
# Analysis of a collection of ChIPseq experiments with the ChromHMM program

**Mar González Ramírez**
Màster Bioinformàtica i Bioestadística
TFM-Estadística i Bioinformàtica

**Enrique Blanco García**
**Carles Ventura Royo**

06/06/2017

## FITXA DEL TREBALL FINAL

| | |
|---|---|
| **Títol del treball:** | Analysis of a collection of ChIPseq experiments with the ChromHMM program |
| **Nom de l'autor:** | *Mar González Ramírez* |
| **Nom del consultor/a:** | *Enrique Blanco García* |
| **Nom del PRA:** | *Carles Ventura Royo* |
| **Data de lliurament (mm/aaaa):** | *06/2017* |
| **Titulació o programa:** | *Màster Bioinformàtica i Bioestadística* |
| **Àrea del Treball Final:** | TFM-Estadística i Bioinformàtica |
| **Idioma del treball:** | *Anglès* |
| **Paraules clau** | *Chromatin segmentation methods, Polycomb, epigenetics* |


**Abstract (in English, 250 words or less):**

The volume of NGS information that is available for the scientific community to study gene regulation is constantly growing, making the analysis of such data more complex. Here we explored the use of chromatin segmentation methods, such as ChromHMM, to solve the problem in particular scenarios. Thus, we obtained different chromatin segmentation models focusing on the Polycomb group of proteins (PcG) and post-translational histone modifications in mouse Embryonic Stem Cells (mESC). We aimed to answer biological questions such as the meaning of intergenic H3K27me3 regions, or the PcG subunits complexity. Moreover, since PcG regulates gene expression during development we applied the same approach during CardioMyocite (CM) differentiation, giving a first insight into this process. The goal of this comparison should be identifying those regions which evolve to a different state during differentiation from mESC to CM. Finally, we concluded that the potential of the chromatin state models lies on the study of those states which help to solve these biological questions, rather than building a comprehensive model which describes the whole epigenomic complexity.

# Index

## List of figures

## List of supplementary figures

## List of tables

# 1. Introduction

## 1.1 Context and justification of the work

The volume of NGS (New Generation Sequencing) information that is available for the scientific community is constantly growing, which makes the analysis of such data more complex. Chromatin segmentation methods are one alternative to overcome this problem. ChromHMM (Ernst and Kellis 2012) is the standard tool in the field, which in fact, has been used widely by the ENCODE consortium to generate chromatin segmentation maps (Ernst et al. 2011). However other methods have also been developed with the same purpose.

Chromatin segmentation methods in general, take as input multiple sources of ChIPseq data and generate a division of the genome into segments that belong to a specific chromatin state. The signature of chromatin states is defined by the particular configuration of ChIPseq features that are present on each segment (Figure 1). In the case of ChromHMM, state emissions are defined by probability to find a specific mark in the regions which belong to that particular state (Ernst and Kellis 2010). Moreover, it is important to take into account that one gene can be segmented into different states as can be seen in Figure 1.

In our lab we are focused on the epigenetic events involving cancer, differentiation and development. Many different ChIPseq data of histone modifications and Polycomb group of proteins (PcG) have been generated to study gene regulation in these scenarios. Therefore, we would like to extract novel knowledge among all that data using chromatin segmentation methods.

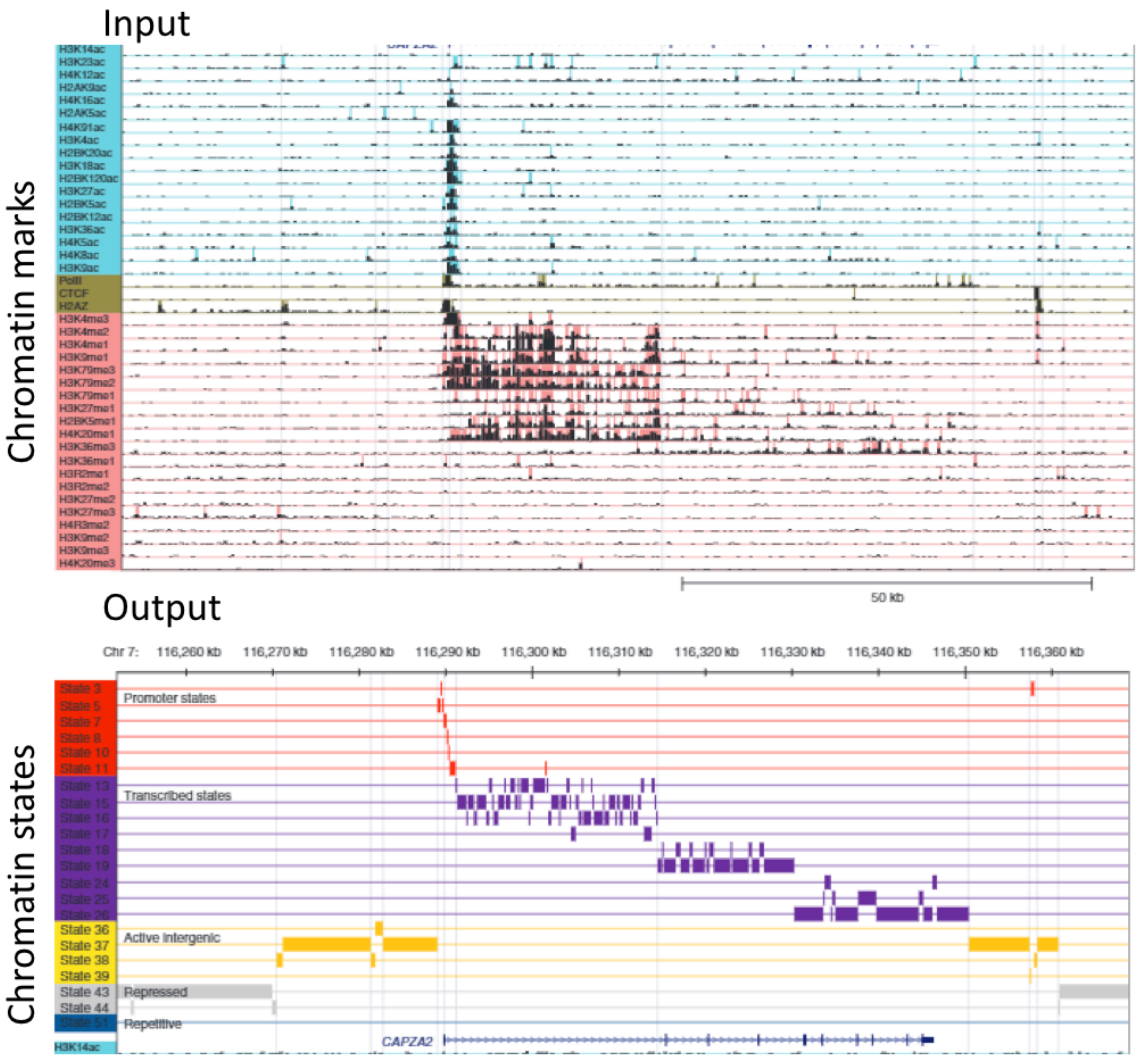



***Figure 1*: Input and output of ChromHMM.**

(Top) This program analyses multiple NGS experiments.

(Bottom) The genome is segmented into different states according to the NGS data.

(adapted from Ernst and Kellis 2010).

Each cell type is governed by a specific gene expression programme. Gene regulation occurs due to the combined action of transcription factors and epigenetic marks, such as DNA methylation and post-translational histone modifications. Two well-known histone modifications are trimethylation of lysine 4 in histone H3 (H3K4me3), which is associated to transcriptionally active genes, and trimethylation of lysine 27 in histone H3 (H3K27me3), which decorates the promoters of transcriptionally inactive genes. The proteins responsible of depositing these two marks are Trithorax Group of proteins (TrxG) and Polycomb Group of proteins (PcG) respectively (Figure 2).

Despite PcG complexes deposit repressive marks, in Embryonic Stem Cells (ESC) they also colocalize with the active mark H3K4me3 in certain promoters called bivalent promoters (Voigt, Tee et al. 2013, Harikumar and Meshorer 2015). These promoters belong to developmental and lineage-specific genes (Voigt, Tee et al. 2013, Harikumar and Meshorer 2015). During differentiation, only the correct lineage-specific genes will become active, and therefore, will be marked only by H3K4me3, and all the other lineage-specific genes will remain silent, and thus decorated by H3K27me3 alone (Figure 2).

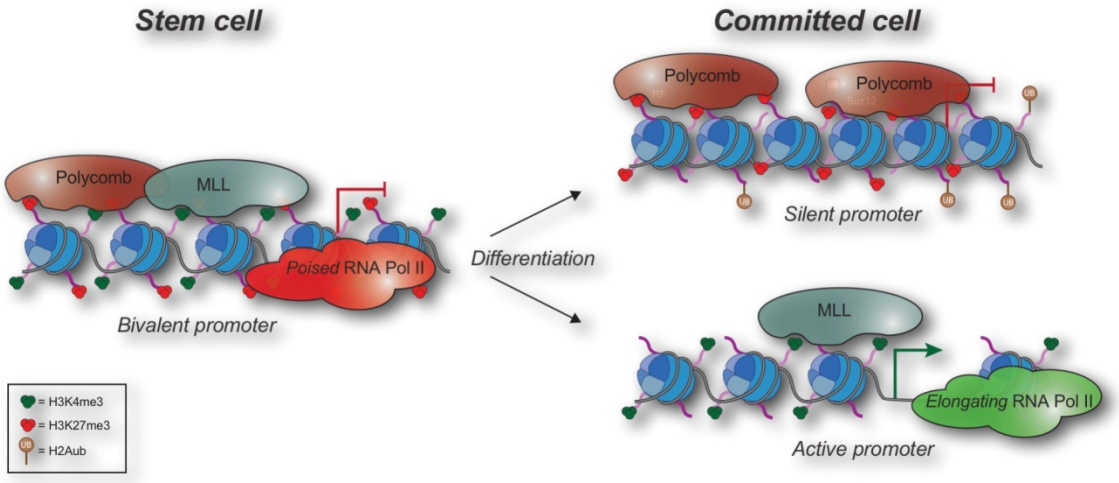



***Figure 2:*** **Gene regulation during differentiation**.
(Left) Bivalent genes are ready to be expressed. Their promorer is marked by H3K4me3 and H3K27me3.
(Right) Along cell differentiation, some genes retain PcG-H3K27me3 to be silent or TrxG-H3K4me3 to be expressed.
(adapted from Di Croce and Helin 2013)

PcG are classified into two main complexes, which are Polycomb Repressive Complexes 1 and 2 (PRC1 and PRC2), and regulate gene expression during stem cell fate decisions in early embryonic development (Aranda, Mas et al. 2015, Pasini and Di Croce 2016). Moreover, their misregulation has been linked to cancer (Pasini and Di Croce 2016).

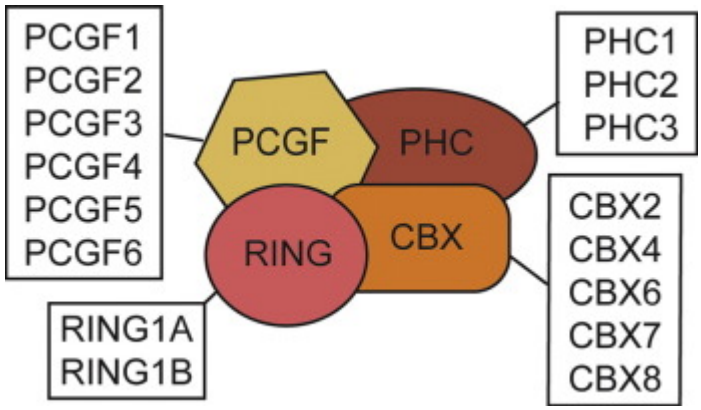



***Figure 3*: PRC1 core components ad their distinct subunits**.
(adapted from Connelly and Dykhuizen 2017).

PRC1 composition is diverse and can be divided into canonical and non-canonical complexes. The canonical PRC1 (Figure 3) contains four core components which are RING1A/B (which is the catalytic subunit), PCGF1-6, PHC1-3, and a member of the CBX family (Connelly and Dykhuizen 2017). The two main complexes of PRC1 present in mouse embryonic stem cells (mESC) are the canonical PRC1-Cbx7 and the non-canonical PRC1-Rybp, which have different biological function (Morey, Aloia et al. 2013).

On the other hand, PRC2 (Figure 4) composition is less complex. Its core components are EZH1/2 (which is the catalytic subunit), SUZ12, EED, and RBBP/7, and its variability comes from the associated factors which modulate its activity (Vizan, Beringer et al. 2015). Two associated PRC2 factors present in mESC are Jarid2 and Epop (previously known as C17orf96), which bind to PRC2 core components and are mutually exclusive (Beringer, Pisano et al. 2016). In fact, PRC2-Jarid2 and PRC2-Epop have different function and different activity, being the first one more repressive (Beringer, Pisano et al. 2016).

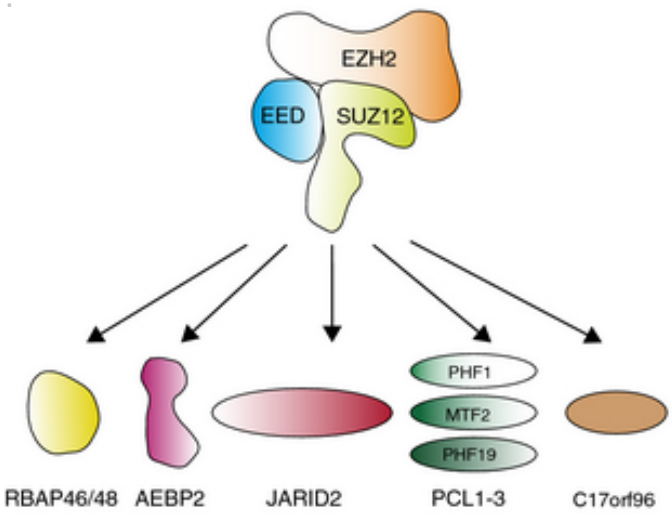



***Figure 4*: PRC2 core components and its associated factors.**
(adapted from Vizán 2015)

The catalytic activity of PRC1 is the monoubiquitination of lysine 119 in histone H2A (H2AK119ub1) and the catalytic activity of PRC2 is H3K27me3 (Morey and Helin 2010). Both complexes colocalize to maintain gene repression, which consists in PRC2 depositing H3K27me3 which serves as a docking site for the canonical PRC1, and then PRC1 deposits H2AK119ub1 (Pasini and Di Croce 2016). However, other non-canonical PRC1 complexes such as PRC1-Rybp, have been shown to be recruited in a PRC2-independent manner (Tavares, Dimitrova et al. 2012).


***Figure 5:* Process of differentiation from Embryonic Stem cells to Cardiomyocytes**.
(adapted from Wamstad, Alexander et al. 2012).

An example of differentiation can be found in Figure 5. Embryonic Stem Cells (ESC) express pluripotency markers such as *Nanog* and *Sox2*, however, during differentiation to CardioMyocytes (CM) these genes become silent, and CM markers such as *Tbx5* and *Myh6*, silent in ESC, become active (Wamstad, Alexander et al. 2012, Morey, Santanach et al. 2015).

We aim to detect new relations of histone marks and PcG components, to better understand gene regulation in mESC. Our approach consists in the use of ChromHMM to extract novel knowledge from a collection of ChIPseq samples. However, we will also explore the use of some other methods created to this purpose in addition to ChromHMM. Finally, we will implement a web server to facilitate biologists the accession to this new information.

## 1.2 Objectives of the work

- Objective 1: To explore the existing chromatin segmentation results that are available in different databases.
- Objective 2: To assess the performance of ChromHMM program on a simple scenario.
- Objective 3: To perform the genome segmentation using PRC1/PRC2 components.
- Objective 4: To explore other methods available in the literature: chromstaR, jMOSAiCS, IDEAS, etc.
- Objective 5: To evaluate the integration of two or more cell lines or conditions.
- Objective 6: To design a web server to distribute the results.

## 1.3 Approach and methodology to follow

The method used to generate chromatin segmentation maps is ChromHMM (Ernst and Kellis 2012), since it is the standard on this field. The approach followed consisted in the following steps:

1. Selection of a subset of ChIPseq samples.
2. Mapping the raw data (reads) into the mouse genome.
3. Generating chromatin state models with ChromHMM.
4. Selecting the number of states according to biological interpretation.

5. Working with the complete set of ChIPseq experiments.
6. Mapping the raw data (reads) into the mouse genome.
7. Generating chromatin state models with ChromHMM.
8. Selection of the number of states according to biological interpretation.

9. Assessment of the chromatin state maps generated by different methods or using different parameters.

10. Analysis of the most interesting states of the selected chromatin segmentation maps.
11. Implementation of a PHP web server to distribute the resulting models.

Steps 9, 10 and 11, are performed for all the models generated in the previous steps (1 to 8).

As an alternative approach, we could have also selected another method to develop this project. Despite we focused on ChromHMM, we also tried to test other available methods.

The methodology used during this project is explained in more detail during Section 2. Methodology.

## 1.4 Work Plan

All the landmarks planned at the beginning of this project, have been accomplished. They are listed below, according to the objective they belong:

- Objective 1: To explore the existing chromatin segmentation results available in different databases.
    - Description of the available data.
- Objective 2: To assess the performance of ChromHMM program on a simple scenario.
    - Basic model of chromatin segmentation in mESC using a small subset of histone marks.
    - Standard model of chromatin segmentation in mESC using a complete set of histone marks.
    - Models of chromatin segmentation using different parameters.

- Objective 3: To perform the genome segmentation using PRC1/PRC2 components.

    - Basic model of chromatin segmentation in mESC using a small subset of histone marks and a component of PRC1 or PRC2.
    - Basic model of chromatin segmentation in mESC using a small subset of histone marks and a component of PRC1 and a component of PRC2.
    - More complex model of chromatin segmentation in mESC using three subunits of PRC1 and three subunits of PRC2.
    - Complex model of chromatin segmentation in mESC using a small subset of histone marks, and three subunits of PRC1 and three subunits of PRC2.
    - More complex model of chromatin segmentation in mESC using a complete subset of histone marks, and three subunits of PRC1 and three subunits of PRC2.

- Objective 4: To explore other methods available in the literature: chromstaR, jMOSAiCS, IDEAS, etc.

    - Basic model of chromatin segmentation in mESC using a small subset of histone marks for chromstaR.
    - Complex model of chromatin segmentation in mESC using three subunits of PRC1 and three subunits of PRC2.
    - Complex model of chromatin segmentation in mESC using three subunits of PRC1 and three subunits of PRC2 using different parameters.

- Objective 5: To evaluate the integration of two or more cell lines or conditions.

    - Basic models of chromatin segmentation in two conditions (mESC and CM) using a small subset of histone marks.

