Alma Mater Studiorum · Università di Bologna

Dipartimento di Fisica e Astronomia "Augusto Righi"
Corso di Laurea in Fisica

Radial organization of GC content: a computational analysis of nuclear 3D genomic structures in different human cell types

Relatrice:
Dott.ssa Alessandra Merlotti

Presentata da:
Samuele Bassi

Anno Accademico 2024/2025

Abstract

Il metodo GPSeq introdotto nel 2020 ha permesso di misurare sperimentalmente come, in alcune linee cellulari umane, il contenuto di G e C del genoma aumentasse avvicinandosi al centro del nucleo. Tale gradiente radiale potrebbe essere alla base di funzioni regolatorie e protettive della trascrizione genica e influenzare la disposizione 3D del DNA. In questa tesi viene proposta un'analisi computazionale volta a studiare la ricostruzione spaziale del contenuto genomico e a stabilire la presenza di tale organizzazione radiale per 9 diversi tipi cellulari umani. La prima di queste è la linea linfoblastoide GM06990, in cui il gradiente è stato osservato mediante GPSeq. La ricostruzione della struttura 3D del DNA è stata realizzata attraverso l'utilizzo del software MOGEN, che, tramite un processo di ottimizzazione, restituisce le coordinate spaziali dei vari frammenti che compongono la sequenza. In particolare, la ricostruzione avviene sulla base dei dati di contatto Hi-C per ciascun tipo cellulare, i quali offrono una misura della vicinanza spaziale relativa dei diversi frammenti del genoma. Andando a calcolare la frazione media di contenuto G e C all'interno di gusci sferici concentrici, è stato possibile eseguire un fit lineare sui dati ottenuti per verificare la presenza del gradiente genomico. Le ricostruzioni 3D prodotte hanno evidenziato come le tipiche caratteristiche topologiche dell'organizzazione genomica siano confermate, sebbene con eccezioni dovute all'estrema dinamicità della struttura e agli effetti causati dal rumore dei dati utilizzati. Nella linea linfoblastoide è stata verificata la presenza di un lieve gradiente di contenuto GC, compatibile solo con altri due tipi cellulari. Nei rimanenti, il gradiente è risultato essere molto lieve o assente.

Contents

1. Introduction

In the last few decades, the interest in characterizing the 3D intranuclear structure of the human genome has grown exponentially due to its profound implications for gene expression and gene regulation. A great number of experimental protocols aimed at assessing the *in situ* position of specific genomic *loci* were developed with the intent of advancing knowledge in this direction, which was only vaguely explored. However, no experiment was capable of extracting genome-wide information from cellular lines until the development of the Hi-C experiment [12], which, by dividing the linear genome sequence into different bins of fixed resolution, provided maps indicating the probability of spatial closeness between any bins across the whole sequence in a specific cellular type. Hi-C data were a crucial step forward in verifying the 3D spatial features of the genome. The results of these experiments were fundamental in mapping the long-range interactions between sequence-wise distant bins, aiding in the identification of various clusters of genome sequences, such as the topologically associated domains, as well as in characterizing the bicompartmental behavior of chromosomes divided into euchromatin and heterochromatin. Having achieved a clearer idea of the overall characteristics of the intranuclear genomic conformation, further advancement required the analysis of the radial arrangement of genomic *loci*, motivated by the various properties that differentiate the topology of the center of the nucleus from the periphery. An experimental protocol to assess radial distances from the nuclear lamina was developed in 2020 [7]. This experiment, called GPSeq, evidenced in two different cellular types that the content of the nucleotides G and C was proportional to the distance of the DNA sequences from the peripheral region. Out of all the genomic and epigenomic differences between the internal and external regions of the nucleus, the anticorrelation between the GC content and the distance from the center is outstanding because of the functional properties it could underlie: for example, the *bodyguard* hypothesis [9] infers a more internal position for euchromatin that is generally GC rich; moreover, the intranuclear location of G and C nucleotides could be important for the regulation of transcription factors and splicing processes [3]; lastly, the biophysical properties of the GC content could play a non-negligible role in the organization of the three dimensional structure of the genome [3]. In order to further analyze the radial properties of GC content, it is now necessary to perform a large number of experiments on different cellular types to assess whether this gradient is always observed and, eventually, in which cellular types. However, conducting

experiments comes with a cost in both money and time. As a matter of fact, the latest experimental protocol published for GPSeq [29] states that approximately two weeks could be the time necessary to go from the preparation of the samples to the production of the libraries describing the radial positioning of the DNA fragments. The possibility of having a priori knowledge, even if rough, about the characteristics of the GC radial trend of the cellular types to be analyzed could be vital for choosing the cellular types to be studied and saving time and materials for their study. These reasons motivate the development of the analyses proposed here to assess the presence of a GC gradient in different cellular types. In particular, the method is used to analyze the GM06990 cellular line (lymphoblastoid) also studied in GPSeq [7], where the radial GC gradient was observed. Afterward, it is applied to other cellular types that are subsequently compared to the previous one. In order to achieve the result, it is necessary to first obtain a virtual reconstruction of the 3D structure of the genome for every cellular type. There are different computational tools that provide this type of reconstruction, but the best results in the visualization of the gradient of the benchmark cellular line were observed in the structures generated by MOGEN [26], a software that, starting from the Hi-C data of a given cellular type, is capable of providing a reliable and genome-wide representation of the spatial positions of the bins.
In this chapter, we present the motivations underlying the study of the radial organization of GC content, as well as the experiments and tools concerning the general field of genomic research and this work.

1.1 DNA function and structure

In order to keep a biological organism alive, it is necessary for cells to follow precise functional instructions. These instructions are conserved within each cell, and if they are of eukaryotic type, they are located precisely within the nucleus, a region of the cellular space delimited by a membrane. In particular, they are encoded into deoxyribonucleic acid (DNA). DNA is the most fundamental organization of biological information in living organisms: it is a polymeric molecule contained inside the nucleus of eukaryotic cells, and it is formed by two strands of monomers called nucleotides. These are composed of a sugar, deoxyribose, linked to a phosphate group and one of four possible nitrogen-containing bases, known as nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The double strand is formed by the bond of A with T or G with C in a complementary way, so the type of nucleotide on one strand always implies the nucleotide bonded to it on the other side. Instead, the nucleotides form the strand by a link between the phosphate group and deoxyribose. The two strands follow a double helix path, and

their ends are characterized by different terminations: the extremities of every finite chain are always composed of a phosphate group at the beginning and a deoxyribose at the end of one strand, and a deoxyribose at the beginning and a phosphate group at the end of the other strand; the phosphate group end is called 5', and the deoxyribose end is called 3'.

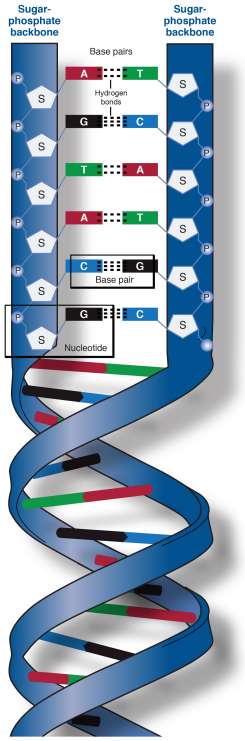

Figure 1.1: A schematic representation of the structure of the two strands of DNA and the double helix. In this case, the end on the upper left is 5', while the upper right end is 3'. Image from [18].

The instructions stored within the DNA mostly concern the production of proteins that are responsible for the execution of the cell's functions. In particular, the four different types of nucleotides are comparable to the letters of an alphabet that encodes the genetic information for the construction of various types of proteins through the necessary amino acids, along with the instructions for the regulation of their creation. However, there is an intermediate passage between the protein-coding sequences of nucleotides and the production of proteins: an enzyme called RNA polymerase creates a copy of the DNA sequence with RNA (which differs by a single nucleotide from DNA, i.e., uracil (U) instead of thymine (T)). This process is referred to as transcription. Usually refined by splicing to achieve the final form, the RNA copy, called messenger RNA (mRNA), is then used

by ribosomes to create proteins. Moreover, not all the sequences of nucleotides are directly responsible for this RNA production. As a matter of fact, there are sequences that are important for the regulation of such production based on the typology of the cell, specifying when, in which types of cells, and in what quantity to build a certain RNA sequence corresponding to a specific protein. Among these regulatory sequences, the promoters are sequences from which RNA polymerase starts the transcription of a specific sequence, and the enhancers are sequences that, when linked with specific proteins called transcription factors (TFs), can increase the production of specific transcripts. In particular, an enhancer is composed of different binding sites where TFs can bind and execute the regulatory operations encoded in the sequence. This means that these particular sequences are not actively transcribed into mRNA, but they are still an important regulatory part of the process of transcription. The sequences univocally associated with a protein-coding process are called genes. The parts of the sequence that are instead protagonists of regulatory processes are excluded from this definition but constitute an important instrument for the expression of the genes. Both are physically formed by a sequence of different nucleotides. The complete collection of information in the DNA (the whole sequence) is called the genome.

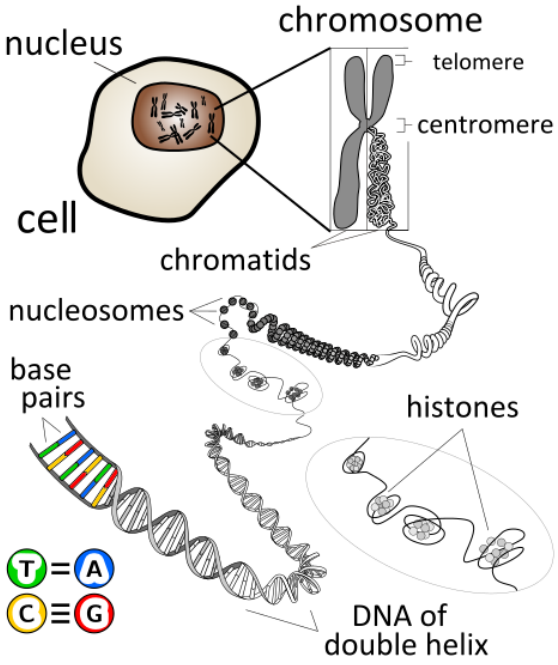

Figure 1.2: A brief figurative summary about the various levels of organization of the DNA. Image from [11]

As one can imagine, the length of the genome is quite impressive (2 meters), considering that it has to be packed in a 6 μm diameter nucleus. This packaging is achieved through a compact folding of the DNA filament. In eukaryotic cells, the double helix of DNA is wrapped around proteins called histones, which are connected to each other through the DNA linker, forming a complex similar to beads on a string. The structure composed of the double strand folding around eight histones is called a nucleosome. The chain formed by the various nucleosomes is referred to as chromatin, which is organized into architectures called chromosomes (see paragraph 1.4 for a slightly more detailed description of how chromatin folds to create chromosomes); in human cells, the genome is organized into 22 pairs of homologous chromosomes and a pair of sex chromosomes (X or Y). Obviously, to complete the process of transcription, it is necessary for the sequence to unravel and be accessible to RNA-polymerase, losing compactness locally and thus changing its 3D structure.

1.2 Hi-C experiment and data

Beyond the scale of the chromosomes, the knowledge of the three dimensional organization of the genome inside the nucleus during the interphase is still a recent and debated topic. Indeed, the most popular representation of the human genome is surely the list of chromosomes printed on paper and separated from one another. However, this is far from the reality of the actual disposition of chromosomes, which are organized in a more complex topology. In particular, the locations of the chromosomes and the structure of the chromosomes themselves are necessarily dynamic: the chromatin needs to be unfolded in order to be transcribed or repaired, and in some cases, it is necessary to bring two linearly distant sequences of the genome spatially close to each other to activate specific productions. For example, the protein CTCF is responsible for establishing contact interactions between promoters and enhancers, which may be distant in the linear sequence, causing an increase in the production of proteins encoded by a specific gene. The reconstruction of the 3D structure of the genome is thus a difficult but interesting task: having a more precise idea of the locations of various chromosomes and specific sequences, particularly those that, even if linearly distant in the sequence of the reference genome, are nevertheless spatially close inside the nucleus, can be of great relevance for gene expression and regulation [23].
A fundamental progress in this direction is represented by the Hi-C experimental method [12], which can be applied to cellular lines in order to gather their Hi-C data, that are used in this work to create the various 3D structures of the genome (see paragraph 2.2). The main intent of the experiment was precisely to obtain knowledge about short-

and long-range interactions between genomic loci and, generally, about the intranuclear genome architecture. This is accomplished through chemical techniques, which, followed by massively parallel sequencing (MPS), provide a numerical estimate of the spatial proximity between two sections of the DNA.
Briefly summarizing the method, the information about the closeness of different parts is conserved by creating a covalent chemical bond between them using formaldehyde (a process called cross-linking). Afterwards, the genome is divided into sections through its digestion by a restriction enzyme (*HINDIII* or *Ncol* in [12]), which leaves their ends with a 5' overhang on one of the two strands composing the filament; these are then filled with the correct nucleotides marked with a biotinylated residue. The ends of the cross-linked sections are then ligated under particular dilute conditions that favor the process; the junction between them is highlighted due to the presence of the biotin marker. By purifying the fragments from the formaldehyde bond and shearing them again into smaller chunks to prepare them for MPS, it is possible to select the ones containing biotin using streptavidin beads. The selected chunks are thus formed by two different parts divided by the biotin marker: these come from two different sections of the sequence that were in spatial proximity when the chemical bonds were created. By performing massive parallel sequencing on these two parts, it is possible to backtrack to the sections to which they belong in the reference genome, which is the genome-wide sequence of nucleotides mapped for one strand in the direction 5' to 3' (the other one is consequently implied). Therefore, it is possible to state that two particular chunks of the genome, possibly distant in the sequence, were in contact.

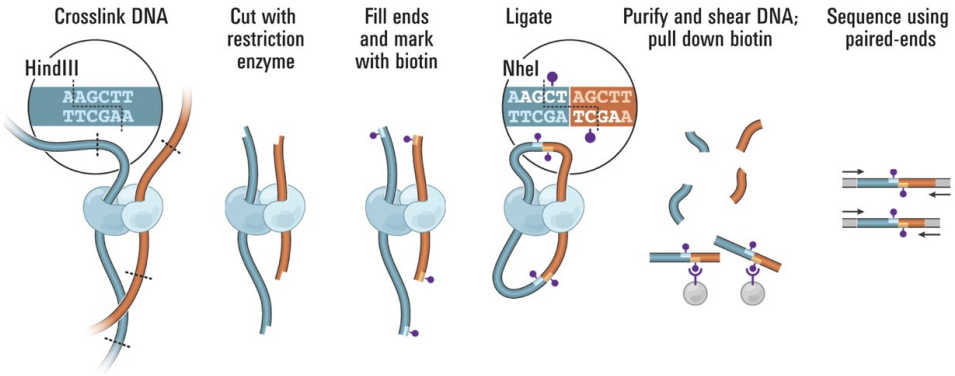

Figure 1.3: An image from [12] representing the process of the Hi-C experiment.

The MPS used in [12] was implemented by an Illumina Genome Analyzer; the core characteristic of this analyzer is that chunks of chromatin are multiplied to form millions of copies having the same sequence. The DNA to be analyzed is made single-stranded, and the first base of the sequence is attached to a complementary base that has fluorescence

depending on the type of nucleotide. At this point, lasers are passed over the sequences created, making the particular fluorescence associated with the base emit light of a color that changes with the type of nucleotide; the signal is obviously amplified due to millions of copies of the sequence previously created. The light is registered by a camera that recognizes the color and therefore the first base of the sequence. The process is then repeated until all the sequence has been registered. This whole Hi-C process is executed on cellular lines containing millions of single cells, leading to millions of possible reads. The Hi-C data are then structured by dividing the genome into conventional bins, choosing a resolution (for example, in [12] and in this work, the resolution of the bins is one megabase (Mb), i.e., 10^6 nucleotides). It is possible to assign to each pair of bins the number of Hi-C contacts collected among the ensemble of cells from the previously described experiment between sections contained in these specific bins. These data (bin i, bin j, and the number of contacts m_{ij} between i and j, called the interaction frequency) compose the Hi-C dataset for the cellular line analyzed. A higher value of contact m_{ij} implies that bins i and j have been measured in spatial proximity with greater frequency across the ensemble of cells; therefore, they can be considered closer to each other in a potential representation of the 3D structure. A visualization of the Hi-C data can be provided by the contact matrix, where different values of m_{ij} are organized in a matrix and highlighted with a heatmap.

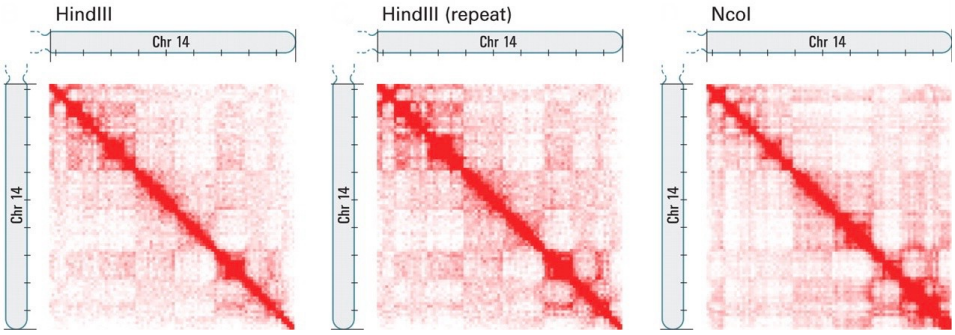

Figure 1.4: An image from [12] representing the contact matrix for the bins of chromosome 14 resulting from three different Hi-C experiment: two executed with the enzyme *HindIII* and one with the enzyme *NcoI*.

It is evident that the experimental process can also lead to incoherent and noisy lectures caused by systematic effects [28]. In particular, the process of massively parallel sequencing is not free from errors, and there is a chance that the analyzed sequences match different sections of the reference genome. Moreover, the three dimensional structure of the genome is intrinsically dynamic, which implies that areas in contact within

one cell could be far apart in another cell of the same type, increasing the difficulty in interpreting the dataset. For these reasons, the number of contacts between bins is usually treated through normalization in order to extract valuable information more easily, and data can be preprocessed to prevent noisy information from being analyzed.

1.3 MOGEN

Hi-C data provide insight into the relative spatial distance between different parts of the genome. Indeed, there is an anticorrelation between the contacts m_{ij} and the spatial distance of bins. This information can be used to deepen the knowledge of the three dimensional structure of the genome inside the nucleus for different cell types. As a matter of fact, it is possible to imagine that a suitably designed algorithm could be used to analyze the Hi-C data and, based on it, produce a virtual reconstruction of the structure of the genome by assigning three dimensional coordinates to the bins. Many tools have been developed to accomplish this intent, but in many cases, the structure produced is reliable for single chromosomal structures and not for an accurate representation of the whole human genome. This is caused by difficulties in managing the noisy interchromosomal contacts of the Hi-C data. MOGEN [26] is a computational tool developed precisely with the aim of overcoming these obstacles and reproducing a reliable representation of the human genome, taking as input the genome-wide Hi-C data. For the sake of this work, a trustworthy reconstruction of the three dimensional structure of the genome would be fundamental in order to proceed with the radial organization of GC content.

