but the major association remains the one with the corresponding state

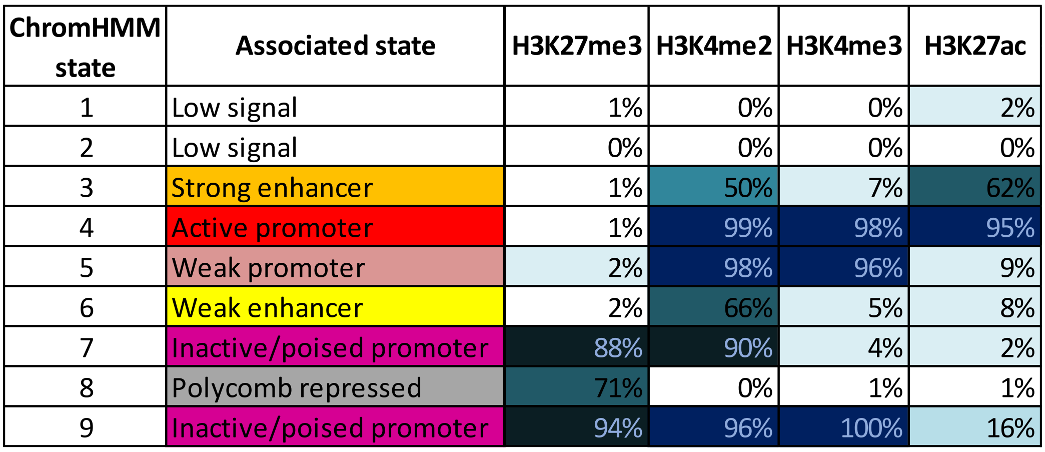
4.2 H1-hESC

With K562 I used an input very similar to those in the Encode study. For this comparison with H1-hESC cells I decided to reduce the number of modifications to further validate the tool and better understand how it works. Histone modifications were not randomly selected: I opted to analyze those that permits us to identify the main chromatin states based on biological knowledge. H1-hESC are human embryonic stem cells of which a ChromHMM study was already performed with 8 modifications and one transcriptional factor as input. This study is uploaded on Encode (https://genome.ucsc.edu/cgibin/hgTrackUi?g=wgEncodeBroadHmm &db=hg19) where it can be downloaded and then used as reference for the comparison. My goal was to find out what results were obtained starting from a reduced number of input data and also to understand if these results were comparable with those of Encode obtained with a very different input. Therefore, I used only 4 modifications: H3K27me3, H3K27ac, H3K4me2, H3K4me3. Bam files with ChIP-Seq data regarding these modifications on these cells have been downloaded from Encode in the same webpage (https://genome.ucsc.edu/encode/dataMatrix/encodeChipMatrixHuman.html) from which I downloaded K562 data.

4.2.1 ChromHMM results

The procedure to generate chromatin states is the same used for K562. However, since I used a smaller number of modifications I also reduced the range of the number of states to be generated. The analysis therefore focuses on combinations of 4 to 10 states. I do not expect to find all the states present in the Encode study because the input is very different but I expect that there will be a good correspondence for the few states identified.

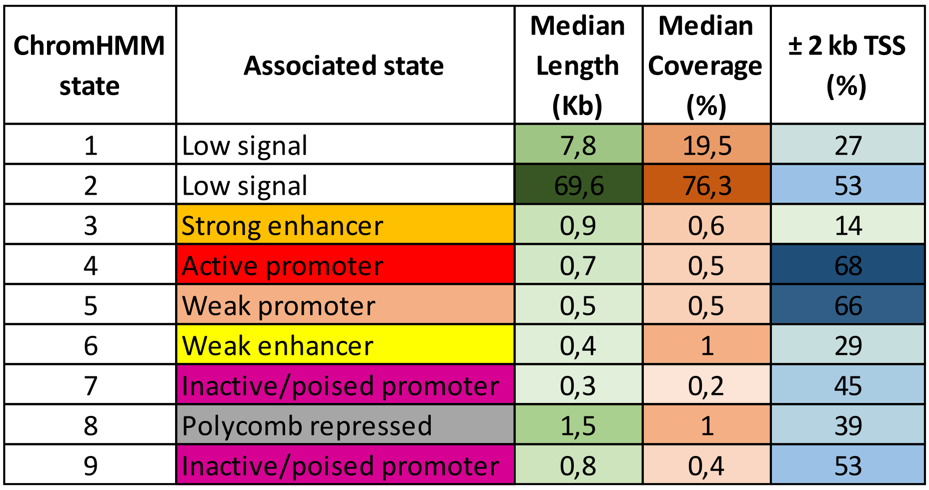
Once the models have been generated, the frequency matrices with Excel were also generated, providing insights into which combination produces the best results. Comparing the distribution of the modifications with the reference ones (Figure 5) I associated a known state to each state generated by ChromHMM to evaluate the tool's ability to recognize the main ones. Just as occurred for K562, models with too few states are not very informative, so they were discarded. Despite of the few input modifications, models with a high number of states were well defined. Thus, as a reference for subsequent analyses I opted to select the 9-state model (Table 4).