- More complex models of chromatin segmentation in two conditions (mESC and CM) using a complete set of histone marks.

- Objective 6: To explore the distribution of the results on a web server.

  - PHP web server.

- Objective extra: To develop a set of scripts to compare two models and to calculate genome coverage.

  - Script to bin the bed files of the segmentation models.
  - Script to compare two chromatin segmentation models.
  - Script to calculate the genome coverage of each segmentation state.

The work plan followed during this work is described in Table 1.

***Table 1:*** **Work plan.**

| | 01/03 -15/03<br>2 weeks | | 16/03 – 22/03<br>1 week | 23/03 – 30/03<br>1 week | 31/03 – 05/04<br>1 week | 06/04 – 12/04<br>1 week | 13/04 – 19/04<br>1 week | 20/04 -26/04<br>1 week | 27/04 – 03/05<br>1 week | 04/05 -10/05<br>1 week | 11/05 – 24/05<br>2 weeks | | 25/05 – 06/06<br>2 weeks | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pla de treball | | | | | | | | | | | | | | |
| Obj 1: Study the state of the art. | | | | | | | | | | | | | | |
| Obj 2: Set up ChromHMM. | | | | | | | | | | | | | | |
| Obj 2: Obtain models. | | | | | | | | | | | | | | |
| Obj 2: Script to compare models. | | | | | | | | | | | | | | |
| Obj 3: Obtain models. | | | | | | | | | | | | | | |
| Obj 4: Explore literature to select methods. | | | | | | | | | | | | | | |
| Obj 4: Set up methods selected. | | | | | | | | | | | | | | |
| Obj 4: Generate models. | | | | | | | | | | | | | | |
| Obj 5: Generate models. | | | | | | | | | | | | | | |
| Obj 6: Implement web server. | | | | | | | | | | | | | | |
| Escriure memòria | | | | | | | | | | | | | | |
| Preparar presentació | | | | | | | | | | | | | | |
| | **PAC1** | | **PAC2** | | | **PAC3** | | | | | **PAC4** | | **PAC5** | |

## 1.5 Brief summary of the products obtained

The products which will appear during this document are:

1. Chromatin segmentation model of histone marks using ChromHMM
2. Chromatin segmentation models of histone marks using ChromHMM under different configurations
3. Chromatin segmentation model of three histone marks and two PcG subunits using ChromHMM
4. Chromatin segmentation model of six PcG subunits using ChromHMM
5. Chromatin segmentation models of six PcG subunits using chromstaR using different parameters
6. Web server implemented in PHP
7. Python script to bin one segmentation
8. Python script to compare two segmentations
9. Python script to calculate genome coverage
10. Chromatin segmentation model of two cell lines using ChromHMM

## 1.6 Brief description of the other chapters of the document

2. Methodology: description of the data and software used during the progress of this project.

3. Results: detailed description of the most relevant results obtained during the progress of this project.

4. Discussion: discussion of the results obtained, possible future work and future implications of chromatin segmentation maps.

# 2. Methodology

## 2.1 Samples

We gathered from the literature multiple ChIPseq experiments from mESC and CM. Table 2 shows the list of samples and the bibliographic source.

***Table 2:*** **List of ChIPseq samples, number of reads and reference.**

| | Name | Number of reads | Reference |
|---|---|---|---|
| **ChIPseq** | H3K4me3_mESC | 26395535 | Beringer (2016) |
| | H3K36me3_mESC | 28971448 | Beringer (2016) |
| | H3K4me1_mESC | 53852420 | Ernst et al. 2011 |
| | H3H27ac_mESC | 51528628 | Ernst et al. 2011 |
| | H3K27me3_mESC | 33884883 | Beringer (2016) |
| | H3K9me3_mESC | 17903118 | Stevens 2017 |
| | Ring1B_mESC | 9223409 | |
| | Cbx7_mESC | 12789858 | |
| | Rybp_mESC | 24176115 | |
| | Suz12_mESC | 15977198 | Beringer (2016) |
| | Epop_mESC | 26817288 | Beringer (2016) |
| | Jarid2_mESC | 26404717 | Beringer (2016) |
| | IgG_mESC | 103412062 | Beringer (2016) |
| | H3K4me3_CM | 66591524 | |
| | H3K36me3_ CM | 76841333 | |
| | H3K4me1_ CM | 86200273 | |
| | H3H27ac_ CM | 89569073 | |
| | H3K27me3_ CM | 74159031 | |
| | WCE_CM | 73299056 | |

Moreover, we gathered the following list of ATACseq and RNAseq experiments from the literature (Table 3).

*Table 3:* **List of ATACseq and RNAseq samples, number of reads and reference.**

| | | | |
|---|---|---|---|
| **RNAseq** | mESC | 198143783 | Beringer (2016) |
| | CM | 289983719 | |
| **ATACseq** | mESC | 22045164 | Mas et al. submitted |

# 2.2 ChIPseq analysis

## 2.2.1 Mapping

The sequence reads were mapped to the mm9 version of the mouse genome with the BOWTIE (Langmead, Trapnell et al. 2009) software, setting the option -m 1, which eliminates those reads which align in more than one region. The ChIPseq profiles were obtained using the function buildChIPprofile from Seqcode (Blanco et al. in preparation).

## 2.2.2 Peak calling

The peak calling was performed using MACS (Zhang, Liu et al. 2008) with the option --shiftsize 100, which shifts tags to their midpoint. The peak calling was done for the samples in Table 4, to test ChromHMM using peaks instead of reads.

*Table 4:* **Number of peaks of several ChIPseq experiments.**

| Treatment | Peaks |
|---|---|
| H3K4me3_mESC | 32433 |
| H3K36me3_mESC | 27182 |
| H3K4me1_mESC | 99142 |
| H3H27ac_mESC | 40620 |
| H3K27me3_mESC | 7703 |
| H3K9me3_mESC | 21850 |

### 2.2.3 Down-sampling

The ChIPseq data on histone modifications in mESC (H3K4me3, H3K36me3, H3K4me1, H3K27ac, H3K27me3 and H3K9me3) were all down-sampled to 17 million reads to test the effect of having the same number of reads when constructing a chromatin segmentation model with ChromHMM. The number of reads was selected according to the experiment with less reads, H3K9me3, which had 17,903,118 reads.

## 2.3 RNAseq

### 2.3.1 Mapping

The pair-end sequence reads were mapped to the mm9 version of the mouse genome with TopHat (Trapnell, Pachter et al. 2009) setting the options --mate-inner-dist 100, which is the expected mean distance between mate pairs, and -g 1, which eliminates those reads which align in more than one region. The RNAseq profiles were obtained using the function buildChIPprofile from Seqcode.

### 2.3.2 Calculating FPKMs

The FPKMs of each gene in the RefSeq catalogue (O'Leary, Wright et al. 2016) of the mouse genome were calculated using Cufflinks (Trapnell, Williams et al. 2010), setting the option --max-bundle-frags 5000000, which specifies the maximum genomic length for the bundles.

## 2.4 ATACseq

### 2.4.1 Mapping

The pair-end sequence reads were mapped to the mm9 version of the mouse genome with BOWTIE (Langmead, Trapnell et al. 2009), setting the options -m 1, which eliminates those reads which align in more than one region, --chunkmbs 2000, which sets the megabytes of memory a thread can use to store path descriptors and -X 2000, which specifies the distance between valid pair-end alignments. Mitochondrial chromosome contamination was removed. The ATACseq profile was obtained using the function buildChIPprofile from Seqcode.

## 2.5 Seqcode

Seqcode (Blanco et al. in preparation) software was used to interpret the states of the chromatin segmentation models. Pie charts representing the genome distribution of every state segments were constructed using the function genomeDistribution. Moreover, the function recoverChIPlevels was used to calculate the normalized number of reads of ChIPseq data, RNAseq data and ATACseq data. Seqcode was also used to extract the lists of genes overlapping with every state, using the function matchpeaksgenes. The overlapping was

defined by segments located in the region between 2.5Kb upstream of the gene and its end.

Seqcode was also used to construct the profiles of ChIPseq data, ATACseq data and RNAseq data using the function buildChIPprofile.

# 2.6 ChromHMM

## 2.6.1 Generate a model

Two steps are needed in ChromHMM to obtain a chromatin segmentation model: first, the program binarizes the NGS data into non-overlapping bins of the same size, and second, it generates the model.

We used the function BinarizeBam for the first step, since our input data is in BAM format (Li, Handsaker et al. 2009), a binary format to store sequencing data. The function needs as a parameter a tab delimited text file specifying the cell line or condition in the first column, the name of the mark in the second column, the file in the third column, and the control for the sample in the fourth column (optional). We used the default parameters which segment the genome in bins of 200 bp. When the input data were peak files in BED format, the function used to binarize the data was BinarizeBed.

The function to obtain the segmentation model is LearnModel, which was used under the default parameters. The program uses a multivariate Hidden Markov Model (HMM) to model the observed combination of NGS data and define the chromatin states (Ernst and Kellis 2010). The number of states needs to be specified.

A summary of these two steps can be found in Figure 6, which includes the commands and an example of two tab delimited text files for one cell line and two cell lines.

**1. Binarization of the data**

```
> java -mx2000M -jar $CHROMHMM BinarizeBam $MOUSE9/ChromInfo.txt 02_map_files/table.txt BinarizedFiles
```

```
mESC    H3K27me3    H3K27me3_mESC.bam    IgG.bam
mESC    H3K4me3     H3K4me3_mESC.bam     IgG.bam
mESC    H3K36me3    H3K36me3_mESC.bam    IgG.bam
mESC    Suz12       Suz12.bam_mESC       IgG.bam
```

```
mESC    H3K27me3    H3K27me3_mESC.bam    IgG.bam
mESC    H3K4me3     H3K4me3_mESC.bam     IgG.bam
mESC    H3K36me3    H3K36me3_mESC.bam    IgG.bam
CM      Suz12       Suz12_mESC.bam       IgG.bam
CM      H3K27me3    H3K27me3_CM.bam      WCE.bam
CM      H3K4me3     H3K4me3_CM.bam       WCE.bam
CM      H3K36me3    H3K36me3_CM.bam      WCE.bam
CM      Suz12       Suz12_CM.bam         WCE.bam
```

**2. Obtaining the model**

```
> java -mx2000M -jar $CHROMHMM LearnModel BinarizedFiles ModelFiles 4 $MOUSE9
```

*Figure 6:* **Summary of the ChromHMM steps in the command line interface.**

Only under the following three conditions the fourth column of the tab delimited file was set to empty:

- When introducing RNAseq and ATAcseq inside of the model, since no control is available for this type of data. However, IgG was still used as a control for the ChIPseq data used to construct this model.
- When testing the use of peaks instead of reads as input for ChromHMM, since the control is necessary during the previous peak calling step.
- When the absence of a control was tested.

## 2.6.2 Selection of the number of states

The general procedure to select the appropriate number of states for a set of samples consists in:

1. Generating several models with different number of states: Section 2.6.1 is repeated several times obtaining models from 2 states to 15 states.
2. Evaluating the resulting models: expression (RNAseq), accessibility (ATACseq) and epigenomic information (ChIPseq) were evaluated for

the chromatin segmentation states. The resulting segmentations were evaluated using Seqcode to generate genome distribution pie charts (function genomeDistribution), to calculate signal strength (function recoverChIPlevels), and to obtain the lists of genes overlapping each state (function matchpeaksgenes).

3. Selecting the most biologically meaningful/interesting model according to the previous step of evaluation: the number of states of model is selected according to its suitability to study the biological question proposed, the concordance with previous knowledge on the marks function, etc.

# 2.7 chromstaR

chromstaR (Taudt et al. unpublished) needs sorted BAM files to obtain a chromatin segmentation map. We used sortBam function from Rsamtools (Morgan et al. 2017) with this purpose. The models were obtained using the function Chromstar, setting the bin size to 200 bp and the model to full. The function also needs a tab delimited file (experiment.table) specifying in each column the file, the mark, the condition, the replicate, pair-ends ore single-ends, and the control file. An example of the experiment.table can be found in Figure 7.

| File | mark | condition | replicate | pairedEndReads | controlFiles |
|---|---|---|---|---|---|
| Cbx7_sorted.bam | Cbx7 | mESC | 1 | FALSE | IgG_sorted.bam |
| Epop_sorted.bam | Epop | mESC | 1 | FALSE | IgG_sorted.bam |
| Jarid2_sorted.bam | Jarid2 | mESC | 1 | FALSE | IgG_sorted.bam |
| Ring1B_sorted.bam | Ring1B | mESC | 1 | FALSE | IgG_sorted.bam |
| Rybp_sorted.bam | Rybp | mESC | 1 | FALSE | IgG_sorted.bam |
| Suz12_sorted.bam | Suz12 | mESC | 1 | FALSE | IgG_sorted.bam |

***Figure 7:*** **Example of experiment.table.**

chromstaR was run setting the rest of the parameters to default, but also the parameters in Table 5 were tested. The default parameters are: eps.univariate = 0.1, eps.multivariate = 0.01, and the number of states is all the theoretically possible combinations of marks.

*Table 5:* **Combination of parameters tested in chromstaR.**

| | **Number of states** | **eps.univariate** | **eps.multivariate** |
|---|---|---|---|
| Test 1 | 10 | Default | Default |
| Test 2 | 6 | Default | Default |
| Test 3 | 3 | Default | Default |
| Test 4 | 6 | 0.01 | Default |
| Test 5 | 6 | 0.2 | Default |
| Test 6 | 6 | Default | 0.001 |
| Test 7 | 6 | Default | 0.1 |

# 2.8 Enrichr

The functional analysis of the lists of genes overlapping with each state was performed using Enrichr (Kuleshov, Jones et al. 2016).

# 2.9 UCSC Genome Browser

The UCSC Genome Browser (Karolchik, Hinrichs et al. 2007) was used to visualize the resulting chromatin segmentations, NGS profiles, etc. All the screen shots which appear during this document were taken from UCSC Genome Browser.

# 2.10 Web server

A webserver has been implemented in PHP to distribute the resulting chromatin segmentation models of the mouse genome. All the information was precomputed before to accelerate the server.

## 2.11 Scripts

All the scripts have been written in Python, and run in an operative system MAC OS X Sierra, processor 2,8 GHz Intel Core i5, and 16 GB of RAM.

# 3. Results

## 3.1 Generation of a reference histone model in mESC

In order to study the performance of ChromHMM in mESC, we obtained a chromatin segmentation model of six histone marks, which we used as a reference. The data used to generate the model are referenced in Table 2. The set of marks used was:

- H3K4me3: activation mark found in promoters.
- H3K36me3: activation mark found in gene bodies.
- H3K4me1: activation mark found in promoters and enhancers.
- H3K27ac: openness mark found in promoters and enhancers.
- H3K27me3: repressive mark found in promoters.
- H3K9me3: repressive mark characteristic of heterochromatin.

Once we obtained a nine state model, we considered to label each state according to its biological interpretation:

- State 1: Intermediate enhancer.
- State 2: Active enhancer.
- State 3: No mark (active).
- State 4: No mark (inactive).
- State 5: Active promoter.
- State 6: Bivalent promoter.
- State 7: Poised enhancers/genes.
- State 8: Active body.
- State 9: Heterochromatin.

From now on, we will explain in detail the meaning of each state of this model.

Each state is defined by the probability to find each of the histone marks in that state, which is represented in Figure 8A. State 1 contains a high probability to find H3K4me1, which is in concordance of an intermediate enhancer, whereas state 2 also has a high probability to find H3K27ac, which is in agreement with an active enhancer state. States 3 and 4 are unmarked states, which cover the most part of the genome, specially state 4, which covers 60% of it (Figure 8B). State 5 has a high probability to find the marks one would expect to find in active promoters, which are H3K4me3, H3K4me1 and H3K27ac, whereas state 6, besides the marks of state 5, also has a high probability to find H3K27me3, in agreement with a bivalent promoter state. State 8 is mostly enriched in H3K36me3, a mark characteristic of genes being transcribed and state 9 is enriched in H3K9me3 a mark of heterochromatin.

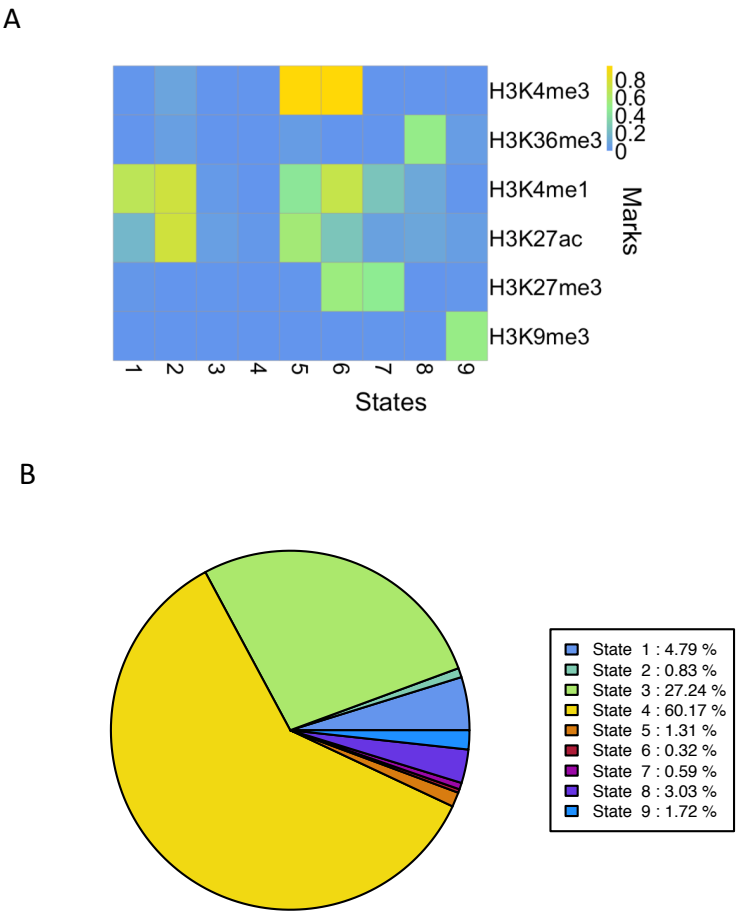



***Figure 8:*** **State emissions and genome coverage of the histone model.**

(A) Emissions of every state of the histone model represented in probability to find a specific mark in each of the states. Yellow means high probability to find the mark, green means mild probability and blue means low probability.

(B) Genome coverage of every state of the model represented in percentage.

The genome distribution also confirms the previous biological interpretation of the segments of each state (Figure 9). State 1 and 2 are located mostly in intergenic regions and introns, in agreement with being enhancers. States 5 and 6 mostly overlap with promoters, in concordance of active and inactive promoters, respectively. Moreover, state 8 mostly overlaps with genes, in agreement with being an active body state.