The solution implemented by MOGEN follows a restraint-based method: the contacts between the bins are converted into restraints that need to be respected in the generated structure. In particular, if the contacts between the two bins are low (under a specific threshold), they are considered non-contacts, and the spatial distance between the two bins in the structure needs to be at least equal to or greater than a specific value. On the other hand, if the contacts are higher than the same threshold, the restraint on the bins will require them to be spatially closer than that specific distance. The coordinates of the bins are then found through an optimization process that tries to satisfy as many restraints as possible. It is possible to set the maximum distance between in contact bins or, analogously, the minimum distance between non contact ones (d_c), as well as the threshold. The restraints have a major priority of being satisfied by the structure if they are implied by a non-contact or when F_{ij}, the normalized interaction frequency, is really high in comparison to the threshold because, in that case, the probability of the two

bins being in contact is effectively high. A fundamental parameter of the reconstruction is, therefore, the threshold: values of Hi-C data that exceed it are considered contacts by the algorithm; conversely, values below it are deemed non contacts. As mentioned before, one of the challenges of reconstruction is dealing with noisy data. The Hi-C dataset is the result of an ensemble of millions of cells; therefore, considering the dynamic structure of the genome and possible systematic errors, almost all the megabases have a non-zero value of contact with each other. Low values of the interaction frequency usually indicate sporadic contacts that should be considered noise rather than significant closeness in space between the *loci*. The threshold allows us to consider these low values as non-contacts, partly solving the problem. Ideally, the threshold could be considered unique for all types of contacts because different values would introduce a bias in the analysis; however, interchromosomal contacts are rarer than intrachromosomal ones [26], which are easily found in the data. Therefore, intrachromosomal contacts should generally have a higher threshold because the frequency of interactions between any fragments in the chromosome is normally much higher than that of interchromosomal bins. The thresholds for interchromosomal and intrachromosomal contacts are therefore usually different. Alternatively to setting the thresholds, it is possible to set the percentage of contacts that the algorithm needs to consider, and the thresholds will consequently be computed by MOGEN. For the same reasons mentioned above, it is necessary for the intrachromosomal percentage to be higher than the interchromosomal one.
Quantitatively, the coordinates of the bins are approximated by optimizing the following scoring function:

$$
\begin{aligned}
F_n = & \sum_{\substack{\text{contacts} \\ (i,j):\,|i-j|\neq 1}} \left(W_1[\text{chr1},\text{chr2}]\ \tanh(d_c^2 - d_{ij}^2)\,\frac{F_{ij}}{\text{totalIF}} + W_2[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{ij}^2 - d_{\min}^2)}{\text{totalIF}} \right) \\
& + \sum_{\substack{(i,j):\,|i-j|=1 \\ \text{chr1}=\text{chr2}}} \left(W_1[\text{chr1},\text{chr2}]\,\text{IF}_{\max}\,\frac{\tanh(d_{a,\max}^2 - d_{ij}^2)}{\text{totalIF}} + W_2[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{ij}^2 - d_{\min}^2)}{\text{totalIF}} \right) \\
& + \sum_{\substack{\text{non-contacts} \\ (i,j):\,|i-j|\neq 1 \\ \text{chr1}=\text{chr2}}} \left(W_3[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{\max,\text{intra}}^2 - d_{ij}^2)}{\text{totalIF}} + W_4[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{ij}^2 - d_c^2)}{\text{totalIF}} \right) \\
& + \sum_{\substack{\text{non-contacts} \\ (i,j):\,|i-j|\neq 1 \\ \text{chr1}\neq\text{chr2}}} \left(W_3[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{\max,\text{inter}}^2 - d_{ij}^2)}{\text{totalIF}} + W_4[\text{chr1},\text{chr2}]\,\frac{\tanh(d_{ij}^2 - d_c^2)}{\text{totalIF}} \right).
\end{aligned}
$$

where:

- totalIF is the total sum of the F_{ij}
- IF_{max} is the maximum value of the F_{ij}
- d_{ij} is the distance between bin i and bin j;
- d_{min} is the minimum distance between bins in contact;
- $d_{a,max}$, $d_{max,intra}$, and $d_{max,inter}$ are the maximum distances for, respectively, adjacent bins, intrachromosomal bins, and interchromosomal bins;
- $W_i[\text{chr1}, \text{chr2}]$ are weights with a specific value for every pair of chromosomes.

All four of these last kinds of parameters, together with d_c that was already mentioned, can be set by the user (see paragraph 2.2).
The scoring function is designed to accomplish all the previously described requests after normalizing and preparing the Hi-C data, setting the thresholds for dividing contacts and non contacts, and setting the maximum and minimum distances. The scoring function is initially computed with a random initialization of the coordinates of the bins. Starting from these values, the function is optimized by gradient-ascent, iteratively modifying the coordinates until the scoring function stops increasing, meaning the maximum has been reached. The expression above is composed of four sums. For each one of them, the scoring function increases if the terms of the argument increase. It is also noteworthy that tanh is an increasing function that assumes values in [-1, 1]. In all cases, the scoring function is increased if the distance between bins respects the general vincula d_{min}, $d_{max,intra}$ and $d_{max,inter}$. The first sum concerns non adjacent bins in contact. The scoring function increases if the distance between these bins is less than d_c, as it was our intent to satisfy the contact restraint. Moreover, it increases if the interaction frequency between the bins is higher: in this way, the optimization prioritizes the satisfaction of the constraints with higher F_{ij}. The same logic is applied to the second sum, which considers the adjacent bins that, by definition, are always in contact. The only difference is the use of $d_{a,max}$ instead of d_c and IF_{max} instead of F_{ij}. The third and fourth sums consider the non-contact bins, respectively for the same and different chromosomes. Contrary to before, the scoring function increases if the distance between the bins is greater than d_c. Finally, there are four sets of weights (W_1, W_2, W_3, W_4) that can be used to modify the value of the scoring function and increase or decrease the prioritization of the associated constraints for each pair of chromosomes. Each weight has a value valid for every pair

of different chromosomes (the value is the same for every pair of chromosomes in order to avoid any bias towards a specific couple) and a value that can be different for every individual chromosome. Introducing the concept of contact and non contact scores, which are the percentages of contact and non contact constraints respected by the structure. The weights modify the characteristics and percentages of contacts and non contacts respected for each pair of chromosomes and within the same chromosome:

- $W_1[\text{chr1}, \text{chr2}]$ are the weights corresponding to the constraint of the maximum distance between bins that are in contact; increasing them gives priority in the scoring function to the satisfaction of these constraints; therefore, they increase the score of contacts in the final structure between 1 and 2;

- $W_2[\text{chr1}, \text{chr2}]$ are the weights corresponding to the constraint of the minimum distance between bins that are in contact. Increasing these weights prevents these bins from being too close in space by suitably prioritizing the satisfaction of the minimum distance constraints; they are useful for the visualization of the structure, but they do not particularly affect the values of contact and non contact scores;

- $W_3[\text{chr1}, \text{chr2}]$ are the weights corresponding to the constraint of the maximum distance between bins that are not in contact. Their increment prevents the structure from being too spread out;

- $W_4[\text{chr1}, \text{chr2}]$ are the weights corresponding to the constraint of the minimum distance between bins that are not in contact. Increasing them prioritizes the satisfaction of non contacts; therefore, the percentages of non contacts are increased. They are the complementary weights with respect to W_1;

1.4 General 3D features and organization of the human genome

The Hi-C technique is just one of the several approaches for the characterization of the 3D structure of the genome. For example, fluorescence in situ hybridization (FISH) is a technique used to identify genomic *loci* in the interphase nucleus. In particular, after fixing one strand of the genome to a slide, it is possible to link the complementary nucleotide sequence to the one of interest by marking it with a fluorescent material in order to localize it *in situ*. Various findings about the 3D disposition of chromosomes have been achieved through the analysis of data and results from these techniques.
Hi-C libraries offer the possibility to compute the probability of intrachromosomal or

interchromosomal contact as a function of the linear distance of the bins of the chromosome; the results in figure 1.5 for the GM06990 cell line (lymphoblastoid) show that, even for linear distances beyond 200 Mb, the intrachromosomal probability of contact is still substantially greater than the average interchromosomal one, confirming the presence of distinct and non-intermingling chromosomal territories within the nucleus.

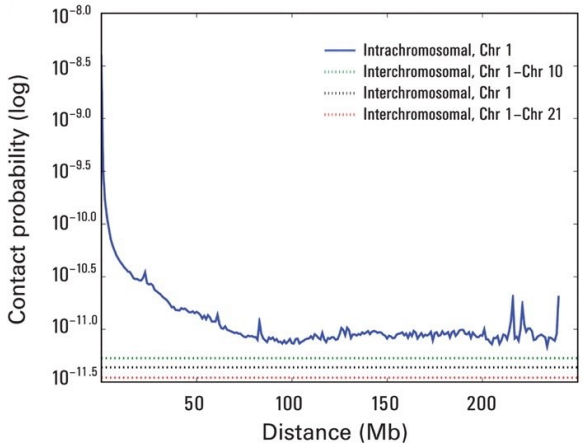

Figure 1.5: A graph from [12] plotting the logarithm of the probability of intrachromosomal contact for chromosome 1 and the average probability of interchromosomal contact between bins of chromosome 1 and chromosome 10, the genome-wide average probability of interchromosomal contact for bins of chromosome 1 and the average probability of interchromosomal contact between bins of chromosome chromosome 1 and chromosome 21 in function of the linear distance between bins of chromosome 1 for a lymphoblastoid cell line. It is evident how the intrachromosomal contact probability decreases with linear distance, eventually reaching a plateau before the 100 Mb of linear distance, but remains particularly high in comparison to the average probabilities of interchromosomal contact, that are constant among all of the bins of chromosome 1.

Through FISH experiments, it is possible to investigate the position of sequences belonging to different chromosomes, showing the differences in the localization of chromosomal territories. As a matter of fact, large chromosomes (from 1 to 15 and X) tend to be placed in the periphery of the nucleus, while small chromosomes are usually located in the center. This is also confirmed by the values of Hi-C interchromosomal contacts, which are usually higher between small chromosomes, consistent with the fact that they are colocalized in the center of the nucleus. In addition, from figure 1.5, it is possible to observe that the average probability of interchromosomal contact between chromosome 1 and chromosome 10 is slightly higher than between chromosome 1 and chromosome 21, confirming that large chromosomes generally do not cluster in the center. An eye-catching exception is that chromosome 18, even though small in dimensions, is usually found near the periphery of the nucleus [6].

In addition to these considerations, Hi-C data open the possibility for an in-depth exploration of the interactions of genomic *loci* and their spatial organization on a subchromosomal level; for example, the visualization of this matrix reveals a plaid pattern for intrachromosomal contacts that implies a possible division of each chromosome into two compartments called A and B as confirmed by PCA [12]

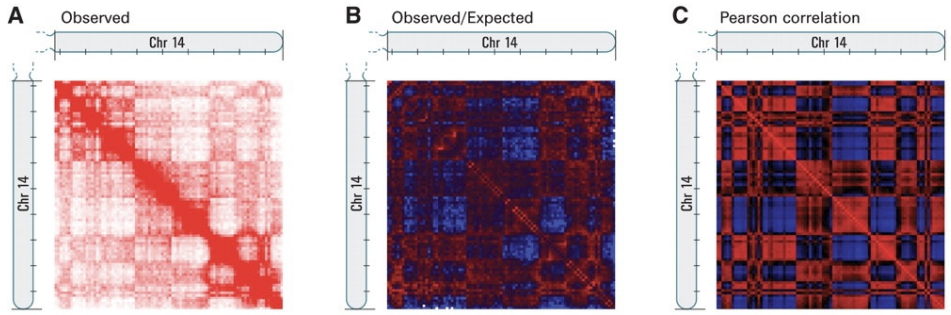

Figure 1.6: An image from [12] plotting matrices of chromosome 14 where the compartmental plaid pattern of subchromosomal structure is evidenced: A) the contact matrix of figure 1.4; B) the same contact matrix where entries are normalized with the expected contacts; red represents a value of contacts higher than expected, conversely blue lower; (C) Pearson correlation matrix that assigns values in the range [-1 (blue), 1 (red)] based on the correlation between the values of contacts of the *loci*.

The difference between the two compartments concerns the state of the chromatin: compartment A is associated with gene-rich chromatin that is actively transcribed, called euchromatin. On the other hand, compartment B is for the most part composed of heterochromatin, which is less associated with protein-coding sequences and more with regulatory functions [13]. It is also possible to observe from Hi-C data that interchromosomal contacts usually appear higher between regions that belong to the same compartment. As a matter of fact, studies imply that, in general, heterochromatin sequences are positioned towards the nuclear periphery, while euchromatin resides mostly in the center [3].

As one can imagine, for every chromosome, the intrachromosomal probability of contacts follows a decreasing trend (figure 1.4). Therefore, the average intrachromosomal probability of contact obviously has a decreasing trend as well. A finer analysis evidences that in sections of length between 500 kb and 7 Mb, which is the usual range of lengths of the sections associated to A or B compartments, the average probability exhibits the power-law scaling $\frac{1}{s}$, where s is the linear distance in the sequence. This suggests a particularly accurate model to describe the folding of the strands: the fractal globule, a structure that can be associated with polymeric clusters. This is a non-equilibrium structure characterized by the features of a Peano curve; in particular, it lacks knots. The functional

interpretation of this peculiarity is obviously that within the phase of transcription, the DNA can be unfolded without causing complications. This model is compatible with the polymeric structure of DNA and the beads on a string chain characterizing chromatin.

1.5 GPSeq

When approaching the description of the spatial features of the genome, it is possible to notice that many of them differ depending on their radial position. For example, the anticorrelation between chromosome size and distance from the center [12] [7], as well as the tendency of active euchromatin to reside in the central regions, in contrast with heterochromatin [3]. Due to the fact that euchromatin is associated with a high density of genes and high transcriptional activity, while heterochromatin is associated with lower values [16], gene density also seems to decrease with distance from the center of the nucleus [3] [17]. It is also well-known that small chromosomes are enriched in genes compared to the larger ones [21]; indeed, all of these properties seem related: small chromosomes are identified as gene-rich and are predominantly characterized by euchromatin (apart from chromosome 18 [6] [14]), while large chromosomes generally have opposite characteristics. Although not universally observed, these features generally suggest a probable overall conformation of the nucleus. In particular, these observations may suggest a possible functional interpretation: heterochromatin could occupy a more external position to operate as a shield against possible radiation damage for the more internal euchromatin [9], which could possibly lead to mutations in the sequence of the genome and subsequently result in pathological behavior by the affected cell. In this case, it would be interesting to understand why there are exceptions to this characteristic, in which cellular types these features are most frequently observed, and whether non-human cells also show these peculiarities. For all these reasons, the development of a method to investigate the radial properties of the genome intranuclear organization, i.e., how the various elements of the DNA vary with distance from the center of the nucleus, is therefore interesting and important in order to deepen knowledge about the spatial disposition of DNA. An experimental protocol to explore the nucleus in this direction was developed in 2020: it is called genomic loci positioning by sequencing (GPSeq) [7], and it is defined as a genome-wide method to obtain distances of fragments of DNA sequences from the nuclear lamina.

The method is based on a slow and gradual digestion of the intranuclear genome by an enzyme (*HINDIII*, as in [12]) that, starting from the periphery, moves towards the center, fragmenting the DNA, which has previously been cross-linked in the same way described in paragraph 1.2 to permeabilize the nucleus. The enzyme *HINDIII* behaves

as it does in the Hi-C experiment; therefore, the fragments are left with an overhang on one strand. The progression in the state of digestion is then observable using a Y-form adapter that links with each fragment through the complementary nucleotides to those left unpaired. This adapter is then detected using fluorescent nucleotides that ligate with the two arms of the adapter. In this way, as the enzyme progresses through the center, it is possible to observe a fluorescent circular band that, starting from the periphery, progressively covers the entirety of the nucleus. This particular fluorescent detection is called YFISH. Afterward, adapters that enable next-generation sequencing are ligated to the fragments. The MPS can be executed for different incubation times of the enzyme, leading to the creation of corresponding different libraries. To assess the radial positioning of the genomic *loci*, it is necessary to identify the radial estimate that shows the highest correlation with the 3D DNA FISH data regarding the cell. This estimate is derived from the variation in restriction probability within a genomic window across different digestion times and is referred to as the GPSeq score.

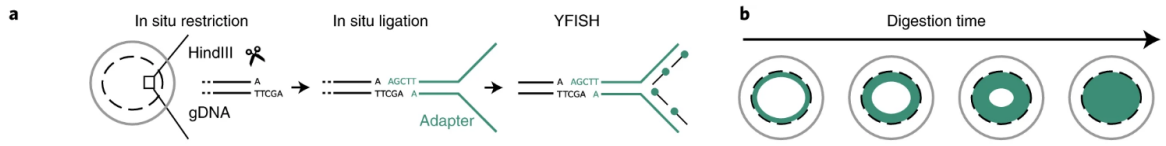

Figure 1.7: a figurative representation of the core passages of the GPSeq experiment from [7].

Studying the genome conformation of the cell line HAP1 through the GPSeq experiment, it is possible to observe that the correlation between the size of the chromosomes and the distance from the nucleus is weakly verified, confirming that the radial position of the chromosomes is not always based on the previously established rules. The same observations apply to both gene expression and gene-density. Noticeably, the highest correlation with GPSeq score as an estimate of radiality was achieved by the content of nucleotides G and C inside the genome, which anticorrelate with the distance from the nuclear lamina. This remarkable consideration is confirmed by the fact that actively transcribed regions, small chromosomes and especially high gene-density areas [27] are usually associated with a high content of GC. As a matter of fact, GC content is commonly used to estimate the genome fragments with a high presence of genes. However, at a resolution of 1 Mb, a model that includes GC content alone cannot accurately predict the radial locations of all the genomic elements throughout the nuclear space. The most accurate model for the task included GC content, gene density, gene expression and chromosome size. Also, euchromatin and heterochromatin, though predominantly confirming their patterns of internal and external positioning respectively, seem to follow

unique and more complex patterns in many cases when analyzed at a finer subcompartmental resolution. The radial polarization between active and inactive chromatin seems to be probabilistic and is not always confirmed in single cell examples [3].

1.6 GC content, radiality and functionality

So far, several properties of the spatial configuration of the genome have been indicated as examples of a peculiar dependence on radiality. Chromosome size, location of euchromatin and heterochromatin, transcriptionally active and inactive chromatin, and gene-density; lastly, GC content. All of these features are correlated, but the cause-and-effect relations between them remain vague and unclear. In this specific work, the intention is to concentrate on the peculiarities of the radial trend of the intranuclear GC content. This choice is justified by the fact that a possible peculiar GC arrangement along the nuclear radius could be crucial for explaining various intranuclear processes, as well as the long-discussed 3D structure of the human genome [3].
As explained in [3], a decreasing radial trend in GC content is observed not only in human cells [7] but also in other vertebrates, such as birds, reptiles, and amphibians. It is noteworthy that GC-predominant sequences have different physical properties from *loci* that predominantly contain A and T. As a matter of fact, the former are subject to fewer fluctuations in their thermodynamic properties at equilibrium and are characterized by increased flexibility. These properties contribute to the fact that the intranuclear stability of the genome is locally increased when the sequence is largely made up of guanine and cytosine. Polymeric simulations show that this difference in physical features is a possible reason for the gradient of GC content, as flexible areas of the polymer tend to colocalize in the center due to entropic forces. Obviously, the property of flexibility is fundamental for the unraveling of the DNA filament. As a matter of fact, transcriptionally active and accessible regions are usually associated with a high GC content. Flexibility is also important when it comes to the dynamical movements of the sequences. For all these reasons, it is not surprising that actively transcribed genes are generally GC-rich. Another important feature of many GC-rich sequences is the high transcriptional activity: as a matter of fact, it is proposed that the high GC content of certain *loci* could favor the creation of transcriptionally accessible clusters through phase separation [10] [24]. Being mostly located in the center, it is unlikely to observe phase separation in euchromatin regions because of the lower density of nucleosomes; thus, a higher GC content could act as compensation [3].
The causes of the emergence of a GC gradient might be related to the specific GC density throughout the sequence and the physical features that lead to GC-rich sequences in the

center, as previously explained. If these suggestions are trustworthy, possible exceptions to the establishment of the gradient could be associated with a pathological spatial disposition of DNA. The functional implications of the radial features remain unclear. As a matter of fact, there is not a certain answer regarding the validity of the bodyguard hypothesis mentioned in 1.5 [7]. As a matter of fact, there are several examples of single-nucleotide mutations associated with peripheral heterochromatin [15] [22], but there are also mutations, such as gene fusions, usually observed in the most internal nuclear regions [4]. However studies on specific functionalities of the GC gradient confirm the importance of a possible decreasing arrangement of this content [3]. For example, RNA-based splicing mechanisms operate on the pre-mRNA to create the mRNA in a different way based on the predominance of AT or GC. There are mechanisms that operate towards the periphery and mechanisms that operate towards the center of the nucleus. An odd radial disposition of GC and AT content consequently varies the splicing processes [25]. Moreover, the results of the GPSeq study[7] show that TF-binding sites are located throughout the genome based on their GC content trend. This implies that the corresponding TFs could also be distributed along a similar radial gradient [3]. Additionally, the process of linking with the sites is affected by a strong correlation in GC content between the binding sequences and the closest linearly outside sequences of the regulatory section. A plausible hypothesis is that TFs could consider the GC content of the sections, and based on that, they could distribute in an analogous way to operate on transcriptional regulatory activities [3].
Further studies are requested in order to advance knowledge about these possible properties. This work is located within this research framework.

