Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

Hi-C DATA SPECTRAL ANALYSIS: SynHi-C MAPS FOR A CASE STUDY WITH ShRec3D ALGORITHM AND VR

Supervisor:
Prof. Daniel Remondini

Co-supervisor:
Dr. Alessandra Merlotti

Submitted by:
Daniele Bruno

Academic Year 2022/2023

Alla nonna Rina,
in memoria

Abstract

Hi-C matrices are milestones for the qualitative and at the same time quantitative study of genome folding, its organization into chromosomal territories, compartments and topological domains. Here we introduce and discuss the synHi-C, a method for synthetic Hi-C data production. It arises from the possibility of characterizing the signal-to-noise ratio starting from a spectral analysis on different types of Hi-C data at different resolutions (1 Mb and 100 kb). Through the spectral analysis, the signal component has been identified, consisting of isolated and scattered eigenvalues even at a great distance from the origin, and the noise component, which follows the Wigner's semicircle law centered in zero, identified through the simulation of random symmetric matrices. By adding the essential matrix (essHi-C) consisting of the sum of the projectors associated to the signal component, with the one reconstructed starting from the projectors of the random component after an eigenvalues reshuffling, it is possible to obtain a potentially vast amount of synthetic matrices. After testing the spectral analysis on the gold standard cell line GM12878, this innovative method has been applied to a real case study consisting of two cases (235 and 295) of a rare prion disease and two controls (LM and MB), demonstrating how not only the intrinsic biological properties of the Hi-C maps, given by the essHi-C component, are enhanced, but also that the statistical properties of the introduced fluctuations are unbiased, reflecting the non-specific component. The validation of these results has been obtained through different methods including the use of scatter plots between synHi-C and original matrices to identify their correlation, the ShRec3D algorithm to verify the coherence between the spatial folding structures of the chromatin after a proper Procrustes analysis and finally through their visualization in Blender and the Virtual Reality (VR) 3D simulated environment inspection.

Contents

Chapter 1

Introduction

Our understanding of the three-dimensional structure of chromosome folding has been significantly shaped by chromosome conformation capture-based methods (3C). Among these, a main role is played by the Hi-C experiments, which are able to capture the nearby interaction fragments in the whole genome. The experiments generate Hi-C maps enclosing the spatial distances between all possible pairs of loci in the genome. The analysis of the Hi-C maps also allows to characterize the interaction patterns, showing the existence of different levels of spatial organization of the genome: the chromosome territories, the chromatin compartments and the topologically associated domains (TADs). These maps have also been studied with spectral analysis methods, which allow to highlight the biological features that emerge from otherwise hidden patterns. In this thesis we want to probe a more in-depth analysis through the use of spectral analysis methods to explore the different components of the characteristic spectrum of Hi-C matrices. We will examine both the GM12878 cell line, a gold standard for the study of Hi-C maps, and a real case study consisting of two cases (235 and 295) of a rare prion disease and two controls (LM and MB). The main objective will be to generate synthetic data that faithfully reproduce both the biological and statistical properties of the Hi-C data considered. To validate these analyses, known tools such as scatter plots will be used, but also novel algorithms such as ShRec3D capable of providing the spatial coordinates relating to the three-dimensional conformation of the chromosomes, as well as the use of software such as Blender and visualization in a Virtual Reality (VR) environment.

Chapter 2

Exploring the 3D organization of genomes

2.1 DNA and chromatin fiber

The deoxyribonucleic acid or DNA is the macromolecule whose most important function lies in the ability to carry genes, the information that specifies all the proteins that make up an organism, including information about when, in which cell types and in what quantities each protein should be produced. In eukaryotes, the DNA in the nucleus is broken down into a set of different chromosomes. For example, the human genome is distributed across 24 different chromosomes. Each of them consists of a single very long linear DNA molecule associated with proteins that fold and pack the thin DNA thread into a more compact structure. To get an idea of the dimensions involved, just think that each human cell contains about 2 meters of DNA if stretched from one end to the other; however, the nucleus of a human cell, which contains the DNA, is only 6 μm in diameter. This is geometrically equivalent to packing 40 km of extremely fine wire into a tennis ball! The complex task of DNA packaging is performed by specialized proteins that bind to and fold it, generating a series of loop and ring-like structures that provide ever higher levels of organization, preventing the DNA from becoming an unmanageable tangle. Unexpectedly, although DNA is folded very tightly, it is packed in a way that allows it to become readily available to the many enzymes in the cell that replicate it, repair it, and use its genes to make proteins. In fact it is important to keep in mind that the chromosome structure is dynamic. Not only do chromosomes

globally condense in accordance with the cell cycle, but different regions of chromosomes in interphase (cell cycle phase) condense and decondense as cells gain access to specific DNA sequences for gene expression, DNA repair and replication. The packaging of chromosomes must therefore be done in a way that allows rapid, localized and on-demand access to DNA. The complex of DNA and proteins is called chromatin (from the Greek chroma, i.e. colour, due to its staining properties).

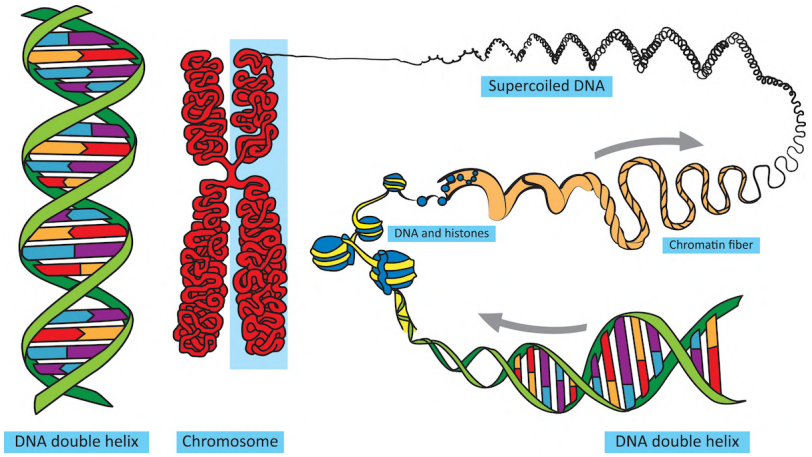


Figure 2.1: The hierarchical structure of DNA folding into chromatin: from the double helix to the chromosome scale [11].

From a chemical point of view, the DNA is a complementary, oriented, spiralized and informational double polynucleotide chain. The genetic information resides in the order of the sequential arrangement of the nucleotides, monomeric units of the nucleic acid polymers, i.e. DNA and RNA, which is translated into the corresponding amino acids (building blocks of protein) exploiting the genetic code. During the cell division process, the genetic information is duplicated (DNA replication), in order to transmit the genetic information to the following cellular generations. A gene is defined as a segment of the DNA containing instructions for a specific protein production. On each chromosome they are arranged in a particular sequence, each having a specific location on the chromosome (called locus). The human genome contains approximately $20000 - 25000$ genes.