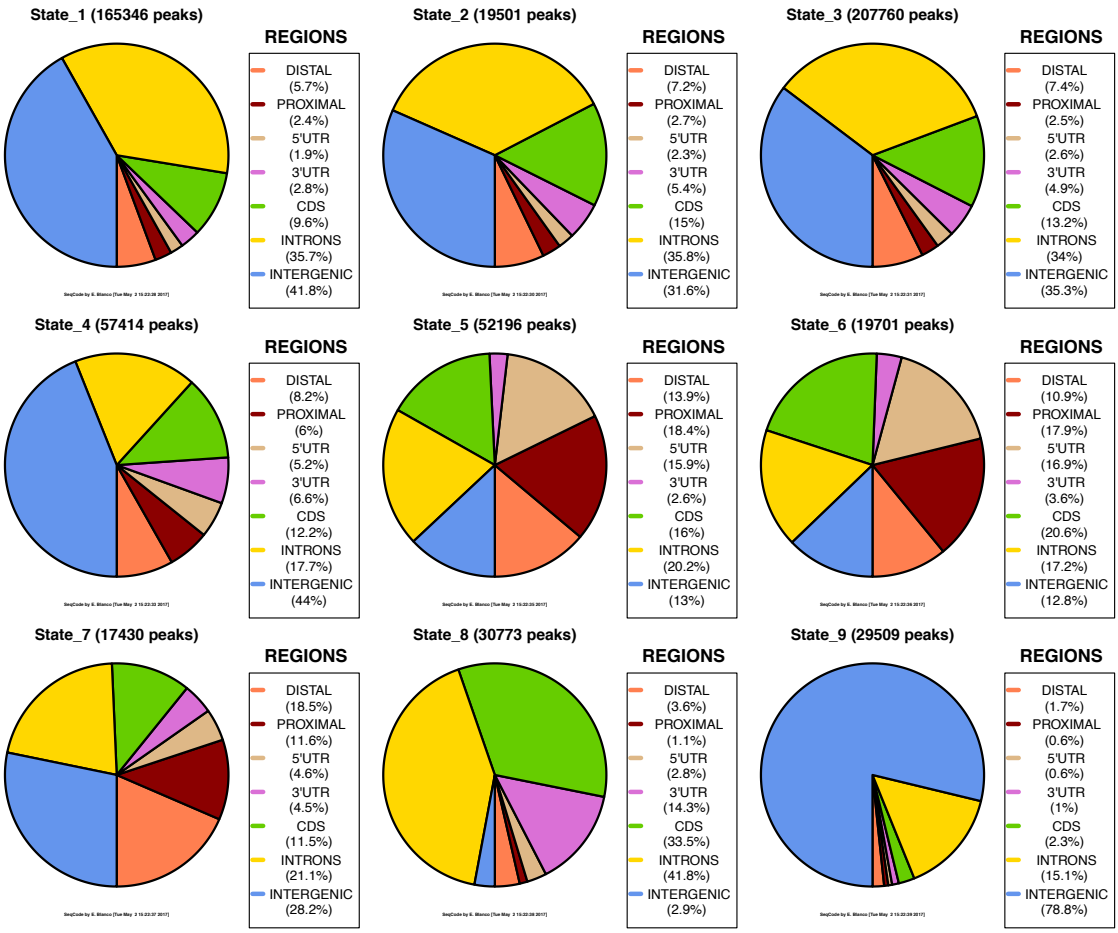



***Figure 9:*** **Genome distribution of the segmentation.**

For each state we show the percentage of fragments overlapping with each region. DISTAL region is the region within 2.5 Kbp and 0.5 Kbp upstream of the TSS. PROXIMAL region is the region within 0.5 Kbp and the TSS. UTR is UnTRanslated sequence. CDS is the protein CoDing Sequence. INTRONS are intronic regions. INTERGENIC is the rest of the genome. TSS is the Transcription Start Site.

Moreover, we confirmed that the ChIP levels of every mark in the segments of every state (Figure 10), were in agreement with their emissions (Figure 8A). For example, H3K4me3 is higher in the promoter states (states 5 and 6), and H3K36me3 is higher in the active body state (state 8). Moreover, H3K4me1 is higher in the enhancer states and the promoter states (states 1, 2, 5 and 6), and there is also a bit of H3K4me1 in the poised enhancer state (state 7). H3K27ac is higher in the active promoter state and the active promoter state (states 1 and 5 respectively), whereas H3K27me3 is higher in the bivalent promoter state and the poised enhancer state (states 6 and 7 respectively). Finally, state 9 (heterochromatin state) has the highest H3K9me3 level.

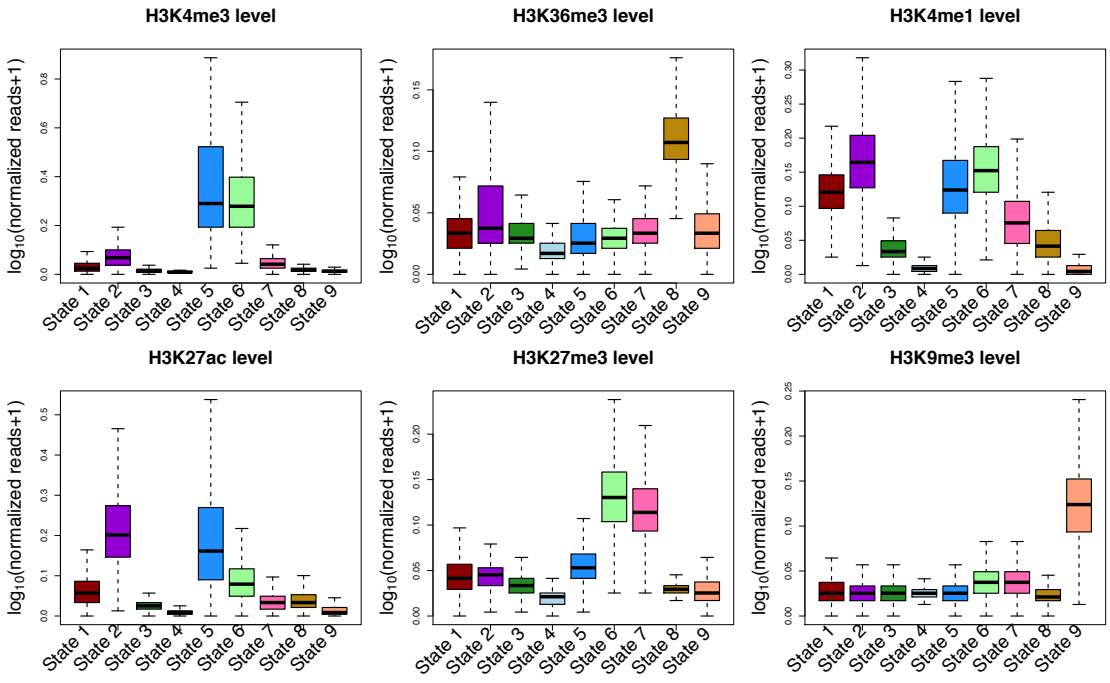



***Figure 10:*** **ChIP levels in the segmentation of every state of the histone model.**
Distribution of the number of reads of each of the marks used to generate the histone model in every segment of each state in the model. The numbers of reads are normalized by the total number of reads of each experiment.

Next, we checked chromatin accessibility using ATACseq data. We confirmed that the promoter states (states 5 and 6), specially the active promoter state (state 5) were those which had the highest ATAC level (Figure 11A). Moreover, state 2, corresponding to active enhancers, had higher ATAC level than the rest of states but the promoter states.

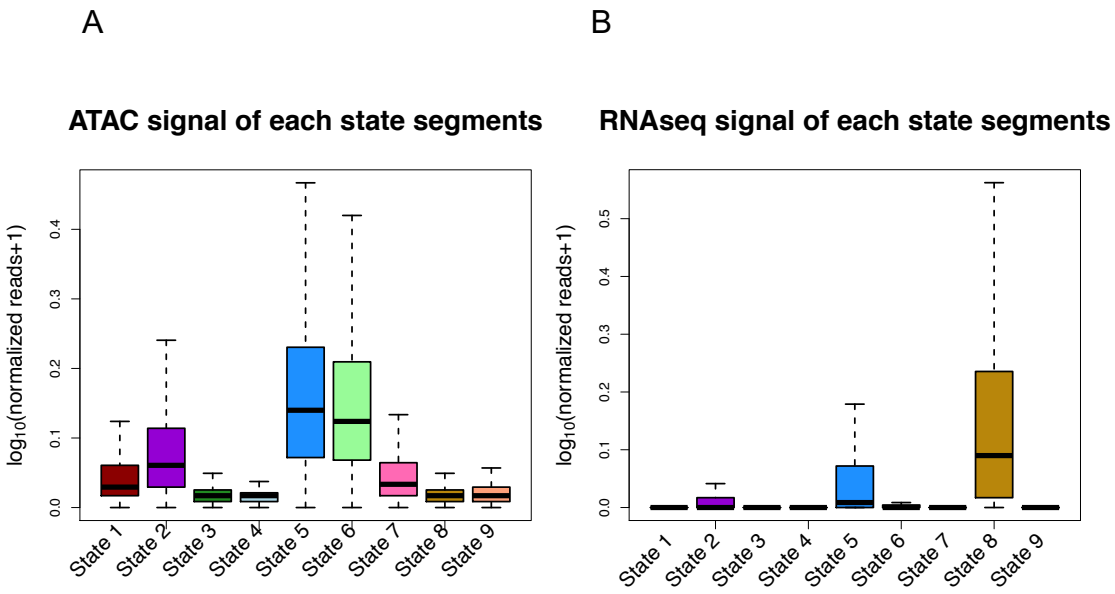



***Figure 11:*** **ATAC and RNA levels of the segmentation of every state in the histone model.** Distribution of the normalized number of reads of ATACseq and RNAseq in the segments of every state of the histone model.

We confirmed, using RNAseq data, that the state with the highest expression was the active gene body state (state 8), as can be seen in Figure 11B. Of note, we observed in state 5 a moderate expression, which can correspond to the promoter state, since the active marks located in the promoter usually go some base pairs downstream the TSS, and sometimes even they cover the whole gene, assigning it to the active promoter state (Figure 12C). Moreover, we also observed a bit of expression in state 2 which in fact might correspond to active enhancers expression. Active enhancers might be translated to eRNAs (enhancer RNAs), thus, we could see a bit of expression in state 2.

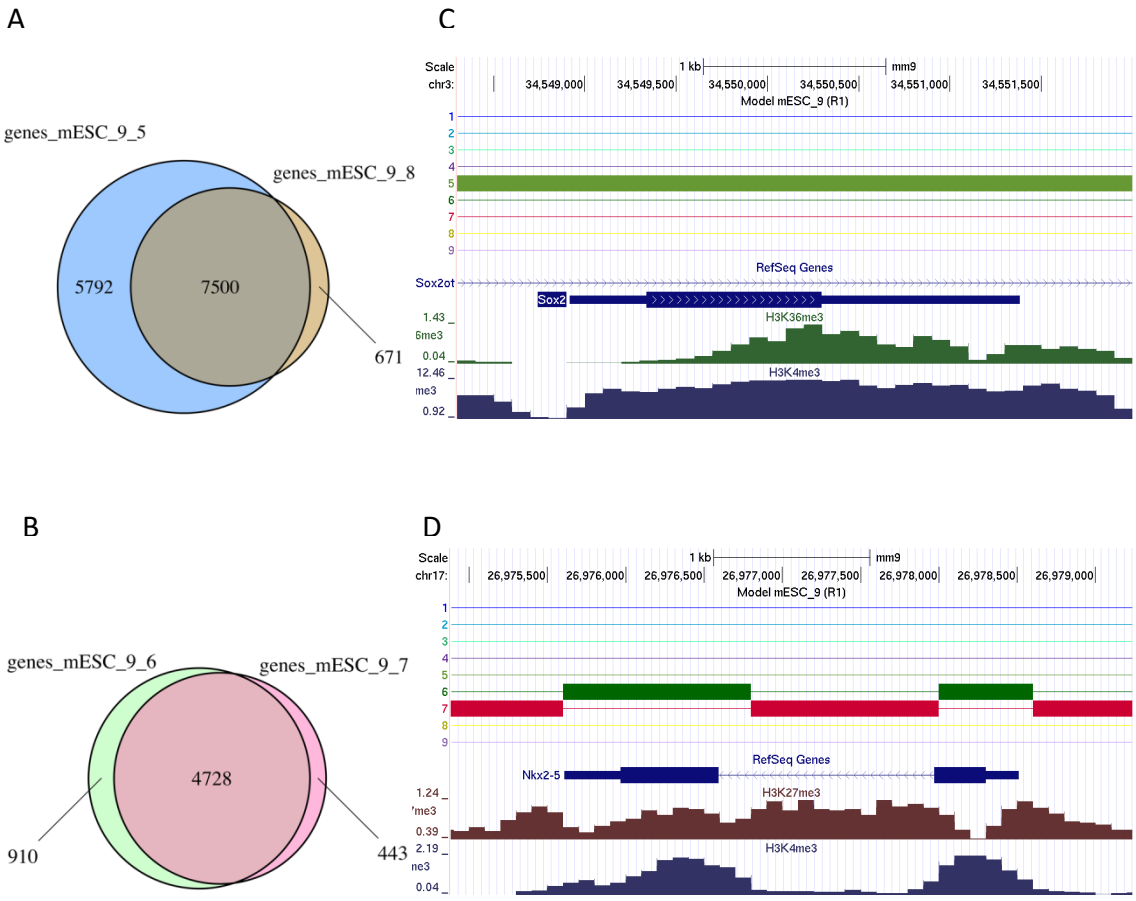



***Figure 12:*** **Effect of broad domains when assigning states in the histone model.**

(A) Target gene overlapping between state 5 (active promoters) and state 8 (active gene body).

(B) Target gene overlapping between state 6 (bivalent promoters) and state 7 (poised genes/enhancers).

(C) Example of a gene only marked by the active promoter state (state 5).

(D) Example of a gene marked by state 6 and 7 (bivalent promoters and poised genes/enhancers respectively).

Comparing the target genes of states 5 and 8, we see that those with an active promoter do not always have an active gene body (Figure 12A). In mESC, pluripotency markers such as Sox2, have broad domains of H3K4me3, which leads to ChromHMM assigning the whole gene to an active promoter state even though they also have H3K36me3 in the gene body (Figure 12C). Comparing the target genes of states having H3K27me3 (states 5 and 7), we see that they have most of their target genes in common (Figure 12B). This is due to H3K27me3 having broader domains than H3K4me3, thus state 6 segments usually have flanking segments of state 7 (Figure 12D).

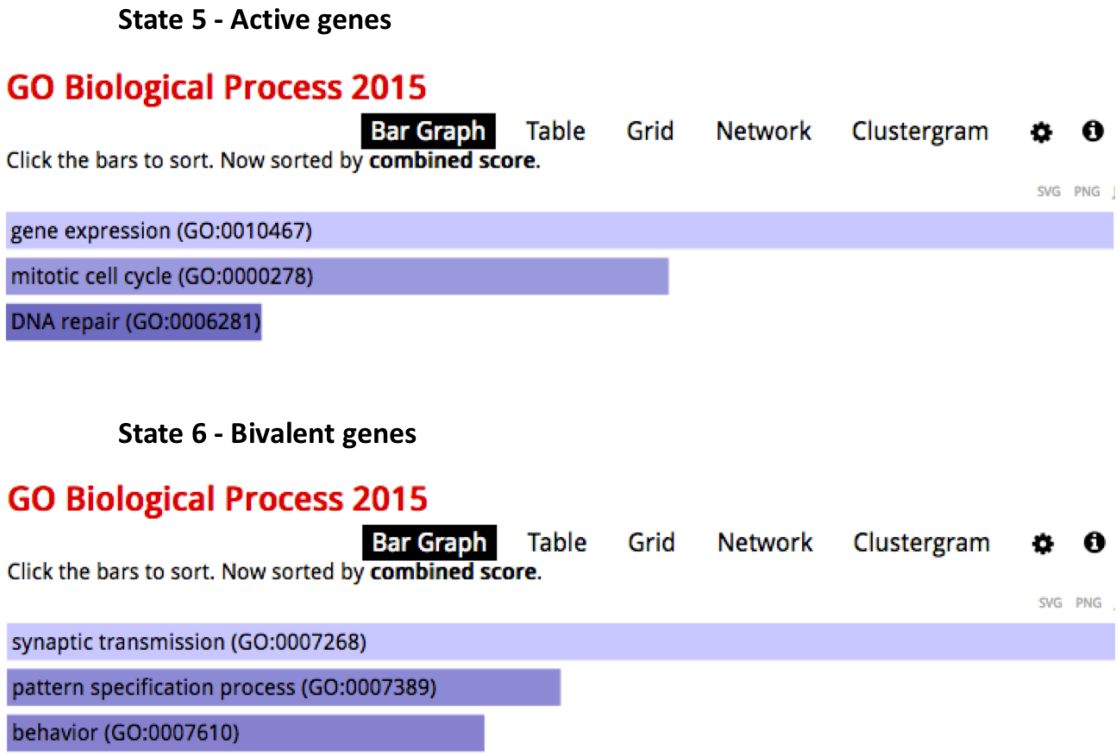



*Figure 13:* **Functional analysis of target genes of states 5 and 6.**
Functional analysis of genes overlapping with segments of state 5 and functional analysis of genes overlapping with segments of state 6.

We next performed a functional analysis of the target genes in states 5 and 6, corresponding to active promoters and bivalent promoters respectively (Figure 13). The most significant biological processes found among both lists of target genes were as expected, the genes from state 5 are enriched in categories such as gene expression, mitotic cell cycle and DNA repair. On the other hand, the genes from state 6 are enriched in categories such as synaptic transmission, pattern specification and behaviour.

Then, we focused on enhancers, wiling to answer if our enhancer states (states 1, 2 and 7) were overlapping with known enhancers. We used a published enhancer list in mESC (Schoenfelder, Sugar et al. 2015) which contains enhancers classified into three categories: active (contains H3K4me1 and H3K27ac), intermediate (contains H3K4me1 alone) and poised (contains H3K4me1 and H3K27me3). We first confirmed that their enhancer definition fitted into our data (Figure 14).