ChromHMM state	Associated state	H3K27me3	H3K4me2	H3K4me3	H3K27ac
1	Low signal	1%	0%	0%	2%
2	Low signal	0%	0%	0%	0%
3	Strong enhancer	1%	50%	7%	62%
4	Active promoter	1%	99%	98%	95%
5	Weak promoter	2%	98%	96%	9%
6	Weak enhancer	2%	66%	5%	8%
7	Inactive/poised promoter	88%	90%	4%	2%
8	Polycomb repressed	71%	0%	1%	1%
9	Inactive/poised promoter	94%	96%	100%	16%

Table 4: The emission table of the states generated by ChromHMM. Despite the few input changes the tool identifies seven different known states. The missing states are those identified by the presence of not chosen modifications, so their absence is consistent with the study. Main functional elements as promoters, enhancers and repressors are identified then the analysis allows to characterize the biological activity of the cell despite the low number of inputs

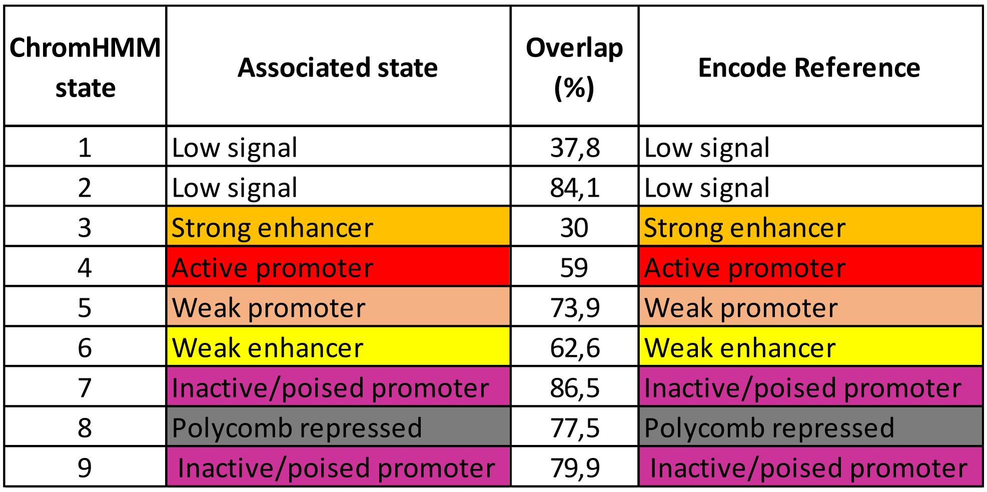
Compared to the previous analysis there are fewer known states, but this is caused by the lack in input of the modifications that characterize them. It is instead interesting to observe how, despite the low number of inputs, ChromHMM manages to find many of the main states accordingly. Moreover, in these results is present the "Inactive/poised promoter" state that does not appear in K562 results and this increases the value of the data obtained. Furthermore, analysis of Coverage, Median Length and ± 2 kb TSS were also conducted for this cell line data. Results are summarized in Table 5, they are similar to those of K562 analysis so even with a very different and limited input, the genome is correctly subdivided into states. Similarly to previous results, two "Low signal" states have high value in all three columns because these states are associated to every DNA region with unrecognized pattern so they contain a wide variety of elements. Moreover, in the ± 2 kb TSS column the highest value is for the two "Promoter" states (active and weak) and this is a result that matches with the biological role of these regions. The new states identified as "Inactive/poised promoter" have an elevate ± 2 kb TSS value too and this confirms their correct identification.



ChromHMM state	Associated state	Median Length (Kb)	Median Coverage (%)	± 2 kb TSS (%)
1	Low signal	7,8	19,5	27
2	Low signal	69,6	76,3	53
3	Strong enhancer	0,9	0,6	14
4	Active promoter	0,7	0,5	68
5	Weak promoter	0,5	0,5	66
6	Weak enhancer	0,4	1	29
7	Inactive/poised promoter	0,3	0,2	45
8	Polycomb repressed	1,5	1	39
9	Inactive/poised promoter	0,8	0,4	53

Table 5: This table summarizes the results of the genome analysis for each state. The values for some states are similar to those obtained for K562 therefore even with a smaller number of modifications and stopping at a lower number of states we can have correct results. The most significant data is the high value in ± 2 kb TSS column for all the states associated to a promoter region

Similarly to K562 analysis, the next control has been the comparison with the results of the analysis on the same cell line made with ChromHMM on Encode. Using *bedtoools intersect* I related each state of my analysis to those of Encode to verify the correlation between the states I associated and the correspondents on Encode. Table 6 summarizes these relationship percentages. In this case the percentages are very high, and this confirms the consistency of the results. Graphically I can see how the corresponding states fit together with the graphs made with *Alluvial*, which show how most of the regions associated with a state in my analysis are associated to the same state in Encode partition. Some examples of these graph are shown in Figure 8 to highlight the great correspondence between associated states.