2.2 3C-based methods

The spatial organization of DNA within the cell is a much studied problem in biology and bioinformatics. Understanding the three-dimensional structure of the chromatin strand in which DNA is organized is very important for exploring the structure of chromosomes, their interactions and functionality. In fact, chromosomes are some of the most complex structures in the cell: the molecular composition of the chromatin fiber is highly varied along its length, and the fiber is intricately tangled in three dimensions. In order to map the local structure of chromatin, momentous efforts are being made to analyze the complex of DNA-associated proteins (histones) and their modifications along the chromosomes. Such studies have led to obtain the genomic locations of genes and regulatory elements that are active in a given cell type and have begun to discover complete sets of functional elements of the human genome and of several model organisms. Only over the last decade, a number of molecular and genomic approaches have been developed to study the three-dimensional chromosome folding at increasing resolution and throughput. They are all based on Chromosome Conformation Capture (3C, henceforth). These methods make it possible to determine how frequently any pair of loci in the genome are in sufficient physical proximity (probably in the range of 10-100 nm) to become cross-linked. Chromosome Conformation Capture is a high-throughput methodology which can be used to analyze the overall spatial organization of chromosomes and to investigate their physical properties at high resolution. In 3C-based methods, cells are crosslinked with formaldehyde to covalently link chromatin segments that are in close spatial proximity. Afterwards, chromatin is fragmented by restriction digestion or sonication. Crosslinked fragments are then ligated to form a unique hybrid DNA molecules. Finally, the DNA is purified and analyzed (figure 2.2).

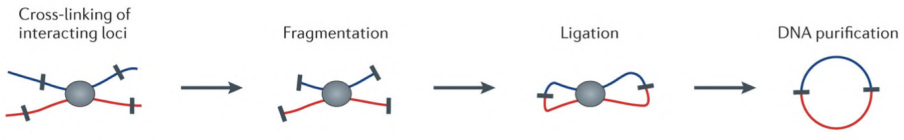


Figure 2.2: Schematic representation of the Chromosome Conformation Capture methodology. From left to right: formaldehyde cross-linking, fragmentation through digestion, intramolecular ligation, and DNA purification [7].

The different 3C-based methods only differ in the way hybrid DNA molecules, each corresponding to an interaction event of a pair of loci, are detected and quantified. In classical 3C experiments single ligation products are detected by the polymerase chain reaction (PCR) one at the time using a pair of locus-specific primers to direct DNA elongation toward each other at opposite ends of the sequence being amplified. Given that 3C is a one-by-one interaction, it is necessary to check one pair of the chromatin region at a time. Thus, it becomes challenging to explore all the genome and most of the 3C analysis usually cover only tens to several hundred kb. To actually see whether far away chromatin regions are interacting with each other the 4C (Circularized Chromosome Conformation Capture) technique has been developed. It uses inverse PCR to generate genome-wide interaction profiles for single loci, so as to be able to understand how one particular chromatin region interacts with all the nearby regions. Another 3C-based approach is the 5C (Carbon Copy Chromosome Conformation Capture), which combines 3C with hybrid capture approaches to identify up to millions of interactions in parallel between two large sets of loci (figure 2.3). Unlike 4C approaches, genome-wide but anchored on a single locus, 5C analysis typically involve Mbs regions in which all the possible pairwise interactions are inspected. Therefore, the 5C methodology is considered as a many-by-many interaction approach thanks to which it allows to explore two sets of hundreds to thousands of restriction fragments to interrogate up to millions of long-range interactions. The latters can cover up tens Mb that can be contiguous or scattered over loci of interest throughout the genome.

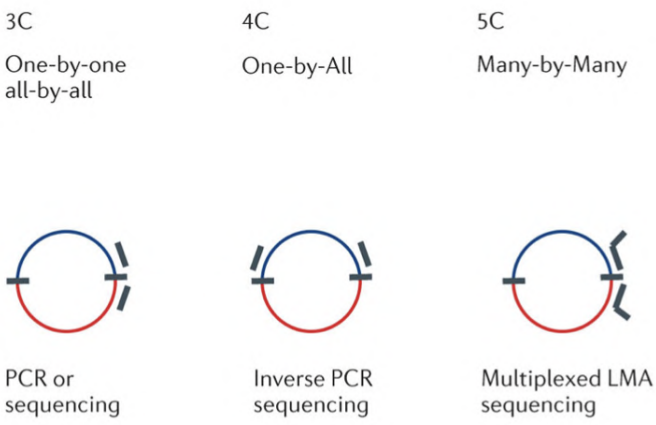


Figure 2.3: Different ligation product detection analysis for the 3C-based methods. **Left:** traditional 3C method quantifies interactions between a single pair of genomic loci (one-by-one) and their ligated fragments are detected using PCR with known primers, that is why this technique requires the prior knowledge of the interacting regions. **Middle:** 4C method captures interactions between one locus and all other genomic loci (one-by-all). It involves a second ligation step, to create self-circularized DNA fragments, which are used to perform inverse PCR. Inverse PCR allows the known sequence to be used to amplify the unknown sequence ligated to it. In contrast to 3C and 5C, the 4C technique does not require the prior knowledge of both interacting chromosomal regions. **Right:** 5C method detects interactions between all restriction fragments within a given region (many-by-many), with this region's size typically no greater than a Mb. This is done by ligating universal primers to all fragments and by using the ligation-mediated amplification (LMA) for amplifying selected DNA sequences [7].

2.3 Hi-C method

The amount of genome-wide 3C-based data is rapidly growing and presents great and exciting opportunities in the field of computational modeling and 3D genome interpretation. In particular, there has been a surprising development in recent times for the production of high-resolution Hi-C (High-throughput Chromosome Conformation Capture) data. Hi-C is a method that adapts the 3C-based approach to probe the three-dimensional architecture of the whole genomes by coupling proximity-based ligation and enabling purification of their products followed by massively parallel sequencing. Knowledge of Hi-C data analysis methods and of the ways available to carry out each analysis step is assuming increasingly importance, as the number and variety of Hi-C data sets increase. Unlike the 3C-based techniques, that require the choice of a set of target loci and do not admit unbiased genome-wide analysis, Hi-C allows

unbiased identification of chromatin interactions across an entire genome. To sum up, the traditional Hi-C experimental procedure consists of six steps:

1. **crosslinking** of cells by formaldehyde;
2. **digestion** of DNA with a restriction enzyme, e.g. HindIII, that leaves the ends free;
3. **labeling** of the ends with biotin;
4. **ligation** of the cross-linked (chimeric) fragments at very low DNA concentration. Under such conditions, ligation of chimeric fragments, which is intramolecular, is strongly favored over ligation of random fragments, which is intermolecular;
5. **shearing** and **purification** of the resulting DNA, streptavidin **pull-down** of biotinylated ligation products;
6. paired-end read **sequencing** of the extracted fragments.

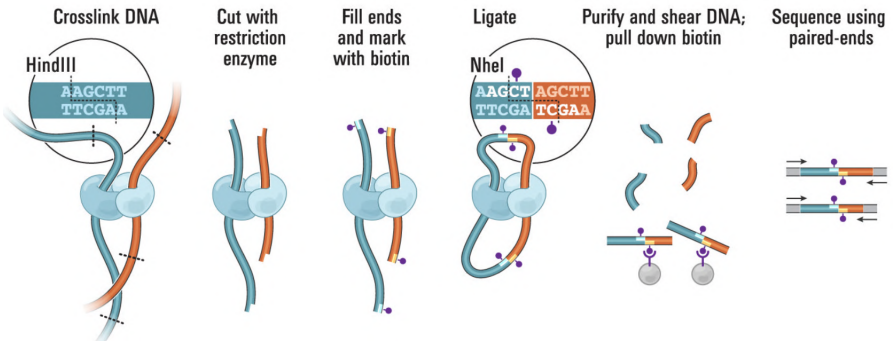


Figure 2.4: Synthetic scheme of the traditional Hi-C experimental technique. From left to right: cells undergo crosslinking with formaldehyde, which produces covalent bonds between spatially adjacent chromatin segments (DNA fragments are represented in blue and red; proteins that mediate these interactions are colored light blue and cyan). Chromatin undergoes digestion by a restriction enzyme (HindIII). Restriction sites are highlighted with dashed lines. The resulting free-ends are bound by biotin-labeled nucleotides (purple dots) and the ligation is performed under dilute conditions to create chimeric molecules. The NheI restriction site is then created (see insert). DNA purification takes place and the biotinylated junctions are isolated with streptavidin and identified by paired-end sequencing [32].

