PLACE HOLDER

Maurizio Gilioli

Contents

1 ABSTRACT

2 INTRODUCTION

2.1 Chromatin as the information center of a cell

[1] All the living organisms possess DNA, which is the main molecule through which information are passed from a generation to another. Chromatin is contained inside the nucleus in an ordered manner[1] . DNA is wrapped around histones forming nucleosomes. Throughout the report, I will call the nucleomes beads, which is a term that underlines the spherical shape of the DNA-histone complex. DNA and histones are subjected to different modifications; among those, methylation is the most important modification involving DNA. Methylation in mammals occurs in specific sites of the genome, called CpGs, where a cyto- sine is connected directly to a guanine. Methy- lations of regulatory elements have been im- plicated in determining cell identity and chro- matin structure[?,2] . CTCF is a protein conserved in eukariotes and is ubiquitous in mammalians[3] . It contains a Zync-finger which binds to DNA. The act of binding is performed in cooperation with cohesins, and causes the folding of the chromatin.

2.2 ATAC-sequencing and CTCF sequencing

ATAC-sequencing is a technology that allows for the identification of open- chromatin regions[4,5]. In order to work, it requires the addition of TN-5, a hyper-active transposase. The latter is preloaded with sequencing adapters[5] to induce a contempourary reaction of fragmentation and ligation of the pieces released, in a process called segmentation. The obtained adapted fragments are then amplified and sequenced. Once the reads are generated, a peak-calling algorithm (in our case MACS-2[6]) is used to determine which portions of the genome present ATAC peaks, and areas where there are significant enrichments of aligned reads with respect to the background. A significant enrich- ment of reads is possible only in accessible regions, which are generally also the most active ones and with available sites for transcription factors binding. CTCF data, named in chapter ..., were obtained through a classical CHIP-sequencing.

2.3 Chromatin Coarse-Graining, chromatin as a polymeric fluid

The Young's modulus (E) of a chain is the extent to which a solid material (or a polymeric fluid, in this case) can be deformed As Robert Hook noticed, the following is valid

$$\sigma = E\frac{\Delta l}{l} \tag{1}$$

Where l represents the length of the chain and Δl the deformation $\sigma^?$.

The entanglement length (N_e) corresponds to the Young's modulus that is experimentally found in the plateau region where a force starts to produce irreversible deformations in a chain.

It is also defined as the "the average number of monomer units along the chain between two nearest effective cross-links." and is related to the ability of chains to form knots between each other[?] .

2.3.1 Persistence length of a polymer chain

The persistence length (PL) of a polymer represents the degree of bendability of the chain. During this project, the persistence length is shown in equation 6.

$$PL = lk_{CG}/b_{CG}/2 \tag{2}$$

With the idea of following a "journey" on the polymer chain, the average angle that you obtain at a contour length s is as in equation 3.

$$\langle \cos \theta(s) \rangle = \exp\left(-\frac{s}{l}\right) \tag{3}$$
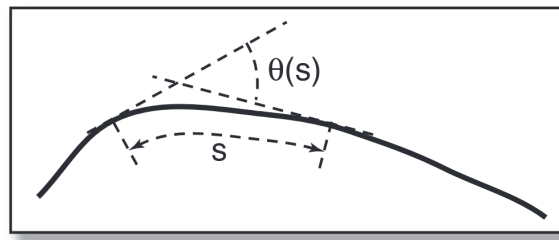

Figure 1. examplecaption

2.4 ChromHMM allows the sequential characterization of DNA regions

ChromHMM is a tool which helps in the annotation of genomic DNA by using epigenomic information[7] . The way it learns chromatin states signatures by using a multivariate hidden Markov model: In each genomic position, it returns the most probable chromatin state (segments) and other useful information, such as the emission/transition parameters of the states, the abundance of the states at the TSS (Transcriptional Starting Site), at the TES (Transcriptional Ending site), and other important portions of the genome (CPG islands, exons, genes)[7,8].

The package works through two functions in particular, which are the following[7] :

1. ***BinarizeBam***: converts a set of bam files of aligned reads into binarized data files in a specified output directory, which can then be used as input to LearnModel. When using this command, it has to be specified the bin size, that in the case of this project was set to be 200 bps.