2. Methods

Having introduced the setting and the main protagonists of the work, we now turn to the description of the methods used during the analysis, starting with a brief summary of the research.
The purpose of the analysis is to compute the GC content fraction as a function of the distance from the nucleus center for different cell lines and to compare it to an example, which is a cell line already provided as an example of structure in MOGEN files [26]. In particular, the benchmark is the GM06990 lymphoblastoid cell line. This particular cell line was already used in [12] and by MOGEN developers to respectively test the experimental Hi-C method and the quality of the reconstruction of the three dimensional structure. Moreover, in [7], it is confirmed that a particular radial chromatin arrangement of this line, comparable to the cell line HAP1 where a GC gradient was identified, has been observed. This lymphoblastoid line thus shows a GC gradient, as can be verified by executing the analysis that will be described.
The starting point is VC (Vanilla Coverage) normalized Hi-C data downloaded from a public database called *Genome Structure Database* [19]: these have a resolution of a megabase (i.e. 10^6 nitrogenous bases, which are A, T, G, C). The data that have been analyzed and used are from these cell lines:

- Hi-C from brain pericyte;
- Hi-C from brain microvascular endothelial cell;
- Hi-C from astrocyte of the spinal cord;
- Hi-C from DLD1, colorectal adenocarcinoma cell line
- Hi-C from ACHN, kidney adenocarcinoma cell line;
- Hi-C from SK-N-MC, neuroblastoma cell line;
- Hi-C from endometrial microvascular endothelial cells;
- Hi-C from endothelial cell of hepatic sinusoid.

After a process of data cleaning, which will be discussed in paragraph 2.1, it is possible to use MOGEN to reconstruct the 3D shape of the genome inside the nucleus based on the

Hi-C contacts, assigning each 1-Mb genomic bin a coordinate in arbitrary units. Meanwhile, it is possible to compute the fraction of GC content residing in each megabase from databases that offer the entire sequence of nucleobases of the human genome. Having assigned the GC content to all the megabases, it is necessary to partition the structure into different concentric shells. At this point, it is possible to compute the mean fraction of GC content between the megabases inside each shell. These values are indicative of the trend that the fraction assumes along the distance from the center of the nucleus and can be used to compare the structures with each other and with the benchmark one. All the analyzes were performed using *Python* as the programming language and *NumPy* as the principal computing library.

2.1 Preprocessing

As mentioned in 1.2, the most common representation for Hi-C contacts is the contact matrix, where the (i,j) element has the value of the interaction frequency between the i-th and j-th megabases. It is possible to highlight different contact values using a scale of colors. Using the logarithm base 2 of the contacts can help visualize the differences between the values more effectively, thereby highlighting the chromosomal structures, which are characterized by a high contact value between the megabases belonging to that chromosome, as well as possible interchromosomal zones that are near each other. That is the choice pursued in this study.
To facilitate better reconstruction, Hi-C contacts were preprocessed by deleting the megabases corresponding to contact values with other megabases that are affected by the noisy effects of the experiment. In particular, the criterion used to exclude the megabases from the analysis consists of computing the sum of all the contacts (using logarithm base 2) of a megabase with the others in absolute value. If the resulting sum is zero, the megabase is deleted. This allows us to exclude entire packs of megabases from the analysis, assuming the same value (1) of contacts for all of the *loci*: these megabases do not provide any valuable information for the analysis and can be visualized on the contact matrix because they are represented as rows and columns of the same color corresponding to the value $log_2(1) = 0$; neglecting them is useful to exclude unimportant megabases from the processing performed by MOGEN.
Moreover, the analysis was limited to the 22 chromosomes and chromosome X. Chromosome Y and mitochondrial DNA were excluded because the Hi-C contacts regarding these are usually too noisy.

2.2 Reconstruction using MOGEN

After preprocessing the Hi-C contacts, they were submitted to MOGEN to build the structure. Having read all the contact values between megabases, MOGEN uses its algorithm to assign a coordinate value in arbitrary units to every megabase. These are condensed into a PDB (Protein Data Bank) file, a format that is commonly used to represent polymeric macromolecules in three dimensional space, as explained in [8]; this type of file format contains the spatial coordinates of each monomer composing the sequence. Indeed, megabases are the monomers in this case, and they are associated with the coordinates computed by MOGEN. In order to extract information from this file, the function *PDBparser* from the package PDB belonging to the library BioPython was used. This function takes the file and returns a structure object, where it is possible to access the various coordinates.
Apart from the pdb file, MOGEN produces two text files: one containing a brief summary of the construction and another, called evaluation file, that includes the scores of contacts and non contacts respected by the construction for every pair of chromosomes.
A large amount of work was spent verifying whether the structures created by MOGEN were well-built. As a matter of fact, a flawed structure that doesn't reflect the well-known topological properties of the intranuclear genome would render the entire work invalid. Firstly, it is important to introduce the principal characteristics of the genome's layout and how the contact and non contact scores can indicate that these peculiarities are satisfied in the reconstruction; then, it is fundamental to understand how MOGEN parameters should be set in order to obtain the most reliable structure that numerically satisfies the characteristics previously introduced. Finally, it is possible to verify that the structures generated with such parameters are credible and to question whether the peculiarities of the genome's layout within the cellular nucleus are respected by visualizing them.
With the same approach applied below, common features of the human genome, such as the ones described in [5], were used to validate the Hi-C experiment in [12]. The most important characteristic a credible structure should possess is that the scores of interchromosomal non-contacts should always be higher than the corresponding contacts. Indeed, it is well-known that single chromosomal areas are defined within the nucleus, implying that chromosomes must not intermingle with each other; even when they are close, the interaction occurs while maintaining the individual chromosomal structures. This property translates into the previous affirmation considering the scores of the evaluation file: the non contact percentage between different chromosomes should be much higher than the contact percentage. If this request is accomplished, then the visualiza-

tion of the structure should verify that chromosomes do not intermingle.
It is fundamental for this request to be satisfied for the largest chromosomes that occupy the most external regions of the nucleus, i.e., chromosomes in the range between 1 and 15, and usually chromosome 18, which is commonly positioned in the periphery, even if it occupies a lower volume in comparison to the others. Small chromosomes occupying the center of the structure are subject to a less rigid evaluation of the scores; as a matter of fact, the intent is to highlight the interactions between different chromosomes by maximizing the contact percentages between them as much as possible while keeping the chromosomes distinct. For this reason, it is possible to create a good structure while still having one or two pairs of small chromosomes intermingling in the central region.
Another important feature deriving from these considerations is that the intrachromosomal contact scores need to be higher than the interchromosomal ones. This assures that the single chromosomal structures occupy a compact subregion of the nucleus without excessively spreading. It is more difficult for larger chromosomes to show a high value of contact percentage because of their dimensions. However, there are parameters that can be accurately set to obtain a good score even for the largest chromosomes.
At last, a feature of the disposition of chromosomes inside the nucleus is, as mentioned before, the position of chromosome 18 and the difference between large and small chromosomes in intranuclear positioning. As a matter of fact, in [26], this criterion is used to evaluate the goodness of the reconstructed structure. Large chromosomes (from 1 to 15 and chromosome X) tend to be placed in the periphery of the nucleus, while small chromosomes are usually located in the center [7] [12]. The structures usually need to respect this difference in placement between broader and more compact chromosomes; therefore, contact scores between different small chromosomes must generally be greater than those of other chromosomes due to the compactness in the center. It is still important to check whether the position of chromosome 18 in the structures is peripheral or not. These last considerations are general, but it is possible to observe exceptions in structures regarding both large and small chromosomes. If these exceptions are confirmed by the Hi-C data, the structure can still be considered reliable. For example, if the reconstruction shows chromosome 18 lying in the center and the contact matrix effectively presents a high value of Hi-C contacts with the small chromosomes in the center, then the violation of the usual feature is not worrying. It is necessary to execute this type of check. It can also be interesting to observe and compare where these features are respected and where they are not.
A fast and reliable method of understanding whether the percentages in the evaluation file respect the previous requests is to compare them with those of the cellular line provided as an example in the MOGEN files, which serve as our benchmark. As a matter of

fact, the reconstruction has the features previously shown; therefore, it can be used as a benchmark to see if the newly generated structure conforms to the paradigms illustrated earlier. This is the procedure used during the work and will be explained in full detail below.

MOGEN parameters, which are thoroughly presented in the *Supplementary Material* of [26], play a core role in the reconstruction; the algorithm is entirely based on them. In addition to the parameters already described in paragraph 1.2, it is obviously necessary to specify the number of chromosomes (23 in this case) and the megabases corresponding to each one of them, the learning rate for the gradient ascent, and the maximum number of iterations.

MOGEN has a set of parameters ready to use. It is fundamental to understand whether these parameters can be used for every cell line or if they need to be modified in order to obtain a more reliable structure. To answer this question, a series of reconstruction tests with different sets of parameters for a single cell line were performed, and the resulting evaluation files were compared to the example one. The comparison consists of computing the euclidean distance of every non contact and contact score between the files:

$$d(S) = \sqrt{\sum_{i=1}^{23} \sum_{j \geq i}^{23} (P_{ij}(S) - P_{0ij})^2}$$

where S is the structure to be compared, the sum is indexed over the chromosomes that vary from 1 to 23 (which is equivalent to chromosome X), $P_{ij}(S)$ is the score of non contacts or contacts respected from the structure S between chromosome i and chromosome j, and P_{0ij} is the corresponding percentage from the lymphoblastoid cell line. The two distances (one computed with contact percentages and the other with non contact percentages) provide a measure of the difference between the generated structure and a well organized genome reconstruction, such as the one used as a benchmark.

It is also possible to visualize, using a scale of colors, the matrices of contact and non contact scores of the structures: these are formed by the values P_{ij} (P_{0ij} for the lymphoblastoid cell line). Obviously, we expect the structures with a low $d(S)$ to be more visually similar to the benchmark one. To verify the consistency of ready-to-use MOGEN parameters, 32 structures of the same cell line (brain microvascular endothelial cell line) were generated with different sets of parameters established by following the instructions in the Supplementary Material of [26] and the parameters already in use. Summarizing the logic of the setting of MOGEN parameters:

- As advised in the supplementary material of MOGEN, all the W1 values were set to 1, and only the W_4 values were modified in order to change contact and

non contact percentages. Indeed, W_1 and W_4 have opposite effects on the values; therefore, it is useful to use only one of these sets while keeping the other fixed at the ineffective value of 1. It is usually sufficient to set the weights W_4 accurately in order to achieve a reliable reconstruction.

- Generally, W_4[chr a $\neq$ chr b] needs to be less than 1 because Hi-C data are usually low for megabases of different chromosomes: non contacts between them are therefore usually respected by the structure, resulting in a high non contact percentage in general; in order to maximize the score of contacts between them, it is useful to prioritize the contribution in the scoring function by setting this weight to a low value.

- Regarding the weights W_4 [chr a = chr b], they are usually set higher than one for opposing reasons. As a matter of fact, intrachromosomal megabase contacts are usually respected, especially for the smallest chromosomes. Therefore, it is necessary to prioritize intrachromosomal non contacts.

These are the most important suggestions to follow. Computing the distances of every structure from the contact and non contact percentages of the benchmark cell line offers the possibility of establishing a ranking of the various reconstructions: the parameters that generate the closest structure could be considered the best parameters for the reconstructions. It is subsequently possible to generate the structure of the other cell lines with them and compute their distance from the benchmark percentages.
It is possible to verify that the properties of the genomic structure are satisfied by visualizing the generated structures and highlighting the various chromosomes. A simple visualization of the structure, with a different color corresponding to each chromosome, should be enough to understand if the main single chromosomal structures are respected. However, in order to precisely analyze the position of the chromosomes inside the nucleus, violin plots could be more effective: a violin plot illustrates the distribution of the distances of the megabases from the center based on the chromosome to which they belong; it is possible to plot the different distributions for each chromosome near each other in order to compare them. The visualization of the structure was performed using *MatPlotLib*, while the violin plots were created with the functions offered by *Seaborn*.

2.3 Gradient computation

At this point, the reliability of the generated structure permits the start of the analysis of the disposition of the base content inside the nucleus.

First of all, it is necessary to obtain the fraction of chromosomal content for every megabase. This can be achieved using a database containing the entire sequence of the bases of the human genome divided by chromosome. Different versions of genome sequencing are available in freely accessible databases. In our case, hg18 is the one used for the benchmark cell line, while hg19 is used for all the others. The possible differences between older and newer versions of genome sequencing will be discussed in the following paragraphs.
In practice, the sequence is an extremely long string that consists of five letters: A, T, G, C, and N. The first four obviously refer to the four famous nitrogenous bases composing DNA. The letter N is instead used to indicate an unknown base. The sequence is encoded in a special file format called fasta. This type of file is a format commonly used for the representation of long nucleotide or peptide sequences. The library SeqIo has come in handy to process this type of file and extract the string containing valuable information. Now, it is possible to subdivide the sequence into chunks of a megabase and compute the fractions of AT, GC, and N contained in each of these pieces. Note that the sum of these fractions must obviously add up to one.
Having assigned a value to the content of the bases for each megabase, it is time to start analyzing the radial features of their disposition using the coordinates $\vec{x}_i$ of the i-th megabase in arbitrary units obtained from MOGEN. Having assumed that the arrangement of the genome inside the nucleus is spherical, even if discrete and not continuous in our case, it is necessary to fix an origin, which is the point that can be used to compute the radial distances of the various megabases. The choice falls on the center of mass of the structure, computed as the mean position of the megabases:

$$\vec{x}_0 = \sum_{i=1}^{N} \frac{\vec{x}_i}{N}$$

It is then possible to compute the distance of the megabases from the center: $r_i = |\vec{x}_i - \vec{x}_0|$. The computation of the radial GC, AT and N fractions is now available. Dividing the space into concentric shells, with the center always placed in the center of mass, the mean of the AT, GC, and N fractions inside each shell is easily computed: it is a mean of the fractions on the megabases residing in the specific shells. Indicating with R_i the external radius of the i-th shell and considering η_i as the number of megabases contained in the shell, i.e., whose distance from the center r satisfies the requisite $R_{i-1} < r \leq R_i$,

the mean GC content inside it is given by

$$GC_i \quad = \quad \sum_{\substack{j=1 \\ R_{i-1}<r_j\leq R_i}}^{\eta_i} \frac{gc_j}{\eta_i}$$

where gc_j is the fraction of GC content within the j-th megabase.
Note that the sum of the means of the various contents within the same shell should always add up to one: indeed, considering that the sum of the fractions of the same megabase is $gc_i + at_i + n_i = 1$, and the definition of the number of megabases inside the i-th shell is

$$\eta_i \quad = \quad \sum_{\substack{j=1 \\ R_{i-1}<r_j\leq R_i}}^{\eta_i} 1$$

it is implied that:

$$GC_i + AT_i + N_i \quad = \quad \sum_{\substack{j=1 \\ R_{i-1}<r_j\leq R_i}}^{\eta_i} \frac{gc_j + at_j + n_j}{\eta_i} = \quad \frac{1}{\eta_i} \quad \sum_{\substack{j=1 \\ R_{i-1}<r_j\leq R_i}}^{\eta_i} 1 = 1$$

At this point, plotting the fractions GC_i, AT_i, and N_i on a graph as a function of the distance from the center can be extremely helpful for grasping the radial trend they exhibit. In particular, the x-axis coordinate assigned to the contents of the i-th shell is the mean of the distances from the center assumed by the megabases within that shell:

$$\bar{R}_i \quad = \quad \sum_{\substack{j \\ R_{i-1}<r_j\leq R_i}}^{\eta_i} \frac{r_j}{\eta_i}$$

To ensure a physically realistic measure, it is necessary to account for the uncertainty that needs to be assigned to the mean values of the contents inside the shells and the mean distance of the megabases. Using the gaussian error for the mean would lead to an underestimate of the errors because the values of GC content composing the distribution in each shell are not mutually independent; in fact, it was mentioned in paragraph 1.4 that compartments A and B of a chromosome tend to have closer spatial interchromosomal distances with compartments of the same type. Compartment A is associated with euchromatin, which is usually GC rich. Therefore, high GC content areas could cluster together in 3D space. Thus, the GC values of close megabases are correlated. At this

point, the choice has fallen on a simple standard deviation of the values:

$$\sigma_i(GC) = \sqrt{\frac{1}{\eta_i - 1} \sum_{\substack{j \\ R_{i-1} < r_j \leq R_i}} (gc_j - GC_i)^2}$$

This is a statistically reliable estimate of the error on both the x and y values and offers a measure of the intrinsic variability of the content. Obviously, an analog form of the standard deviation can be expressed for AT and N contents inside the shell, as well as for the radial distances of the megabases The simplest manner to organize the shells that subdivide the space for our analysis is by choosing an initial external radius for the first shell R_1 and keeping the thickness of the shells constant by initializing the other external radii as $R_j = jR_1$. However, this isn't optimal in this case; in fact, the volume of the shell increases as the two radii increase in value; this results in an excessive difference between the number of megabases contained in the first shell, which has a lower volume than the others and therefore generally contains fewer megabases, i.e., not enough to conduct a reliable statistical analysis. A more appropriate method is to require that the shells have the same volume. This allows for a higher number of megabases in the first shell while keeping the number of megabases across the shells roughly high enough to run a statistic. In addition, it is still a method that is directly related to the distance from the center. Another method used consists of building the various shells with the aim of keeping the number of megabases inside each one constant. A double analysis based on different constructions for the shells is anyway important to verify the consistency of the results and check that the results are not influenced by a biased choice of the shells.
Let us assume the intent is to divide the structure into N_{shells} shells with this characteristic. In order to perform the previous computations, it is necessary to have an expression for the external radius of each shell. In this case, it can be useful to compute the volume of the genome structure, which is given by $V = \frac{4}{3}\pi R_{max}^3$, where $R_{max} = max_i\{r_i\}$ is defined as the value of the highest distance from the center of mass assumed between the megabases. Then, the constant volume that the shells need to have is $V_c = \frac{V}{N_{shells}}$. For the first shell, assuming $R_0 = 0$, this implies:

$$V_c = \frac{4}{3}\pi R_1^3 \Rightarrow R_1 = (\frac{3}{4\pi}V_c)^{\frac{1}{3}}$$

For the second shell $V_c = \frac{4}{3}\pi(R_2^3 - R_1^3)$. By inserting the previous expression for V_c, it is possible to compute the value of R_2:

$$V_c = \frac{4}{3}\pi R_2^3 - \frac{4}{3}\pi R_1^3 \Rightarrow V_c = \frac{4}{3}\pi R_2^3 - V_c \Rightarrow R_2 = (\frac{3}{4\pi}2V_c)^{\frac{1}{3}}$$

In general, the expression for the external radius of the i-th shell is given by:

$$R_i = (\frac{3}{4\pi} i V_c)^{\frac{1}{3}}$$

Therefore, by deciding the number of shells to achieve, this expression easily computes the external radii of the various shells to keep the volume constant among them.
If we define the operation $int(r)$ as rounding down the real number r to the nearest integer, and the operation $a\%b$ as the remainder of the division between the integer numbers a and b, the outer radius for the shells at constant volume is determined by setting $R_0 = 0$ and choosing the minimum value R_i greater than R_{i-1} (the outer radius of the previous shell) that satisfies:

$$\eta_i = \sum_{\substack{j=1 \\ R_{i-1} < r_j \leq R_i}}^{\eta_i} 1 = \begin{cases} int(\frac{N}{N_{shells}}) + 1 & \text{if i} \in \{1, 2, ..., (N\%N_{shells})\} \\ int(\frac{N}{N_{shells}}) & \text{if i} \in \{(N\%N_{shells}) + 1, ..., N_{shells}\} \end{cases}$$

where N is the total number of megabases. It is also necessary to specify that if $N\%N_{shells} \neq 0$, the number of megabases per shell can actually assume two values: $int(\frac{N}{N_{shells}})$ and $int(\frac{N}{N_{shells}}) + 1$. However, the value of $\frac{N}{N_{shells}}$ will be high enough to make this variation negligible.
These expressions finally allow the computation of all the quantities previously described for both types of shells. It is possible to visualize the positions of the various megabases in three dimensional space, highlighting them with a color map depending on their GC fraction. These should offer a first suggestion regarding a possible radial arrangement of the content. However, in order to execute a more precise analysis, graphs representing the points formed by the couples $(GC_i, \bar{R}_i)$, $(AT_i, \bar{R}_i)$, $(N_i, \bar{R}_i)$ have been visualized. These are useful for comparing the trends of the different contents based on radiality and for proceeding into a deeper analysis, which will be explained in detail in the next section.