Formaldehyde is an organic compound that allows DNA crosslinking, i.e. the covalent bond between segments of chromatin next to each other in the 3D structure. In the process called restriction digestion, a restriction enzyme breaks down the DNA structure at a specific sequence of bases (usually 4-6 bases), called the restriction site. Biotin is a water-soluble vitamin that is linked to nitrogenous bases to mark them and streptavidin is a protein with a very high affinity for biotin, thus allowing the selection of the molecules that contain it. In general, sequencing can be performed starting from a single end of the DNA strand to be sequenced (single-end reads) or starting from both ends and continuing in opposite directions (paired-end reads). The resulting DNA sample contains ligation products consisting of fragments that were originally in close spatial proximity in the nucleus, marked with biotin at the junction. Thus, the Hi-C method is really able to capture the nearby interaction fragments in the whole genome (all-by-all), of course depending on the depth of sequencing. A Hi-C library is created by shearing the DNA and selecting the biotin-containing fragments with streptavidin beads. The library is then analyzed using massively parallel DNA sequencing, producing an archive of interacting fragments.

2.3.1 Hi-C contact maps

In general, the Hi-C procedure generates a Hi-C map, i.e. a genome-scale sequencing library that allows the measurement of 3D distances between all possible pairs of loci in the genome. The Hi-C map consists of a contact list among chimeric DNA fragments produced by the Hi-C experiment. By segmenting the linear genome into loci of a fixed size, i.e. a sequence of non-overlapping windows of equal sizes, the Hi-C map can be represented as a genome-wide contact matrix M, where the entries M_{ij} are the number of contacts observed between loci i and j. A contact is defined as a pairing between reads (short sequences of synthetic DNA that are produced during the sequencing reaction) which is not excluded by the elimination of duplicate reads (which correspond to unligated fragments) or which do not align uniquely to the genome. The contact map can be visualized as a heatmap (figure 2.5), whose inputs are called pixels. An interval refers to a set of consecutive loci and the contacts between two intervals therefore shape a rectangle or a square (block) in M. The window size is also referred to as the resolution of the Hi-C matrix and

it is defined as the size of the loci used to construct the contact matrix. To increase the resolution by a factor of n, one must increase the number of reads by a factor of n^2. The most common resolutions are 1 Mb and 100 kb.

2.3.2 Genome compartments and TADs

Using the genome-wide Hi-C technique, which map all the interactions in a genomic region of interest or in complete genomes, the genome has been shown to be partitioned into several organizational levels: chromosome territories, compartments and domains. The result of the first chromosome folding were obtained by analyzing the trend of the average intrachromosomal contact probability for pairs of loci and different chromosomes separated by a certain genomic distance (distance in base pairs along the nucleotide sequence). It has been seen that the function of contact probability decreases monotonically with genomic distance on every chromosome, suggesting polymer-like behavior in which the three-dimensional distance between loci increases with increasing genomic distance. Even at distances greater than 200 Mb, it is always much greater than the average contact probability between different chromosomes, which implies the existence of chromosome territories.

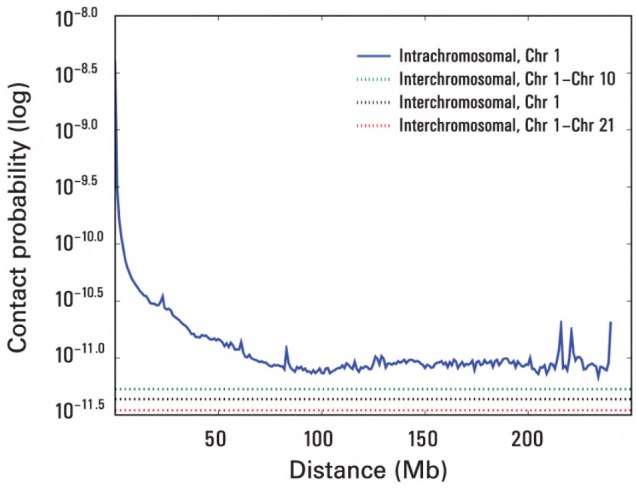


Figure 2.6: The presence and organization of chromosome territories. Probability of contact decreases as a function of genomic distance on chromosome 1, eventually reaching a plateau around 90 Mb (continuous blue line). The level of interchromosomal contact (black dashes line) differs for different pairs of chromosomes; loci on chromosome 1 are most likely to interact with loci on chromosome 10 (green dashes line) and least likely to interact with loci on chromosome 21 (red dashes line). Interchromosomal interactions are depleted relative to intrachromosomal interactions. Adapted from [32].

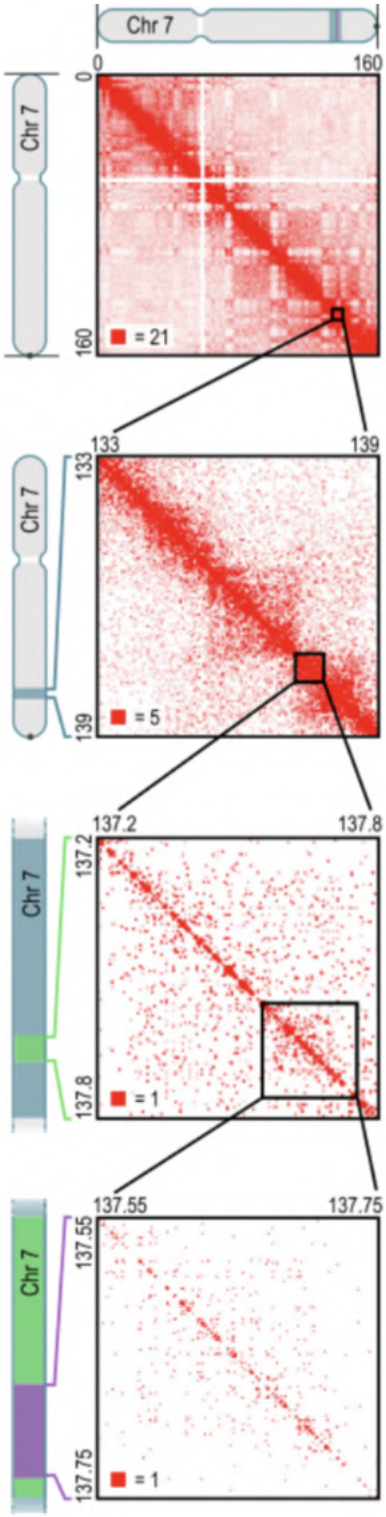


Figure 2.5: Contact map for the chromosome 7 of the lymphoblastoid cell line GM06990. From top to bottom: increasing zoom. Adapted from [32].

Moreover, inter- and intra-chromosomal contact maps for mammalian genomes have revealed a plaid pattern of interactions that can be approximated by two distinct compartments A (active) and B (inactive) that alternate along chromosomes with a characteristic size of about 5 Mb each.