2. ***LearnModel***: takes a set of binarized data files, learns chromatin state models, and by default produces a segmentation, generates browser output with default settings, and calls OverlapEnrichment and NeighborhoodEnrichments with default settings for the specified genome assembly. A webpage is a created with links to all the files and images created .

The results obtained are shown in chapter 5.1.

3 Aim of the project

The project is part of the thesis whose aim is to predict matrices of contact of chromatin through the results obtained with molecular coarse-grained simulations of 2 Mb portions of the chromatin. The scope of the report is also to gather opinions and useful feedback to improve the thesis work, which will continue for another two months.

4 METHODS

All the simulations were performed making use of the simulation software LAMMPS[?,9] .

4.1 Data used for the project

The data of CTCF and ATAC for the IMR90 cell line included in the paper written by Jimin and colleagues in 2023 about the Origami simulation tool[10] were used for the project (see table 1).

Cell-Type	CTCF ChiP-seq	ATAC-seq
IMR90	ENCSR000EFI	ENCSR200OML

Table 1. Table referring to the data used for the analysis. All of them were used for the training also of the Origami[10] model. The written entries are for the ENCODE database[11].

The analyzed region was ANPEP in IMR90, taking into account the bases between the 89'000'000th and the 91'000'000.

4.2 Finding enriched states in 5 kbs long beads

To find the enriched states, the procedure in process 1 was followed. The fold change for a state in a bin was found by taking the ratio of the proportion in the specific chromosome and in the bin. Once done that, the state with the highest fold-change was assigned to the bin.

Algorithm 1: Finding enriched states in 5-kb long bins

Result: Enriched states

$n = 0$;

forall *chromosomes* **do**
    │ Find proportions of the states in the chromosome

foreach *bin* **do**
    │ Calculate bin proportions for each state Compute the fold changes Assign the state with the highest fold change

Generate .bed files Visualization in IGV of regions of interests /* each bin is 5 kb long   */

4.3 The Model

The objective was to create a polymer model with a coarse-graining resolution of 5000 bp. The region of interest was ANPEP, it had a length of 2000000 bp and was considered in human genomes. The model creation was done by using code written by Marco Di Stefano. The total Genome length was obtained from the UCSC genome browser[12] . All the simulations were performed making use of periodic boundaries (chapter

All the simulations were performed making use of the *run_lammps* function inserted in the TADPHYS package. The *bond_style* that was set for LAMMPS[13] was the the *fene* bond style whose potential is written in equation 4.

The FENE potential is a finite extensible nonlinear elastic potential and is generally used for polymer models. The first term is attractive, the second Lennard-Jones (LJ) term is repulsive. The first term extends to R_0, the maximum extent of the bond. The second term is cutoff at $2^{1/6}\sigma$, the minimum of the LJ potential[13] . The K is an energy/distance² measure

$$E = -0.5KR_0^2 \ln\left[1 - \left(\frac{r}{R_0}\right)^2\right] + 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6\right] + \varepsilon \tag{4}$$

The following specifications were made for the FENE interactions:

1. 30.0: Maximum force the bond can withstand. It represents the stiffness or strength of the bond.

2. 1.5: Maximum extension of the bond. This is the maximum distance at which the bond can be stretched.

3. 1.0: Equilibrium bond length. This is the ideal or equilibrium distance between the bonded particles.

4. 1.0: Bond force constant. It affects how quickly the potential energy increases as the bond length deviates from the equilibrium length.

To allow for the excluded-volume interactions, a simple Lennard-Jones potential was included, written as in equation

$$E_{LJ} = 4\varepsilon \left[\left(\frac{r_0}{r} \right)^{12} - \left(\frac{r_0}{r} \right)^{6} \right] \qquad\qquad r < r_c$$

$$= 4\varepsilon \left[\left(\frac{2^{1/6}\sigma}{r} \right)^{12} - \left(\frac{2^{1/6}\sigma}{r} \right)^{6} \right] \qquad\qquad r < r_c \qquad\qquad (5)$$

Three parameters are set:

1. ε (energy unit): 1.0, Note that σ is defined in the LJ formula as the zero-crossing distance for the potential, not as the energy minimum at $r_0 = 2^{1/6}\sigma$.