ChromHMM state	Associated state	Overlap (%)	Encode Reference
1	Low signal	37,8	Low signal
2	Low signal	84,1	Low signal
3	Strong enhancer	30	Strong enhancer
4	Active promoter	59	Active promoter
5	Weak promoter	73,9	Weak promoter
6	Weak enhancer	62,6	Weak enhancer
7	Inactive/poised promoter	86,5	Inactive/poised promoter
8	Polycomb repressed	77,5	Polycomb repressed
9	Inactive/poised promoter	79,9	Inactive/poised promoter

Table 6: Summary table of the comparison with the Encode reference where the percentage of correlation between corresponding states is shown. Except for first "Low signal" state and "Strong enhancer" state all the other correlations are higher than 50% so these results confirm the correct performed association.

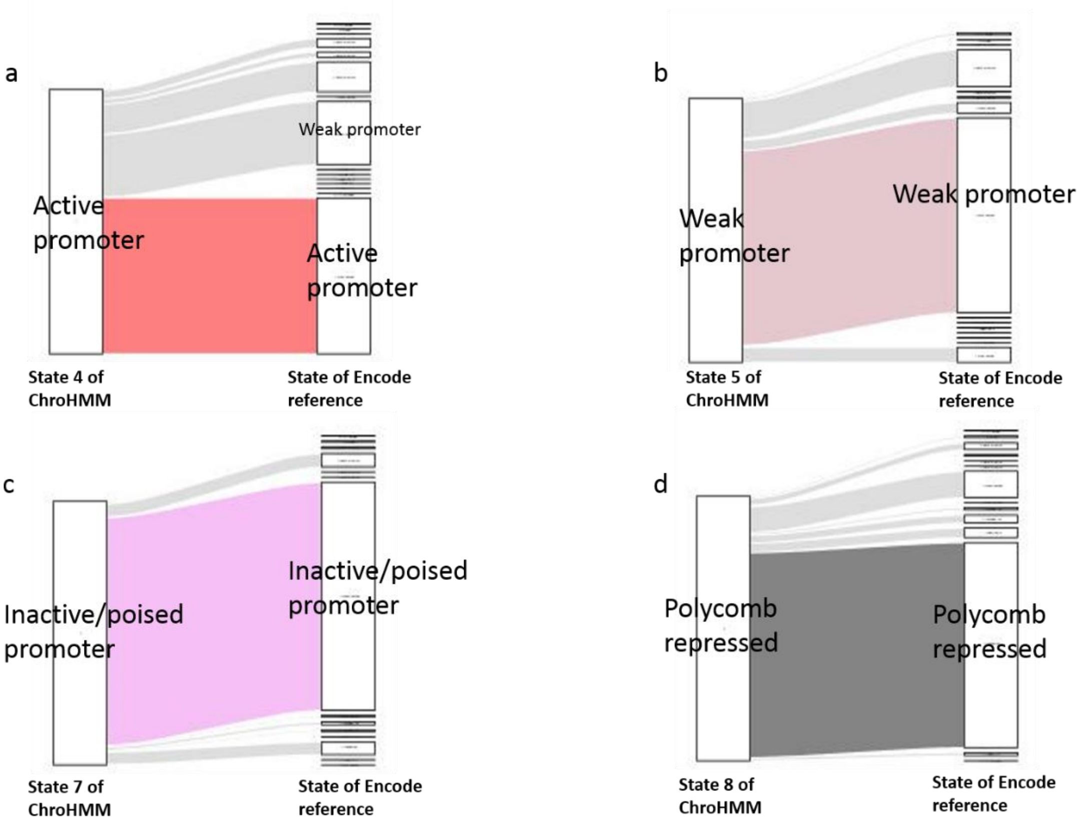


Figure 8: Four Alluvial graphs showing the correspondence between retrieved states in my analysis and the Encode reference. In each graph the major overlap is colored to highlight the states that are more similar. From these results the good quality of the performed analysis is observed: the states that match most are those that have the same biological function.“Active promoter” state (**a**) has the greatest association with the corresponding state but he is quite fragmented: however the second major association is with a promoter state so the difference between the results is not excessive. “Weak promoter” state (**b**) is more fragmented but the prevalent association is the correct one. Finally, “Inactive/poised promoter” state (**c**) and “Polycomb repressed” state (**d**) show a large degree of overlap.