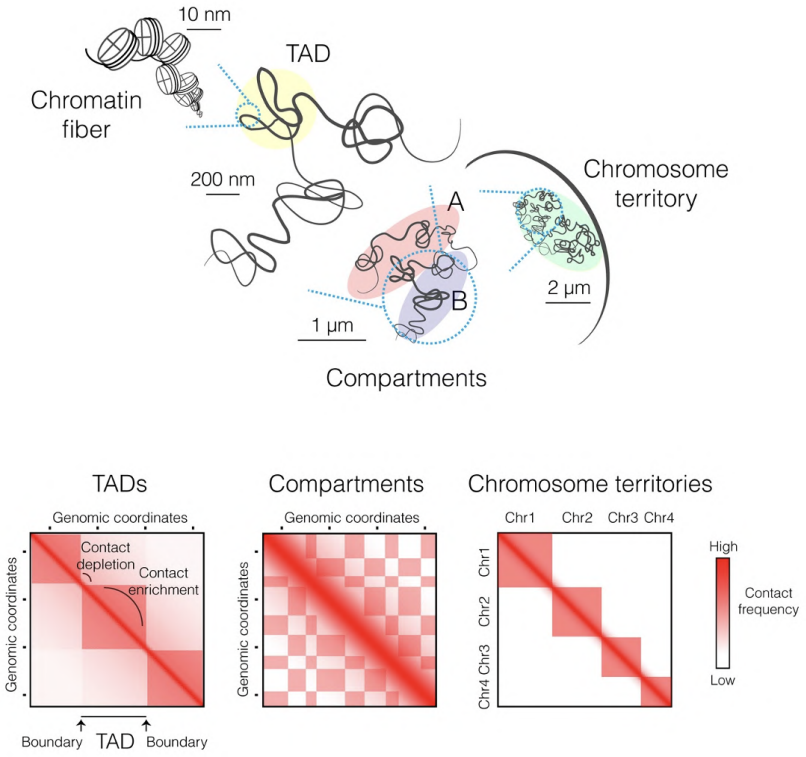


Figure 2.7: Relation between DNA three-dimensional structures with Hi-C contact maps at different scales. **Top:** Schematic view of chromosome folding inside the nucleus. The finest layer of chromatin folding is at the DNA-histone association level, forming nucleosomes organized into the chromatin fiber (10 nm). Chromatin is packed at different nucleosome densities depending on gene regulation and folds into higher-order domains (200 nm) of preferential internal interactions (TADs). At the chromosomal scale (1 μm), chromatin is segregated into A and B compartments of interactions. Individual chromosomes occupy their own space within the nucleus, forming chromosome territories (2 μm).
Bottom: Schematic representation of Hi-C maps at different genomic scales, reflecting the different layers of higher-order chromosome folding. Genomic coordinates are indicated on both axes, and the contact frequency between regions is represented by a red scale. **Left:** TADs appear as squares along the diagonal enriched in interactions, separated by contact depletion zones delimited by TAD boundaries. **Middle:** At the chromosomal scale, chromatin long-range interactions form a characteristic plaid pattern of two mutually excluded A and B compartments. **Right:** Intrachromosomal interactions are dominant compared to interchromosomal contacts, consistent with the formation of individual chromosome territories [47].

The A and B compartments preferentially interact with the corresponding one throughout the genome. Thus, the Hi-C map imply that regions tend to be closer in space if they belong to the same compartment than if they do not. Taken together, these pieces of observations confirm the spatial compartmentalization of the genome inferred from Hi-C. As the resolution of the data increased, domains of smaller dimensions were found, from which it was assumed that the compartments are divided into condensed structures hundreds of kb in size referred to as Topologically Associating Domains (TADs). They can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status. Loci located within these TADs tend to frequently interact with each other, but much less frequently with loci located outside their domain. This feature enabled researchers to identify TADs throughout the human genome by analyzing lower resolution, but genome-wide, Hi-C interaction maps. These analysis showed that TADs are universal building blocks of chromosomes and that the human and mouse genomes are each composed of over 2000 TADs covering over 90% of the genome.

Chapter 3

Materials and methods

The Hi-C data are allocated into a square symmetric matrix M, where each entries M_{ij} stands for the total number of read pairs sequenced (contacts) between loci i and j. Moreover, it corresponds to a matrix which is non-negative and with the dominant values located along the main diagonal. In fact, considering two loci belonging to the same chromosome, the maximum possible distance between them is equal to the length of the DNA that separates them. Therefore, two loci i and j that are close along the 1D chromatin chain tend to have a high M_{ij} count, regardless of the 3D conformation of the chromosome. Furthermore, segments from chromosome centrometric regions cannot be uniquely mapped due to the presence of repeated sequences along the chromosomal strand. Thus within each Hi-C matrix there are zero-valued default lines. The latters are usually removed since they are non-informative. Finally, it is always true that $M_{ij} \geq 0$ since the entries encoded the contact counts between pairs of loci. A Hi-C matrix can therefore be naturally associated to the genome through a graph, where vertices are defined by binned loci in the genome, and the edge weight between a pair of loci is proportional to their contact frequency. The case $M_{ij} = 0$ implies that the nodes are not connected. For example the TADs are strongly connected graph components having strong intra-connections and weak inter-connections (figure 3.1).

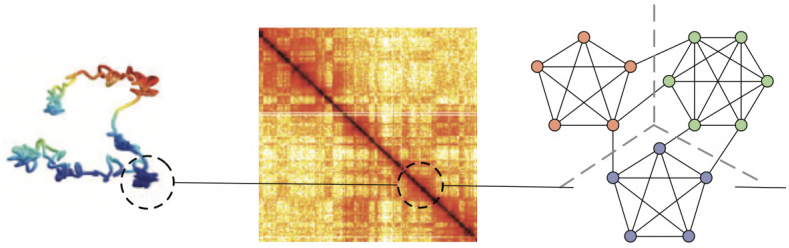


Figure 3.1: Illustration of topologically associated domains represented in different modalities. From left to right: physical structures of locally compact regions, diagonal blocks in a Hi-C map and graph model of the contact architectures [2].

Quantitatively studying the properties related to the spatial conformation of chromosomes in the cell nucleus is possible thanks to the spectral graph theory applied to the Hi-C matrix. The most relevant concepts of graph theory and the spectral decomposition of the associated matrices are reviewed below.

3.1 Graphs and adjacency matrices

We define a undirected graph as the ordered pairs of sets $\mathcal{G}(V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ is a finite set of vertices (or nodes) with cardinality N and E is an edge set consisting of paired vertices of the form $\{v_i, v_j\}$, with $i \neq j$ to avoid loops. The order of a graph coincide with its number of vertices $|V|$, while the size is its number of edges $|E|$. Moreover, the degree of a given vertex $d(v_i)$ is the number of edges that are connect to it. The edges (or links) of a graph define a symmetric relation on the vertices, called the adjacency relation. Specifically, two vertices v_i and v_j are adjacent if $\{v_i, v_j\}$ is an edge. A graph may be fully specified by its adjacency matrix $\mathbf{A}(\mathcal{G})$ (or simply $\mathbf{A}$ for notation convenience), which is an $N \times N$ square matrix encoding the adjacency relationships in the graph $\mathcal{G}$, where the entries $\mathbf{A}_{ij}$ specifying the number of connections from vertex i to vertex j. In particular, for a undirected graph, the adjacency matrix $\mathbf{A}(\mathcal{G})$ is defined as

$$\mathbf{A}_{ij} = \begin{cases} 1 & \{v_i, v_j\} \in E \\ 0 & otherwise \end{cases}$$

Therefore $\mathbf{A}(\mathcal{G})$ in this special case of undirected graph, is a binary matrix with zeros on its diagonal, because an edge in this kind of graph cannot start and end at the same vertex. The adjacency matrix is also symmetric, meaning

that $\mathbf{A}_{ij} = \mathbf{A}_{ji}$, in fact all of its edges are bidirectional. In this context it is possible to express the degree of a vertex as

$$d(v_i) = \sum_{j \in N} \mathbf{A}_{ij}$$

More generally, instead of considering binary connections between pairs, weights can be assigned to each edge such that

$$\mathbf{A}_{ij} = \begin{cases} w_{ij} & \{v_i, v_j\} \in E \\ 0 & otherwise \end{cases}$$

where w_{ij} are real numbers used to characterize the connection strengths. A weighted graph is therefore a special type of labeled graph in which the labels are numbers. The following figure 3.2 shows an example of a weighted undirected graph.