2. σ (distance unit): 1.0

3. LJ cutoff (distance unit): 1.12246152962189

The steps performed will be described through the following dot list:

The computation of the parameters for Coarse Graining: Both parameters for the CG model (table **??**) and the CG model were calculated. The number of bp wrapping around a bead in the FS method was considered to be 150, while instead the linker portion was considered to be of length 50 bp (table **??**). The thickness of a bead (of a nucleosome) was taken as equal to 10 nm, while instead the default Kuhn length was set to 50 nms. The genome densities ($\rho_{FS} = \rho_{CG} = 0.012$ bp/nm^3) are imposed to be the same for the FS and the CG model.

. The beads and bonds have all the same length, consequently, the contour length is exactly the product between the number of beads and the size of the beads in the FS and in the CG model. The DNA content was 2000000 bp. To find the parameters for the CG model, the DNA content of the Kuhn segments (Dlk_{CG}) were tuned to match the desired value of DNA content in CG beads (v_{CG}).

Property	Formula	Value
c	*const.*	19
v_{FS} (DNA content of a monomer in b.)	*const.*	150+50 bp = 200 bp
b_{FS} (Diameter of a bead in nm)	*const.*	10 nm
lk_{FS} (Kuhn length of the chain in FS)	*const.*	50 nm
ρ_{FS} (Genome density)	*const.*	0.012 bp/nm^3
N_{FS} (Number of monomers to represent the chromosome)	$\frac{DNAcontent}{v_{FS}} * ncopies$	30000 mon.
N_{FS}^k (Number of Kuhn lengths of the chain)	$\frac{N_{FS}*b_{FS}}{lk_{FS}}$	6000 k. l.
ρ_{FS}^k (Genome density in Kuhn lengths)	$\frac{\rho_{FS} \cdot b_{FS}}{v_{FS} \cdot lk_{FS}}$	1.2e − 05 1/nm³
L_{FS} (Polymer contour length)	$N_{FS} * b_{FS}$	300000 nm
Le_{FS} (Entanglement length of the chain in nm)	$lk_{FS} * \left(\frac{c}{\rho_{FS}^k * lk_{FS}^3} \right)^2$	8022.22 nm
Number of monomers in a Kuhn length FS	lk_{FS}/b_{FS}	5 mon.
Blk_{FS} (Bead content of a Kuhn length FS)	$(lk_{FS} \cdot b_{FS})/v_{FS}$	2.5 nm^2/bp
Dlk_{FS} (DNA content of a Kuhn length FS)	$(lk_{FS} \cdot v_{FS})/b_{FS}$	1000 bp

Table 2. Parameters calculated for the Fine Scale (FS) model

Property	Formula	Value
c	*const.*	19
v_{CG} (DNA content of a monomer in b.)	*const.*	5000 bp
Dlk_{CG} (DNA content of a Kuhn length CG)	*tuned const.*	33791 bp
ϕ_{CG} (Volumetric density of the chain in the CG model for IMR90 cell-type)	*const.*	0.1
ρ_{CG} (Genome density in bp/nm³)	*const.*	$0.012\ \mathrm{bp/nm^3}$
b_{CG} (Diameter of a bead in nm)	$\sqrt{\left(\sqrt{\frac{Dlk_{CG}}{Blk_{FS}}}\right)/\rho_{CG}\cdot\frac{6}{\pi}\cdot\phi_{CG}}$	43.0155 nm
lk_{CG} (Kuhn length of the chain in CG)	$\sqrt{Dlk_{CG}*Blk_{FS}}$	290.65 nm
Number of monomers in a Kuhn length CG	lk_{CG}/b_{CG}	6.75687 mon.
N_{CG} (Number of monomers to represent the chromosome)	$\frac{\mathrm{DNAcontent}}{v_{CG}}*\mathrm{ncopies}$	1200 mon.
$side_{CG}$ (size of the cubic simulation box)	$\frac{(N_{CG}\cdot v_{CG}/\rho_{CG})^{1/3}}{b_{CG}}$	18.4515 nm
N_{CG}^{k} (Number of Kuhn lengths of the chain)	$(N_{CG}*b_{CG})/lk_{CG}$	177.597 k. l.
ρ_{CG}^{k} (Genome density in Kuhn lengths bp/nm)	$\frac{\rho_{CG}\cdot b_{CG}}{v_{CG}\cdot lk_{CG}}$	3.55194e-07 ᵇᵖ/nm
L_{CG} (Polymer contour length)	$N_{CG}*b_{CG}$	51618 nm
Le_{CG} (Entanglement length of the chain in nm)	$lk_{CG}*\left(\frac{c}{\rho_{CG}^{k}*lk_{CG}^{3}}\right)^{2}$	1379.51 nm