4.3 Combined analysis K562 + H1-hESC

One possibility offered by ChromHMM is to generate chromatin states giving as input the modifications on more than one cell line. This should improve the quality of the results because the tool has more elements to use. To do this I set as input the BAM files of the same 4 modifications (H3K27me3, H3K27ac, H3K4me1, H3K4me3) for the two cell lines already analyzed.

4.3.1 ChromHMM results

In this type of analysis, the same files as the previous analyses will be generated, however BED files containing the division of the genome into states are autonomously generated for both K562 and H1-hESC. As previously described, chromatin states generated via ChromHMM were labeled according the reference table. In this study the preferred model was the one featuring 9 different states. The result is shown in Table 7 in which, similar to H1-hESC results, all the main states are identified despite the low number of inputs, confirming the tool efficacy.

ChromHMM state	Associated state	H3K27me3	H3K4me1	H3K4me3	H3K27ac
1	Inactive/poised promoter	91%	38%	27%	4%
2	Polycomb repressed	17%	0%	0%	0%
3	Low signal	0%	0%	0%	0%
4	Low signal	0%	1%	0%	1%
5	Strong enhancer	2%	93%	19%	55%
6	Weak enhancer	2%	52%	1%	6%
7	Strong enhancer	11%	98%	98%	93%
8	Active promoter	1%	7%	98%	93%
9	Weak promoter	5%	42%	81%	9%

Table 7: The emission table of the states generated by ChromHMM by combining both K562 and H1-hESC data. The identified states are the main known states and are clearly distinguished.

With the new BED files generated by this combined analysis I performed the analysis of Coverage, Median Length and ± 2 kb TSS for both cell lines (Table 8). Analyzing the results, it is evident how the differences are even more marked than in the analyses on single cell lines. The “Active Promoter” state in both cells stands out in the column of proximity to TSS, followed by the other promoter regions: this reinforces the quality of the results. Both “Low signal” states have high values in all the columns making these results are similar to those obtained for K562 and H1-hESC analyzed individually.

		K562			H1-hESC		
ChromHMM state	Associated state	Median Length (kb)	Median Coverage (%)	± 2 kb TSS (%)	Median Length (kb)	Median Coverage (%)	± 2 kb TSS (%)
1	Inactive/poised promoter	1,2	0,3	22	2,7	1,1	51
2	Polycomb repressed	12,5	13,2	32	3,0	1,3	34
3	Low signal	92,3	60,7	41	78,8	47,6	48
4	Low signal	6,1	19,5	23	9,3	44,5	22
5	Strong enhancer	0,6	1,1	18	0,6	0,6	15
6	Weak enhancer	0,8	3,2	16	0,7	3,8	13
7	Strong enhancer	0,8	0,9	42	0,5	0,2	53
8	Active promoter	1,1	0,6	73	0,9	0,3	81
9	Weak promoter	0,6	0,5	47	0,7	0,6	68

Table 8: The table summarizes the results of the genome analysis for both K562 and H1-hESC. The values are similar in the two lines so both samples are a good point of reference

Analyses of comparison with BED files of the Encode analysis were also performed. Table 9 shows these results: in most cases there is an elevated correlation between correspondent states, which is not influenced by the simultaneous use of the two samples (as compared to single sample analysis). There are three values less than 50% for K562 and four for H1-hESC while in the results of only H1-hESC there are only two states with a similar value. In some states there are significant differences between the two lines: for examples in the “Inactive/poised promoter” state the percentage difference is equal to 33.4% and for the second “Strong enhancer” state this difference is equal to 45.3%. Therefore, the combined use of multiple samples does not improve the quality of the results.

ChromHMM state	Associated state	K562 Overlap (%)	H1-hESC Overlap (%)	Encode Reference
1	Inactive/poised promoter	18,0	51,4	Inactive/poised promoter
2	Polycomb repressed	51,1	48,7	Polycomb repressed
3	Low signal	92,7	91,5	Low signal
4	Low signal	35,3	62,4	Low signal
5	Strong enhancer	41,4	30,4	Strong enhancer
6	Weak enhancer	52,5	42,0	Weak enhancer
7	Strong enhancer	71,0	25,7	Strong enhancer
8	Active promoter	87,9	75,9	Active promoter
9	Weak promoter	50,7	68,0	Weak promoter

Table 9: In this table the main results of the analysis with *bedtools intersect* are summarized with the correlation percentages between corresponding states for both cell lines. It is interesting to observe how in some states there is a great difference in the overlap changing the cell line

5. ChromHMM analyses on public data

5.1 Liver

The first sample I used to test ChromHMM on public data with no available ChromHMM results are liver cells. BAM files with ChIP-Seq data regarding many modifications were downloaded by Encode (https://www.encodeproject.org/matrix/?type=Experiment&status=released&files.file_type=fastq&files.file_type=bam) and I used data of 4 modifications because with previous analyses I have already verified the correct tool activity when this type of input is used. Selected input modifications are: H3K27me3, H3K27ac, H3K4me1, H3K4me3.