2.4 Fit

Deducing the results of the analysis through the visualization of the structure and the graphs is important for gaining a preliminary idea of the behavior exhibited by the contents of the bases along the distance from the center of the nucleus. However, taken alone, it only allows one to make suppositions and superficial comparisons between the different structures and the cellular line of example. It would be necessary to quantify numerically, in some way, the behavior of the GC, AT, and N fractions for reconstruc-

tion in order to make precise comparisons and establish with scientific certainty whether gradients are present inside the nucleus. An optimal procedure to accomplish this is to execute a fit of a model on the data; this allows us to understand whether the points we see on a generic graph respect the initial hypothesis. In this case, wanting to verify if a non-null gradient is present, a good choice would be to perform a linear fit $y = a + bx$ of the data and verify if the angular coefficient b is non-vanishing, comparing it across the various structures, especially with the one taken as an example, which effectively shows a radial distribution of GC content.
Usually, the method used for this purpose is the classic linear regression. This analysis computes the parameters a and b of the line by minimizing the sum of the vertical distances between the line and the points, as explained in [20], as a result of the maximum likelihood applied to this specific context. Effectively, if N points are considered to be gaussianly distributed with constant variance around their real values on the y-axis, modified by random systematic effects and independently from one another, the probability that a specific point is aligned with the hypothesis of linearity, i.e., that it assumes a value in the interval Δy around y_i, should be:

$$P(y_i|model) \quad \alpha \quad \exp(-\frac{1}{2}(\frac{y_i - bx_i - a}{\sigma}))\Delta y$$

The probability that the entire dataset follows a linear trend is then easily defined as the product of all these factors for each point. It is then possible to use the result of Bayes' theorem:

$$P(model|dataset) \quad \alpha \quad P(dataset|model)P(model)$$

where P(model) is the probability that the model fits the data previously among all the possible models. It is possible to assume that this probability is uniform; therefore, the most probable model is the one that maximizes P(dataset—model), which is equivalent to minimizing the negative logarithm of it, resulting in the expression:

$$\sum_{i=1}^{N} \frac{(y_i - bx_i - a)^2}{\sigma^2} - Nlog(\Delta y)$$

Indeed, Δy is constant and σ is too, implying that the minimization of the previous expression is equivalent to minimizing the vertical distances between the points and the line. In this way, it is possible to compute the best values for parameters a and b, leading to the formulas for linear regression.
In our case, the variance is not constant for all the points because they are associated with generally different errors. In this case, the correct minimization should consider the

different variances, leading to the expression:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - bx_i - a)^2}{\sigma_i^2}$$

which is the χ^2 function, an error-weighted version of the previous expression. Its minimization with respect to the parameters a and b implies the formulas for linear regression with different errors on the y-axis.
It is, moreover, possible to compute the errors for the parameters, easily considering the propagation of the uncertainties of the points:

$$\sigma_a = \sum_{i=1}^{N} \sigma_i \frac{\partial a}{\partial y_i} \qquad \sigma_b = \sum_{i=1}^{N} \sigma_i \frac{\partial b}{\partial y_i}$$

Unfortunately, all the previous passages have a precise prerequisite: it is necessary to be able to consider the errors of the x-axis coordinate of the point (in our case, the standard deviation of the distances from the center of the megabases inside the shell in this case) negligible in comparison to the y-axis errors (in our case, the standard deviation of the GC, AT, or N fractions of the megabases inside the shell). This is not compatible with our dataset, as the errors on the distance from the center are of the same magnitude as the errors on the fractions. This means it is impossible to use the most common linear regression for the work. As a matter of fact, the variance of the linear combination $(y_i - a - bx_i)$ of the two independent variables y_i and x_i is $\sigma^2(y_i - a - bx_i) = \sigma_{y_i}^2 + b^2\sigma_{x_i}^2$. Therefore, it is necessary to consider this weight for every point, leading to the following:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - bx_i - a)^2}{\sigma_{y_i}^2 + b^2\sigma_{x_i}^2}$$

The minimization of this expression, at least for parameter b, is not trivial like the previous ones and requires particular numerical methods, such as finding the errors of the parameters.
Fortunately, a method implemented by *odrpack* comes in handy. It is a general expansion of the previously described method for linear regression, called ODR (orthogonal distances regression). This procedure is based on the same core reasoning as the previous one, but instead of minimizing the χ^2 based on only vertical distances, the minimization concerns the actual orthogonal distances between the points and the line. In this case, it is not necessary to neglect the errors on the x-axis. It is possible to show that the minimization implemented in this general case is equivalent to the minimization of the

last χ^2 in the case of a linear function.
The library implements two methods for the ODR. One method is employed when the function f chosen for fitting can express the relation between the variables y and x in an explicit way, as in the case of $y = bx + a$. The second method is used if f expresses the relation implicitly, for example, in the equation of an ellipse. Obviously, the procedure that is interesting in this case is the first one.
The algorithm for minimization is presented in the reference guide of the 1992 version ([2]) and has remained unaltered until the latest 2024 version. In particular, let us assume that the true values y_i^* of the respondent quantity are related to the x_i^* through the function f, which is the one used for the fit. The function can be nonlinear and depends on a set of parameters $\vec{\beta}^*$, so the relation can be expressed as $y_i^* = f(x_i^*, \vec{\beta}^*)$. Now, the observed values are y_i and x_i, which differ from the true values due to random effects. In particular, by defining the two distances $\epsilon_i^* = y_i^* - y_i$ and $\delta_i^* = x_i^* - x_i$, it is possible to rewrite the previous relation as $y_i = f(x_i + \delta_i^*, \vec{\beta^*}) - \epsilon_i^*$. The problem of ODR consists of approximating the true set of parameters $\vec{\beta}^*$ by finding the parameters $\vec{\beta}$ such that the sum of the orthogonal distances between each point $(x_i, y_i) = (x_i, f(x_i+\delta_i, \vec{\beta})-\epsilon_i)$, where ϵ_i is the vertical distance between the points and the curve, while δ_i is the horizontal one, and the curve $f(x, \vec{\beta})$ is minimized. Indeed, ϵ_i and δ_i are approximations of ϵ_i^* and δ_i^*. The quantity that should be minimized, given the constraint that $\epsilon_i = y_i - f(x_i + \delta_i, \vec{\beta})$, is:

$$\chi^2 = \sum_{i=1}^{N} \epsilon_i^2 + \delta_i^2 = \sum_{i=1}^{N} (y_i - f(x_i + \delta_i, \vec{\beta}))^2 + \delta_i^2$$

If x_i and y_i have different errors associated, the previous quantity is easily generalized using weights :

$$\sum_{i=1}^{N} w_{\epsilon_i}(y_i - f(x_i + \delta_i, \vec{\beta}))^2 + w_{\delta_i}\delta_i^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i + \delta_i, \vec{\beta}))^2}{\sigma_{y_i}^2} + \frac{\delta_i^2}{\sigma_{x_i}^2}$$

The minimization of this function, with respect to δ_i and the parameters $\vec{\beta}$, is possible using a trust region Levenberg-Marquardt method, as described in [1].
Computing the errors of the parameters is similarly impossible in general, and it is necessary to resort to approximations. In particular, let us pose $g_i(\vec{\beta}, \delta_i) = w_{\epsilon_i}[f(x_i + \delta_i, \vec{\beta})-y_i]$, $g_{i+N} = w_{\delta_i}\delta_i$, and $\boldsymbol{\theta} = \begin{pmatrix} \vec{\beta} \\ \tilde{\delta} \end{pmatrix}$, where $\tilde{\delta}$ is a vector containing all the components δ_i. Then we can observe that $\chi^2 = \sum_{i=1}^{2N} g_i(\boldsymbol{\theta})$ and the corresponding Jacobian can be written as

$$\boldsymbol{J}_{ij} = \frac{\partial g_i}{\partial \theta_j}$$

Defining $\boldsymbol{\Omega} = diag\{w_{\epsilon_i}, w_{\delta_i}, i \in \{1, 2, ..., N\}\}$ and

$$\boldsymbol{\sigma}^2 = \frac{\boldsymbol{g}(\boldsymbol{\theta})^T \boldsymbol{\Omega} \boldsymbol{g}(\boldsymbol{\theta})}{N + p}$$

where $\boldsymbol{g}$ is the function column vector containing all the g_i, and p is the number of parameters, it is possible to define the covariance matrix as

$$\boldsymbol{V} = \boldsymbol{\sigma}^2 [\boldsymbol{J}^T \boldsymbol{\Omega} \boldsymbol{J}]$$

Errors on the parameters are then computed as the square root of the diagonal elements of this matrix estimated by *odr_fit* . However, it is necessary to provide the algorithm in advance with a more or less realistic estimate of the two parameters a and b. Good starting values can significantly decrease the computational cost of the algorithm and can sometimes be fundamental in choosing the correct solution between two possible minimum values. In order to do this, a procedure from [20] was applied, consisting of executing a normal linear fit to a particularly processed version of the data. In particular, it is necessary to scale the value of the fractions of GC content by the factor given by the rate of the standard deviation between the GC content values across the shell and between the corresponding distances from the center: $f = \frac{\sigma_y}{\sigma_x}$ where $\sigma_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2}$ with $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ and σ_x is analogous. After this, the x_i and the y_i of the data have exactly the same variance. It is now possible to apply the linear fit using the weights

$$w_i = \frac{1}{\sigma_{x_i}^2 + \sigma_{y_i}^2} = \frac{1}{2\sigma_{x_i}^2}$$

This operation provides two parameters a_{scaled} and b_{scaled} that need to be rescaled back by dividing by f in order to be used as the initial guesses to proceed with the ODR. The linear fit was executed using the function *polyfit* from NumPy.

At this point, using *odr_fit* with the starting guesses, the parameters a and b are obtained along with their associated errors. It is possible to use the parameter b to verify the presence of the gradients and to compare different cellular lines with each other and with respect to the benchmark one.

A measure of the quality of the executed fit is also necessary in order to trust the credibility of the parameters. The function used for this purpose, following instructions from [20], is the survival function implemented by the chi2 package *chi2* from SciPy. This function is defined as one minus the cumulative function computed from the χ^2 probability distribution function with the degrees of freedom (N - p, in our case N - 2). It can be said that it represents the probability of a value of χ^2 being higher than

or equal to the observed value, commonly referred to as the p-value. If the p-value is too low, then it is possible that the proposed model is not valid for the dataset because of a too high χ^2: the differences between the model and the data are too large to be attributed to gaussian fluctuations. On the other hand, if the p-value is close to one, which would correspond to an ideally perfect situation where there is no possibility of a better estimate of the parameters, it could be that the model fits the dataset perfectly, or it may be due to an overestimation of the errors that increases the range of tolerance for the correspondence between the model and the data.

2.5 hg18 and hg19 reference genome versions

Before all the passages discussed above, a preliminary comparison was performed between the hg18 and hg19 versions of the reference genome regarding GC content within various chromosomes to identify hypothetical differences that could explain the incompatibility between the obtained results. As a matter of fact, the first version was used in the analysis of the lymphoblastoid cell, where the gradient is observable, while the second one was used in the other cell types. The hg19 version is newer than hg18. Hence, we may presume a difference in GC content between the two versions, which would be a problem for the comparison between the various cellular lines and the example one: there could be bases identified by the letter 'N' in the hg18 version that were later identified as G, C, A, or T in the hg19 version; thus, they are no longer ignored when computing the GC fraction. In addition, there could also be changes affecting the known bases. Hence, we may presume a difference in GC content between the two versions: there could be bases identified by the letter 'N' in the hg18 version that were later identified as G, C, A, or T in the hg19 version; thus, they are no longer ignored when computing the GC fraction; in addition, there could also be changes affecting the known bases. A difference such as this in the sequencing of the genome could lead to different results between the example cell line and the others, rendering a fair comparison between the cell types and the example one very difficult.
In order to understand if an eventual discrepancy between the two versions could be non-negligible, the fraction of GC content for each chromosome and both versions was computed directly from the sequences of the databases providing them. Subsequently, a similar computation was performed by calculating the average GC content over megabases within each chromosome after preprocessing. that is to say, after dividing sequences into megabases, removing the megabases characterized by none or only one Hi-C contact, and obtaining the GC fraction of the remaining ones, the average GC content over megabases belonging to the same chromosome was computed. Such a procedure was

performed for both versions, and the resulting comparison is really useful for detecting important differences, given that the gradient is computed as a mean of the GC fraction over the megabases within the same shell, where the ones excluded by preprocessing are not considered.

3. Results and discussion

In this section, the main goal is to illustrate the results of the analysis described in the previous chapter. Starting from the very beginning, it is necessary to verify the effects of the preprocessing phase on Hi-C data of the cell lines that are going to be analyzed by visualizing the contact matrix before and after the selection of the megabases. Afterwards, it is fundamental to assess the best MOGEN parameters for the generation of the three dimensional genome, comparing the distances $d(S)$ between the structures generated with different parameters for the brain pericyte cell line and establishing a ranking among this sample. It is moreover possible to verify that the structures generated are reliable by visualizing the various chromosomes and the violin plots representing the radial distribution of their megabases. The last important check is constituted by the comparison between the hg18 and hg19 versions of the human genome, which is fundamental to assess that any eventual differences in the gradient between the benchmark cell line and the others are not caused by discrepancies due to the two sequences used. In the end, it will be time to observe the GC fraction contained in the megabases and comment on the graphs and the results of the linear fit of the various radial contents (AT, GC and N).

3.1 Contact matrices and data preprocessing

In figure 3.1, it is possible to observe the visualization of the Hi-C data; in particular, the logarithm base two of the various interaction frequencies between the megabases for every cell line apart from the benchmark one, whose coordinates were already provided by MOGEN. The highest values of the contacts between intrachromosomal bins are well highlighted by the formation of squares along the diagonal; each corresponds to the interaction frequencies within a specific chromosome, and they progressively decrease in size proportionally to chromosome dimensions until chromosome X. However, the matrices are characterized by noisy lines of contacts with a value of 0, which are deleted during the preprocessing phase. In figure 3.2 it is finally possible to distinguish the plaid pattern typical of the two compartments A and B associated with euchromatin and heterochromatin [12]. With finer observation, the astrocyte, adenocarcinoma, endometrial and brain endothelial cell lines seem to have a generally brighter color, indicating a roughly higher value of interaction frequencies. It is generally observable that there is a more

incisive increase in the contacts in proximity to the smaller chromosomes, a sign of their typical central intranuclear location. On the other hand, it is also possible to observe a general decrease in the correspondence of the megabases of chromosome 18, a possible clue to its external placement, with differences among cell lines regarding the contrast with other values.

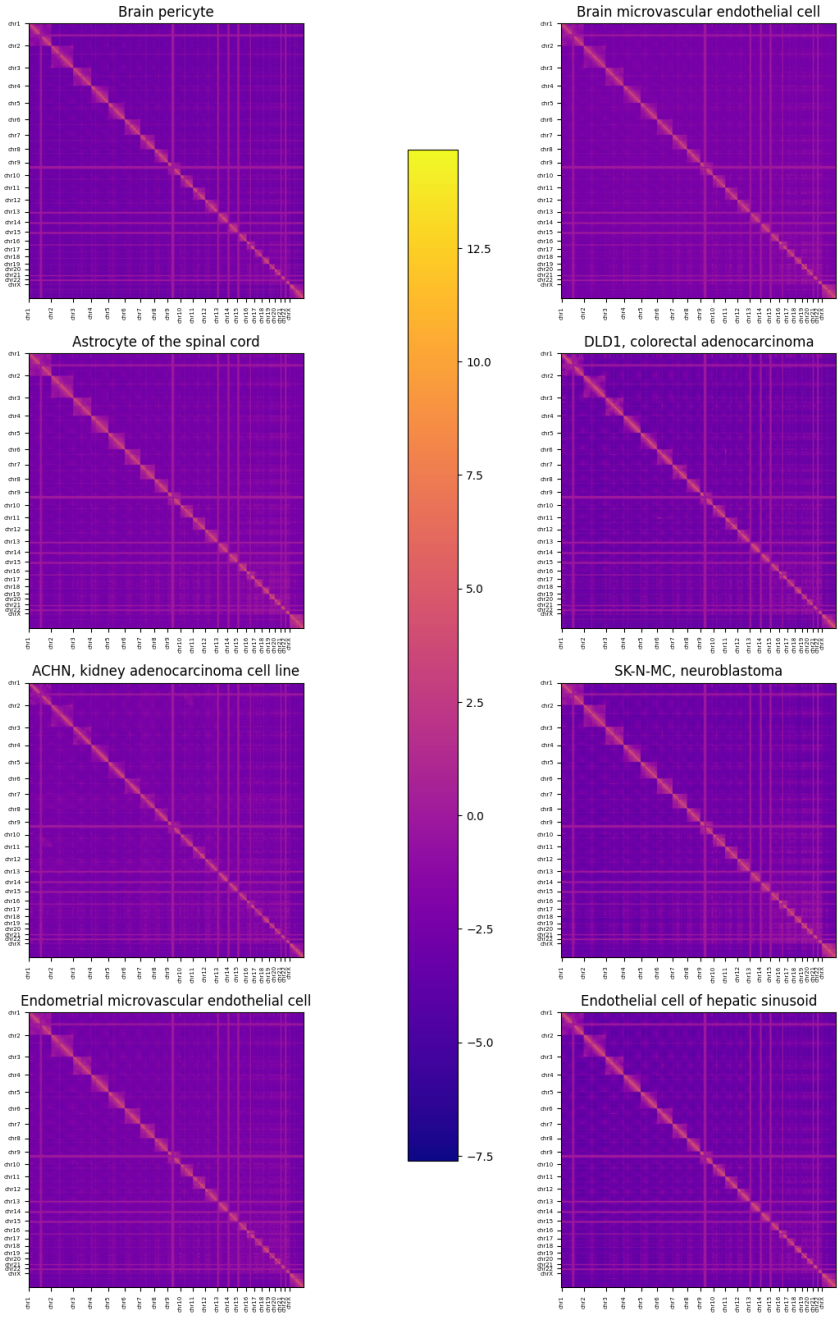

Figure 3.1: Unprocessed contact matrices of the different cell lines. The heatmap is based on the $log_2(F_{ij})$, where F_{ij} is the normalized value of interaction frequency between bin i and bin j. By definition, the contact matrices are square and symmetric.

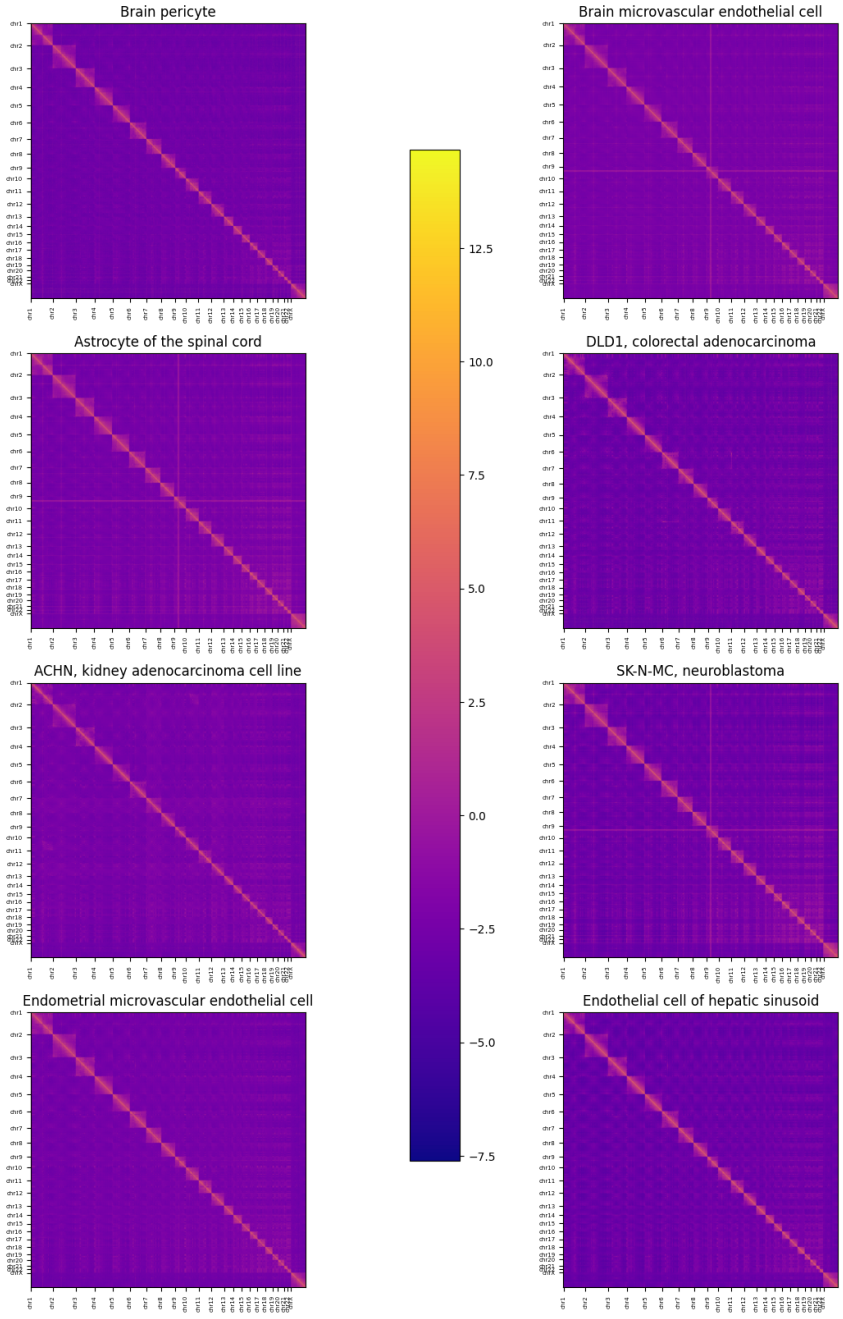

Figure 3.2: preprocessed contact matrices of the different cell lines. As in figure 3.1, the heatmap is based on the $log_2(F_{ij})$.