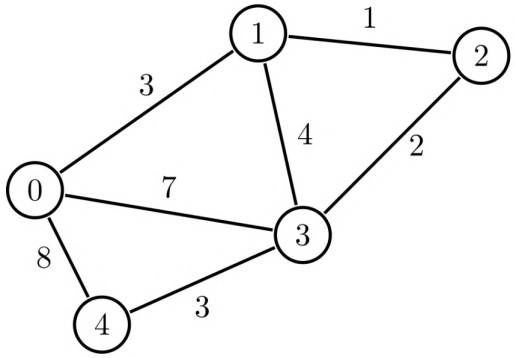


Figure 3.2: Example of a weighted undirected graph with 5 nodes and 7 edges. Each edge has a weight. For example, the edge $\{0, 3\}$ has the weight 7 and the edge $\{1, 2\}$ has the weight 1 [21].

Therefore the adjacency matrix encoding the graph structure $\mathcal{G}$ is

$$\mathbf{A}(\mathcal{G}) = \begin{pmatrix} 0 & 3 & 0 & 7 & 8 \\ 3 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 2 & 0 \\ 7 & 4 & 2 & 0 & 3 \\ 8 & 0 & 0 & 3 & 0 \end{pmatrix}$$

3.2 Spectral properties of symmetric random matrices

The Hi-C matrix can be interpreted as a weighted adjacency matrix for a undirected graph. As we have seen, such a matrix is symmetric, so all the spectral properties relating to symmetric matrices are satisfied. In particular, symmetric matrices have purely real eigenvalues. This can be easily demonstrated recalling the hermitian scalar product in $\mathbb{C}^n$ to be

$$\langle \mathbf{x}, \mathbf{y} \rangle = \bar{\mathbf{x}}^T \mathbf{y} = \sum_{i=1}^{n} \bar{x}_i y_i$$

where the overline is the complex conjugate and the vectors $\mathbf{x}$ and $\mathbf{y}$ are thought of as columns with n components. Then, if $A\mathbf{x} = \lambda\mathbf{x}$, the following equations hold

$$\bar{\lambda} \langle \mathbf{x}, \mathbf{x} \rangle = \langle \lambda\mathbf{x}, \mathbf{x} \rangle = \langle A\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, \lambda\mathbf{x} \rangle = \lambda \langle \mathbf{x}, \mathbf{x} \rangle$$

from which $\lambda = \bar{\lambda}$ since $\langle \mathbf{x}, \mathbf{x} \rangle$ is always a real positive number for each array $\mathbf{x} \neq 0$. Now we considered the so-called Gaussian orthogonal ensemble (GOE), whose elements are symmetric matrices with entries drawn from a Gaussian distribution with zero mean and variance ρ^2. On average, the salient spectral properties of the elements belonging to the GOE are related to the distributions follow by their eigenvectors and eigenvalues. Precisely, the orthonormal eigenvectors sample uniformly the surface of a unit $(N-1)$-sphere, where N is the linear size of the matrices. The generic component x of any eigenvector follows the same Gaussian distribution

$$p(x) = \sqrt{\frac{N}{2\pi}}\, e^{-\frac{Nx^2}{2}}$$

with zero mean and variance equal to $\frac{N}{2}$. The largest eigenvalue is a real random variable, and the vector of all eigenvalues ordered by size is an $\mathbb{R}^N$-valued random variable. In some ways, the theory of eigenvalues of random matrices mimics that of real random variables, and the following distribution plays the role of the Gaussian

$$p(\lambda) = \begin{cases} \frac{2}{\pi\Lambda^2}\sqrt{\Lambda^2 - \lambda^2} & -\Lambda < \lambda < \Lambda \\ 0 & otherwise \end{cases}$$

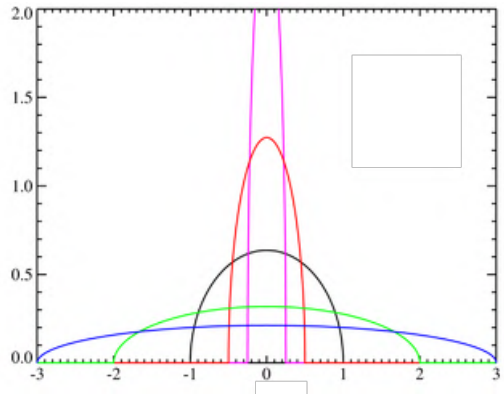


Figure 3.3: Wigner's semicircle law for different values of the parameter Λ ranging from 0.25 to 3 [50].

which depends on the parameter Λ defining the support of λ to be the interval $[-\Lambda, \Lambda]$. The distribution $p(\lambda)$ is referred to as the Wigner's semicircle law (figure 3.3), even though strictly speaking is a scaled semicircle, i.e. a semi-ellipse. Wigner showed that as N approaches to infinite, the eigenvalues distribution of the GOE converges to the semicircle law (figure 3.4). The idea is that if you sample a random matrix from the GOE and then pick one of its eigenvalues at random, the resulting distribution will depend on N and converge to the semicircle as $N \to \infty$. It was quickly recognized that random matrices exhibit universality, i.e. ensembles with differently distributed entries have same limiting eigenvalue distribution.

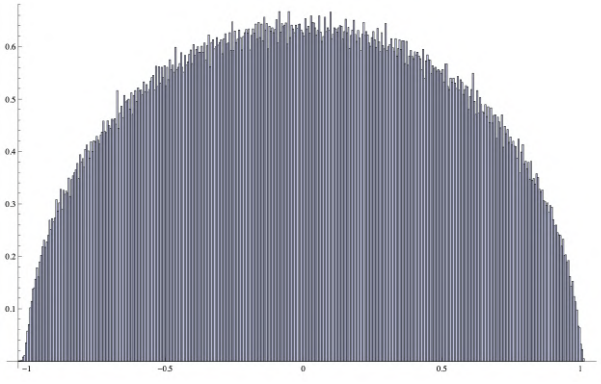


Figure 3.4: A histogram plot of the normalized eigenvalues for 500 matrices, each 400×400. The entries are chosen independently from the standard normal $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ [16].

3.3 Synthetic data

Synthetic data are artificially produced information rather than generated by real-world events. They are created through algorithms and are used as a stand-in for production test datasets to validate mathematical models and to train machine learning (ML) models that captures the patterns in a real dataset. While collecting high-quality data from the real world is difficult, expensive, and time-consuming, synthetic data technology allows users to quickly, easily, and digitally generate data in the amount they want, customized to their needs.