5.1.1 ChromHMM results

I generated models from 6 to 15 states, and after that I checked the frequency matrices, to investigate association between known states and those generated by the tool. The association is more difficult than in the previous cell lines because the numeric correspondence between my combinations and reference ones defined by Encode are less consistent and already at combinations with low number of states some of them do not clearly match with a known state. Consequently I cannot use a 9-state model (as in H1-hESC and in combined study) because of the low number of different states identified: I have to use as reference the model with 11 states (Table 10) in which there is a better distribution of the modifications patterns and more different states are present. Some of these are clearly identifiable (as "Polycomb repressed" or "Active promoter"), while for some other unknown states the association is difficult and not unique; however, despite of the low input number, I can associate every generated state with a known state.

ChromHMM state	Associated state	H3K27 me3	H3K4 me1	H3K4 me3	H3K27 ac	Median length (kb)	Median coverage (%)	± 2 kb TSS (%)
1	Strong Enhancer	1%	13%	1%	48%	1,0	1,0	16
2	Strong enhancer	0%	96%	11%	89%	0,8	1,4	14
3	Low Signal	0%	1%	0%	0%	6,0	17,8	21
4	Weak enhancer	0%	76%	1%	4%	0,7	3,0	15
5	Weak promoter	6%	89%	86%	7%	0,5	0,5	50
6	Strong Enhancer	1%	100%	99%	98%	0,6	0,8	51
7	Active Promoter	1%	3%	96%	91%	0,7	0,5	67
8	Weak Promoter	1%	0%	55%	0%	2,3	0,9	23
9	Low Signal	0%	0%	0%	0%	68,2	58,4	46
10	Low Signal	3%	0%	0%	0%	31,0	15,4	47
11	Polycomb repressed	39%	17%	2%	1%	2,0	0,4	45

Table 10: 11 state model summary table. Here all the associations with known states are reported with also the results of the numerical analysis concerning genome occupation. Also in this case higher ± 2 kb TSS value is those of the "Active promoter" state and this is coherent with his biological role

To be sure of the correctness of these associations I performed the same statistical studies on the genome executed for the previous cell lines and following controls reinforce the meaning of these pairings. The three "Low signal" states have high values in all the columns, the highest ± 2 kb TSS value is associated to "Active promoter" state and also "Weak promoter" state has an elevated value.

5.1.2 Gene ontology validation

To assess the quality of the obtained results, I relied on gene ontology analysis: for a gene list I found all the annotation information related to many different arguments as biological process, molecular function, cellular component, pathway and many others. This is a really significative analysis because it allows verifying if the regions that I have associated with a determined state really have this role. To perform this study, I used a list in which the TSS positions and their associated gene name were reported. Comparing this file with the BED file obtained from ChromHMM containing the association between every region and its chromatin state, it was possible to obtain a list with the names of the associated genes for every state. Different states need different analysis because some elements as Promoters or Polycomb repressed regions are close to genes they act on, while enhancers regulate genes distant from their position. Then for "Active promoter" and "Polycomb repressed" states my focus was on annotation of overlapping genes with a minimum overlap of 30%, via the *bedtools intersect* function. This function compares two BED files and shows the results of the overlapping regions. Instead enhancers do not act on nearby genes so for Enhancer states I used *bedtools closest* function to report the nearest correspondence with TSS regions with a maximum distance of 100 kb. Using these functions, I generated lists of genes associated with different regions which were used for the gene ontology studies. Gene ontology studies were carried out using ToppGene, a web portal for gene list enrichment analyses that helped me to retrieve many different information about calculated chromatin states. Gene lists previously mentioned were set as input in ToppGene and analyses were performed using default parameters. ToppGene reports several results tables (Biological Process, Pathway, Molecular Function and so on). For the purpose of my validation, results featured in the "Biological Process" and "Pathway" sections of obtained results were further inspected. Table 11, 12 and 13 report the results of these analyses with biological processes and pathways; it also shown their p-value (calculated by ToppGene) which demonstrates the significance of the data.