These data are used for the reconstruction with MOGEN. Even after preprocessing, astrocyte, brain endothelial and neuroblastoma cell types show two noisy contact lines corresponding to some of the bins of chromosome 9. The sum of the absolute values in these lines is not zero, although they likely do not provide reliable data.

3.2 MOGEN parameters validation

As explained in paragraph 2.2, it is now necessary to compare the 32 reconstructions of the brain pericyte cell to the structure of the lymphoblastoid cell line provided by the MOGEN authors, aiming to discover the ideal set of parameters for creating all the other virtual representations. The structures were partly created by varying the parameters between those already used and those advised in the supplementary material or the paper ([26]) of MOGEN. The details about how the parameters were varied can be consulted in the appendix. Observing the non contact score and contact score matrices (Figure 3.3, Figure 3.4), it is first possible to see that they are roughly complementary to one another, respectively. Therefore, the observations on one are conversely applicable to the other. The non contact score matrix of the benchmark structure shows high scores for all the larger chromosomes and chromosome 18. In particular, they are generally higher than the diagonal values, which correspond to the intrachromosomal non contact scores; the lowest values are associated to the areas of the smaller chromosomes. All of these features have been mentioned in paragraph 2.2 as examples of a reliable structure. Observing the trials for the brain pericyte now, it is evident that the majority can be discarded as they do not reproduce any of these characteristics. As a matter of fact, their matrices do not have any common patterns with the benchmark. Other structures, such as 11, 13, 15, 25, 30 and others, show a contrast between the values of chromosome 18 and its closer chromosomes. However, they do not have other reliable scores, especially for chromosome X, which is often low in non contact scores (high in contact scores), while in the reference structure, its scores are oppositely set. The only structure that is not too far from having similar features to the benchmark is structure 1.
The suggestions can be numerically confirmed by the computation of the distances $d(S)$ of the trials, both in contact and non contact scores, as described in paragraph 2.2. A ranking between the structures can be established by sorting these distances in ascending order with their respective structures, as reported in Table 3.1. As expected, structure 1 is the "closest" to the benchmark structure for both contact and non contact scores. The corresponding parameters for the reconstruction are exactly the ones provided by the MOGEN repository.

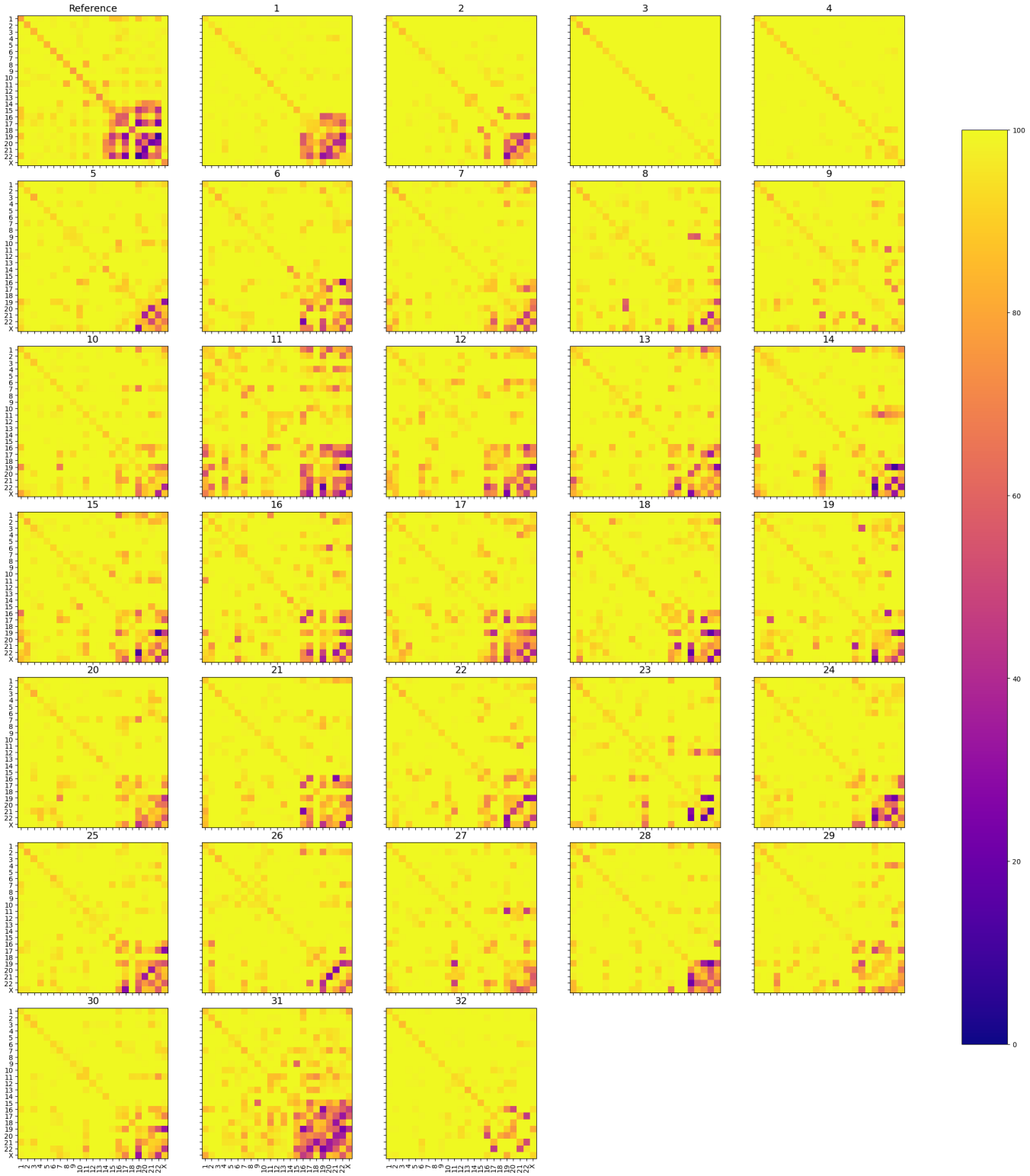

Figure 3.3: Non contact scores matrices of the 32 structures created for the brain pericyte cell line with different parameters and the reference one, which is from the lymphoblastoid reconstruction. The non contact score between chromosome i and chromosome j is the percentage of non-contact constraints between bins respectively belonging to chromosome i and j satisfied by the reconstruction. The heatmap is based on these values.

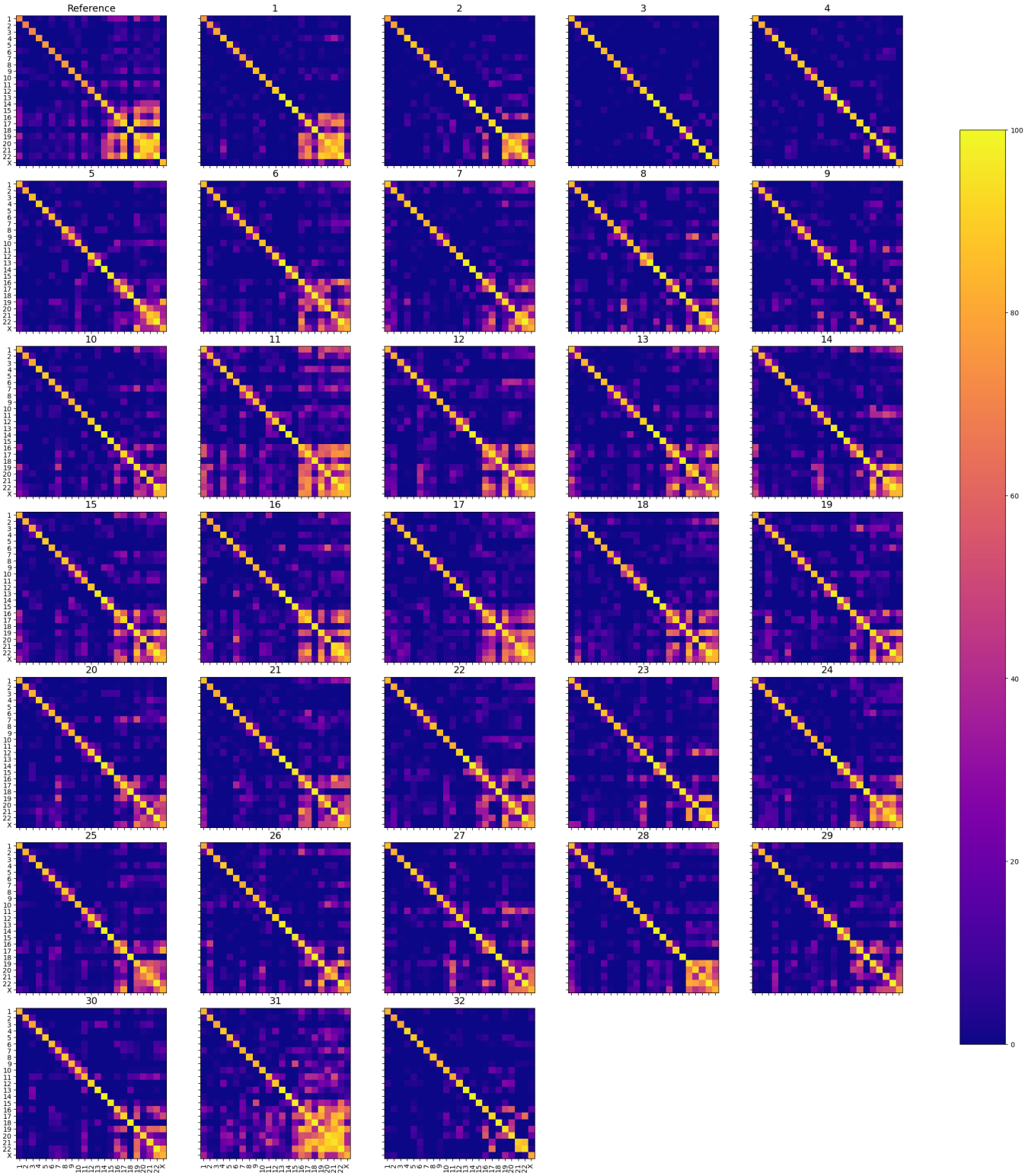

Figure 3.4: Contact scores matrices of the 32 structures created for the brain pericyte cell line with different parameters and the reference one, which is from the lymphoblastoid reconstruction. The contact score between chromosome i and chromosome j is the percentage of contact constraints between bins respectively from chromosome i and j satisfied by the reconstruction. Analogously to Figure 3.3, the heatmap is based on these values.

$d_c(S)$	S
228.36	1
261.19	17
266.68	2
266.80	25
282.12	24
282.33	28
283.86	22
285.06	13
287.77	30
287.80	15
290.12	31
290.20	21
293.12	16
298.30	7
298.57	5
299.14	18
301.78	10
304.71	26
305.28	12
305.61	6
307.25	27
307.37	14
309.92	32
311.55	29
311.81	20
320.98	23
327.62	8
328.20	19
328.38	4
332.10	9
332.41	3
340.86	11

$d_{nc}(S)$	S
176.37	1
206.39	2
211.95	31
213.06	25
222.09	17
224.16	24
225.16	28
227.46	26
228.60	22
234.92	13
235.90	30
236.73	15
236.86	6
238.92	5
239.06	32
239.10	12
239.49	16
240.62	18
243.36	7
243.81	20
244.12	29
245.57	27
246.86	21
248.54	10
252.76	23
253.15	4
254.80	3
254.95	9
258.34	14
259.48	8
260.82	19
272.53	11

Table 3.1: Rankings of the distances of the various structures in ascending order computed as defined in 2.2. $d_c(S)$ is defined as the euclidean distance of the contact scores of the structure S from the contact scores from the benchmark structure. $d_{nc}(S)$ is the analogous quantity for the non contact scores.

The parameters of structure 1 were subsequently used to generate all the structures of the various cell lines. The visualization of their contact and non contact score matrices in Figure 3.5 and Figure 3.6 provides insightful information on the three dimensional reconstructions.

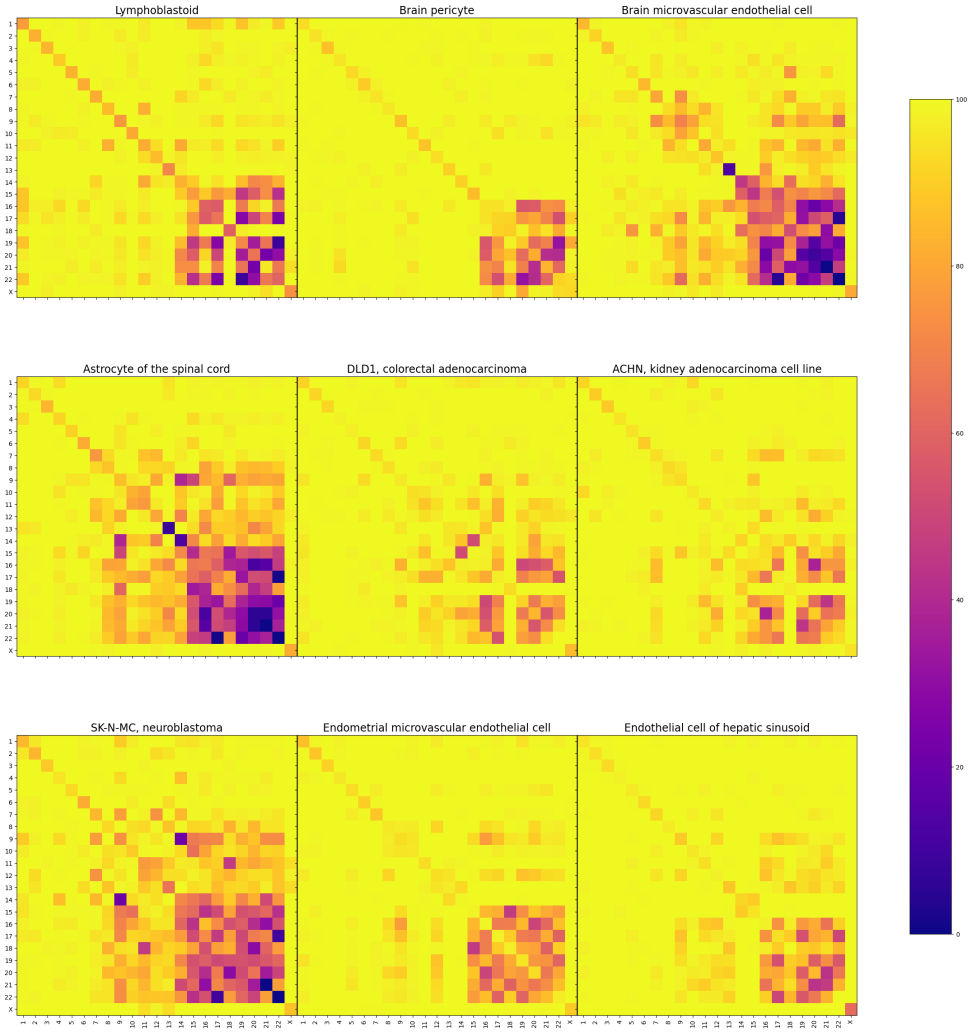

Figure 3.5: Non contact scores matrices of the structures created for the various cell lines.

In comparison to the lymphoblastoid cell, brain pericyte cell, colorectal and kidney adenocarcinoma cells, and the endothelial cells of hepatic sinusoid are the most similar, revealing a clear contrast between the lines for chromosome 18 and the small chromosomes, even if their non contact scores are generally higher than in the reference. The brain microvascular endothelial and astrocyte lines show lower non contact scores for the small chromosomes, including chromosome 18. Lastly, in the neuroblastoma and endometrial endothelial cells, it is almost impossible to see differences between chromosome 18 and the others, with non contact scores gradually increasing in proportion to

the reduction of the chromosome dimensions. These differences between the structures will be evident when visualizing the violin plots (figure 3.8). It is necessary to highlight the fact that the neuroblastoma, the astrocyte and the brain microvascular endothelial cells show similar patterns in both the contact and non contact scores. In particular, the contact scores regarding chromosome 9, which showed the noisiest data for these lines in paragraph 3.1, share an unusually high value.

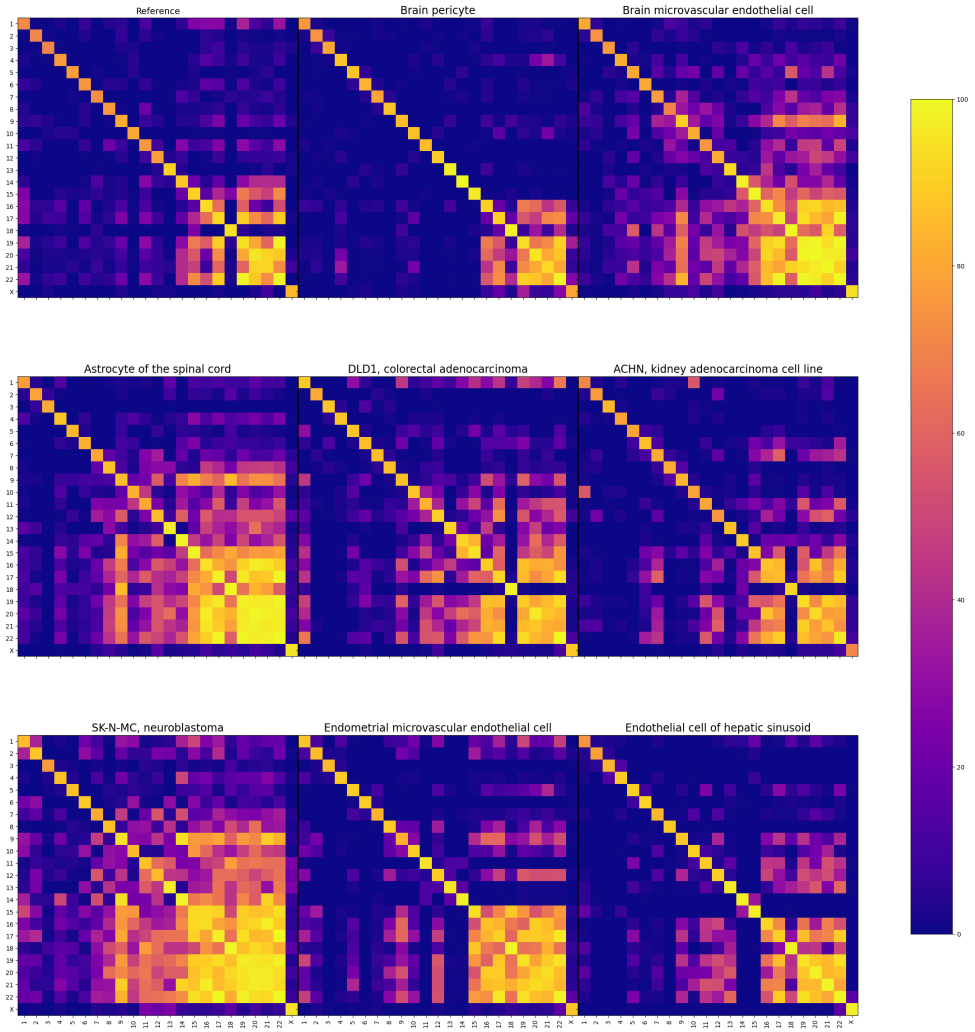

Figure 3.6: Contact scores matrices of the structures created for the various cell lines.

The expectations are confirmed by table 3.2, which represents the distances of the scores of the reconstruction from the benchmark structure in ascending order. As a matter of fact, the triad of cells previously mentioned is the "farthest" from the lymphoblastoid cell in both rankings, while the kidney adenocarcinoma cell and the brain pericyte are the closest.