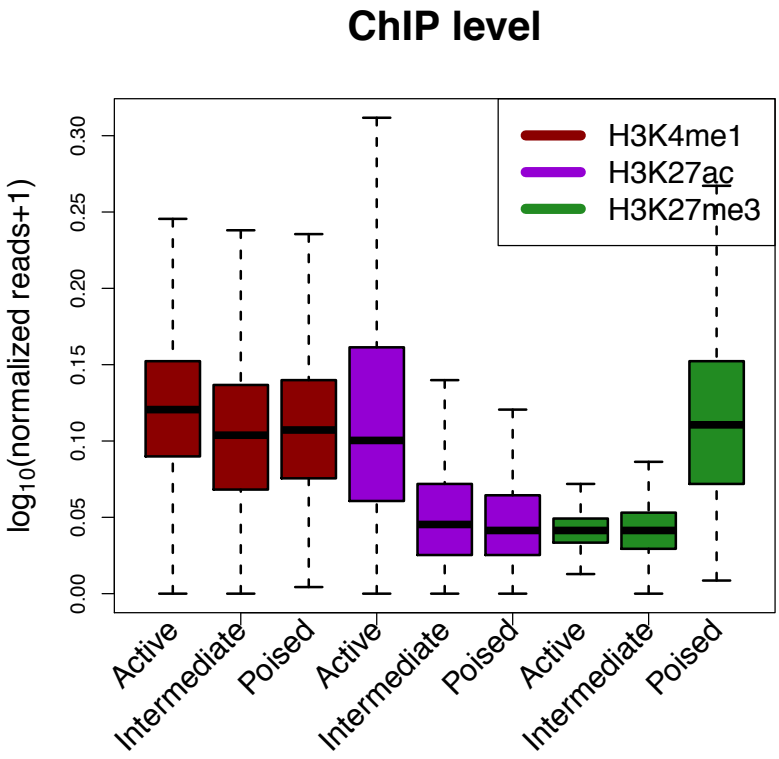



***Figure 14:*** **Confirmation of the enhancer classification in mESC.**

Distribution of the normalized number of reads of H3K4me1, H3K27ac and H3K27me3 in the three categories (active, intermediate and poised) of the collection of mESC enhancers. Active enhancers are the only category which contain high levels of H3K27ac, whereas poised enhancers are the only category which contain high levels of H3K27me3.

We next calculated the percentage of each type of enhancers overlapping with our states (Table 6). We observed that most of the poised enhancers (59%) were overlapping with state 7. State 2 was overlapping with more active enhancers (41%) than intermediate (5%) or poised (2%). Despite state 1 was mostly overlapping with active enhancers (78%), we assigned it to intermediate enhancers since the overlapping was also high (68%). Moreover, as we observed in Figure 10, this state had a high signal of H3K4me1 and a low signal of H3K27ac, matching the intermediate state definition, whereas state 1 had a high signal of H3K4me1 and an also high signal of H3K27ac, matching the active state definition.

*Table 6:* **Overlap of the three categories of enhancers with segments of every state.**
Percentage of enhancers of each category overlapping with segments of every state.

| State | Active Enhancer | Intermediate Enhancer | Poised Enhancer |
|---|---|---|---|
| **1** | 78% | 68% | 29% |
| **2** | 41% | 5% | 2% |
| **3** | 75% | 72% | 37% |
| **4** | 3% | 8% | 11% |
| **5** | 15% | 5% | 5% |
| **6** | 1% | 1% | 28% |
| **7** | 1% | 2% | 59% |
| **8** | 11% | 5% | 1% |
| **9** | 4% | 1% | 6% |

We also observed that the unmarked state 3 was highly overlapping with all types of enhancers: 75% of the active enhancers, 72% of the intermediate enhancers and 37% of the poised enhancers. Moreover, we also observed that this same state was also overlapping with the most part of the genes (Table 7), even though its genome coverage (27.24%) was lower than the genome coverage of the other unmarked state (60.17%), as we can observe in Figure 8B. These observations suggested that state 3 might be an active unmarked state containing active genes and enhancers, whereas state 4 might be an inactive unmarked state.

*Table 7:* **Number of genes overlapping with each state in the histone model.**

| State | Genes* |
|---|---|
| 1 | 15744 |
| 2 | 5545 |
| 3 | 18232 |
| 4 | 9469 |
| 5 | 13292 |
| 6 | 5638 |
| 7 | 5171 |
| 8 | 8171 |
| 9 | 3346 |

* Segments located in the region between 2.5Kb upstream of the gene and its end.

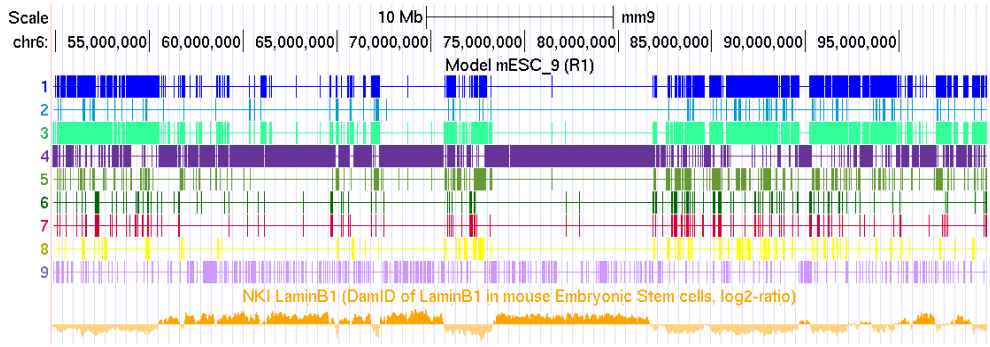



*Figure 15:* **Comparison of the histone model segmentation and the DamID of LaminB1 in mESC.**

Moreover, looking at the UCSC track NKI Nuclear Lamina Associated Domains (LaminB1 DamID) of mESC, we observed that state 3 pattern seemed to correlate with the absence of LaminB1, whereas state 4 pattern seemed to correlate with the presence of LaminB1 (Figure 15). We confirmed our observations calculating the number of LADs and no LADs overlapping states 3, 4 and 9. In agreement with our conclusions, state 4 was mostly overlapping with LADs, in concordance with its assignation to an inactive state lacking genes and enhancers, while state 3 was mostly overlapping no LADs, in concordance with its assignation to an active state containing genes and enhancers. Moreover, the heterochromatin state was also mainly overlapping LADs as expected for a state containing H3K9me3 (Figure 10).

*Table 8:* **Number of LADs and no LADs per state.**

| State | LAD | No LAD |
|---|---|---|
| 3 | 1578 | **3253** |
| 4 | **3834** | 2535 |
| 9 | **2379** | 1489 |

# 3.2 Parameter selection in ChromHMM

We tested the histone chromatin segmentation model using different ChromHMM parameters (Table 9) in order to evaluate the changes in the final model. We explored the use of peaks or reads, IgG as a control or no control, different number of reads or same number of reads, and the possibility to introduce expression and accessibility data inside of the model.

*Table 9:* **Parameters tested in ChromHMM.**

The selected parameters are highlighted in grey.

| **PARAMETER** | OPTION1 | OPTION2 |
|---|---|---|
| **Input files** | Reads | Peaks |
| **Control** | IgG | No control |
| **Number of reads** | Different number of reads | Same number of reads |
| **Expression data (RNAseq)** | Outside the model | Inside the model |
| **Chromatin accessiblity (ATACseq)** | Outside the model | Inside the model |

The model obtained with peak files, contains a new state (state 6) which contains H3K27ac alone. Rather than a real state, it seems to be an artefact of the peak calling. Moreover, it does not distinguish the two unmarked states. (Figure S1).

In the model with no control, the heat map of emissions contains higher levels of background, making the model more difficult to interpret. A new state appears (state 2), which seems to be an artefact of the background since it contains H3K9me3 together with active marks such as H3K36me3 and H3K27ac. (Figure S2).

Downsampling the model to the same number of reads, seems to have a counter effect in H3K27me3, since the model does not seem to detect properly the state belonging to bivalent promoters. The bivalent state of this new model seems to correspond to state 7, however, this state is also overlapping with the active promoter state (state 5) of the default histone model. (Figure S3).

Introducing RNAseq data to the model does not seem to add valuable information, since it divides the active gene body state into two states, which seem to be the exons (state 9) and the introns (state 8). Therefore, it seems better to leave this interpretation outside the model. Next, we can use such data to interpret the final models. On the other hand, introducing ATACseq gives rise to a new state, composed by intergenic peaks of ATACseq not overlapping with any other mark. This state could be interesting to study, nevertheless, we decided again to keep ATACseq outside of the model and use this information later to interpret the states. (Figure 16).

From now on, all the models in the document were generated using the parameters selected in Table 9.

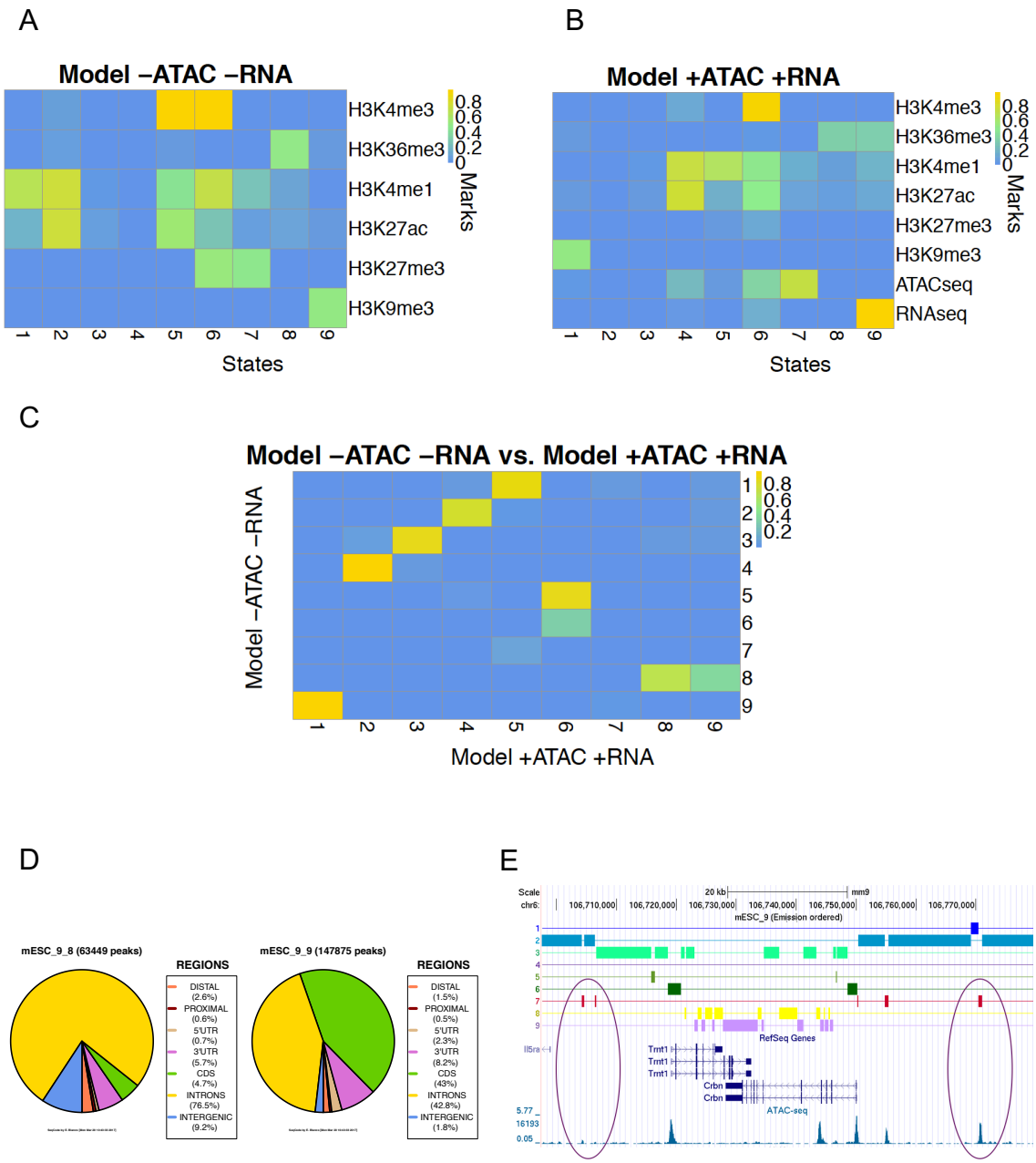



*Figure 16:* **Comparing the use of ATACseq and RNAseq data inside or outside of the model.**

(A) State emissions of the model using ATACseq and RNAseq data outside of the model, represented in probability to find a specific mark in each of the states.

(B) State emissions of the model using ATACseq and RNAseq data inside of the model, represented in probability to find a specific mark in each of the states.

(C) Comparison of the model using ATACseq and RNAseq inside and outside of the model.

(D) Genome distribution of states 8 and 9.

(E) Example of ATACseq peaks which belong to state 7, containing ATACseq only.

# 3.3 Introducing one PRC1 and one PRC2 subunits to the model

Besides bivalent genes, the identification of poised enhancers is also important (Rada-Iglesias, Bajpai et al. 2011). Comparing the overlap of genes containing H3K4me3 and H3K27me3 (Figure 17A) to the overlap of peaks containing these marks (Figure 17B), we observed that almost all the H3K27me3 target genes were also target genes of H3K4me3, whereas the subset of peaks containing H3K27me3 only is bigger. Therefore, we hypothesized that this subset of peaks might be intergenic, and therefore, correspond to poised enhancers.

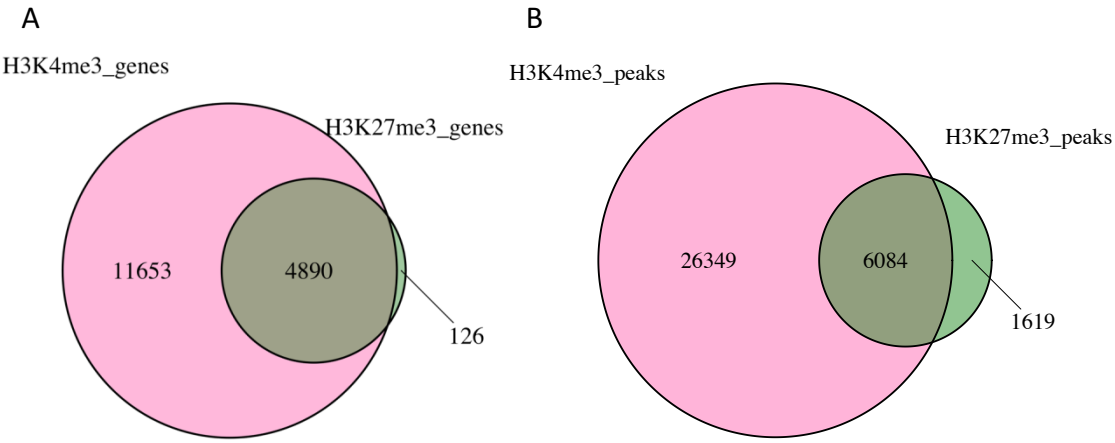



***Figure 17:*** **Overlap of genes and peaks containing H3K4me3 and H3K27me3.**
(A) Overlap of H3K4me3 and H3K27me3 at gene level.
(B) Overlap of H3K4me3 and H3K27me3 at peak level.

The model we obtained to address this question was constructed using three histone marks covering active and inactive genes, and enhancers and poised enhancers (H3K4me3, H3K4me1 and H3K27me3), and one PRC1 subunit (Ring1B) and one PRC2 subunit (Suz12). The data used are referenced in Table 2. We obtained a model of six states, whose emission probabilities, can be found in Figure 18. Since we were interested in poised enhancers which contain H3K27me3, we focused on those states containing this mark, which are states 1 and 6.

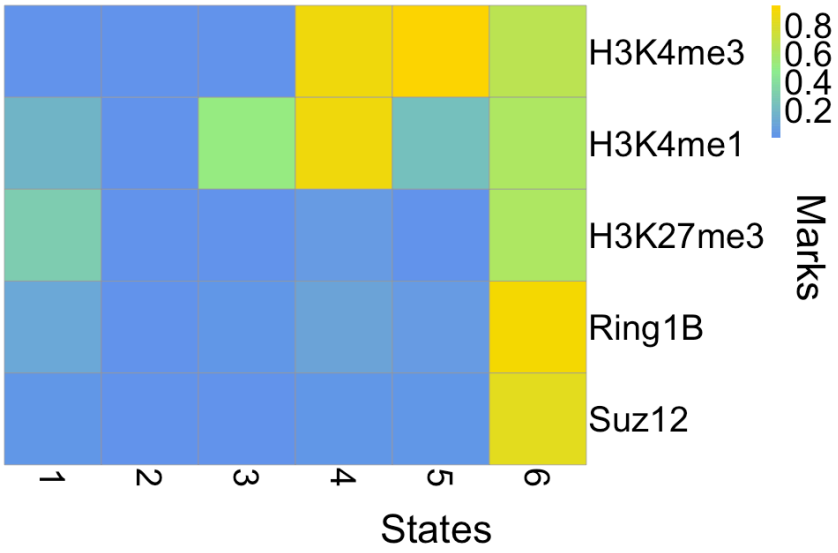



***Figure 18:*** **State emissions of the model introducing one PRC1 and one PRC2.**
Emissions of every state of the model introducing one PRC1 and one PRC2 represented by probability to find a specific mark in each of the states. States 1 and 2 are important to study poised enhancers in mESC.

The genome distribution of the segments of states 1 and 6 (Figure 19), suggests that state 1 might be present in poised enhancers since it is mainly located in intergenic regions and introns, whereas state 6 corresponds to bivalent promoters since its segments are located mostly in promoters. Moreover, state 6 also contains H3K4me3, a characteristic mark of promoters.

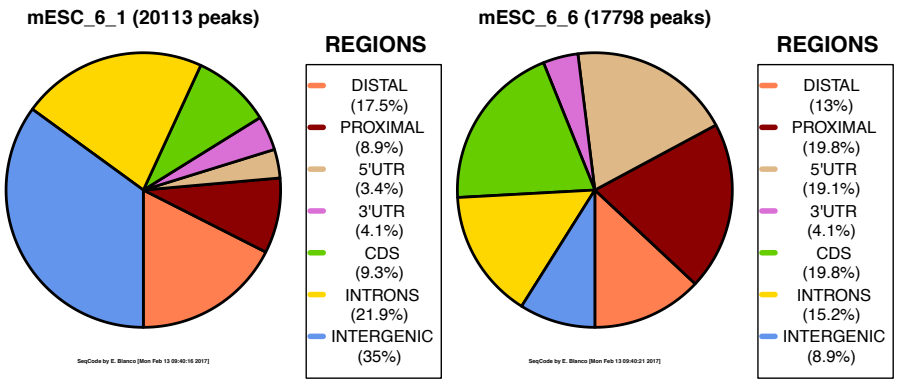



***Figure 19:*** **Genome distribution states 1 and 6 of the model introducing one PRC1 and one PRC2**.
Genome distribution of states 1 and 6, represented as percentage of fragments overlapping with each category. State 1 is enriched in intergenic regions, whereas state 6 is enriched in promoters.

Next, we checked whether intergenic peaks of H3K27me3 were mostly overlapping with state 1 or state 6, and we used as a control intergenic peaks of H3K4me3. We observed that 81% of the intergenic peaks of H3K27me3 were overlapping state 1, whereas roughly half of them (43% of the intergenic peaks) were overlapping state 6 (Table 10). The percentages of intergenic peaks of H3K4me3 overlapping states 1 and 6 were more or less the same, 14% and 13% respectively. These results also suggest that state 1 is more likely to be marking poised enhancers.