- ***ACTIVE PROMOTER***

a

Biological Process	p-value
Positive regulation of RNA metabolic process	1.01E-26
Positive regulation of gene expression	3.24E-25
Positive regulation of macromolecule biosynthetic process	5.33E-25
Regulation of transcription by RNA polymerase II	2.41E-24
Positive regulation of RNA biosynthetic process	2.82E-23
Positive regulation of nucleic acid-templated transcription	3.68E-23

b

Pathway	p-value
Gene Expression	1.04E-11
mRNA Splicing - Major Pathway	6.44E-10
mRNA Splicing	3.27E-9
Circadian Clock	1.10E-8
Metabolism of proteins	2.33E-7
Cellular responses to stress	6.63E-7

Table 11: These tables show the gene ontology results for the gene associate to "Active promoter" state. Most significative biological processes (**a**) and pathways (**b**) are displayed

The table on the left shows the most significant processes associated to genes under the control of active promoters: they are all housekeeping processes [22] and their p-value indicate that the association is strong. The results are similar for the table on the right which shows the pathways associated with active promoters and also these pathways refer to housekeeping ways. These are constitutive processes and pathways that are required for the maintenance of basic cellular function. It is correct that they are related to "Active promoter" regions because this element's function is to activate the genes under its control and all these housekeeping genes need to be constitutively activated.

- ***POLYCOMB REPRESSED***

a

Biological Process	p-value
Neuron differentiation	3.12E-56
Generation of neurons	3.90E-56
Central nervous system development	5.20E-55
Head development	1.55E-43
Brain development	2.79E-43
Central nervous system neuron differentiation	5.67E-43

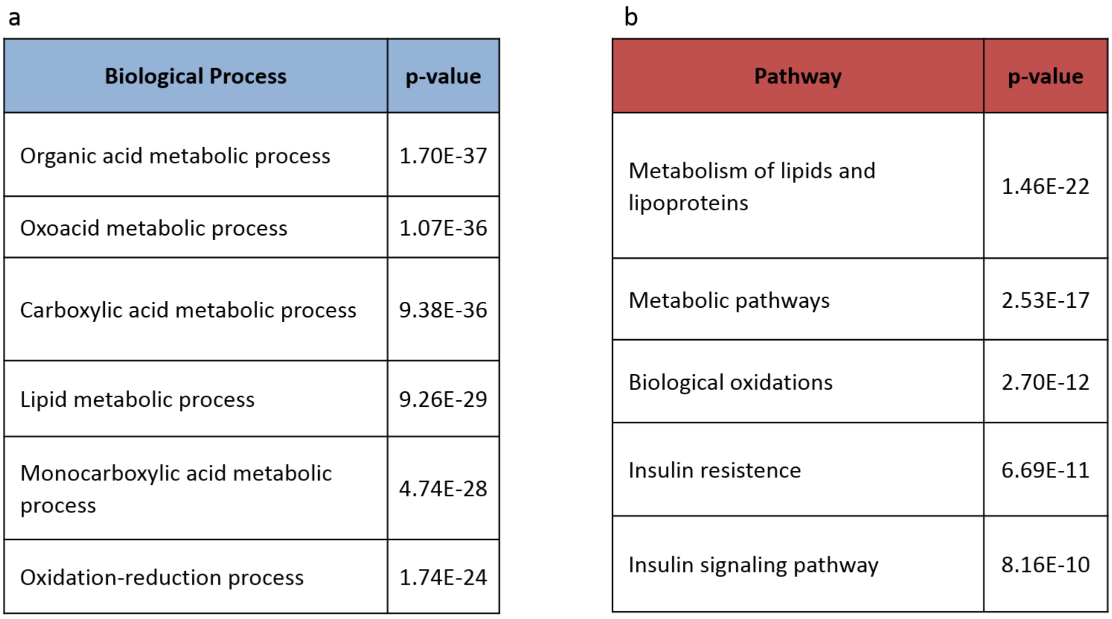
b

Pathway	p-value
Neuronal System	7.11E-25
Cardiac conduction	3.34E-10
Muscle contraction	1.45E-8
Developmental Biology	3.11E-8

Table 12: These tables show the gene ontology results for the gene associate to “Polycomb repressed” state. Most significative biological processes **(a**) and pathways (**b**) are displayed

Both the most significant processes and the most significant pathways associated to genes under Polycomb repressor control have no relation to the cellular activity of the liver. They refer to neuronal, cardiac and muscular function which have to be repressed in liver cells, so it is correct that they are associated with “Polycomb repressed” state.

- ***STRONG ENHANCER***



a

Biological Process	p-value
Organic acid metabolic process	1.70E-37
Oxoacid metabolic process	1.07E-36
Carboxylic acid metabolic process	9.38E-36
Lipid metabolic process	9.26E-29
Monocarboxylic acid metabolic process	4.74E-28
Oxidation-reduction process	1.74E-24

b

Pathway	p-value
Metabolism of lipids and lipoproteins	1.46E-22
Metabolic pathways	2.53E-17
Biological oxidations	2.70E-12
Insulin resistence	6.69E-11
Insulin signaling pathway	8.16E-10

Table 13: These tables show the gene ontology results for the gene associate to “Strong enhancer” state. Most significative biological processes **(a**) and pathways (**b**) are displayed.