$d_c(S)$	S
223.90	ACHN, kidney adenocarcinoma cell line
228.36	Brain pericyte
256.04	DLD1, colorectal adenocarcinoma
278.75	Endothelial cell of hepatic sinusoid
313.25	Endometrial microvascular endothelial cell
333.90	Brain microvascular endothelial cell
403.51	Astrocyte of the spinal cord
519.05	SK-N-MC, neuroblastoma

$d_{nc}(S)$	S
176.37	Brain pericyte
189.11	ACHN, kidney adenocarcinoma cell line
189.32	Endothelial cell of hepatic sinusoid
193.61	DLD1, colorectal adenocarcinoma
198.6	Endometrial microvascular endothelial cell
243.87	Brain microvascular endothelial cell
258.29	SK-N-MC, neuroblastoma
285.44	Astrocyte of the spinal cord

Table 3.2: Rankings of the distances of the structures of the various cell lines in ascending order computed as defined in 2.2. $d_c(S)$ is defined as the euclidean distance of the contact scores of the structure S from the contact scores from the benchmark structure. $d_{nc}(S)$ is the analogous quantity for the non contact scores.

The value of the distance between contact scores is higher than that between non contact scores; this is possible due to the high dynamical state in which chromosomes lie inside the cell nucleus. The variety of contacts that can be shown among different cell lines is broad, whereas the non contacts between the various chromosomes are subject to stricter rules to prevent excessive intermingling. Moreover, the two rankings, even if similar, are not identical: the kidney adenocarcinoma is the most similar to the lymphoblastoid cell in the contact score distance, while the brain pericyte is the most similar in the non contact score distance. This peculiarity is confirmed by the visualizations of the matrices in figures 3.5 and 3.6, showing that the contact score and non contact score matrices are not precisely complementary as we previously assumed. Effectively, the ACHN contact score map generally shows high levels of contacts even for chromosomes outside of the submatrix relative to the small chromosomes, just as in the lymphoblastoid cell. The

brain pericyte cell presents negligible contact scores between larger chromosomes in comparison to the lymphoblastoid and pericyte cells, but the non contact score map overlaps more accurately with the benchmark. Similar considerations are also valid for the couple composed of the colorectal adenocarcinoma cell and the endothelial cell of the hepatic sinusoid, and the couple formed by astrocyte and neuroblastoma cells, which mutually swap positions in the rankings.

3.3 3D structures

After the analysis of the contact and non contact scores, it is possible to visualize the structures generated for the various cell lines to observe their features.

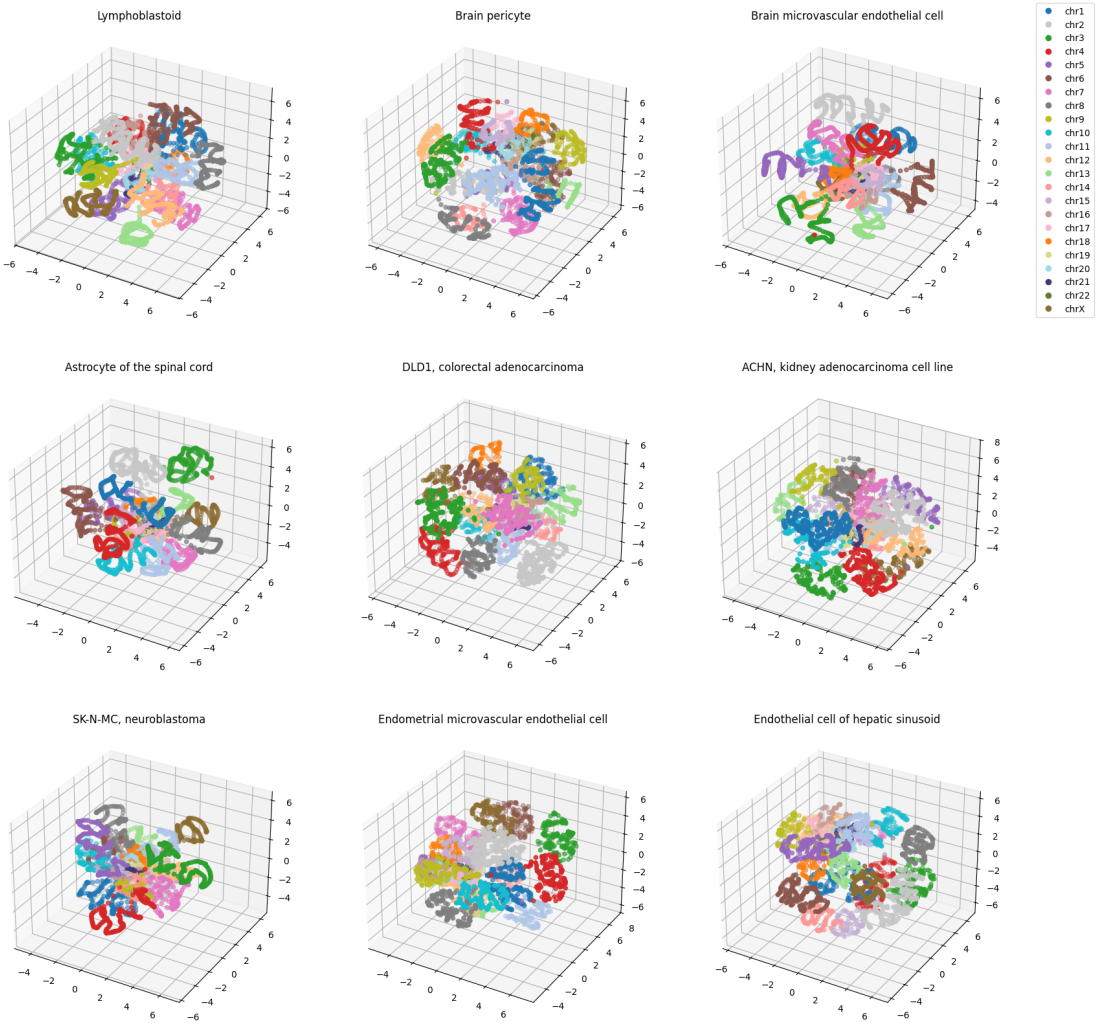

Figure 3.7: visualization of the various structures. Each chromosome is represented with a different color.

Observing the structures in Figure 3.7, it is evident that the single chromosomal regions are well defined: each chromosome occupies a distinct and limited volume, without intermingling, confirming the presence of chromosomal territories, as expected for a good 3D structure reconstruction.
At this point, let us delve deeper into a more precise analysis of the reconstruction. A more effective idea about the position of the various chromosomes is provided by the violin plots in Figure 3.8.

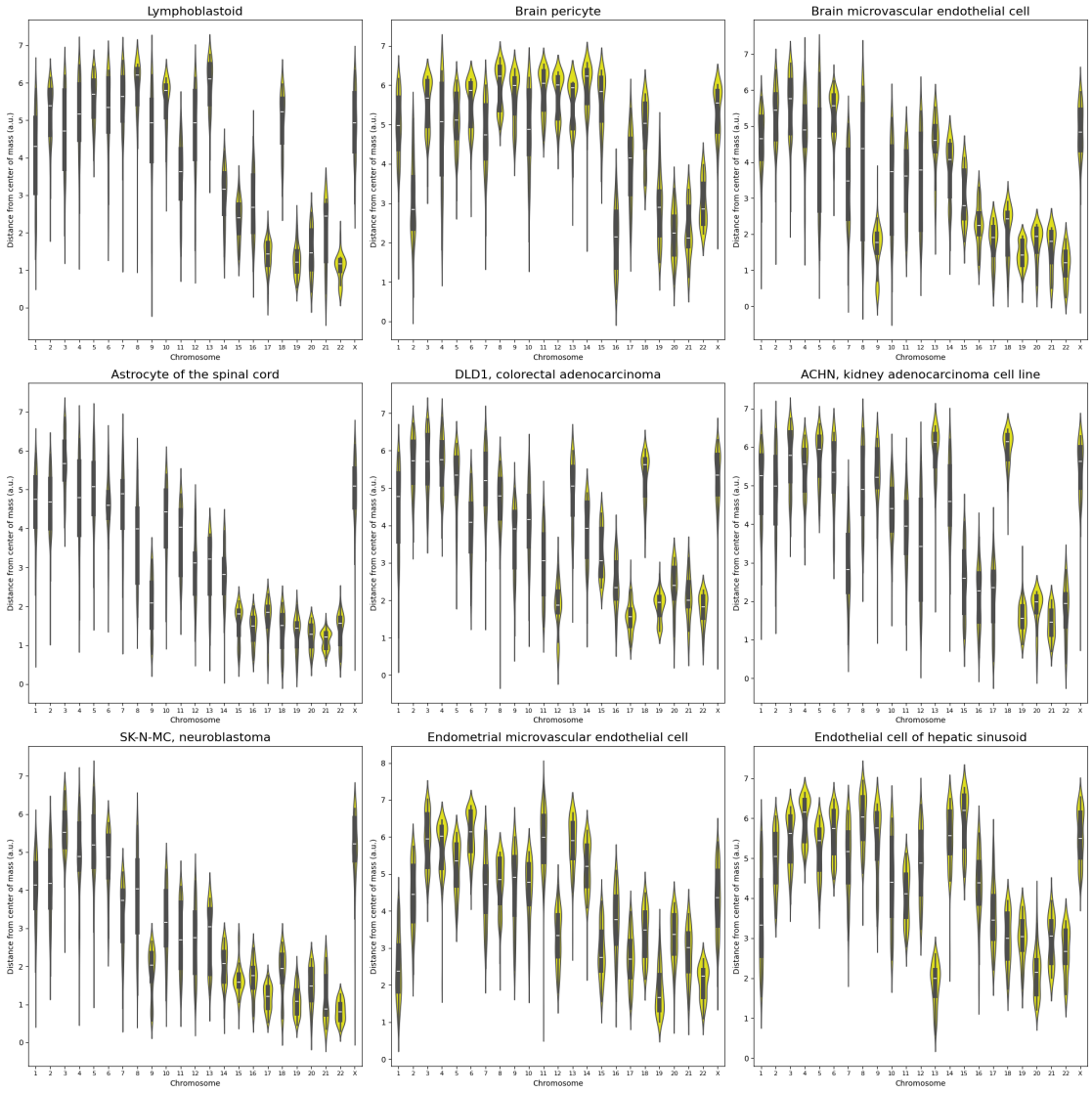

Figure 3.8: Violin plots of the structures representing the distribution of the distances of the megabases from the center for each chromosome.

As can be observed, not all the structures meet the requirement that chromosome 18 should be external, as suggested by the analysis of the contact and non contact score

matrices;

- the benchmark lymphoblastoid, the brain pericyte, the DLD1, and the AHCN cells, which were previously mentioned for their similarities in the contact and non contact scores matrix, follow the usual topology of chromosomal territories inside the nucleus, with some minor exceptions: small chromosomes (16-22, with the exception of chromosome 18) are clustered in the center of the structure, while large chromosomes occupy peripheral positions. It is noteworthy that chromosome 12 in the DLD1 line is distributed mainly in the central region compared to the other structures;

- the endothelial cell line of the hepatic sinusoid has a similar structure to the previous ones; however, chromosome 18 resides in the center of the nucleus, as evidenced by lower non contact score values for this chromosome in figure 3.5. Moreover, chromosome 13 is the closest to the center, a peculiar feature in comparison to its usual position (towards the periphery) and to the other structures visualized here.

- the brain microvascular endothelial cells, the astrocytes of the spinal cord, and the neuroblastoma cell lines have a rather compact structure with some peculiarities; in particular, chromosomes 9 and 18 are mainly distributed in the central region. The central position of chromosome 9 could be an effect of the noisy lines that we showed in paragraph 3.1, as it was already highlighted in paragraph 3.2 for their contact and non contact scores matrices.

- the endometrial microvascular endothelial cell line is singular in comparison to the others; there is not a substantial difference in the positioning of large and small chromosomes. Chromosome 18 lies close to the center. In addition, chromosome 1, which is the largest chromosome, is the most central chromosome here.

In all the plots, it is possible to observe that large chromosomes extend across a large portion of the nuclear radius, with tails of megabases that elongate and reach near the center. This is a particular interaction of elongated regions of chromosomes already observed in the results of the testing of MOGEN[26].

3.4 Comparison between hg18 and hg19

The comparison between the two versions of the reference genome, hg19 and hg18, is fundamental in determining the correctness of the comparison in nucleotide content between the lymphoblastoid line (where hg18 was used) and the others (where hg19 was used). The comparison was performed using the method described in paragraph 2.5 for both AT and GC contents.

Chromosome	GC fraction hg18	GC fraction hg19	GC hg18 (*preprocessing*)	GC hg19 (*preprocessing*)
1	0.3799	0.3773	0.41 ± 0.06	0.41 ± 0.06
2	0.3937	0.3942	0.40 ± 0.05	0.40 ± 0.06
3	0.3874	0.3905	0.39 ± 0.05	0.39 ± 0.06
4	0.3742	0.3755	0.38 ± 0.05	0.38 ± 0.06
5	0.3883	0.3881	0.39 ± 0.05	0.39 ± 0.06
6	0.3876	0.3875	0.39 ± 0.05	0.39 ± 0.06
7	0.3975	0.3978	0.40 ± 0.06	0.40 ± 0.07
8	0.3916	0.3922	0.40 ± 0.06	0.39 ± 0.06
9	0.3539	0.3515	0.40 ± 0.06	0.40 ± 0.08
10	0.4043	0.4029	0.41 ± 0.06	0.41 ± 0.07
11	0.4054	0.4037	0.41 ± 0.06	0.41 ± 0.07
12	0.4017	0.3978	0.40 ± 0.05	0.40 ± 0.07
13	0.3225	0.3198	0.38 ± 0.05	0.38 ± 0.06
14	0.3394	0.3363	0.41 ± 0.05	0.40 ± 0.07
15	0.3421	0.3362	0.42 ± 0.04	0.41 ± 0.06
16	0.3978	0.3910	0.44 ± 0.08	0.43 ± 0.08
17	0.4497	0.4363	0.45 ± 0.05	0.44 ± 0.08
18	0.3902	0.3804	0.39 ± 0.06	0.38 ± 0.08
19	0.4228	0.4564	0.47 ± 0.07	0.5 ± 0.1
20	0.4205	0.4166	0.43 ± 0.08	0.42 ± 0.09
21	0.2975	0.2978	0.39 ± 0.09	0.4 ± 0.1
22	0.3365	0.3264	0.47 ± 0.06	0.46 ± 0.09
X	0.3851	0.3844	0.39 ± 0.04	0.39 ± 0.05

Table 3.3: Comparison between hg18 and hg19 versions of the human genome in terms of the GC content fraction in each chromosome. In particular the second and third columns show the fractions computed directly from the sequences: hence they are pure numbers reported to the fourth decimal in the table. The fourth and fifth columns instead are composed by GC fractions of each chromosome computed averaging over the megabases that resulted significant after the preprocessing of the brain pericyte line, and in order to perform a comparison between them, they were associated with the standard deviation (for the same reasons explained in 2.3) of the GC fraction over the various megabases belonging to the same chromosome.

Chromosome	AT fraction hg18	AT fraction hg19	AT hg18 (*preprocessing*)	AT hg19 (*preprocessing*)
1	0.5301	0.5265	0.57 ± 0.07	0.57 ± 0.08
2	0.5848	0.5853	0.59 ± 0.07	0.59 ± 0.07
3	0.5886	0.5932	0.60 ± 0.06	0.60 ± 0.07
4	0.6050	0.6062	0.61 ± 0.06	0.61 ± 0.08
5	0.5943	0.5941	0.60 ± 0.05	0.60 ± 0.07
6	0.5912	0.5908	0.60 ± 0.06	0.59 ± 0.08
7	0.5782	0.5784	0.59 ± 0.07	0.58 ± 0.08
8	0.5834	0.5840	0.59 ± 0.07	0.59 ± 0.08
9	0.5026	0.4993	0.57 ± 0.08	0.6 ± 0.1
10	0.5680	0.5660	0.57 ± 0.07	0.57 ± 0.08
11	0.5699	0.5676	0.58 ± 0.06	0.57 ± 0.09
12	0.5828	0.5770	0.59 ± 0.05	0.58 ± 0.09
13	0.5147	0.5102	0.60 ± 0.08	0.60 ± 0.08
14	0.4907	0.4862	0.59 ± 0.05	0.58 ± 0.08
15	0.4685	0.4605	0.57 ± 0.04	0.57 ± 0.08
16	0.4903	0.4820	0.54 ± 0.09	0.5 ± 0.1
17	0.5379	0.5218	0.54 ± 0.06	0.53 ± 0.09
18	0.5906	0.5758	0.59 ± 0.08	0.6 ± 0.1
19	0.4514	0.4874	0.51 ± 0.06	0.5 ± 0.1
20	0.5325	0.5275	0.54 ± 0.09	0.5 ± 0.1
21	0.4303	0.4316	0.6 ± 0.1	0.5 ± 0.2
22	0.3648	0.3538	0.51 ± 0.05	0.5 ± 0.1
X	0.5900	0.5888	0.60 ± 0.06	0.59 ± 0.08

Table 3.4: Comparison between hg18 and hg19 versions of the human genome in terms of the AT content fraction in each chromosome, analogous to table 3.3.

It is possible to observe from tables 3.3 and 3.4 that the pure fractions of GC content and AT content are not identical for the two versions. However, such discrepancies are negligible and do not justify the possible differences observed during the computation of the gradient. In support of this, the GC fraction computed for each chromosome after preprocessing is always consistent within the uncertainty, even if the errors are quite large. Such slight differences cannot explain the eventual incoherence of the computed gradients with the benchmark lymphoblastoid line. Note that the compatibility of the two versions in GC content and AT content implies compatibility in N content because of the constraint $GC + AT + N = 1$ for every chromosome.

3.5 Radial GC, AT, N content trends

Finally, it is necessary to discuss the presence of the GC gradient among the various cell types. A first glance at the 3D distribution of GC content is provided by Figure 3.9.

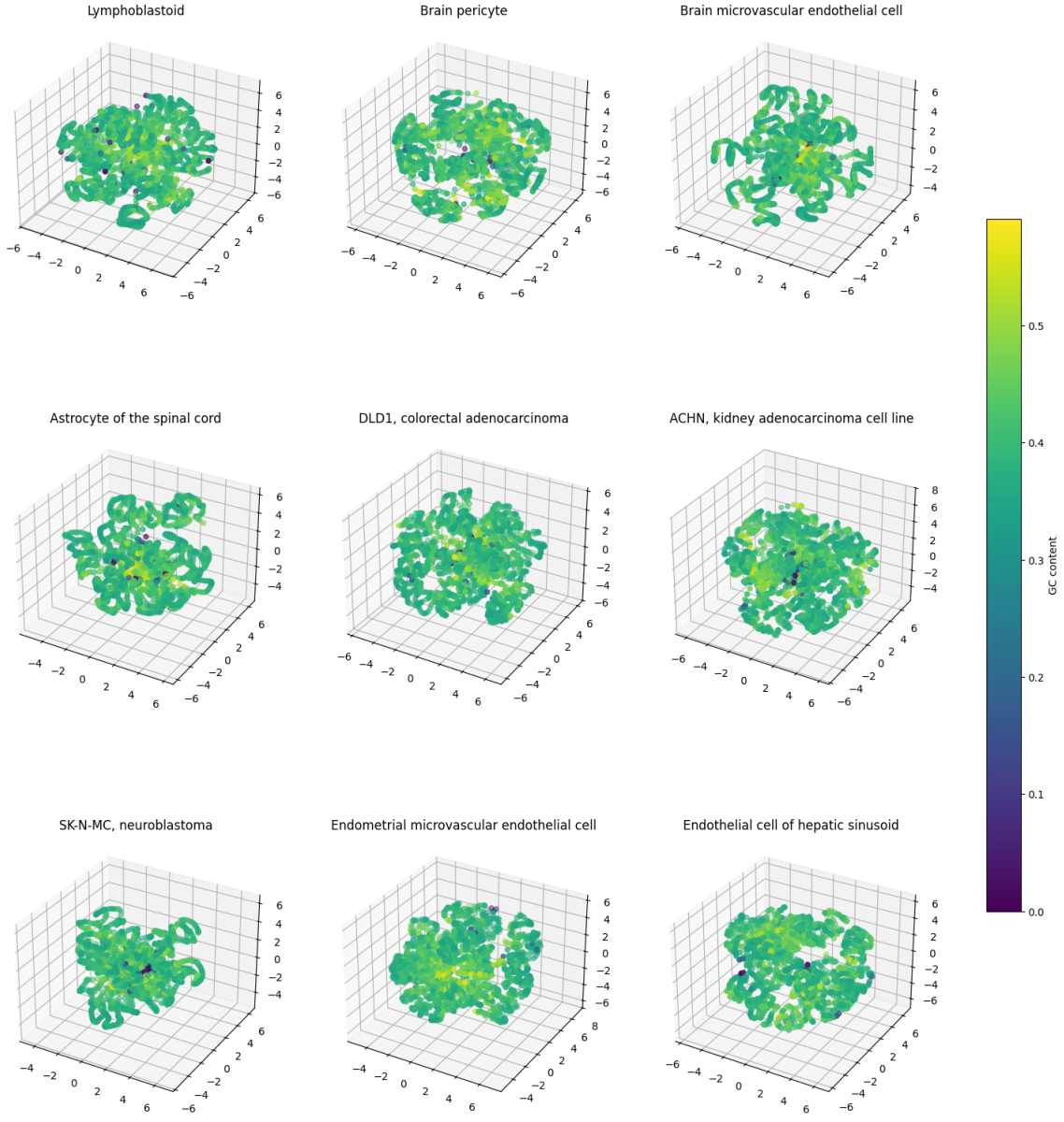

Figure 3.9: visualization of the various structures with a color map based on the GC content of each megabase.