Table 3. Parameters calculated for the coarse-grained (CG) model

Generation of the initial conformation rosettes: Once found the coarse graining parameters, rosettes for 100 replicates were built, with a radius of 12.0 nm inside a cubic box of 300 nm. The particle radius was set to 0.5 nm. In each replicate, three equal chains were built, by setting a different random seed each time. The total number of particles in each chain was of 400 beads.

Finding the optimal pressure: Once the rosettes were made, a decompaction was performed starting from the compacted rosettes. A range of values of pressure was tested from 0.1 to 1 with steps every 0.1. For each replicate, a new random seed was generated and stored. Before attempting the decompression, the minimal energy structure was found by taking into consideration a stopping energy tolerance of $1*10^{-4}$, a stopping tolerance force of $1*10^{-6}$, a maximal number of iterations and evaluations of 100000 steps. The process was long 1000 steps with a duration of 0.001 ps. and the final structure was taken as the decompressed one. persistence length (PL) regulates the potential produced by angle interaction. The optimal pressure was attested at 0.192

$$PL = lk_{CG}/b_{CG}/2 \qquad\qquad (6)$$

Decompaction and relaxation: Once the optimal value for the pressure was found (0.192) for each replicate, other two simulations respectively $5,000,000$ and $25,000,000$ steps long were performed (not after a minimization phase). This time, the step was set to have a temporal length of 0.0012 ps In both the cases MSD values were collected every 100 steps. A frame was dumped every 5000 steps. At the end, the trajectories were collected and analyzed by computing the *RMSD*, *Rg* and the autocorrelation function as described in chapter 4.4. For the sake of simplicity, the collection was accomplished by capturing one frame for every 50,000 steps.

Computing matrices of contact: Once defined the step at which the simulations were considered to be at the equilibrium, some dictionaries produced with *ChromHMM*[7,8] , by using CTCF and ATAC-seq data, were used to define the identity of the beads. During the next steps, we will try to select the best attraction parameters; the stepwise algorithm 3 is being currently tested. All the beads were considered to have a radius of 0.5 long. To quantify attraction potentials, a new Lennard-Jones potential was added (as defined in equation 5). In particular, the cutoff distance r_0 was set to be equal to:

$$r_0 = r_{\mathrm{cutoff}} = \sigma*2.5$$

Once the interaction parameters are set, other 1 second long simulations are performed with the intention of generating contact maps. Only the interactions occurring intra-chain were considered.

4.4 Trajectories analysis
Three types of analysis were performed (all the terms are explained in the Glossary):
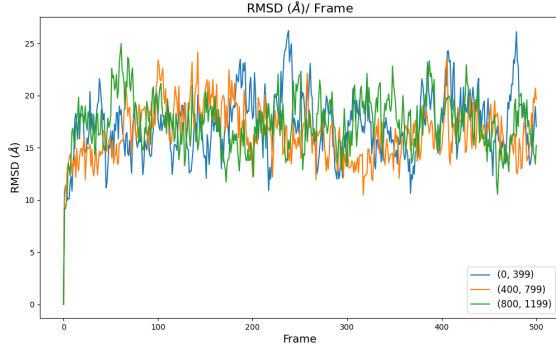
(a) Graph representing the RMSD of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.
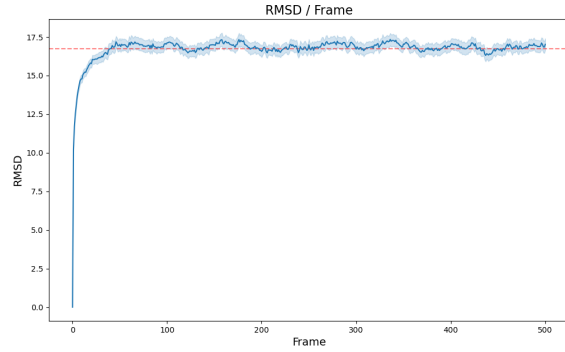
(b) Figure representing the collective behaviour of all the chains of all the 100 replicates. It is possible to observe a plateau at approximately $50 * 50000$ steps. The red dashed line represents the average value.