Both the biological processes and pathways enriched in genes associated to strong enhancers refer to fundamental processes and pathway in liver cell. Metabolic processes are the ones that mostly characterize liver cell activity [23] and also insulin pathways are very specific of these cell type [24]. They are associated to strong enhancers regions and this is correct because enhancers increase the transcription of genes under their control.

5.2 Lung

I performed the same kind of analysis conducted on liver on lung cells to test the tool on more than one sample and to have more insights about its performances. Input files were BAM type of ChIP-Seq experiments and were downloaded from Encode (https://www.encodeproject.org/matrix/?type=Experiment&status=released&files.file_type=fastq&files.file_type=bam) choosing the same four modifications used previously: H3K27me3, H3K27ac, H3K4me1, H3K4me3.

5.2.1 ChromHMM results

Below the results of the analyses conducted on lung cells are presented. In this study the best model proved to be the one with 11 states (Table 14), because previous combinations contained many states which were difficult to associate. Also, informative chromatin states were clearly defined only in models composed by a high number of states. An interesting observation is the appearance of a new chromatin state called “Inactive enhancer” but some unknown states are very difficult to associate with known ones so there are more problems in subsequent analysis. For example, it is possible to individuate only one promoter state while three combinations are “Low signal” states which do not give much information.

ChromHMM state	Associated state	H3K27 me3	H3K4 me1	H3K4 me3	H3K27 ac	Median length (kb)	Median coverage (%)	± 2 kb TSS (%)
1	Low Signal	0%	0%	0%	0%	83,1	58,5	55
2	Low Signal	2%	0%	1%	0%	19,5	13,5	31
3	Polycomb repressed	21%	3%	3%	3%	3,8	3,1	24
4	Inactive enhancer	42%	60%	26%	19%	1,0	0,3	31
5	Weak enhancer	2%	49%	6%	24%	0,6	1,5	12
6	Low Signal	0%	1%	1%	1%	5,1	18,0	21
7	Strong enhancer	1%	73%	8%	95%	0,7	1,5	17
8		0%	5%	1%	27%	0,9	2,4	16
9		2%	16%	15%	22%	0,4	0,5	40
10	Strong Enhancer	4%	87%	96%	92%	0,5	0,2	71
11	Active Promoter	1%	4%	96%	94%	0,9	0,5	78

Table 14: 11 state model summary table. Here all the associations with known states are reported with also the results of the numerical analysis concerning genome occupation

Statistical studies on the genome show that “Active promoter” state is the one with the highest value in the ±2 kb TSS column and this coincides with the results of the previous analyses, also one “Strong enhancer” state has an elevate value in this column and this is different than the studies performed so far.

5.3 Common Myeloid Progenitor

As a further analysis, I generated chromatin states using as sample common myeloid progenitor cells. To perform this study I used BAM files with ChIP-Seq data downloaded from Encode (https://www.encodeproject.org/matrix/?type=Experiment&status=released&files.file_type=fastq&files.file_type=bam) choosing the same four modifications used previously: H3K27me3, H3K27ac, H3K4me1, H3K4me3.

5.3.1 ChromHMM results

The results are very similar to those obtained for lung cells. Also in this study I chose selected the model with 11 states as reference (Table 15) because the previous combinations were not very informative but also in the chose model there are two unknown states that cannot be associated to any known state. There is a great presence of H3K4me1 modification therefore many states are associated with enhancer condition, in fact the new "Inactive enhancer state" is present in two unknow states. Inactive enhancers are DNA regions with enhancer activity which in a certain situation are inactivated to avoid the activation of the genes under their control [25].

ChromHMM state	Associated state	H3K27 me3	H3K4 me1	H3K4 me3	H3K27 ac	Median length (kb)	Median coverage (%)	± 2 kb TSS (%)
1		0%	0%	2%	44%	1,4	1,3	28
2	Active Promoter	0%	0%	96%	95%	1,1	0,6	68
3		0%	0%	77%	10%	0,6	0,3	62
4	Inactive enhancer	40%	0%	62%	8%	0,8	0,1	57
5	Polycomb repressed	39%	0%	0%	0%	3,6	2,3	33
6	Low Signal	0%	0%	0%	0%	22,3	90,4	31
7	Inactive enhancer	43%	82%	2%	0%	1,1	0,2	28
8	Weak enhancer	0%	81%	0%	0%	1,6	4,5	12
9	Strong enhancer	0%	88%	3%	52%	1,0	0,2	25
10	Strong Enhancer	0%	85%	96%	95%	0,7	0,1	68
11	Weak promoter	4%	84%	79%	11%	0,5	0,05	62

Table 15: 11 state model summary table. Here all the associations with known states are reported with also the results of the numerical analysis concerning genome occupation.