The lymphoblastoid line shows an increasing radial arrangement of the GC content, confirming the validity of the method. Apart from the lymphoblastoid cell line, where the gradient has already been measured, there are structures where the content seems to in-

crease towards the center and others where it seems randomly positioned. In particular, the brain microvascular endothelial cell, the astrocyte of the spinal cord, the neuroblastoma, the colorectal adenocarcinoma, and the endometrial endothelial cell appear to show a radial disposition, while in other cell lines, the organization is more uncertain. However, it is possible to observe that, excluding the lymphoblastoid cell line, the majority of the structures show an accumulation of extremely low GC values in the center. These megabases correspond to sequences primarily composed of N's. These *loci* , which are difficult to sequence, are constrained by the reconstruction of MOGEN in the center of the nucleus.

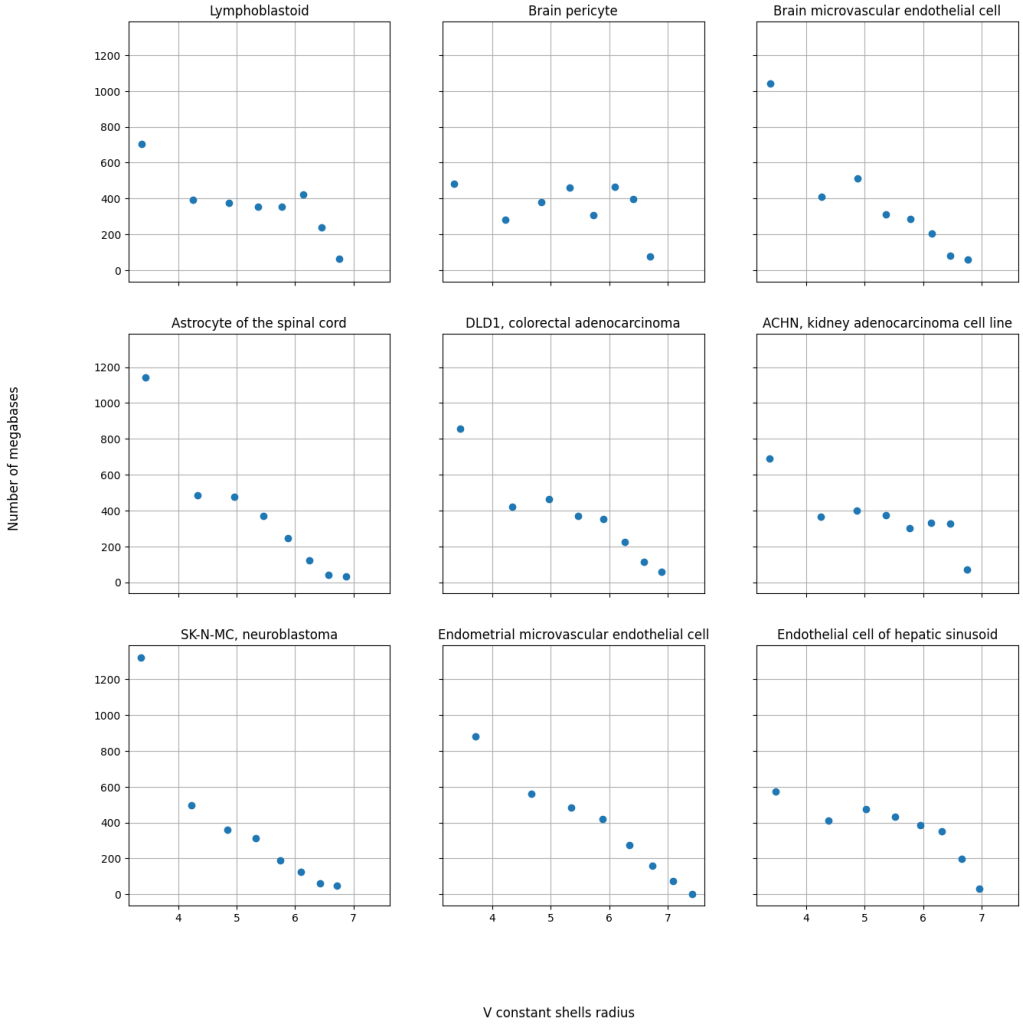

Figure 3.10: Plots representing the number of megabases inside the 8 V constant shells in function of the outer radius for the various structures: the distance between the various points is the thickness of the shells.

The gradients are computed by dividing the structure into different shells; in particular,

the space was divided into 8 shells for the computation of the gradient, keeping the volume of the shells constant, while it was divided into 13 shells for the computation with the number of megabases inside each shell fixed. A higher number of V-constant shells would lead to an insufficient number of entries inside each one, i.e., less than 20 megabases. 13 N-constant shells are instead extremely well populated, and adding other shells would not lead to a more precise mapping of the gradient: the first shell tends to always occupy the same space, while the other shells accumulate towards the end of the radius due to the fact that the volume of the shells increases.

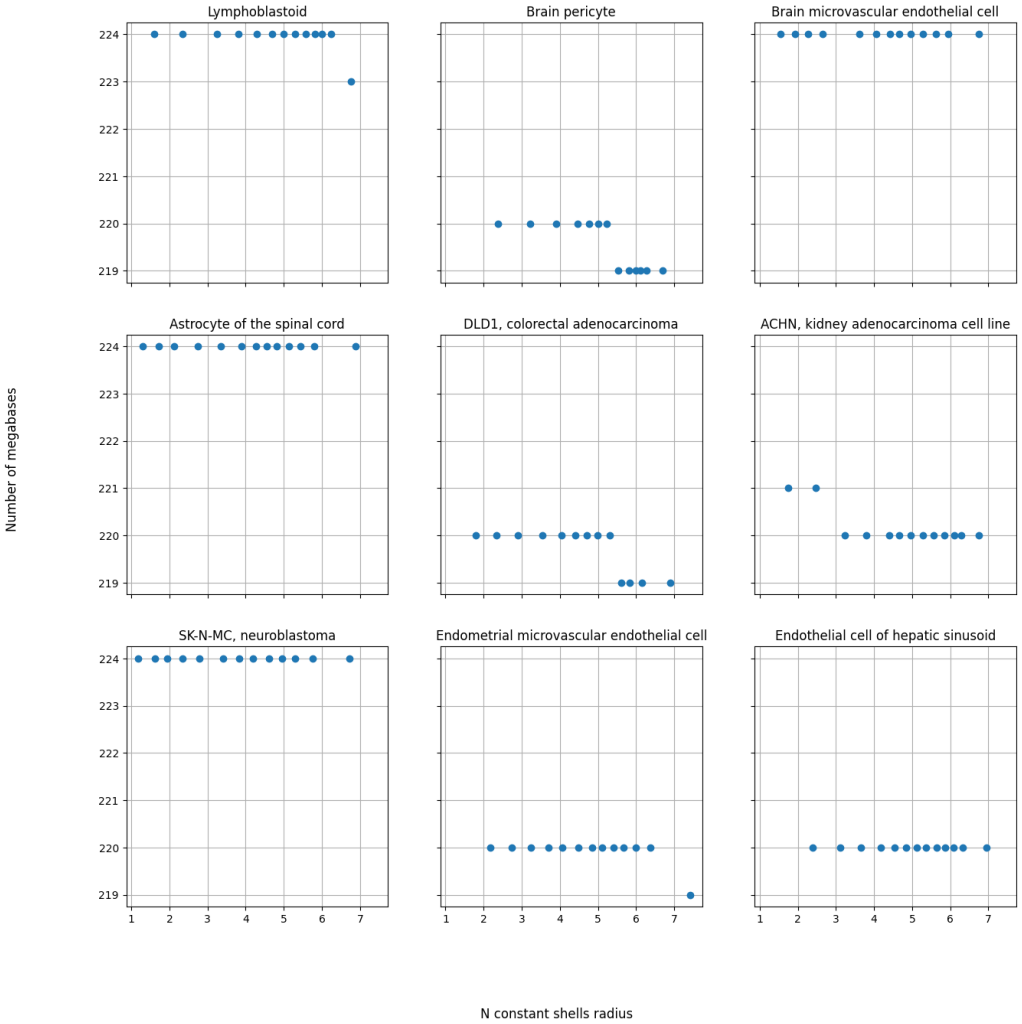

Figure 3.11: Plots representing the number of megabases inside the 13 N constant shells in function of the outer radius for the various structures: the distance between the various points is the thickness of the shells.

Observing the plots of the megabases inside the 8 shells at constant volume in figure 3.10, it is noticeable that the disposition of the megabases in the shells is variable across

the different cell types but has also consistent patterns. The first shell is the thickest and most populated: this occurs because the density of the megabases in space increases when approaching the center of mass, at least until a certain point. On the other hand, the last shell is always the emptiest: in order to keep the volume of the shell constant, it is necessary to reduce the thickness. There are cell types where the population decreases as one gets farther from the center; for example, in the endometrial microvascular endothelial cell; on the other hand, some cells have different peaks of density, such as the ACHN cell line. However, only the brain pericyte line has a roughly uniform distribution of the number of megabases in the various shells, at least until the last one, while other cell lines experience a significant jump between the number of megabases in the first shell and the rest. It is possible to observe that the astrocyte of the spinal cord line, together with the neuroblastoma cell line, is characterized by a compact structure, as already mentioned in paragraph 3.4: the majority of the megabases reside in the first shell (more than a thousand). Another important fact is that the last shell of the endometrial microvascular endothelial cell is extremely poorly populated (3 megabases), as it is practically overlapping with 0 in the graph. This shell will be excluded from the execution of the linear fit, as it does not contain enough megabases for a correct statistical analysis. Regarding the shells with a constant number of megabases, each contains roughly 220 Mb. It is interesting to observe that the thickness is usually lower for the shells that are in the middle, and for most of the cell types, the shells accumulate towards the periphery before the last shell (figure 3.11). A reason behind this is that, as was briefly mentioned above, a higher average radius implies a higher volume for the shells at a given thickness; therefore, in the latest shells, less thickness is required to reach the fixed number of megabases contained. Two exceptions are the neuroblastoma and the astrocyte: as mentioned before, these are compact structures that compensate for the higher volume of the latest shell with a higher density closer to the center of mass of the structure.

Having verified the composition in number of megabases of the various shells, it is now time for the analysis of the graphs and the results of the various fits.

Observing the graphs in figures 3.12, 3.13, it is immediately evident that the points closer to the center of the nucleus are characterized by higher errors. This indicates that, towards the center of the nucleus, there is a non-negligible amount of variation in values among the three types of base contents considered. It is also noticeable that the only values of the N fraction that visibly deviate from zero are in the first shells; this is compatible with the considerations above regarding the confinement of the megabases with extremely low values of GC in the center of the structures.

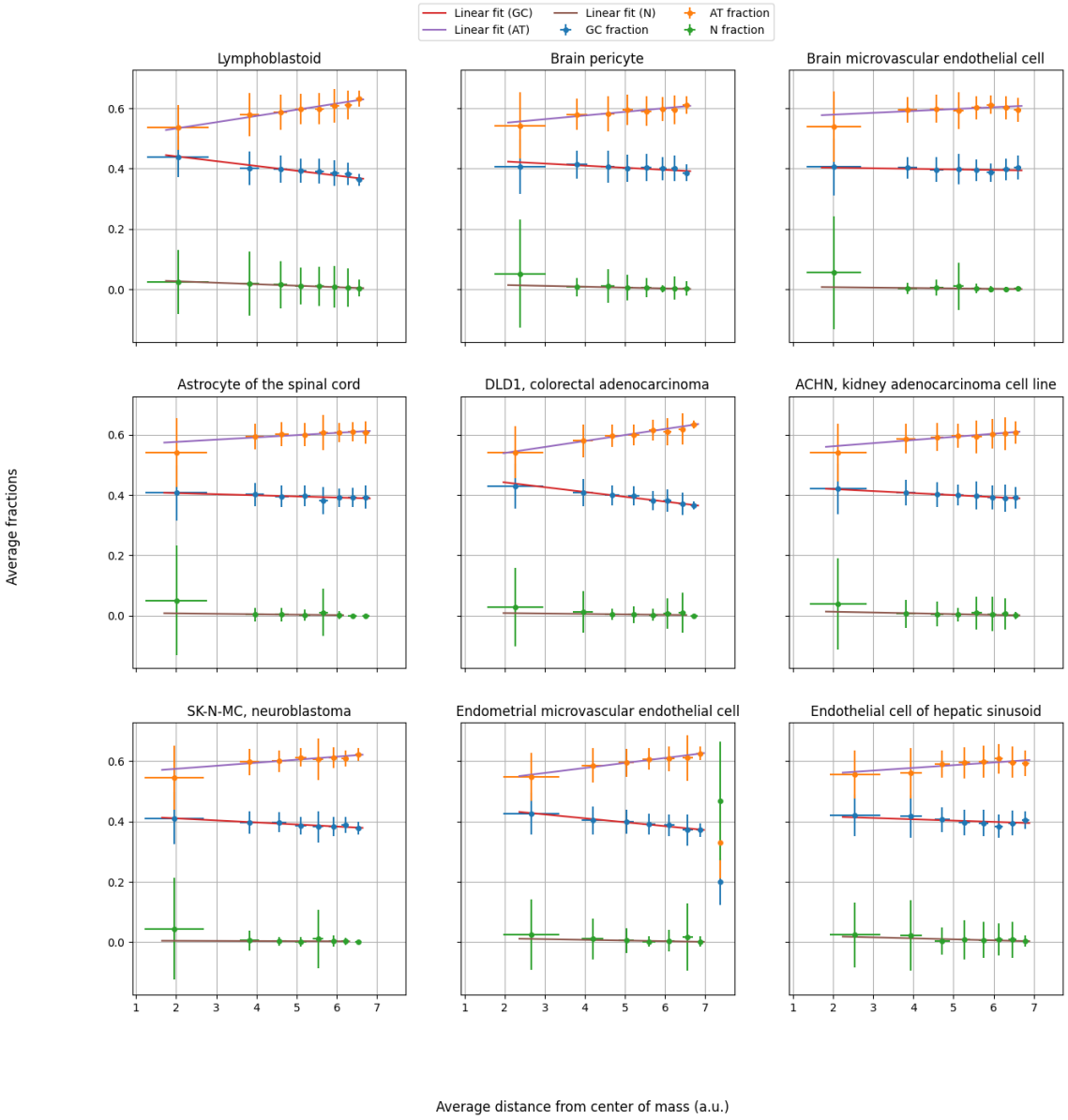

Figure 3.12: Visualization of the average values of GC content, AT content, N content as a function of the average distance from the center of mass for all of the constant volume shells for each structure, along with their regression line $y = a + bx$ obtained via ODR.

The center of the structures, therefore, appears to be populated by values with large variability and by *loci* that are difficult to sequence: the large errors of the first shells represent these features. After the first shells, the errors tend to stabilize at a lower value, except for neuroblastoma, astrocyte, and brain endothelial cells, which are three structures with common peculiarities, as mentioned in the previous paragraphs. These cell types show an increase in errors and, therefore, in the variability of nucleotide content when reaching distances from the center of the nucleus between 5 and 6 a.u. for the first

two and close to 5 for the last one.

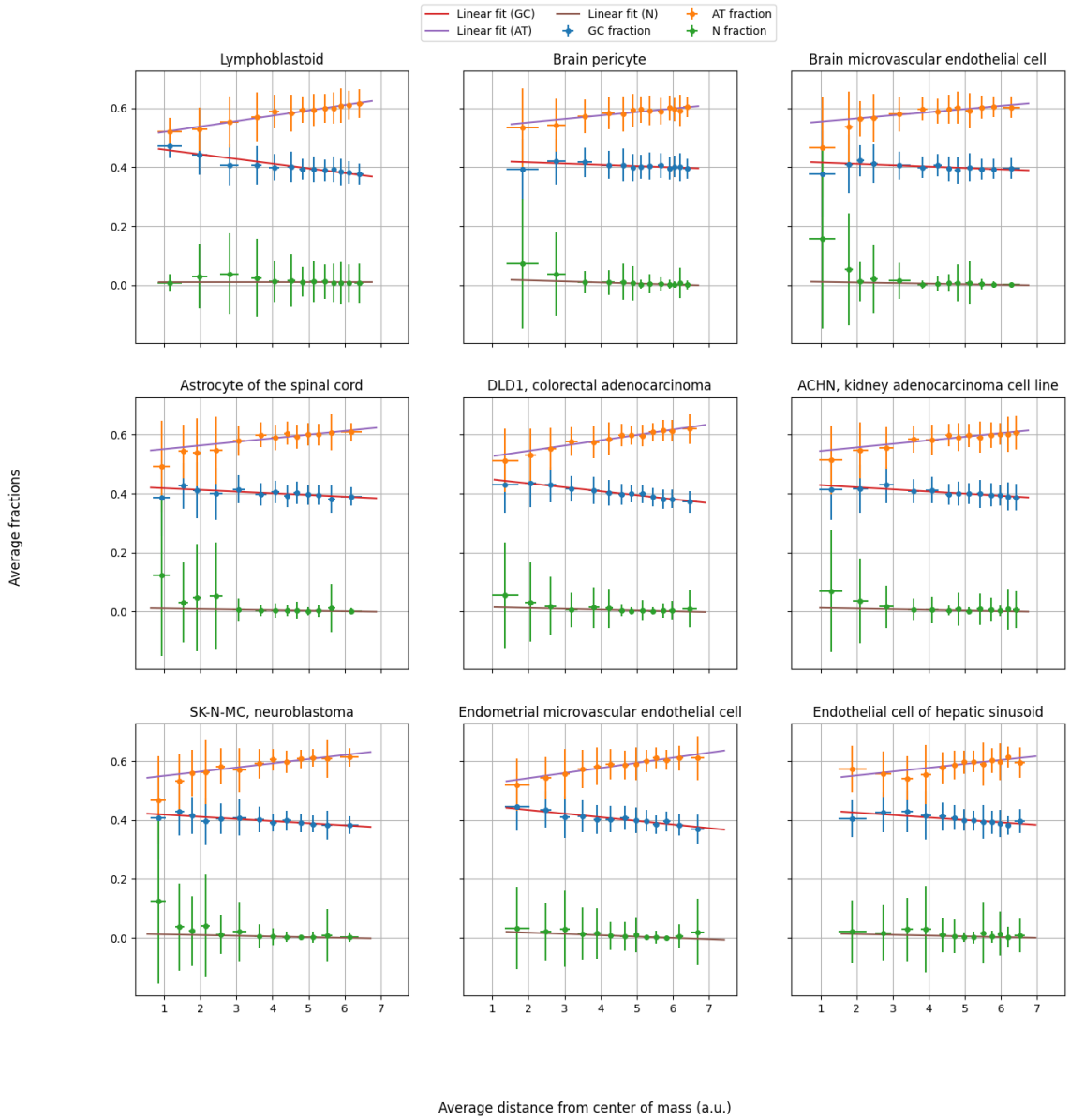

Figure 3.13: Visualization of the average values of GC content, AT content, N content as a function of the average distance from the center of mass for all of the N-constant shells for each structure, along with their regression line $y = a + bx$ obtained via ODR.