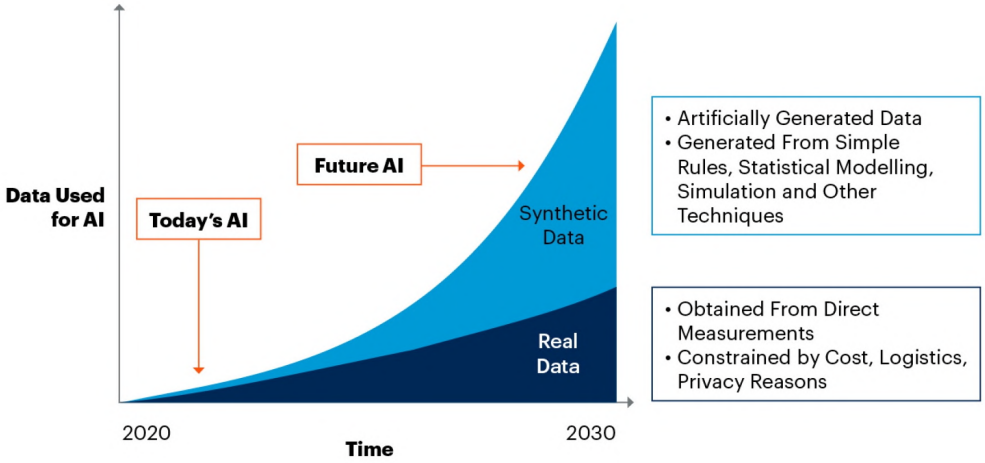


Figure 3.5: Forecast of use of synthetic data for which by 2030 they will completely overshadow real data in AI models becoming the main form of data used in AI [18].

The use of synthetic data is gaining wide acceptance because it can provide several benefits over real-world data. Synthetic data can be artificially generated to mimic the behavior of real-world data, enabling to create a diverse and large amount of training data without spending a lot of time for doing real experiments. This approach can be used to create synthetic data sets that are similar to real data sets in terms of their distribution and variability. Another process for creating synthetic data is to use a random number generator to generate uniform data. As can be seen from figure 3.6, the synthetic data retains the inner structure of the original data, but they are not the same.

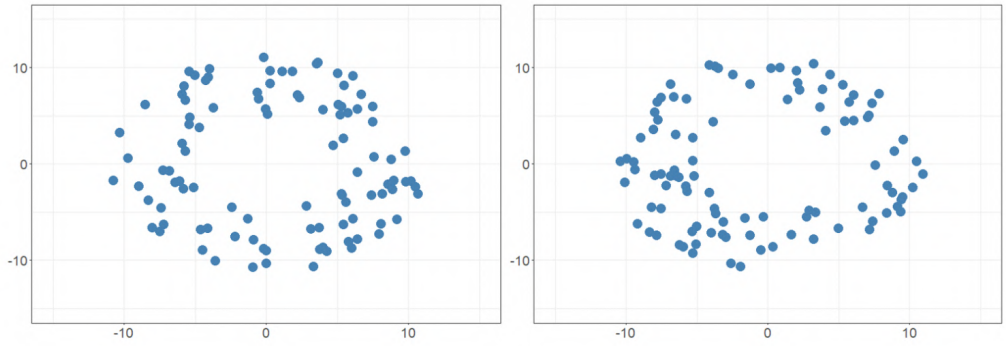


Figure 3.6: Comparison between original (on the left) and synthetic (on the right) set of data in which the structural similarity is well visible [46].

The synthetic data generation has several benefits which we list below.

- **Full user control:** a synthetic data simulation enables complete control over every aspect, satisfying every need. In particular, it is possible to control event frequency, item distribution and many other factors, tailoring the data to certain conditions that can't be obtained with original data. Some examples include controlling the degree of class separations, sampling size and noise level of the considered data set;

- **Cost-effective:** synthetic data can be an inexpensive alternative to real-world data. Of course, synthetic data creation is not free. The main cost of synthetic data is an upfront investment in building the simulation. However, real data enforce timely and financial costs every time a new data set is required or an existing one is revised;

- **Faster production:** synthetic data are not gathered from experimental results, thus it's possible to create a data set more quickly with the right software and simulation. As a result, a great amount of synthetic data can be created in a shorter amount of time.

- **Data privacy:** synthetic data can resemble all important statistical properties of real data without containing any information that could be used to identify the real data. Consequently, this characteristic makes the synthetic data anonymous, eliminating any concern about privacy regulations.

- **Data labeling:** regarding supervised learning tasks, manually labeling a multitude of instances can be time-consuming and error-prone. Syn-

thetically labeled data can be created to improve the model development process increasing labeling accuracy.

3.4 Datasets

The Hi-C data used in this thesis work comes from two distinct datasets. A first analysis concerns the human cell line GM12878 [36]. It is a lymphoblastoid cell line commonly used as surrogates for peripheral blood lymphocytes. It belongs to the numerous set of immortalized cell lines, i.e. continuously growing cells derived from biological samples. The lymphoblastoid cell lines present a low somatic mutation rate in continuous culture, making them the preferred choice of storage for individuals' genetic material. As one of the most reliable, inexpensive, and convenient sources of cells, they have been used by several large-scale genomic DNA sequencing efforts [10] [40]. Particularly, the cell line GM12878 is a popular sample that has been widely used in genomic studies. For example, it is one of three Tier 1 cell lines of the Encyclopedia of DNA Elements (ENCODE) project, which aim is to aid in the integration and comparison of data produced using different technologies and platforms, designating cell types that will be used by all investigators [48] [55]. Apart from this standard reference cell line, we will focus our inspection on very specific primary cells, which are freshly isolated cells from organ tissue and maintained for growth in vitro, without genetic modifications as happens with immortalized cell lines. Primary cell cultures more closely mimic the physiological state of cells in vivo and generate the most biologically relevant data. In particular we used 4 different samples, divided into 2 controls (LM and MB) and 2 cases (235 and 295) which refer to prion disease.

3.5 Essential and synthetic Hi-C maps

The Hi-C experiments, as seen in section 2.3.2, were able to increase our understanding of the structural-functional interplay among chromosomes in the genome. Particularly, Hi-C maps investigation demonstrated that interchromosome interactions are suppressed compared to intrachromosome ones, giving quantitative support to the earlier notion of chromosome territories [5]. Moreover, by examining the dominant eigenvectors of the interaction matrices, it has been revealed the existence of chromatin compartments [32] and TADs, which may form complex nested structures [9],[35]. Relevant observation arise also from comparative Hi-C analysis for data cross-validation with different protocols or resolution, which identify common and statistically significant features of the Hi-C maps [15],[44],[52]. Starting from these observations, spectral analysis methods have been applied to save the robust and significant interactions within the Hi-C maps, enhancing the capability to find out meaningful differences and affinities across matrices. The spectral analysis exploits the information encoded by the eigenvectors and corresponding eigenvalues. However, spectral methods are so far focused only on the first one or two eigenvectors of the Hi-C matrices, which are informative for the chromatin compartmentalization or the first ten as in [17]. Specifically, they shows that most of the Hi-C maps spectrum is compatible with that of random matrices (except for a limited set of eigenvectors with atypically large eigenvalues in modulus), and thus represents a non-specific component shared across chromosomes from different samples. Discounting this part of the spectrum, and retaining only what they term the essential component, it is possible to enhance the definition of chromosomes' architectural features, such as TADs as well as the similarities of replicates and dissimilarities of different cell lines. Finally, they show that essential matrices are stable against variations of the resolution. Starting from the entire Hi-C matrix M, the essential matrix (or essHi-C, for brevity) M^{ess} is defined by the entries

$$M_{ij}^{ess} = \sum_{n=1}^{n^*} \lambda_n x_n^{(i)} x_n^{(j)} = \sum_{n=1}^{n^*} \lambda_n P_n^{ij}$$