Subsequent genome analyses were similar to those performed for lung cells. There were two states with an high ±2 kb TSS value: one of them is "Active promoter" and this is coherent with its biological function while the other is a "Strong enhancer" state and this does not coincide with its role.

6. Conclusions

In conclusion, during my thesis I used ChromHMM, a software to infer chromatin states from ChIP-Seq data. The results obtained in the validation analyses demonstrated the ability of the tool to operate correctly with different setups: both results on K562 and on H1-hESC provided an excellent characterization of cellular activity. Moreover, despite the limited number of inputs in the H1-hESC analysis, ChromHMM was able to sustain high quality in the results and managed to identify the main chromatin states. In addition, comparisons with the analyses performed by Encode proved how ChromHMM managed to divide the genome in states according to expectations (as defined by the gold standard analysis provided by Encode). This comparison showed a very similar association between calculated and expected chromatin states despite different numbers of histone modifications used as input. A further confirmation of the reliability of the results obtained was given by the analysis of proximity with the TSS: in all cases the highest value was associated with the "Active promoter" state, which is an important indicator supporting the partition of the genome operated by ChromHMM. Conversely, analyses performed on public data gave heterogeneous results: on one side, the identification of states from liver cells data provided good results while in the case of lung cells and common myeloid progenitors the analysis resulted more problematic. In fact, in liver cells, despite a reduced number of chromatin modification as input, it was possible to obtain an excellent characterization of the genome, also supported by the gene ontology results. The pathways and biological processes associated with the inferred states were consistent with their function, and the p-values indicated a significant correlation between these elements. In case of liver cells and common myeloid progenitor, it was more difficult to interpret the results, due to the low number of calculated chromatin states and intrinsic difficulty in associating them to known, reliable ones. This might be due to the quality of ChIP-Seq data used as input to ChromHMM. In conclusion, ChromHMM proved to be a tool suitable for the study of chromatin states on various cell data of different type. What emerged is that the tool does not necessarily require a large number of inputs but manages to operate even starting from a few data, making it valuable also for those experiment featuring a limited number of investigated histone modifications.

7. References

[1] Baranello, L., Levens, D., Gupta, A. & Kouzine, F. The importance of being supercoiled: How DNA mechanics regulate dynamic processes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1819**, 632–638 (2012).

[2] Arents, G. & Moudrianakis, E. N. Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. *PNAS* **90**, 10489–10493 (1993).

[3] Clark, D. J. & Felsenfeld, G. Formation of nucleosomes on positively supercoiled DNA. *The EMBO Journal* **10**, 387–395 (1991).

[4] Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology* **15**, 703–708 (2014).

[5] Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

[6] Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).

[7] Villar-Garea, A. & Imhof, A. The analysis of histone modifications. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1764**, 1932–1939 (2006).

[8] Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics* **9**, 179–191 (2008).

[9] Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669–680 (2009).

[10] Sambrook, J. & Russell, D. W. Fragmentation of DNA by Sonication. *Cold Spring Harb Protoc* **2006**, pdb.prot4538 (2006).

[11] Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31–46 (2010).

[12] Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135–1145 (2008).

[13] Jiang, S. & Mortazavi, A. Integrating ChIP-seq with other functional genomics data. *Brief Funct Genomics* **17**, 104–115 (2018).

[14] Jason Ernst *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

[15] Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

[16] Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature *Genetics* **39**, 311–318 (2007).

[17] Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492 (2017).

[18] Ernst, J. & Kellis, M. ChromHMM Version 1.19 *User Manual*. **24**

[19] Project Overview – ENCODE. Available at: https://www.encodeproject.org/help/project-overview/.

[20] Andersson, L. C., Nilsson, K. & Gahmberg, C. G. K562—A human erythroleukemic cell line. *International Journal of Cancer* **23**, 143–147 (1979).

[21] Kim, T. H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

[22] Chang, C.-W. et al. Identification of Human Housekeeping Genes and Tissue-Selective Genes by Microarray Meta-Analysis. *PLoS ONE* **6**, e22859 (2011).

[23] Nguyen, P. et al. Liver lipid metabolism. *Journal of Animal Physiology and Animal Nutrition* **92**, 272–283 (2008).

[24] KruppS, M. N. & Lane, M. D. Evidence for Different Pathways for the Degradation of Insulin and Insulin Receptor in the Chick Liver Cell. 7

[25] Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).