Coming to the core of the work, the results presented in figures 3.12 and 3.13 show that the GC and AT trends vary across cellular types. All the structures present a decrease in GC content and N content and an increase in AT content proportional to the distance from the center of the structure. However, the intensity of the GC gradient differs between cell types: there are graphs that show a roughly constant behavior

of the GC fraction, including structures that, from the first visualization, appeared to be radially organized (such as the astrocyte cell or the brain endothelial cell), while in other reconstructions, the GC fraction gradient appears to be more emphasized (as in the endometrial endothelial cell line and in the colorectal adenocarcinoma line, which are the only two structures visually matching the gradient of the lymphoblastoid cell). The decrease in GC content needs to be counteracted by an increase in AT content and vise versa; especially in the N constant graphs, there are cell lines where the AT content has a gradient while the GC content is not decreasing at the same speed. Assuming the presence of the gradient is insufficient if only one of the two base fractions shows a change along the distance values. As mentioned above, the last shell at constant volume of the endometrial endothelial cell is poorly populated, and as can be observed from the graph, the corresponding points are totally incompatible with the rest of the dataset. These points are not considered in the execution of the fits. Moreover, it is possible to observe that the values of the N content in the last V-constant shell of the neuroblastoma and colorectal adenocarcinoma cells, as well as in the last two V-constant shells of the astrocyte cell, are exactly 0: the megabases in these shells do not contain any N in the sequence; all the unknown nucleotides, as mentioned before, are clustered in the central regions. It is therefore impossible to compute an error for the N content of these shells through the standard deviation, which results in 0. This leads to the fact that in the ODR the weights associated with these values diverge. As a consequence, these points were neglected in the regression of the N content for these specific cells.
The results of all the fits are presented in table 3.5.

Lymphoblastoid

	V const	N const
b(GC)	-0.016 ± 0.002	-0.016 ± 0.002
a(GC)	0.47 ± 0.01	0.475 ± 0.008
Q(GC)	1.00	1.00
b(AT)	0.021 ± 0.002	0.018 ± 0.001
a(AT)	0.492 ± 0.013	0.502 ± 0.004
Q(AT)	1.00	1.00
b(N)	-0.0049 ± 0.0002	0.000 ± 0.001
a(N)	0.036 ± 0.001	0.010 ± 0.004
Q(N)	1.00	1.00

Brain pericyte

	V const	N const
b(GC)	-0.007 ± 0.002	-0.004 ± 0.001
a(GC)	0.44 ± 0.01	0.425 ± 0.008
Q(GC)	1.00	1.00
b(AT)	0.012 ± 0.002	0.012 ± 0.002
a(AT)	0.53 ± 0.01	0.528 ± 0.009
Q(AT)	1.00	1.00
b(N)	-0.003 ± 0.001	-0.004 ± 0.001
a(N)	0.020 ± 0.008	0.023 ± 0.006
Q(N)	1.00	1.00

Brain microvascular endothelial cell

	V const	N const
b(GC)	-0.002 ± 0.002	-0.005 ± 0.001
a(GC)	0.41 ± 0.01	0.420 ± 0.007
Q(GC)	1.00	1.00
b(AT)	0.006 ± 0.004	0.011 ± 0.002
a(AT)	0.57 ± 0.02	0.54 ± 0.01
Q(AT)	1.00	1.00
b(N)	-0.001 ± 0.001	-0.002 ± 0.001
a(N)	0.010 ± 0.005	0.013 ± 0.006
Q(N)	1.00	1.00

Astrocyte of the spinal cord

	V const	N const
b(GC)	-0.004 ± 0.002	-0.006 ± 0.002
a(GC)	0.414 ± 0.009	0.424 ± 0.008
Q(GC)	1.00	1.00
b(AT)	0.007 ± 0.002	0.012 ± 0.003
a(AT)	0.56 ± 0.01	0.54 ± 0.01
Q(AT)	1.00	1.00
b(N)	-0.001 ± 0.001	-0.002 ± 0.001
a(N)	0.011 ± 0.008	0.013 ± 0.006
Q(N)	0.42	1.00

DLD1, colorectal adenocarcinoma

	V const	N const
b(GC)	-0.016 ± 0.001	-0.013 ± 0.001
a(GC)	0.474 ± 0.007	0.462 ± 0.005
Q(GC)	1.00	1.00
b(AT)	0.020 ± 0.002	0.018 ± 0.002
a(AT)	0.50 ± 0.01	0.509 ± 0.008
Q(AT)	1.00	1.00
b(N)	-0.001 ± 0.002	-0.003 ± 0.001
a(N)	0.012 ± 0.011	0.018 ± 0.008
Q(N)	1.00	1.00

ACHN, kidney adenocarcinoma

	V const	N const
b(GC)	-0.007 ± 0.001	-0.007 ± 0.001
a(GC)	0.433 ± 0.003	0.436 ± 0.006
Q(GC)	1.00	1.00
b(AT)	0.010 ± 0.002	0.012 ± 0.002
a(AT)	0.54 ± 0.01	0.53 ± 0.01
Q(AT)	1.00	1.00
b(N)	-0.003 ± 0.001	-0.002 ± 0.001
a(N)	0.018 ± 0.007	0.015 ± 0.008
Q(N)	1.00	1.00

SK-N-MC, neuroblastoma

	V const	N const
b(GC)	-0.007 ± 0.001	-0.007 ± 0.001
a(GC)	0.424 ± 0.008	0.426 ± 0.005
Q(GC)	1.00	1.00
b(AT)	0.010 ± 0.003	0.014 ± 0.003
a(AT)	0.55 ± 0.02	0.54 ± 0.01
Q(AT)	1.00	1.00
b(N)	0.000 ± 0.001	-0.002 ± 0.002
a(N)	0.005 ± 0.007	0.014 ± 0.008
Q(N)	1.00	1.00

Endometrial microvascular endothelial cell

	V const	N const
b(GC)	-0.013 ± 0.001	-0.012 ± 0.002
a(GC)	0.462 ± 0.007	0.459 ± 0.008
Q(GC)	1.00	1.00
b(AT)	0.017 ± 0.001	0.017 ± 0.002
a(AT)	0.511 ± 0.009	0.508 ± 0.009
Q(AT)	1.00	1.00
b(N)	-0.002 ± 0.001	-0.005 ± 0.001
a(N)	0.017 ± 0.009	0.027 ± 0.008
Q(N)	1.00	1.00

Endothelial cell of hepatic sinusoid

	V const	N const
b(GC)	-0.004 ± 0.003	-0.008 ± 0.002
a(GC)	0.42 ± 0.02	0.44 ± 0.01
Q(GC)	1.00	1.00
b(AT)	0.009 ± 0.003	0.013 ± 0.003
a(AT)	0.54 ± 0.02	0.53 ± 0.02
Q(AT)	1.00	1.00
b(N)	-0.003 ± 0.001	-0.003 ± 0.002
a(N)	0.026 ± 0.008	0.02 ± 0.01
Q(N)	1.00	1.00

Table 3.5: Parameter estimates obtained via ODR. The regression model is $y = a + bx$, where a and b denote the intercept and slope. Q is the p-value introduced in paragraph 2.4.

It is possible to observe from these values that the fit between the graphs with shells at constant V and the graphs with shells at constant N yields mostly compatible results; or, at least, the error intervals of the parameters are adjacent. This is fundamental because it implies that the results are not dependent solely on a particular set of shells that could be biased towards the results; rather, they are consistent across different sets. There are only three cases of incompatible parameters, considering the error interval, from a fit for the same cell line and different types of shells: two concern the regression of the N content, specifically in the lymphoblastoid cell and the endometrial microvascular endothelial cell. These are cases where a different set of shells changes the outcome of the fit because the N constant shells sample the internal regions with more precision than the V constant shells. The other case of incompatibility is the b parameter of the

fit for the colorectal adenocarcinoma DLD1, but the error intervals are, anyway, close. Another important observation is that all the p-values of the fits (Q) equal 1, meaning that the data are almost perfectly aligned with the model or that the uncertainties have been overestimated. The fits are obviously not perfect, but due to the decision to use a large error (the standard deviation, paragraph 2.3), the lines created pass through all the points, considering the error bars. So the p-values resulting 1 are a consequence of the overestimation of the errors.
Finally, the results of the linear fits for the different cell types can be compared based on the values reported in table 3.6, to determine whether any of the analyzed samples are compatible with the benchmark.

Cellular lines	a(GC, V const)	b(GC, V const)	a(GC, N const)	b(GC, N const)
Lymphoblastoid	0.47 ± 0.01	-0.016 ± 0.002	0.475 ± 0.008	-0.016 ± 0.002
Pericyte	0.44 ± 0.01	-0.007 ± 0.002	0.425 ± 0.008	-0.004 ± 0.001
Brain endothelial	0.41 ± 0.01	-0.002 ± 0.002	0.420 ± 0.007	-0.005 ± 0.001
Astrocyte	0.414 ± 0.009	-0.004 ± 0.002	0.424 ± 0.008	-0.006 ± 0.002
DLD1	0.474 ± 0.007	-0.016 ± 0.001	0.462 ± 0.005	-0.013 ± 0.001
ACHN	0.433 ± 0.003	-0.007 ± 0.001	0.436 ± 0.006	-0.007 ± 0.001
SK-N-MC	0.424 ± 0.008	-0.007 ± 0.001	0.426 ± 0.005	-0.007 ± 0.001
Endometrial endothelial	0.462 ± 0.007	-0.013 ± 0.001	0.459 ± 0.008	-0.012 ± 0.002
Endothelial sinusoid	0.42 ± 0.02	-0.004 ± 0.003	0.44 ± 0.01	-0.008 ± 0.002

Table 3.6: Values of the parameters a (intercept) and b (slope) for the linear fit of GC content across various cell types.

The values of the b parameters for the various cell lines are close to zero. This is due to the small difference between the GC content in the first and last shells. However, there are cell lines that show a well-defined, softly decreasing trend in the GC content, including the lymphoblastoid cell line that we used as benchmark. For the other cell lines:

- Only the colorectal adenocarcinoma DLD1 and the endometrial microvascular endothelial cells show values of the GC gradient that are compatible with the lymphoblastoid line;
- The brain pericyte line, the kidney adenocarcinoma line ACHN, and the neuroblastoma cell type SK-N-MC show a smaller gradient of GC content;
- The other cell lines have gradient values that are very small, compatible with zero.

4. Conclusions

The computational method used in this work can capture the main features of the 3D genome organization and identify the presence of the GC gradient in the benchmark lymphoblastoid line GM06990, where a decreasing trend in the GC fraction along the nuclear radius has also been experimentally verified. In order to achieve this, an analysis was also conducted on the parameters of MOGEN to build a reliable reconstruction of the genome disposition, concluding that the set already implemented in the repository is the most trustworthy among the 32 different sets used as trials. In addition, a comparison between the hg18 and hg19 versions of the human reference genome was executed in order to compare the results of this benchmark cell type with others, showing that the two versions have negligible differences in nucleotide content across the various chromosomes.

The analysis of the 3D structure and the radial organization for different cell types showed that the arrangement of the intranuclear genome and its nucleotide content is not univocal. The chromosomal territories were well-defined in every structure; however, in some cell types, the typical characteristics of chromosome positioning are not evident, while other structures generated adhere to the usual features of the radial chromosomal arrangement based on the dimensions (small chromosomes clustered in the center and larger chromosomes lying in the periphery). This could show once again that the environment of the intranuclear genome is, in general, extremely dynamical. Also, the majority of the cell types analyzed do not show a specific arrangement of GC content, with a radial trend that is consistent with constant behavior. Only in two of the eight cases considered did the analysis provide a substantial decrease in the GC fraction, compatible with the observations in the lymphoblastoid line. It is thus challenging to deduce whether a particular radial organization has functional meaning, and the reasons why the gradient is only observed in specific cell types remain unknown. As already evidenced in the introduction, to clarify the presence and functional properties of the radial arrangement of the human genome, and to assess whether a particular arrangement can be observed in the cells of other animal species, it will be necessary to execute a large number of experiments on different cell types. Hopefully, the computational analysis performed hereby will be useful for filtering the cases worth further investigation.

However, the implemented analysis greatly relies on the quality of the Hi-C data for the specific line and the capacity of MOGEN to manage noisy contact values; therefore,

possible discrepancies in the results among different cell types could be caused by flaws in the processing of the data for the present reconstruction. To determine with certainty whether the analysis produced realistic results in these cases, it would be necessary to execute an experiment capable of measuring the radial organization of the genome, such as GPSeq. Future studies are therefore needed to assess the compatibility between the developed method and real experiments.

Appendix

We refer to the values of the supplementary material through the abbreviation s.m. and to the values of the article by MOGEN by citing it. Also, the values referred to as standard are the ones preset in the software. The values used to set the parameters of the scoring function for the generation of the 32 structures were:

- Number of chromosomes: 23
- Intra-chromosomal threshold: 0.65 (s.m.) used for all the structures apart from 1, 2, 31, and 32, which are generated using 80%(standard) and structures 26, 27 generated with 0.7.
- Inter-chromosomal threshold: 0.58 (s.m.) used for all the structures apart 1, 2, 31, 32 that are generated using 18%(standard) and structures 26, 27 generated with 0.53.
- Contact distance d_c: 6.0 (always)
- Minimum distance d_{min}: 0.2 (always)
- Maximum distance of 2 adjacent fragments $d_{a,max}$: 1.8 ([26]) apart from structure 1 generated using 1.5 (standard)
- Maximum distance between any 2 fragments of the same chromosome $d_{max,intra}$: 20.0 ([26]) apart from structure 1 generated using 30.0 (standard)
- Maximum distance between any 2 fragments $d_{max,inter}$: 169.0 ([26]) apart from structures 1 and 2 generated using 150.0 (standard)
- Learning rate: 0.001 (always)
- Maximum number of iterations for the optimization: 20000 (always)

For the weights instead:

- all the weights W_1 were always set to 1, as advised in the s.m.;
- all the weights W_2 were always set to 3 (s.m.) apart from structures 1 and 2 generated using $W_2[chr1 = chr2] = 4.0$ (standard) and $W_2[\text{chr1} \neq \text{chr2}] = 2.0$;

- The weight $W_3[chr1 \neq chr2]$ was always set to 0.5 apart from structures 28, 29, 30 generated using $W_3[chr1 = chr2] = 0.05$ (s.m.); the weights W_2[chr1 = chr2] were always set to 5.0;
- The weight $W_4[chr1 \neq chr2]$ for every structure can be seen in the first table below. For structures 1, 2, and 32, the standard value (0.5) already used in MOGEN is applied. From structure 3, the value was set to one and decreased, as advised in the supplementary material. It was varied in the range [0.1, 1] to maximize the total percentages of contact and non contact scores while preventing chromosomes from intermingling, tolerating a maximum of two low interchromosomal non contact scores if between small chromosomes. If the values of interchromosomal non contact scores were too high, the weight was increased; conversely, it was decreased. The weights $W_4[chr1 = chr2]$ were instead varied from their standard values to achieve high and similar values of intrachromosomal contact and non contact scores. They were varied from structures 3 to 9 (see table 1 below). After structure 9, they were kept constant. In structures 1, 2, and 32, they were set to the standard values.

In the end, the standard parameters were used for all the reconstructions of the cell types.

Chromosome	$W_4[chr1, chr2]$ standard	$W_4[chr1, chr2]$ structure 9
1	1.2	1.1
2	1.35	1.1
3	2.2	1.65
4	2.2	2.2
5	2.0	2.0
6	2.6	2.6
7	3.8	3.2
8	3.5	2.9
9	3.3	2.7
10	4.8	4.6
11	4.2	3.5
12	3.8	3.8
13	7.0	7.2
14	6.0	6.4
15	7.0	7.75
16	8.0	7.8
17	6.5	6.8
18	12.0	12.0
19	10.0	10.1
20	11.0	11.1
21	10.0	11.0
22	11.0	11.2
X	3.0	2.4

Table 1: The standard values and the values used after structure 9 for the weights $W_4[chr1 = chr2]$.

Bibliography

[1] Paul T Boggs and Paul T Boggs. "Orthogonal distance regression". In: (1989).

[2] Paul T Boggs et al. "User's reference guide for odrpack version 2.01: Software for weighted orthogonal distance regression". In: (1992).

[3] Britta AM Bouwman, Nicola Crosetto, and Magda Bienko. "A GC-centered view of 3D genome organization". In: *Current Opinion in Genetics & Development* 78 (2023), p. 102020.

[4] Roberto Chiarle et al. "Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells". In: *Cell* 147.1 (2011), pp. 107–119.

[5] Thomas Cremer and Christoph Cremer. "Chromosome territories, nuclear architecture and gene regulation in mammalian cells". In: *Nature reviews genetics* 2.4 (2001), pp. 292–301.

[6] Jenny A Croft et al. "Differences in the localization and morphology of chromosomes in the human nucleus". In: *The Journal of cell biology* 145.6 (1999), pp. 1119–1131.

[7] Gabriele Girelli et al. "GPSeq reveals the radial organization of chromatin in the cell nucleus". In: *Nature biotechnology* 38.10 (2020), pp. 1184–1193.

[8] Jenny Gu and Philip E Bourne. *Structural bioinformatics*. Vol. 44. John Wiley & Sons, 2009.

[9] TC Hsu. "A possible function of constitutive heterochromatin: the bodyguard hypothesis". In: *Genetics* 79 (1975), pp. 137–150.

[10] Kamel Jabbari, Maharshi Chakraborty, and Thomas Wiehe. "DNA sequence-dependent chromatin architecture and nuclear hubs formation". In: *Scientific Reports* 9.1 (2019), p. 14646.

[11] KES47. *Chromosome_en.svg*. `https://commons.wikimedia.org/wiki/File:Chromosome_en.svg`. Wikimedia Commons, CC BY 3.0. 2010.

[12] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *science* 326.5950 (2009), pp. 289–293.

[13] Ryan L. McCarthy, Jingchao Zhang, and Kenneth S. Zaret. "Diverse heterochromatin states restricting cell identity and reprogramming". In: *Trends in Biochemical Sciences* 48.6 (2023), pp. 513–526. ISSN: 0968-0004. DOI: https://doi.org/10.1016/j.tibs.2023.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0968000423000737.

[14] Tom Misteli. "Chromosomes in space and time". In: *Trends in Cell Biology* 12.7 (2002), p. 322.

[15] Sandro Morganella et al. "The topography of mutational processes in breast cancer genomes". In: *Nature communications* 7.1 (2016), p. 11383.

[16] Olivia Morrison and Jitendra Thakur. "Molecular complexes at euchromatin, heterochromatin and centromeric chromatin". In: *International Journal of Molecular Sciences* 22.13 (2021), p. 6922.

[17] Andrea E Murmann et al. "Local gene density predicts the spatial position of genetic loci in the interphase nucleus". In: *Experimental cell research* 311.1 (2005), pp. 14–26.

[18] National Human Genome Research Institute. *Phosphate Backbone*. Public domain. Accessed March 2026. 2024. URL: https://www.genome.gov/genetics-glossary/Phosphate-Backbone.

[19] Oluwatosin Oluwadare et al. "GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data". In: *BMC molecular and cell biology* 21.1 (2020), p. 60.

[20] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[21] Laleh Rafiee, Sasan Mohsenzadeh, and Mostafa Saadat. "Nonrandom gene distribution on human chromosomes". In: (2008).

[22] Benjamin Schuster-Böckler and Ben Lehner. "Chromatin organization is a major influence on regional mutation rates in human cancer cells". In: *nature* 488.7412 (2012), pp. 504–507.

[23] Tom Sexton et al. "Gene regulation through nuclear organization". In: *Nature structural & molecular biology* 14.11 (2007), pp. 1049–1055.

[24] Anisha Shakya and John T King. "DNA local-flexibility-dependent assembly of phase-separated liquid droplets". In: *Biophysical journal* 115.10 (2018), pp. 1840–1847.

[25] Luna Tammer et al. “Gene architecture directs splicing outcome in separate nuclear spatial regions”. In: *Molecular Cell* 82.5 (2022), pp. 1021–1034.

[26] Tuan Trieu and Jianlin Cheng. “MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data”. In: *Bioinformatics* 32.9 (2016), pp. 1286–1292.

[27] Rogier Versteeg et al. “The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes”. In: *Genome research* 13.9 (2003), pp. 1998–2004.

[28] Eitan Yaffe and Amos Tanay. “Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture”. In: *Nature genetics* 43.11 (2011), pp. 1059–1065.

[29] Wing Hin Yip et al. “GPSeq maps the radial organization of eukaryotic genomes along the nuclear periphery–center axis”. In: *Nature Protocols* (2026), pp. 1–72.