Figure 2. RMSD profiles

1. **RMSD**: The Root Mean Square Deviation is calculated by using the *MDanalysis* package[?] and is calculated as follows:

$$RMSD = \rho(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i \left(\vec{x}_i(t) - \vec{x}_i^{REF}\right)^2} \tag{7}$$

   Before performing this type of calculation, the structures were aligned to the first frame (each frame of each replicate was aligned to the first frame of their specific replicate). This type of alignment was done making use of the *AlignTraj* function[?]. In general, the smaller is the difference between two structures, the lower is the value of RMSD. The results are written in graph

2. R_g: The Radius of Gyration computed by *MDanalysis* was calculated as written in equation 8. This quantity is a measure of how the mass of an object is spread out relative to a particular axis of rotation. In general, it tells "how spherical" is an object[?,?].

$$R_g = \sqrt{\frac{\sum_i m_i \vec{r}_i^2}{\sum_i m_i}} \tag{8}$$

3. **Autocorrelation function:** The autocorrelation function can be written as shown in equation 9[14]. Results are shown in

$$r_k = \frac{C_k}{C_0} = \frac{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})(A_{t+k} - \bar{A})}{\frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})^2} \tag{9}$$

   Where $C_k = \frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})(A_{t+k} - \bar{A})$ is the autocovariance function at lag k and $C_0 = \frac{1}{M}\sum_{t=1}^{M-k}(A_t - \bar{A})^2$ is the variance function.

4.5 Algorithms used for comparison

As stated in the methods section 4.3 about the simulations, the attraction potentials are sequentially added to the model in order to improve the predictions. The SCC metric was used to compute the difference between the control contact matrix and the CG derived one (chapter 4.5.1).

Two methods were then considered to add the variables, which are expressed in algorithm 2 and 3. Since the states are differentially populated, as stated in chapter 5.1, it is possible to argue that the largest contribution to the result would be given, in a
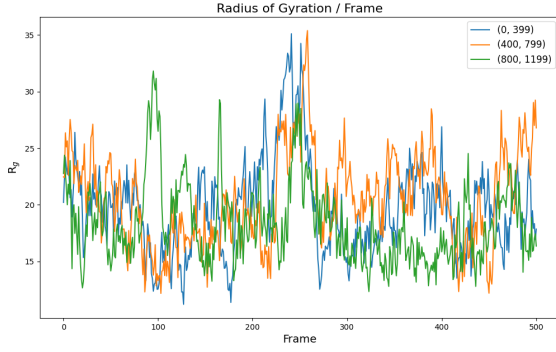
(a) Graph representing the R_g of the chains pertaining to the first replicate. As it is possible to see from the legend, the first, the second and the third chain are represented respectively in blue orange and green.
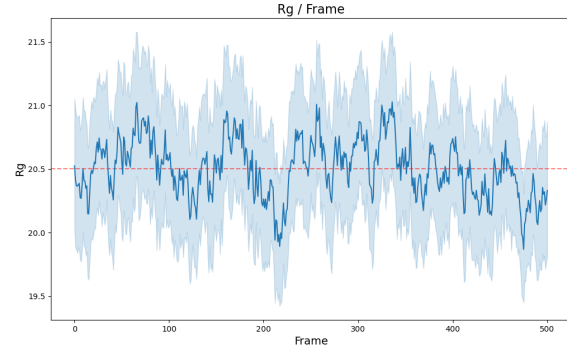
(b) Figure representing the collective behaviour of all the chains of all the 100 replicates. The red dashed line represents the average value.

Figure 3. R_g profiles

hierarchical manner, by the most populated states. As a consequence of this consideration, I could fix the values related to the most prevalent states, before considering those that are less present. This type of mechanism is described in the greedy process of algorithm 2. If that assumption is not accepted, then either you test all the possible configurations, either you find a better way to test the generated models. The algorithm 3 is a stepwise solution which allows to solve partially the problem at the cost of more simulations to perform.