***Table 10:*** **Percentage of intergenic peaks of H3K27me3 and H3K4me3 overlapping with states 1 and 6.**

| State | H3K27me3 intergenic peaks (1930/7703) | H3K4me3 intergenic peaks (9218/32433) |
|---|---|---|
| **1** | 81% | 14% |
| **6** | 43% | 13% |

In order to confirm that state 1 is present in poised enhancers, we calculated the percentage of the published poised enhancers used before, which were overlapping state 1 and 6 (Table 11). We found that poised enhancers were the type of enhancer mostly overlapping with state 1 (51%). However, the overlapping with state 6 was almost the same (45%).

***Table 11:*** **Percentage of poised enhancers overlapping segments of state 1 and 6.**

| State | Active Enhancer | Intermediate Enhancer | Poised Enhancer |
|---|---|---|---|
| **1** | 1% | 3% | 51% |
| **6** | 0% | 1% | 45% |

To understand why the percentages were so similar, we checked the ChIP signal in the published poised enhancers which were overlapping state 1 and 6 (Figure 20). We observed that both subsets exhibited H3K4me1, a mark characteristic of enhancers and promoters. However, state 6 also showed H3K4me3, which is a mark characteristic of promoters but not enhancers, suggesting that the subset of published poised enhancers overlapping state 6 were in fact promoters.

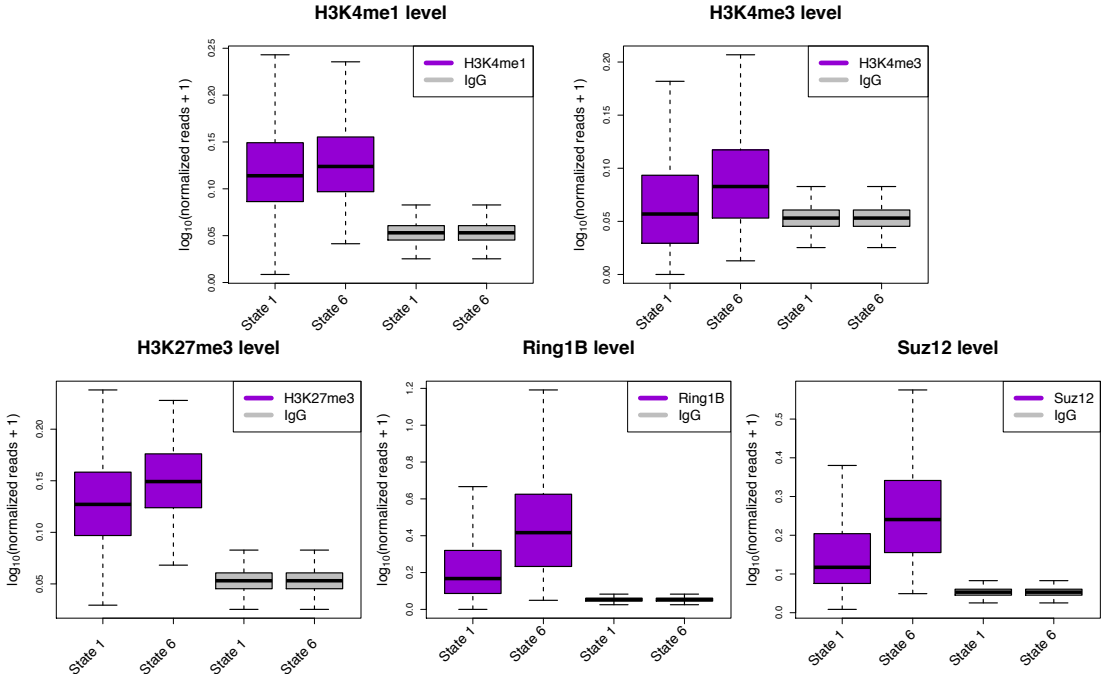



***Figure 20:*** **ChIP levels of poised enhancers overlapping states 1 and 6.**
Distribution of the normalized number of reads of each of the marks used to generate the histone model in the poised enhancers overlapping states 1 and 6.

Moreover, we confirmed that both subsets of enhancers showed H3K27me3. Interestingly, we also found that not only the list of enhancers overlapping state 6 had Ring1B and Suz12, but also the ones overlapping state 1.

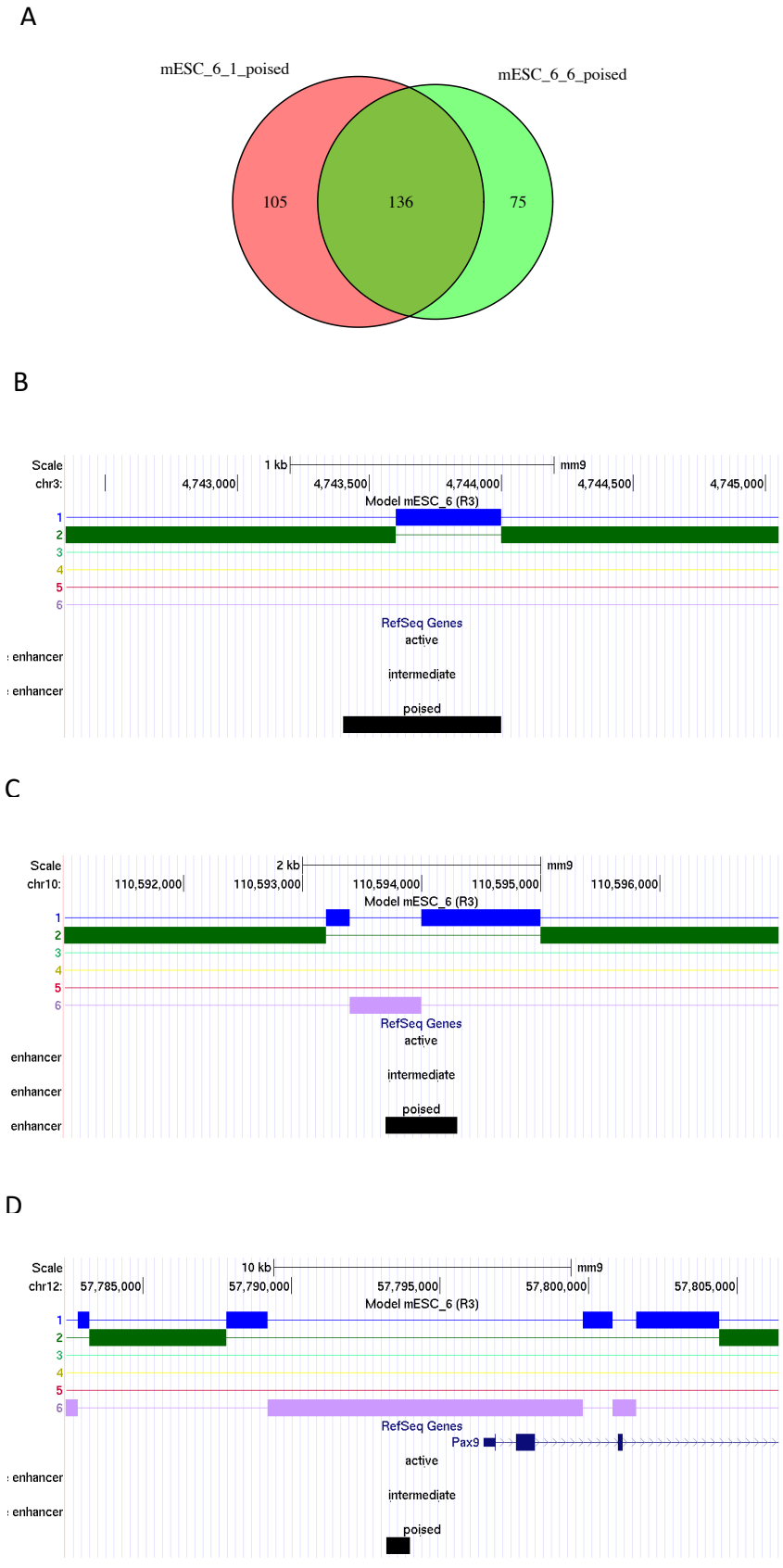



*Figure 21:* **Distribution of poised enhancers overlapping states 1 and 6.**

(A) Overlapping of poised enhancers marked by state 1 and 6.

(B) Example of poised enhancer marked by state 1.

(C) Example of poised enhancer marked by state 1 and 6.

(D) Example of poised enhancer marked by state 6.

We observed that most of the published poised enhancers were shared between states 1 and 6 (Figure 20A). Among the published poised enhancers marked only by state 1, we observed that they were mostly located in intergenic regions (Figure 20B) and the same was happening with those which were marked by both states (Figure 20C). However, as we suspected, those overlapping only state 6 were very close to TSSs, confirming that they might be bivalent promoters instead of poised enhancers.

## 3.4 Working only with PRC1 and PRC2 subunits

In order to study the composition of PRC1 and PRC2 subunits, we obtained a chromatin segmentation model using three PRC1 subunits (Ring1B, Cbx7 and Rybp) and three PRC2 subunits (Suz12, Epop and Jarid2). The data used are referenced in Table 2. The model contains six states, whose emissions can be found in Figure 22. In this model, we focused on PRC2 subunits since it is known that Epop and Jarid2 are mutually exclusive factors of PRC2 (Beringer et al. 2016).

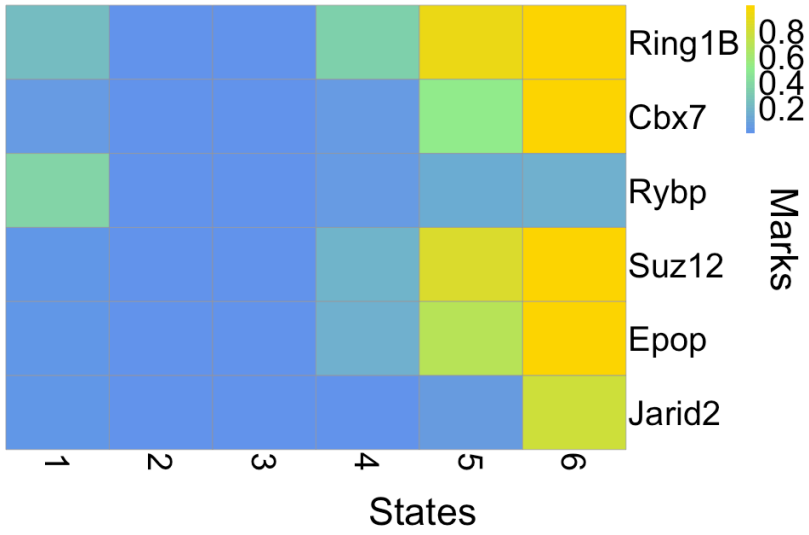



*Figure 22:* **State emissions of the PcG model.**

Emissions of every state of the PcG model represented in probability to find a specific mark in each of the states.

We looked at states 4, 5 and 6, since state 6 seems to have both, EPOP and Jarid2, state 5 seems to have Epop but not Jarid2, and state 4 neither Epop nor Jarid2. We checked the ChIP levels of these two PRC2 subunits in these three states, and whereas Epop was present in all three states, Jarid2 was absent in state 4 (Figure 23).

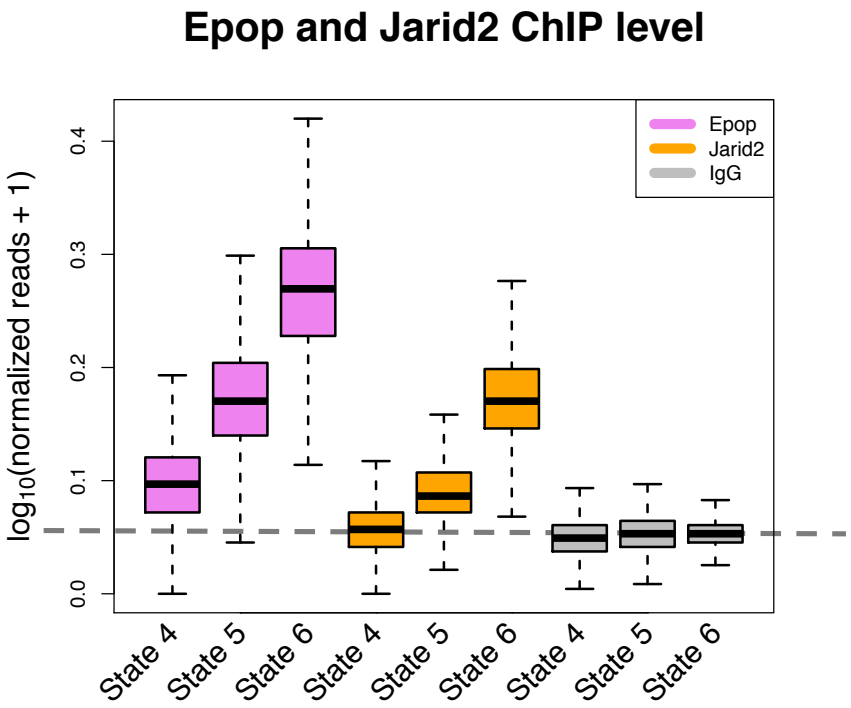



*Figure 23:* **ChIP levels of PRC2 subunits in states 4, 5 and 6.**
Distribution of the normalized number of reads of Epop and Jarid2 in the segments of state 4, 5 and 6. Epop is present in all three states, whereas Jarid2 is not present in state 4.

We next compared our current segmentation model with the previous segmentation of the histone model in Section 3.1 (Figure 23). We observed that whereas state 5 and 6 had most of the bins in common with the active promoter state in the histone model (state 6), state 4 had most of the bins in common with the poised enhancer state in the histone model (state 7). This suggests that maybe PRC2-Jarid2 is not involved in depositing H3K27me3 in poised enhancers.

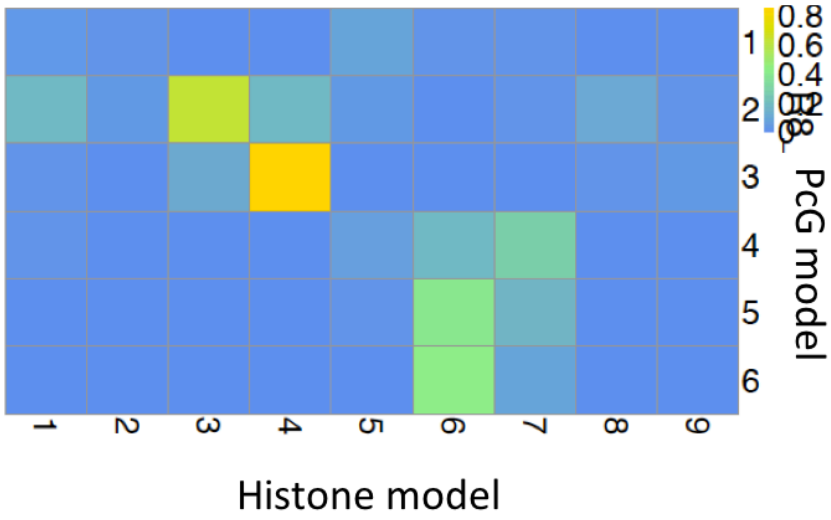



*Figure 24:* **Comparison of the segmentations of the PcG model and the histone model.**

# 3.5 Generating a comprehensive model of the mESC

Once we situated in the previous sections the behaviour of ChromHMM in different scenarios, we obtained a 12 state model using all the histone marks and all the PcG component data used during this document. The data used are referenced in Table 2. The emission probabilities of the states are in Figure 25.

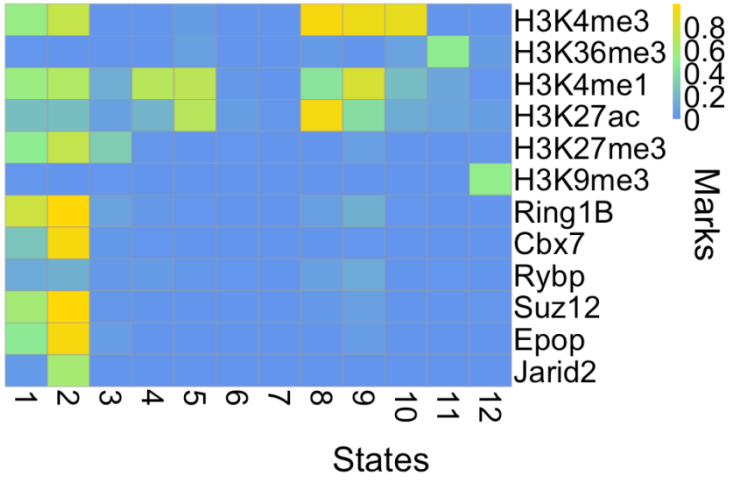



*Figure 25:* **State emissions of the PcG and histone model.**

Emissions of every state of the PcG and histone model represented in probability to find a specific mark in each of the states.

With this model we can study multiple epigenetic factors. For instance, a putative poised enhancer state, which seems to be state 3 according to our previous experience with the other models. The poised enhancer state usually contains H3K27me3 alone, and this is the case with state 3. We checked the genome distribution of the state 3 and as a control, we checked the genome distribution of state 2 which contains also the other PcG subunits besides H3K27me3 (Figure 26). We saw that state 2 segments were mainly located in promoters, whereas state 3 segments were mainly located in intergenic regions. These observations suggest that state 2 might be bivalent promoters and state 3 poised enhancers.

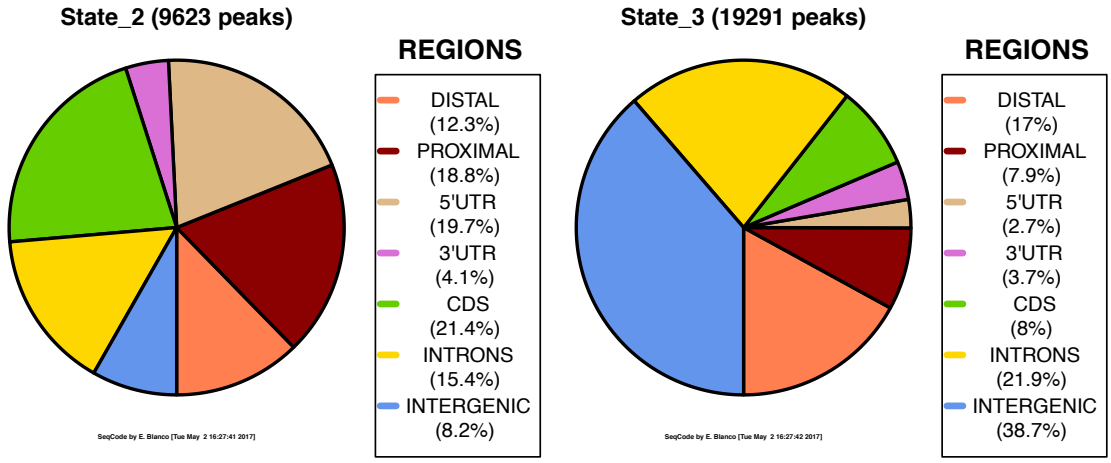



***Figure 26:*** **Genome distribution states 2 and 3 of the PcG and histone model**.
Genome distribution of states 2 and 3, represented as percentage of fragments overlapping with each category.