where $x_n^{(i)}$ is the i-th component of the n-th eigenvector of M and the associated eigenvalue λ_n. With $P_n^{ij} = x_n^{(i)} x_n^{(j)}$ we indicate the projector related to the eigenvector $\mathbf{x}_n$. Therefore, the term $\lambda_n P_n^{ij}$ describe the contribution of

eigenspace $\mathbf{x}_n$ to the overall Hi-C matrix pattern. The eigenspaces are ranked for decreasing modulus of the eigenvalues and the summation is restricted to the top n^* essential spaces. In the present thesis we want to extend the concept of essential matrix by retaining not only the (arbitrarily chosen) ten larger eigenvalues in modulus (as in [17]), but all the eigenvalues belonging to the signal component. The latter is composed by all the eigenvalues that are outside the non-specific random component, which follows the semicircle Wigner's law. Furthermore, here we introduce a new Hi-C map that we term synthetic Hi-C matrix, or synHi-C, which represents a random copy of the original Hi-C map, even though it maintains all the relevant original features as for the essential one. We define the synHi-C matrix as

$$M_{ij}^{syn} = \sum_{n=1}^{n^*} \lambda_n x_n^{(i)} x_n^{(j)} + \sum_{n=n^*+1}^{N} \tilde{\lambda}_n x_n^{(i)} x_n^{(j)} = M_{ij}^{ess} + \sum_{n=n^*+1}^{N} \tilde{\lambda}_n x_n^{(i)} x_n^{(j)}$$

where $\tilde{\lambda}_n$ denotes a random reshuffling of the eigenvalues λ_n, related to the corresponding eigenvectors $\mathbf{x}_n$, selected from the random component of the spectrum. The reshuffling of the eigenvalues corresponds to a permutation of the weights given to each eigenvector and so to each projector which sum make up the entire Hi-C map. In this way it is possible to generate as large as desired numbers of Hi-C matrices, making sure of not loosing any intrinsic feature encoded by the essHi-C matrix. In our case in which the Hi-C data relate to a genetic disease that is very rare and hardly studied at the Hi-C level, the possibility of producing synthetic data capable of enriching the few available data samples will prove to be very useful and therefore be able to make statistics.

3.6 Preprocessing

The Hi-C matrices are therefore seen as adjacency matrices encoding an undirected weighted graph whose nodes correspond to the loci into which the genome is divided and whose links are characterized by a weight given by the value of the entries $M_{ij} \geq 0$. These matrices are called raw and contain the frequencies of contact observed between two distinct loci of the genome. However, to obtain easily viewable and interpretable Hi-C maps, it is useful to perform a preprocessing of the raw matrices acquired directly in the laboratory.

3.6.1 Dinamic range reduction

After the removal of the null-value bands present in the raw matrix as taken in the laboratory, which correspond to uniquely non-mappable regions in the reference genome and therefore to isolated nodes in the view of the Hi-C matrix as an adjacency matrix of a undirected graph, we proceed with a further preprocessing step. Let M be the observed Hi-C matrix of a given chromosome of length L (with unmappable regions removed), and let $\bar{M}$ be a transformed Hi-C matrix with reduced dynamic range

$$\bar{M_{ij}} = \begin{cases} \log(M_{ij}) & M_{ij} \neq 0 \\ 0 & otherwise \end{cases}$$

where the logarithmic function is introduced to reduce the large dispersion of the raw Hi-C matrix data due to the intrinsic drop of counts with increasing genomic distances between pairs of loci along the DNA.

3.6.2 Hi-C data normalization methods

In Hi-C maps, normalization is performed to correct systematic biases, making them more comparable and downstream analysis reliable. In fact, the interaction frequencies of contact matrix contain many unwanted biases that are derived from different systematic deviation in experimental procedures and driven by DNA sequence and different technical variations, such as fragment length, sequence mappability, copy number variations and other unknown factors. It is believed that these biases lead to anomalous variability among Hi-C raw data. In Hi-C analysis workflow, the normalization methods attempt to remove these biases as far as possible in order to preserve the interaction frequen-

cies reflecting the underlying architecture. A large number of these methods are available. Our datasets undergo different types of normalizations: Iterative Correction and Eigenvectors decomposition (ICE), Vanilla-Coverage (VC) and Sequential Component Normalization (SCN). In particular, ICE normalization [24] is an implicit individual-sample approach which attempts to make all bins of contact matrix equally visible using a matrix-balancing strategy. An implicit approach assumes that the cumulative effect of bias is fully captured in the sequencing coverage of each bin. The ICE normalization considers the systematic biases between two bins to be the product of their individual biases and the maximum likelihood solution is obtained by iterative correction procedures for determining the individual biases. Specifically, it seeks iteratively for systematic biases that equalize the sum of counts per bin in the matrix. At each iteration, a new matrix is generated by dividing each cell by the product of the sum of counts in its row times the sum of counts in its column. The process converges to a matrix in which all bins have identical sum. For a raw matrix M of linear size N, the normalized matrix W of entries W_{ij} is iteratively computed for each step n as

$$W_{ij} = \frac{M_{ij}}{\sqrt{\sum_{n=0}^{N} W_{in} \times \sum_{n=0}^{N} W_{nj}}}$$

This normalization has usually a quite strong effect, and visually the matrices look very smooth and regular as can be seen in figure 3.7.

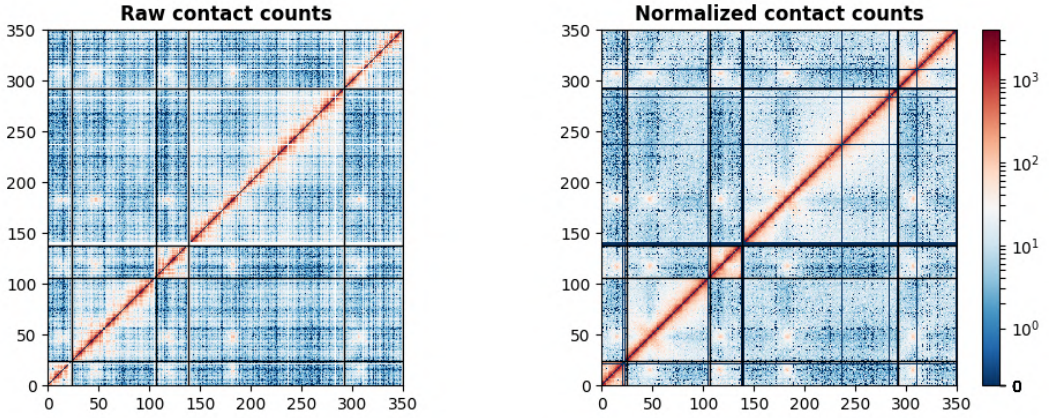


Figure 3.7: Hi-C data comparison between raw and ICE normalization matrices. The sample data set comes from the Python package *iced* which includes the first 5 chromosomes of the budding yeast *S. cerevisiae* [43].

The Vanilla-Coverage normalization is just a variation of the ICE where a single iteration is performed. The Sequential Component Normalization [4] is an implicit approach which can be applied to any genomic contact map and independently from the protocol that was used to generate it. Firstly, SCN normalization will give an equal weight to each restriction fragment in the contact map. Therefore, the fragments with very low number of reads, corresponding to ones that could not be properly detected, are likely to introduce noise in the normalized contact map and have to be removed. A way to identify these fragments is to compute the distribution of reads in the contact map which is roughly gaussian, with a long tail corresponding to low interaction fragments as shown in figure 3.8.