Algorithm 2: Greedy matrix comparison

Result: Best performing greedy model

forall *attraction parameter* **do**

    Construct models by adding the most present attraction parameter to the $(n-1)^{\text{th}}$ step configuration ;

    Compute SCC of the model with respect to the reference ranging among a list of possible values;

    Select the value of the parameter which gives the best results;

    Add that attraction with that coefficient;

return *Best greedy model*;

Another possible way to compare the matrices would be to compute the Spearman correlation coefficient. Ideally, it would be interesting to see if there are cases where the two metrics produce different results, and to understand which of them is better in what cases

4.5.1 the SCC metric

The SCC metric is described in the paper written by Yang and colleagues in 2017[15, 16] . It is a Stratum Adjusted Correlation Coefficient and can quantify the similarity between an Hi-C matrix and another. In general, the most common techniques to use in these situations is either to analyze the matrices by eye, or, in a certainly more precise way, to calculate a Pearson/Spearman correlation coefficient. However Hi-C data have certain unique characteristics, including domain structures (such as topological association domain (TAD) and A/B compartments) and distance dependence. Indeed, the chromatin interaction frequencies between two genomic loci, on average, decrease substantially as their genomic distance increases. Standard correlation approaches do not take into consideration these structures and may lead to incorrect conclusions[15, 16] .

The SCC metric could be seen as a weighted Pearson coefficient, as written in equation 10.

Variables

Algorithm 3: Step-wise process for matrix comparison

Result: Best model
$n = 0$;
Continue = False;
while *Continue == True* **do**
    Continue = False;
    Construct models by adding the most present attraction parameter to the $(n-1)^{\text{th}}$ step configuration ;
    Compute SCC of the model with respect to the reference ranging among a list of possible values;
    Select the value of the parameter which gives the best results;
    if *addition gives better results* **then**
        Add that attraction with that coefficient;
        Continue = True
    forall *state in model* **do**
        Remove that state from the model;
        Vary the value associated to the state with the highest frequency among those remaining;
        Compute SCC of each model with respec to the reference ranging among a list of possible values
    Select the reduced model which gives the best results;
    if *removal gives better results* **then**
        Perform the reduction;
        Continue = True
    $n = n + 1$
return *Best model*;
`/* n is the step number                                                              */`

N_k	$k \in K$	Number of observations in stratum k;
X_k	$k \in K$	Observations in stratum k in matrix X;
Y_k	$k \in K$	Observations in stratum k in matrix Y;
$r_{1k} = \frac{\sum_{i=1}^{N_k} x_{ik} y_{ik}}{N_k} - \frac{\sum_{i=1}^{N_k} x_{ik} \sum_{j=1}^{N_k} y_{jk}}{N_k^2} = E(X_k Y_k) - E(X_k) E(Y_k)$	$k \in K$	Correlation between X_k and Y_k;
$r_{2k} = \sqrt{\text{var}(X_k) \cdot \text{var}(Y_k)}$	$k \in K$	Square root of the product between the variances of X_k and Y_k;
$\rho_k = r_{1k}/r_{2k}$	$k \in K$	Pearson coefficient related to bin k;

Formula

$$\rho_s = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k \tag{10}$$

5 RESULTS AND DISCUSSION

5.1 *ChromHMM* results

In total, 4 states were considered to be present. Two functions in particular were used: BinarizeBam and LearnModel. The data shown in 4.1 were aligned to *hg38* reference genome[17] . Results are shown in image 4; by taking a look to its subfigures, the following considerations could be done:

1. Clear absence or presence of ATAC and CTCF signals could be detected in figure 4a. for this reason, the following states are defined:

   - **State 1**: State without the presence of ATAC and CTCF signal
   - **State 2**: State with ATAC but not CTCF peaks
   - **State 3**: State with the presence of both ATAC and CTCF signal
   - **State 4**: State with CTCF but not ATAC peaks

2. The states 1 and 2 in particular tend to perform transitions towards themselves instead of different states (figure 4b)