We selected those state 3 segments which were intergenic and thus, candidates for being poised enhancers, and check their ChIPseq level of H3K27me3, Ring1B and Suz12, using as a control H3K4me3 and IgG (Figure 27). Therefore, we confirmed that H3K27me3 was present in those regions and H3K4me3, which is a negative marker for enhancers, was not, since its ChIP signal was even lower than the IgG signal. Moreover, we observed that these intergenic segments of state 3 were also decorated by Ring1B and Suz12, confirming that PcG is there depositing H3K27me3.

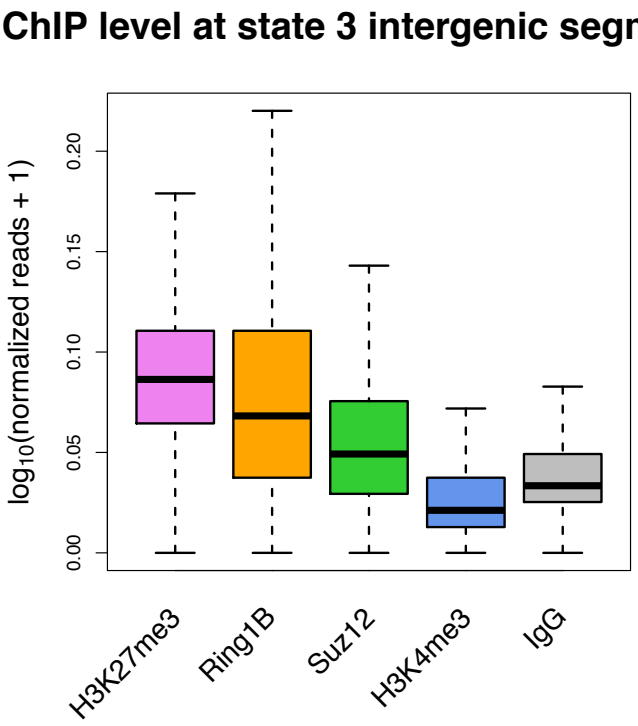



***Figure 27:*** **ChIP levels of PcG in state 3 intergenic segments.**
Distribution of the normalized number of reads of H3K27me3, Ring1B and Suz12 in state 3 intergenic segments. H3K4me3 and IgG levels are used as negative controls.

# 3.6 Web server

We have developed a web server to visualize the models explained during this document. We have also incorporated other meodels that can be interesting. Our web server can be accessed through the following link: http://ldicrocelab.crg.eu/VisualizeChromModels/. In the web server form, one can select the model to be visualized. There is also a description of all the final models available in the web server (Figure 27).

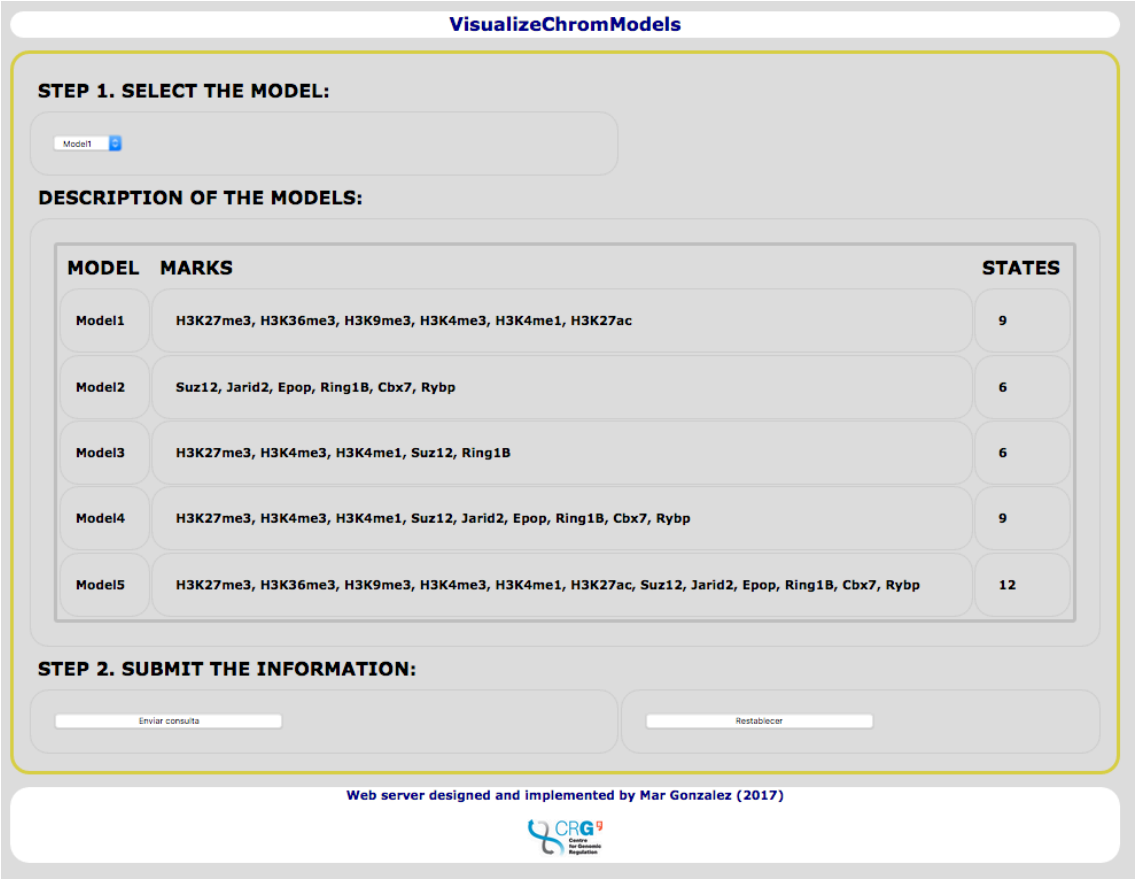
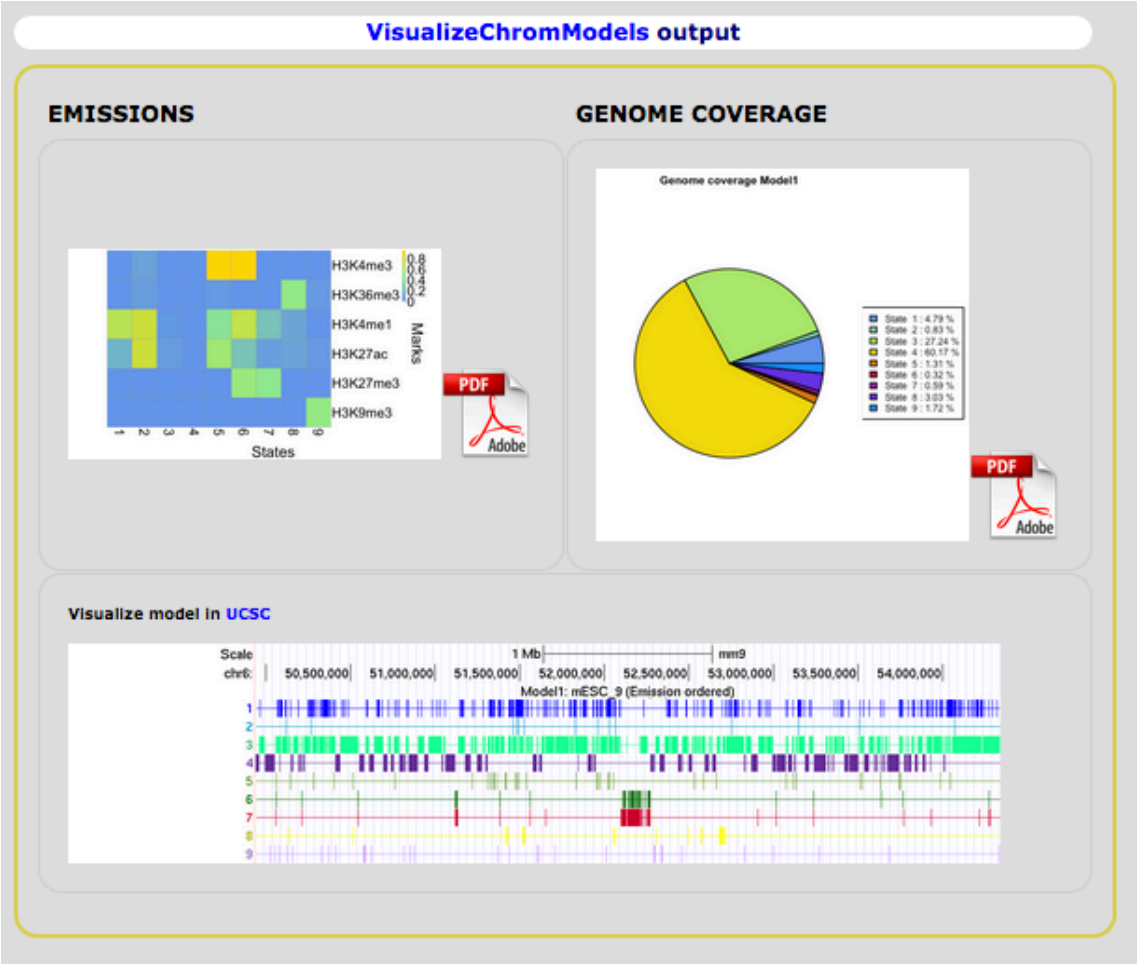



*Figure 28:* **Web server form page.**

All the figures in the web server output can be downloaded in a pdf file. For a particular model, in the first section (Figure 29) , the web server provides a heat map of each state emissions, genome coverage of each state represented in a pie chart and a link to visualize the segmentation of the whole genome in the UCSC Genome Browser (Karolchik, Hinrichs et al. 2007).

A




B

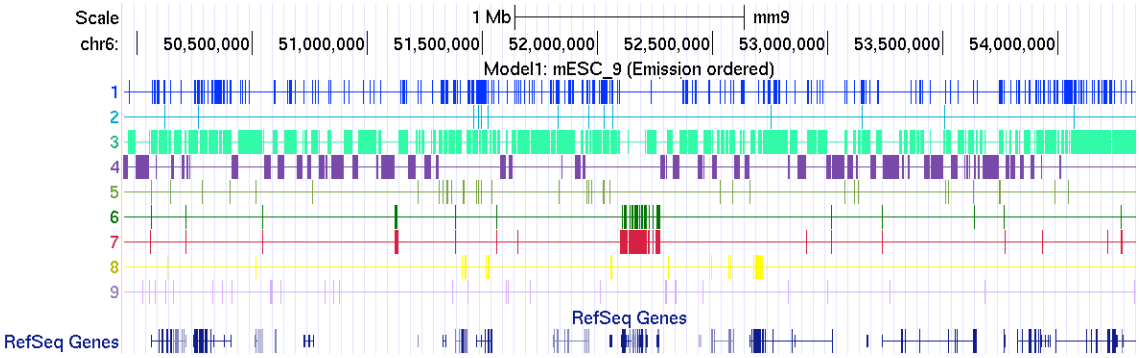



*Figure 29:* **First section of the web server output: the model.**

(A) First section of the web server output.

(B) Region loaded by UCSC link.

In the second section, the web server outputs a pie chart for each state, representing the genome distribution of the segments of this particular state (Figure 30). For instance, in Figure 30, we observe that segments belonging to state 1 of the model are mostly located in intergenic regions or introns.

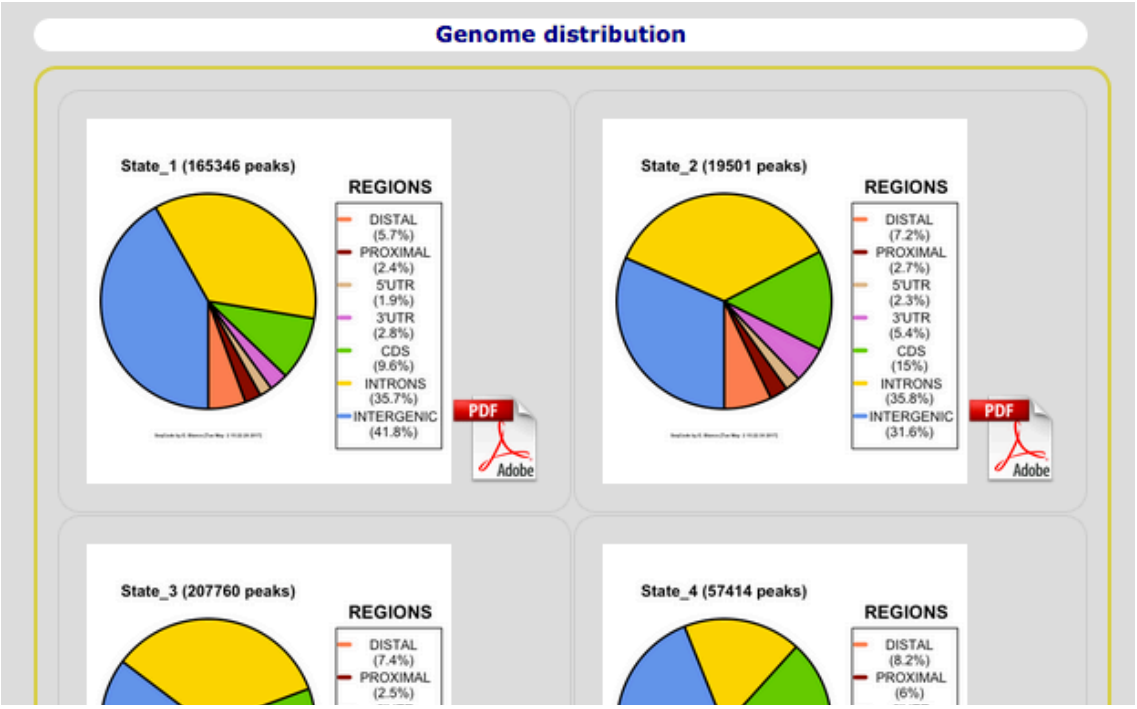



*Figure 30:* **Second section of the web server: genome distribution of the states.**

Finally, the user will find several links to download the bed files of the genome segmentation, and the text files containing the lists of target genes overlapping with each state (Figure 31).


*Figure 31:* **Last section of the web server: segmentation files and lists of target genes.**

# 3.7 Pilot test using two cell lines: mESC and CM

Finally, we aimed to evaluate ChromHMM performance using two cell lines to obtain a model. We generated a 9 state model using five histone marks of mESC and CM, which are: H3K4me3, H3K36me3, H3K4me1, H3K27ac and H3K27me3. The data used is referenced in Table 2. The emissions of the states are in Figure 32.

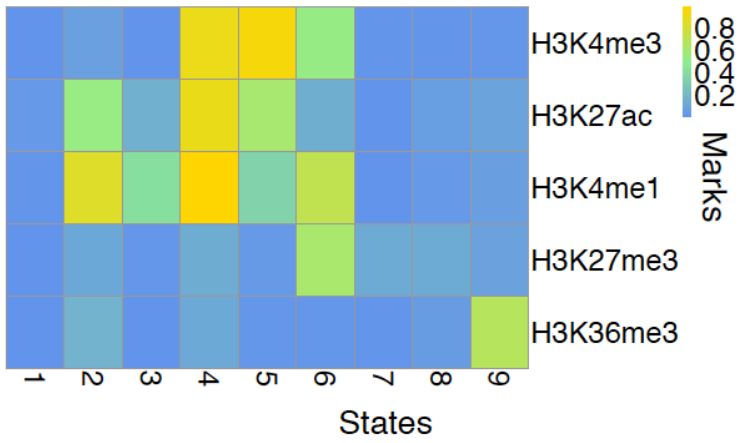



***Figure 32:*** **State emissions of the model using two cell lines.**
Emissions of every state of the model using two cell lines represented in probability to find a specific mark in each of the states.

We compared the chromatin segmentation of mESC and CM in three regions. One of them containing a pluripotency marker (*Nanog*), thus expressed in mESC but not in CM. Another region containing a marker of CM (*Tbx5*), expressed in CM but not in mESC. Finally, a region containing a housekeeping gene (*Polr2b*) expressed in both (Figure 33).

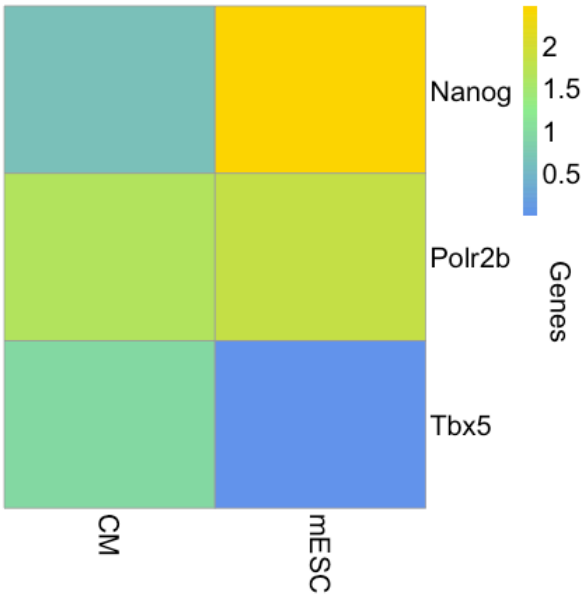



*Figure 33:* **Nanog, Polr2b and Tbx5 expression in mESC and CM.**

In CM, the TSS of *Nanog* is marked by state 6, a repressed state since it contains H3K27me3. Moreover, its gene body is covered by state 8 which is an unmarked state. On the other hand, in mESC, *Nanog* is covered by state 5 which contains H3K4me3, an active mark (Figure 33A). Regarding *Tbx5*, we observed the opposite. Its TSS is marked by state 6 in mESC, whereas in CM is marked by state 4, which is an active state since it contains H3K4me3. Moreover, the gene body of *Tbx5* is marked by state 9 in CM, which is an active state since it contains H3K36me3, whereas in mESC is marked by state 1, an empty state (Figure 33B). Finally, the TSS of *Polr2b* is marked by state 5 and its gene body by state 9 in both cell lines, which are active states (Figure 33C).