3. State 2 (with ATAC) and 3 (with ATAC and CTCF) tend to localize in CpG islands, exons and Transcriptional Starting Sites (figures 4c, 4d, 4e)
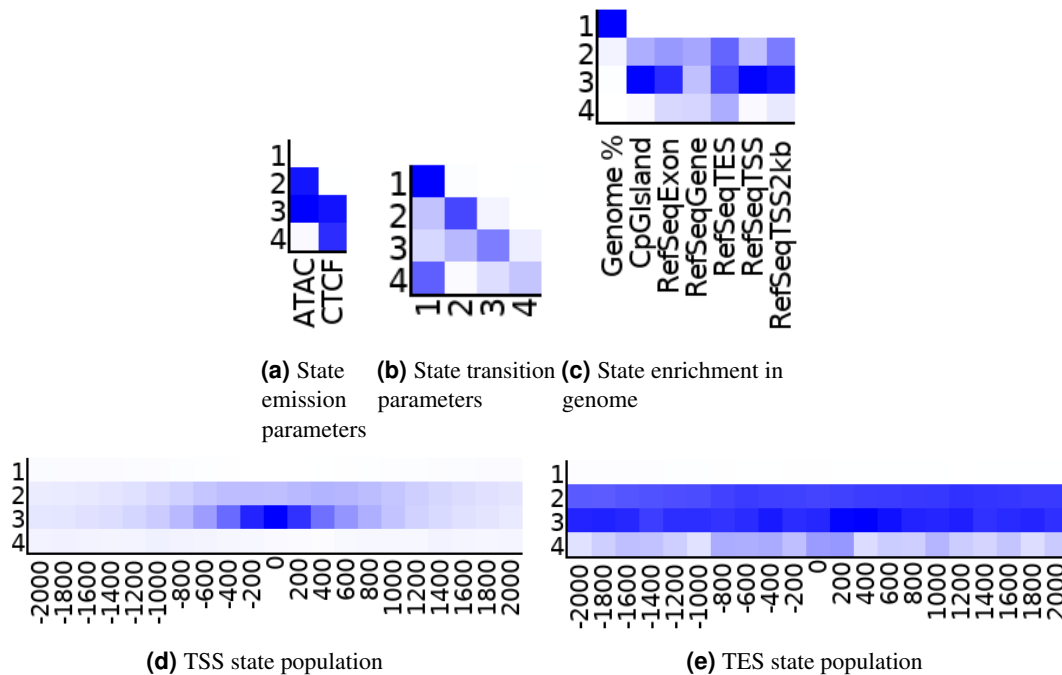

(a) State emission parameters

(b) State transition parameters

(c) State enrichment in genome

(d) TSS state population

(e) TES state population

Figure 4. Results from ChromHMM for the IMR90 replicates

The proportions represented in image 5b were obtained. In general, the proportions calculated taking into consideration the segments are much lower with respect to those calculated with the bins. The reason is that the first state is in general much more present than the other three. The ratios are larger with 5 kbp bins as in fact whenever there are a few occurrences of the rarely present states, those are assigned with a great probability and convert a great number of segments which are not anymore assigned to state 1.
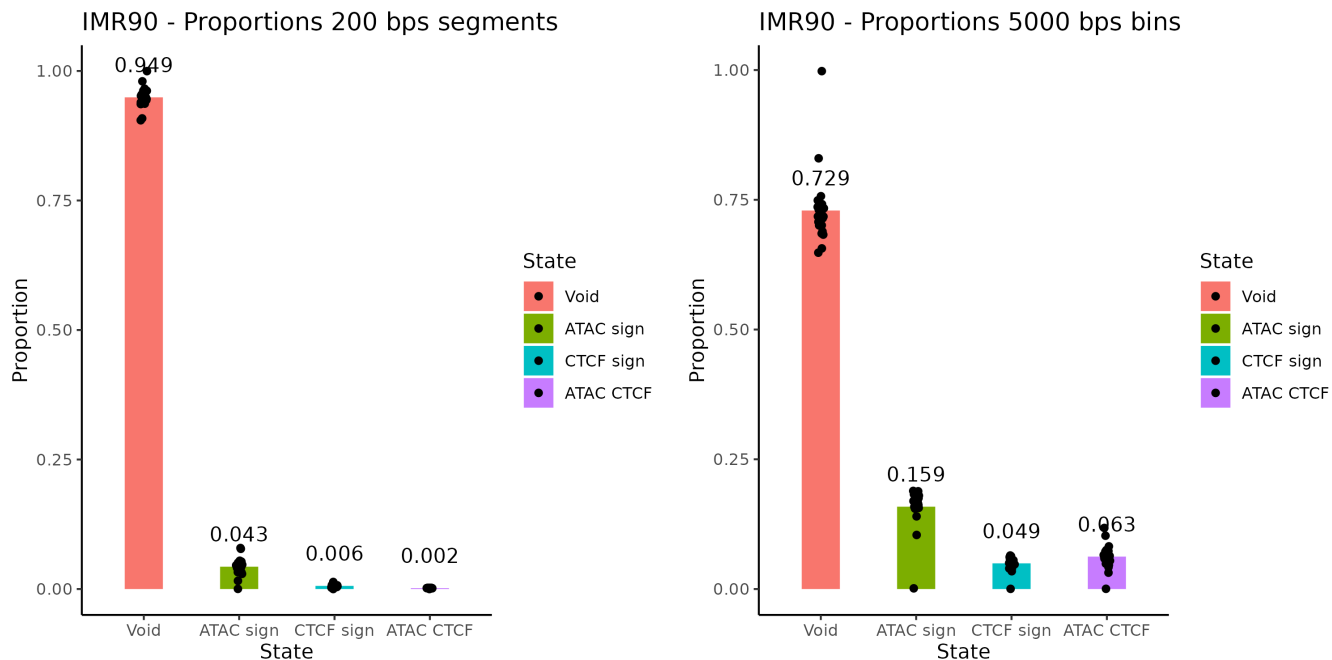
(a) Proportions of states in 200 bps segments. The segments were directly found by *ChromHMM*. Each dot represents the proportion relative to a chromosome.