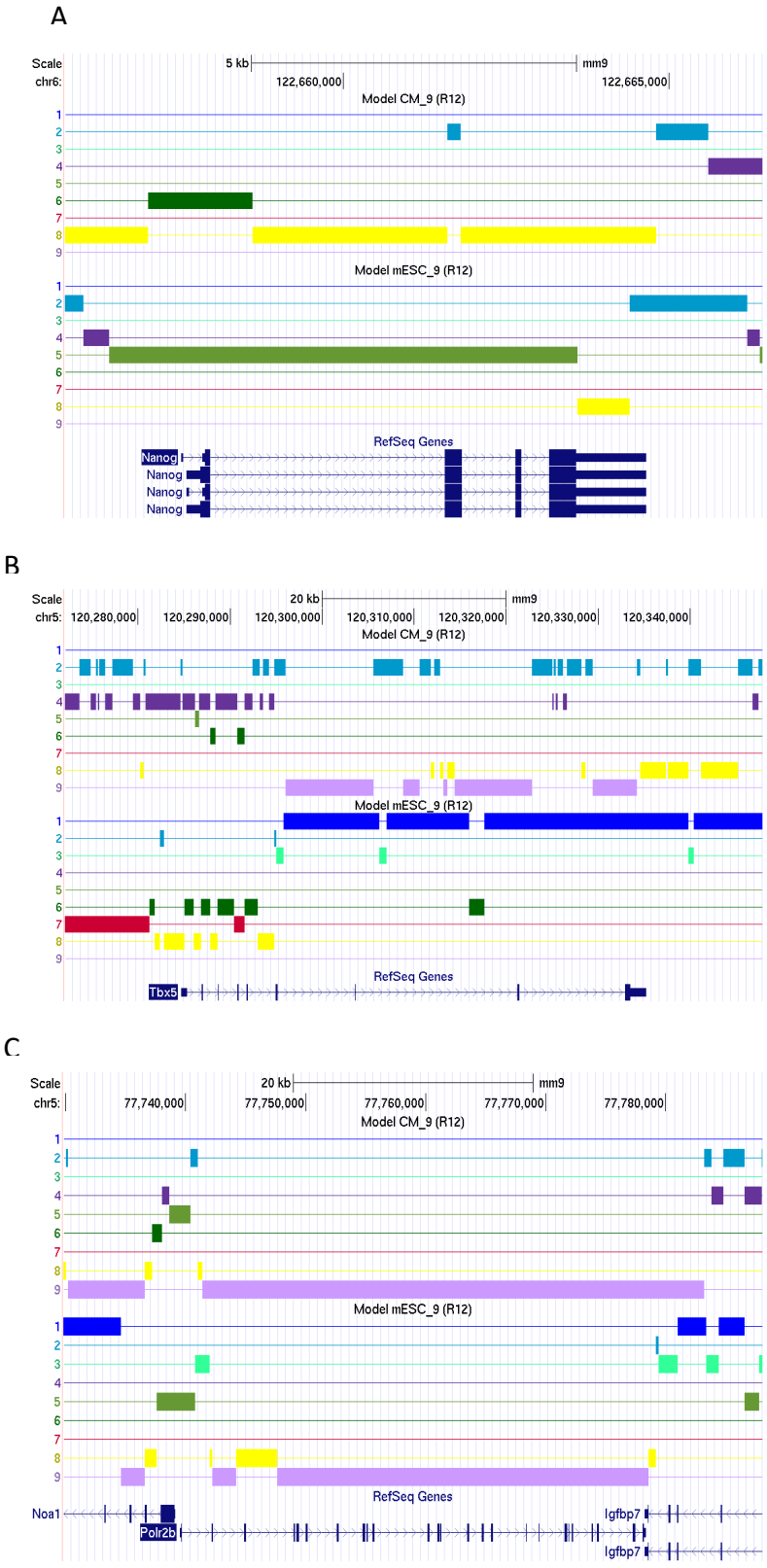



***Figure 34:*** **Genome segmentation of mESC and CM in three different regions.**

(A) Genome segmentation of mESC and CM in the region of the pluripotency marker *Nanog*.

(B) Genome segmentation of mESC and CM in the region of the CM marker *Tbx5.*

(C) Genome segmentation of mESC and CM in the region of the housekeeping gene *Polr2b.*

## 3.8 Other chromatin segmentation methods

We have also tested other segmentation methods besides ChromHMM. In all cases we did not get positive results. We tried jMOSAiCS, but it is no longer available for the current versions of R. We examined IDEAS but it does not work under MAC OS X, and we will need therefore to prepare a virtual machine with enough memory resources. We tested chromstaR but we did not manage to obtain biologically meaningful results using different parameters (Table 5), as the models were unexpectedly decorating expressed genes with repressive states (Figure S4).

## 3.9 Scripts developed for this project

To our knowledge, no scripts are available to perform the comparison of two different chromatin segmentations. Therefore, we developed two scripts in Python with the objective of comparing two different segmentations (Figure 34). First of all, we needed a script to bin the output of ChromHMM, so that we could compare bin by bin two segmentations. The script, which we called BinModels.py, takes as input the segments.bed file, and divides each segment into bins of equal size.

The second script is the one which makes the comparison between the previously binned segmentations. It is called compare2models.py. It calculates the number of common bins between two states from two models and, also the total number of bins in each state. Next, compare2models.py calculates the similarity between two states according to the formula in Figure 34, stores the values into a matrix. Finally, the script outputs a heat map that shows the results of the comparison state by state, in which yellow means high similarity and blue low similarity.

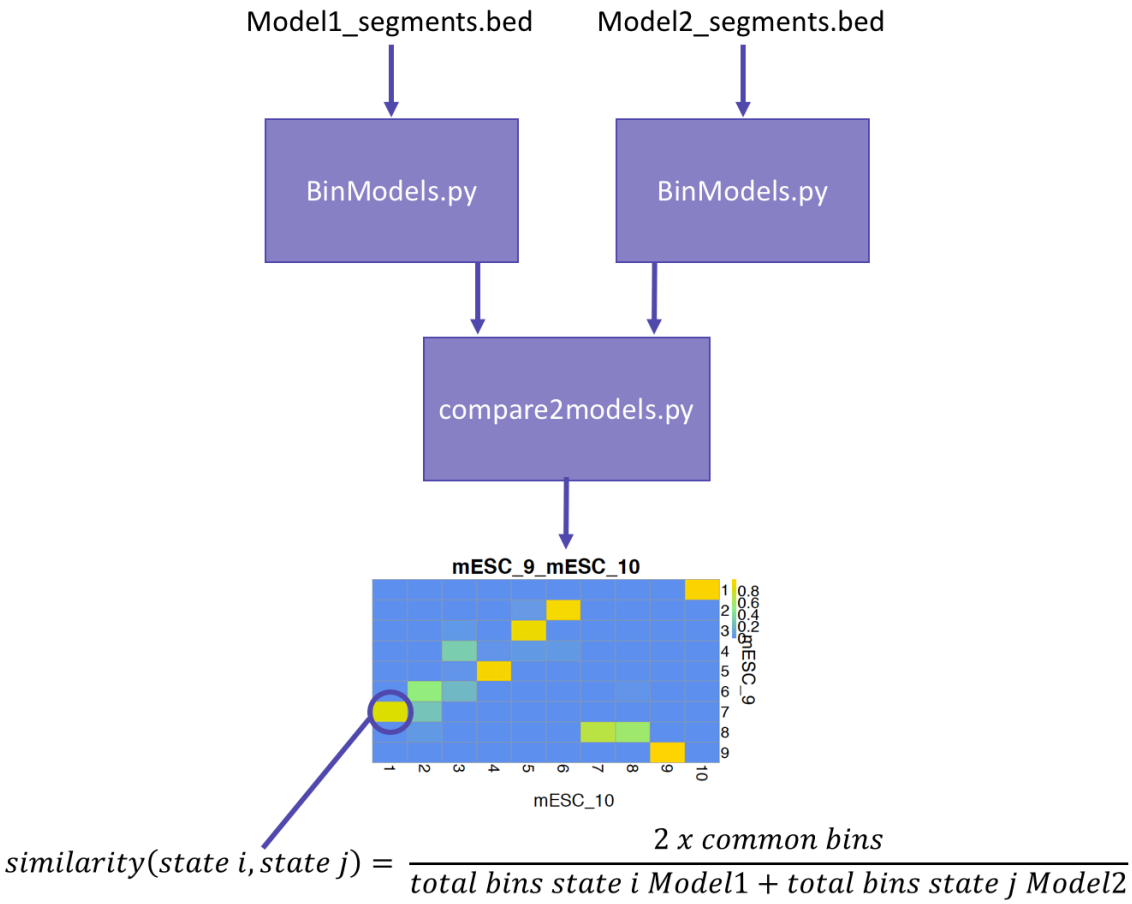



*Figure 35:* **Scripts to compare models.**

Moreover, we also developed a third script named StateCoverage.py, which calculates the genome coverage of every state in the model. The script takes as input a binned segmentation obtained from BinModel.py and calculates the percentage of bins marked by every state.

# 4. Discussion

During this project we have addressed the problem of extracting novel knowledge from the increasing number of NGS data. Here we propose that chromatin segmentation methods, such as ChromHMM, might be a possible solution.

First of all, we have focused on a basic histone model using six histone marks (H3K4me3, H3K36me3, H3K4me1, H3K27ac, H3K27me3, H3K9me3), where we have shown its utility to describe the epigenomic landscape of mESC. We have been able to assign each state to an epigenetic feature such as active promoters, active gene bodies, bivalent promoters, etc. Our results confirmed what was previously known about these histone marks, suggesting that chromatin segmentation methods might be a useful tool to understand the interplay between epigenetic factors and its implication in gene regulation. However, we consider that chromatin segmentation maps have not been extremely used to extract novel information yet.

We have also generated chromatin segmentation maps to address different biological questions, such as the biological meaning of intergenic H3K27me3, or the PRC2 complexity. We have shown that the potential of the models lies on the study of those interesting states which help to solve the biological question proposed, rather than finding a comprehensive model which describes the whole epigenomic complexity. Moreover, we have observed that one of the key potentials of chromatin segmentation maps relies on the comparison between different cell lines or conditions, to understand the epigenetic changes occurring along these transitions.

Despite its potential, no software is available to help interpreting chromatin segmentation models. In this sense, during this project we had the need to compare two chromatin segmentations, thus we developed our own script with this purpose. Moreover, further advances in this field need to be addressed. For

instance, when comparing two conditions, a second output with the list of transitions occurring from one condition to the other is necessary to further study them to understand the differences between both conditions. Another possible improvement after the comparison of two segmentations would be to get a measure of the similarity between the two models besides the similarity between the states. The similarity between models will help during the evaluation step to select the number of states of the model that one would like to study. Moreover, we have shown that independent models constructed from different data can be compared to extract novel knowledge as we did when comparing the histone model and the PcG model.

Besides the need of going further with the technical aspects emerged from this project, the models generated during this time can be also further analysed to extract novel biological knowledge. For example, when studying poised enhancers in the model using three histone marks and two PRCs, we could extract the intergenic and intronic segments of state 1 which are the top candidates for being poised enhancers. We could check their ChIP levels to confirm that they have the characteristic marks of poised enhancers and also ChIP levels from PcG subunits. We could try to assign the enhancers to their target genes, and check if the categories from a functional analysis are related to development as one might expect, thus favouring the hypothesis of being a poised enhancer state.

Moreover, in the PcG model where we studied PRC2 complexity, we could also study PRC1 complexity. In fact, we observed that state 1 had a high emission of Ring1B and RYBP, both components of PRC1, but the emissions of PRC2 components where very low. Moreover, Cbx7 had also a low emission in this state, suggesting that these regions might be targets of PRC1-RYBP and not targets of PRC1-Cbx7. As we know, RYBP and Cbx7 are mutually exclusive (Morey, Aloia et al. 2013) and thus we might have found a state to better understand the functional differences between this two factors. We could do a functional analysis of the target genes of state 1, and also check their expression.

Regarding the model using two cell lines, further progress need to be done. Despite the technical aspects mentioned before, we also noticed that the approach to construct the models could be improved. We observed that some states are masked when obtaining the model for the two cell types at once, for example, a poised enhancer state was missing. In fact, we know that this state exists in mESC since we have seen it in the models generated for mESC alone. One possible explanation is that CM might not have this state, thus masking it from the model. Moreover, this could also be happening the other way around, and the presence of mESC data might be masking CM-specific sates. Possible strategies to solve this problem are to obtain both segmentations independently or to obtain one segmentation for both. This second strategy consists in considering that we only have one cell line, but labelling the marks as mark plus cell line (for instance, H3K27me3-mESC). One additional advantage of this strategy is that we will get also the regions which are transitioning from one state to another, however it will increase considerably the number of states.

Once the technical aspects mentioned during this section are solved, the biological questions one might address are innumerable. One straightforward aspect to study deduced from our experience, is the changes in states from active/repressed regions during differentiation such as from mESC to CM. Since little is known about poised enhancers and their activation during differentiation, chromatin segmentation maps seem to be a good strategy. Moreover, since PcG misregulation has been linked to cancer (Pasini and Di Croce 2016), and many ChIPseq data of cancer cell lines and tissues has been generated, chromatin segmentation models might be a good strategy to understand their epigenetic landscape and unravel the PcG effect during cancer. In fact, it would be interesting to look at differences in epigenetic states between cancer tissues and healthy tissues, being the comparison of their chromatin segmentations a good approach.

Despite we have focused in ChromHMM, all these suggestions might also be applicable to other chromatin segmentation methods. We have tested other methods such as chromstaR and jMOSAiCS. However the results have not been informative in all cases. Many other methods are available, and their

comparison them would be useful to select the best method for each possible application of chromatin segmentation models.

Finally, another issue addressed during this project is how to make all this source of new knowledge accessible to the scientific community. Here we have implemented a web server with this purpose where one can find some of the models generated during this time. Another issue would be how to facilitate the analysis of chromatin state models to the wet lab biologist. For example, it would be interesting to implement another web server to compare chromatin segmentations, which might output a measure of the similarity between the states or the models, the list of transitions between the two segmentations, etc. depending on the user interest.

In conclusion, chromatin segmentation methods seem a promising tool to study gene regulation under different scenarios despite they have been barely used, except for generating preliminary representations. However, further improvements in technical aspects need to be done to extract all the novel information that chromatin segmentation models can offer.

# 5. Conclusions

During this project we have learned how to analyse different types of NGS data, such as ChIPseq, RNAseq and ATACseq experiments. These experiments are used to study gene regulation, thus the amount of data is constantly growing. Therefore we have learned about chromatin segmentation methods which might be a solution to address this problem. Throughout our experience, we concluded that the potential of the models lies on the study of those states which help to solve the biological question proposed, rather than finding a comprehensive model which describes the whole epigenomic complexity.

We have accomplished all the objectives previously specified. However, to do so, we had the need to develop tools to compare chromatin state segmentations, adding a new objective into the work plan. Despite this modification, we could accomplish all the objectives and tasks planned.

During this project, we have noticed the lack of technical resources to study chromatin segmentation maps in detail. Therefore, this is a key line of work for the future. We have moved forward into this aspect, nevertheless further advances need to be done. For example: calculating a measure of similarity between two models, and obtaining the list of regions transitioning from one state towards another one. Another technical aspect to explore in the future, is how to obtain models to compare two cell lines, since we have seen that constructing the model for both at the same might not be the best approach.

Moreover, a key point to keep in mind is how to facilitate the accession to this new information and resources to wet lab biologists. Therefore, a future plan might be to implement different web servers to visualize the novel information and also to analyse and compare different chromatin segmentations.

As also discussed in Section 4, our models could be further analysed. For example: state 1 of the PcG model to study the PRC1 complexity and also we

could dig further into poised enhancers by studying intergenic and intronic segments of state 1 in the model using three histone marks and two PcG subunits.

Moreover, other biological questions can also be addressed using chromatin segmentation methods. For example, as also mentioned in Section 4, this approach would be interesting to study the epigenetic changes that occur during differentiation. For instance, from mESC towards CM as we started to study during this project. Furthermore, it would also be interesting compare the epigenomic landscape of healthy versus cancer tissues.

Finally, different chromatin segmentation methods should be compared in the future. ChromHMM is the standard tool in the field, but other methods might also be useful for different applications.

# 6. Glossary

**ATACseq**: Assay for Transposase Accessible Chromatin with high-throughput sequencing

**Cbx7**: Chromobox protein homolog 7

**ChIPseq**: Chromatin ImmunoPrecipitation assays with high-throughput sequencing

**CM**: CardioMyocyte

**Epop**: Elongin BC and Polycomb repressive complex 2-associated protein

**Jarid2**: Protein Jumonji

**H3K27ac**: acetylation of lysine 27 in histone H3

**H3K27me3**: trimethylation of lysine 27 in histone H3

**H3K36me3**: trimethylation of lysine 36 in histone H3

**H3K4me1**: monomethylation of lysine 4 in histone H3

**H3K4me3**: trimethylation of lysine 4 in histone H3

**H3K9me3**: trimethylation of lysine 9 in histone H3

**mESC**: mouse Embryonic Stem Cell

**Myh6**: Myosin-6

**Nanog**: Homeobox protein NANOG

**PcG**: Polycomb Group of proteins

**Polr2b**: DNA-directed RNA polymerase II subunit RPB2

**PRC1**: Polycomb Repressive Complex 1

**PRC2**: Polycomb Repressive Complex 2

**Ring1B**: E3 ubiquitin-protein ligase RING2

**RNAseq**: RNA sequencing

**Rybp**: RING1 and YY1-binding protein

**Sox2**: Transcription factor SOX-2

**Suz12**: Polycomb protein SUZ12

**Tbx5**: T-box transcription factor TBX5

**TrxG**: Trithorax Group of proteins

# 7. References


Aranda, S., G. Mas and L. Di Croce (2015). "Regulation of gene transcription by Polycomb proteins." Sci Adv **1**(11): e1500737.

Beringer, M., P. Pisano, V. Di Carlo, E. Blanco, P. Chammas, P. Vizan, A. Gutierrez, S. Aranda, B. Payer, M. Wierer and L. Di Croce (2016). "EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells." Mol Cell **64**(4): 645-658.

Connelly, K. E. and E. C. Dykhuizen (2017). "Compositional and functional diversity of canonical PRC1 complexes in mammals." Biochim Biophys Acta **1860**(2): 233-245.

Di Croce, L. and K. Helin (2013). "Transcriptional regulation by Polycomb group proteins." Nat Struct Mol Biol **20**(10): 1147-1155.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Ernst, J. and M. Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization." Nat Methods **9**(3): 215-216.

Harikumar, A. and E. Meshorer (2015). "Chromatin remodeling and bivalent histone modifications in embryonic stem cells." EMBO Rep **16**(12): 1609-1619.

Karolchik, D., A. S. Hinrichs and W. J. Kent (2007). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit 1 4.

Kuleshov, M. V., M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen and A. Ma'ayan (2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." Nucleic Acids Res **44**(W1): W90-97.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Morey, L., L. Aloia, L. Cozzuto, S. A. Benitah and L. Di Croce (2013). "RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells." Cell Rep **3**(1): 60-69.