(b) Proportion of states in 5000 bps long bins. Each dot represents the proportion relative to a chromosome.

Figure 5

5.1.1 Some examples

The following images (image and image) were created by using IGV, a visualization tool[?, ?] .

Figure 6. IGV snapshot of a portion of the ANPEP region analyzed

5.2 Results from model selection

6 CONCLUSIONS

7 Glossary

Bead	The complex formed by the DNA and the histone proteins
TSS	Transcriptional Starting Site
TES	Transcriptional Ending site
FS	Fine Scale
CG	Coarse Graining
k. l.	Kuhn length
mon.	Monomer
PL	persistence length
LJ	Lennard-Jones
FENE potential	Finite Extensible Nonlinear Elastic potential
RMSD	Root Mean Square Deviation
Rg	Radius of Gyration

References

1. Paro, P. D. R., Grossniklaus, P. D. U., Santoro, D. R. & Wutz, P. D. A. Biology of Chromatin. In *Introduction to Epigenetics [Internet]*, DOI: 10.1007/978-3-030-68670-3_1 (Springer, 2021).

2. Shareef, S. J. *et al.* Extended-representation bisulfite sequencing of gene regulatory elements in multiplexed samples and single cells. *Nat Biotechnol* **39**, 1086–1094, DOI: 10.1038/s41587-021-00910-x (2021).

3. Kim, S., Yu, N.-K. & Kaang, B.-K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**, e166–e166, DOI: 10.1038/emm.2015.33 (2015).

4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218, DOI: 10.1038/nmeth.2688 (2013).

5. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518–1552, DOI: 10.1038/s41596-022-00692-9 (2022).

6. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137, DOI: 10.1186/gb-2008-9-9-r137 (2008).

7. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492, DOI: 10.1038/nprot.2017.124 (2017).

8. Chilled House Vibes. Learning Chromatin States from ChIP-seq Data: ChromHMM - Luca Pinello (2015).

9. Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer & Richard Berger. LAMMPS. https://www.lammps.org/#gsc.tab=0.

10. Tan, J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol* **41**, 1140–1150, DOI: 10.1038/s41587-022-01612-8 (2023).

11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: 10.1038/nature11247 (2012).

12. UCSC Genome Browser Home. https://genome.ucsc.edu/.

13. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171, DOI: 10.1016/j.cpc.2021.108171 (2022).

14. Suma, A., Di Stefano, M. & Micheletti, C. Electric-Field-Driven Trapping of Polyelectrolytes in Needle-like Backfolded States. *Macromolecules* **51**, 4462–4470, DOI: 10.1021/acs.macromol.8b00019 (2018).

15. Lin, D., Sanders, J. & Noble, W. S. HiCRep.py: Fast comparison of Hi-C contact matrices in Python. *Bioinformatics* **37**, 2996–2997, DOI: 10.1093/bioinformatics/btab097 (2021).

16. Yang, T. *et al.* HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* **27**, 1939–1949, DOI: 10.1101/gr.220640.117 (2017).

17. Homo sapiens genome assembly GRCh38 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.