Morey, L. and K. Helin (2010). "Polycomb group protein-mediated repression of transcription." Trends Biochem Sci **35**(6): 323-332.

Morey, L., A. Santanach, E. Blanco, L. Aloia, E. P. Nora, B. G. Bruneau and L. Di Croce (2015). "Polycomb Regulates Mesoderm Cell Fate-Specification in Embryonic Stem Cells through Activation and Repression Mechanisms." Cell Stem Cell **17**(3): 300-315.

*Morgan M, Pagès H, Obenchain V and Hayden N (2017). Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.28.0, http://bioconductor.org/packages/release/bioc/html/Rsamtools.html.*

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy and K. D. Pruitt (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic Acids Res **44**(D1): D733-745.

Pasini, D. and L. Di Croce (2016). "Emerging roles for Polycomb proteins in cancer." Curr Opin Genet Dev **36**: 50-58.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn and J. Wysocka (2011). "A unique chromatin signature uncovers early developmental enhancers in humans." Nature **470**(7333): 279-283.

Schoenfelder, S., R. Sugar, A. Dimond, B. M. Javierre, H. Armstrong, B. Mifsud, E. Dimitrova, L. Matheson, F. Tavares-Cadete, M. Furlan-Magaril, A. Segonds-Pichon, W. Jurkowski, S. W. Wingett, K. Tabbada, S. Andrews, B. Herman, E. LeProust, C. S. Osborne, H. Koseki, P. Fraser, N. M. Luscombe and S. Elderkin (2015). "Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome." Nat Genet **47**(10): 1179-1186.

Taudt, A., Nguyen, M.A., Heinig, M., Johannes, F. and Colome-Tatche, M. chromstaR: Tracking combinatorial chromatin state dynamics in space and time. bioRxive. Unpublished (18/10/2016)

Tavares, L., E. Dimitrova, D. Oxley, J. Webster, R. Poot, J. Demmers, K. Bezstarosti, S. Taylor, H. Ura, H. Koide, A. Wutz, M. Vidal, S. Elderkin and N. Brockdorff (2012). "RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3." Cell **148**(4): 664-678.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

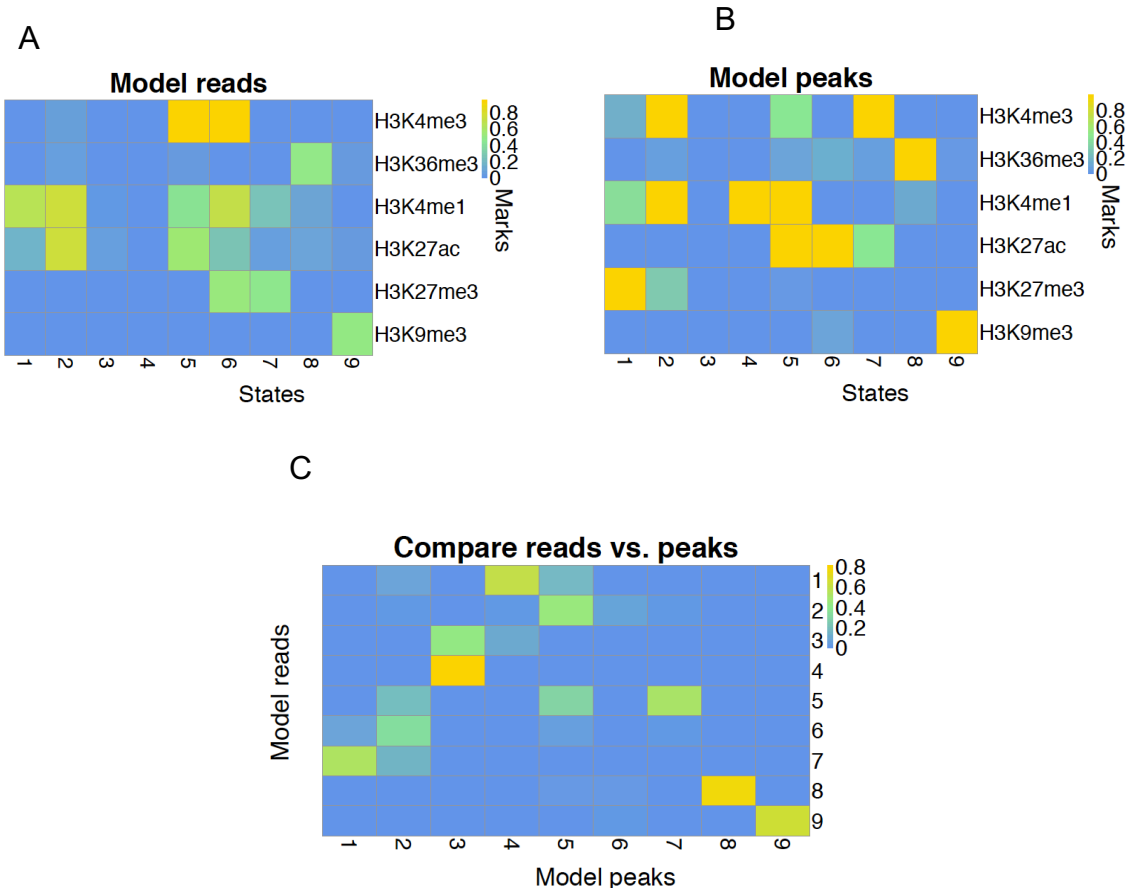
Vizan, P., M. Beringer, C. Ballare and L. Di Croce (2015). "Role of PRC2-associated factors in stem cells and disease." FEBS J **282**(9): 1723-1735.

Voigt, P., W. W. Tee and D. Reinberg (2013). "A double take on bivalent promoters." Genes Dev **27**(12): 1318-1338.

Wamstad, J. A., J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer and B. G. Bruneau (2012). "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage." Cell **151**(1): 206-220.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

# 6. Annexes




D

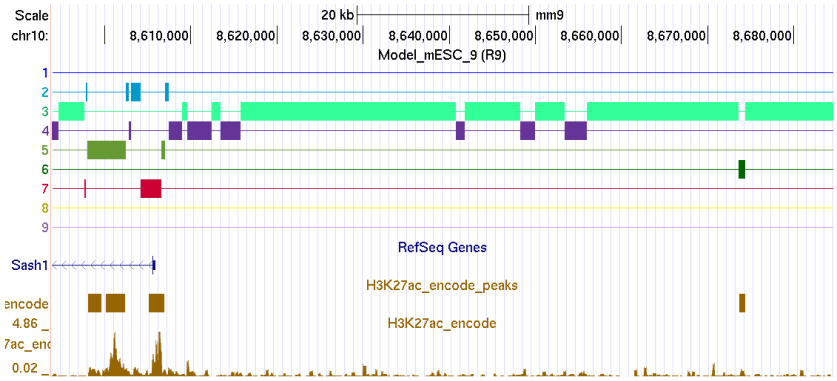



***Figure S1:*** **Comparing the use of peaks versus reads to construct the histone model.**

(A) State emissions of the model using reads, represented in probability to find a specific mark in each of the states.

(B) State emissions of the model using peaks, represented in probability to find a specific mark in each of the states.

(C) Comparison of the model using peaks and the model using reads.

(D) Example of H3K27ac peaks in states 5 and 7, and state 6.

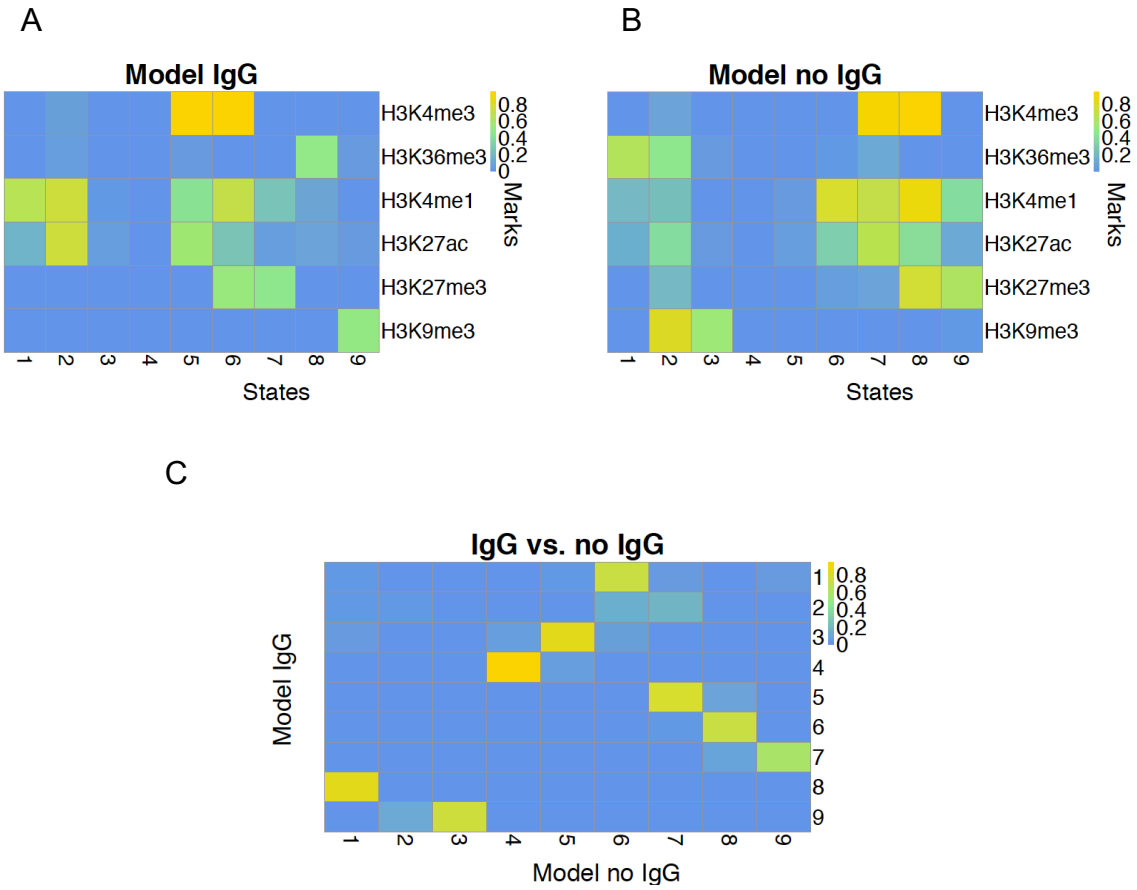



***Figure S2:*** **Comparing the use of IgG as a control or no control.**

(A) State emissions of the model using IgG as a control, represented in probability to find a specific mark in each of the states.

(B) State emissions of the model using no control, represented in probability to find a specific mark in each of the states.

(C) Comparison of the model using IgG as a control and the model using no control.

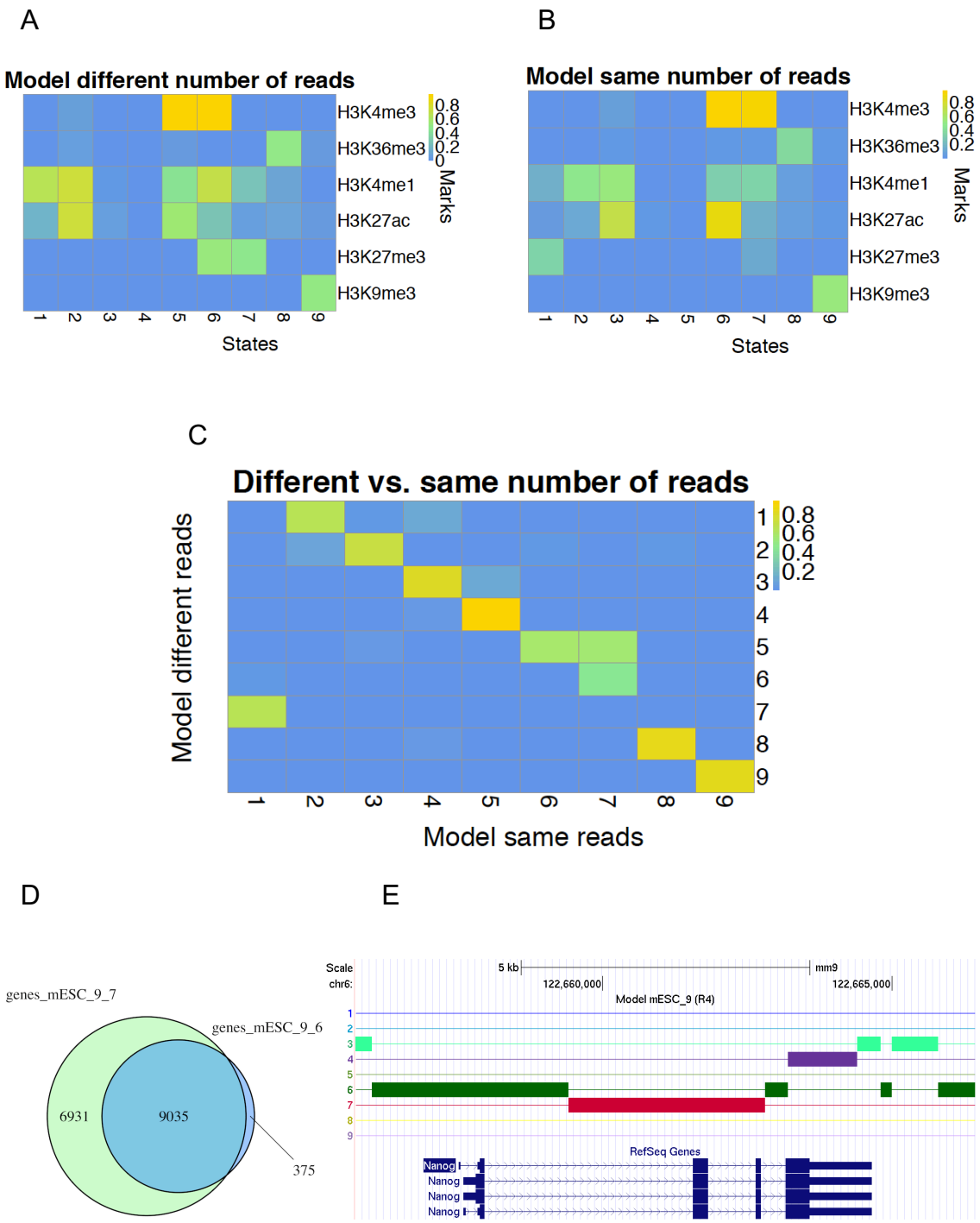



***Figure S3:*** **Comparing the use of different number of reads or same number of reads.**

(A) State emissions of the model using different number of reads, represented in probability to find a specific mark in each of the states.

(B) State emissions of the model using the same number of reads, represented in probability to find a specific mark in each of the states.

(C) Comparison of the models using different number of reads or same number of reads.

(D) Overlapping of states 6 and 7 target genes of the model using the same number of reads.

(E) Example of the pluripotency gene Nanog, which should be marked by state 6 only, in the model with the same number of reads since it is an active gene.

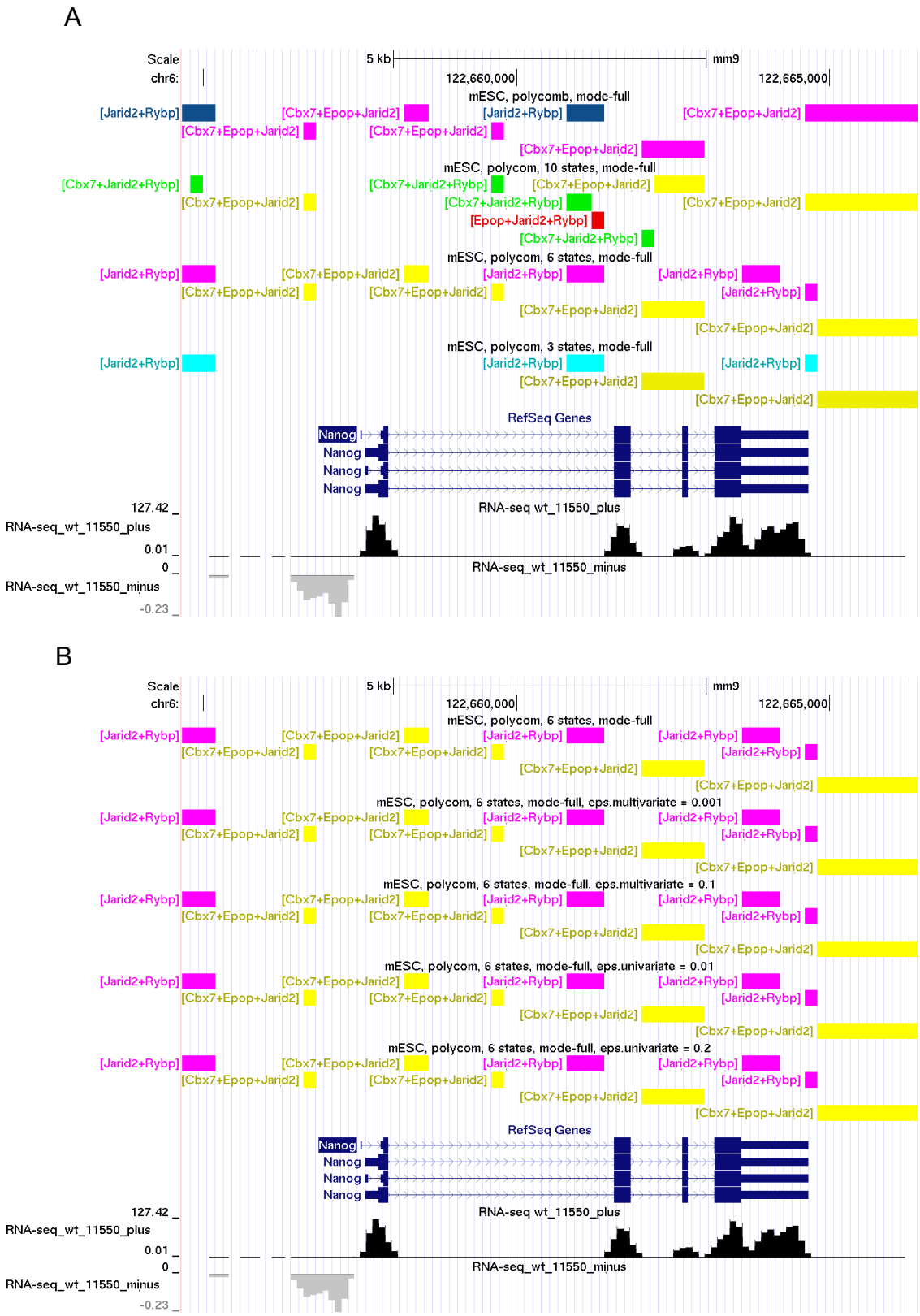



*Figure S4:* **Segmentations by chromstaR using different parameter setting.**

(A) Assessing the use of different number of states.

(B) Assessing different values for eps.univariate and eps.multivariate.