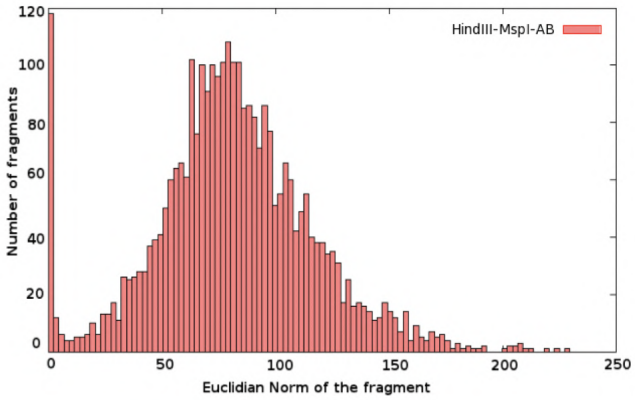
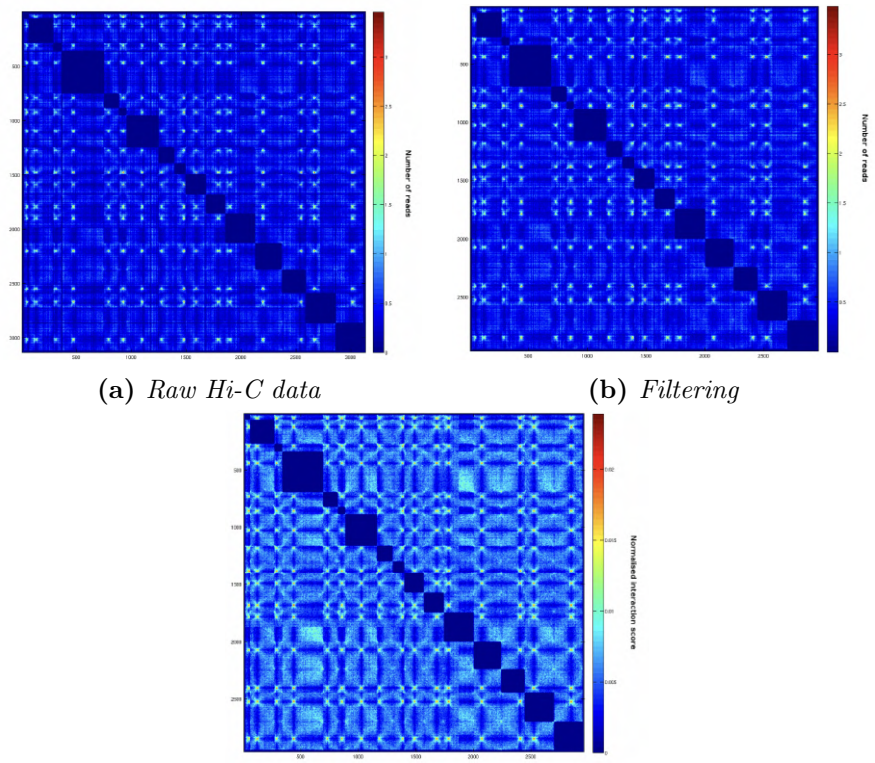


Figure 3.8: Distribution of the norms for each fragment for the interchromosomal interactions in the experiment HindIII-MspI-Conditions AB. It remains low interacting fragments that are removed. [4].

Once low interacting fragments are removed, the second step is to normalize all rows and columns of the contact map to one so that the matrix remains symmetric. This is done through the following procedure. Firstly, each column vector is normalized to one, using the euclidian norm. Then each line vector of the resulting matrix is again normalized to one. The whole process is repeated sequentially until the matrix is symmetric once again, now with each row and each column normalized to one. Usually, two or three iterations are sufficient to insure convergence. This normalization can be viewed as a sequence of extensions and shrinking of interaction vectors so that they tend to reach the sphere of radius one in the interactions space. An example of SCN normalized Hi-C matrix starting from the raw and by filtering its interaction fragments is shown in figure 3.9.



(a) *Raw Hi-C data* **(b)** *Filtering*

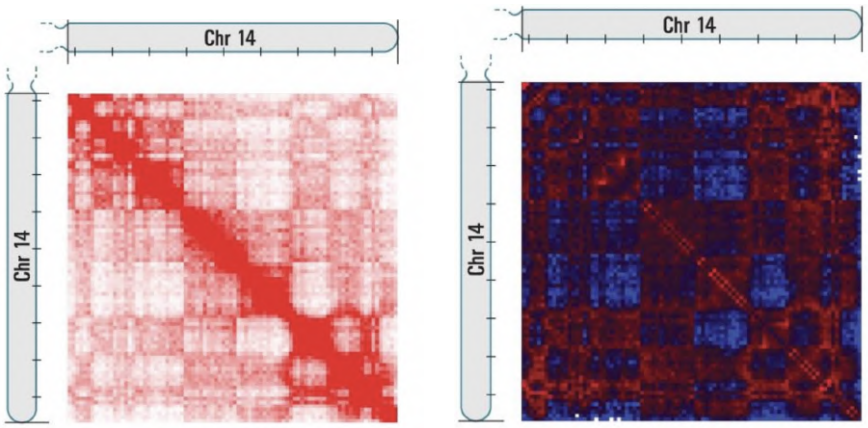
(c) *SCN Hi-C data*

Figure 3.9: Inter-chromosomal contact maps for *S. cerevisiae* at diffent steps of the normalization procedure. The fisrt contact map represents interchromosomal interactions from raw data **(a)**. In the second matrix **(b)**, low interacting fragments have been removed. Finally, the last contact map **(c)** is the matrice obtained after processing with the SCN. The colorbar indicates the number of reads. [4].

A different type of normalization very often used for Hi-C maps, independent from those mentioned so far, is the Observed-over-Expected (OoE) normalization. It is carried out for discounting the overall dependence of the entries an Hi-C matrix M_{ij} on genomic distance $s = |i - j|$ (as seen in figure 2.6). The entries of the OoE normalized matrix W_{ij} are defined as

$$W_{ij} = \frac{M_{ij} \times s}{\sum_{\{(m,n)|s=|m-n|\}} M_{mn}}$$

which corresponds to normalize the entries through the average values of M_{ij} at a genomic distance s. In figure 3.10 are shown the raw and OoE Hi-C maps in the case of chromosome 14. This kind of normalization is particularly convenient to enhance the interchromosomal interaction patterns, which are otherwise suppressed by the chromosome blocks along the main diagonal.



(a) *Raw Hi-C data* **(b)** *OoE Hi-C data*

Figure 3.10: (a): Map of chromosome 14 at a resolution of 1 Mb (1 tick mark = 10 Mb) exhibits substructure in the form of an intense diagonal and a constellation of large blocks (three experiments combined, range: $0-200$ reads).**(b):** The Observed-over-Expected matrix shows loci with either more (red) or less (blue) interactions than would be expected given their genomic distance (range: 0.2–5 reads) [32].

3.7 ShRec3D algorithm

The shortest-path reconstruction in 3D (ShRec3D, henceforth) is a two-step alternative algorithm adapted from network analysis for translating contact maps into distances, followed by a 3D reconstruction. The main goal is to derive three-dimensional chromosomal structures from Hi-C contact maps by combining the shortest-path distance, i.e. the length of the shortest path relating any two nodes on the graph, with classical multidimensional scaling (MDS). The latter find the solution to the mathematical problem of reconstructing a spatial structure from the distances between its elements involving the computation of the first three eigenvectors from an intermediary matrix, the Gram matrix. An essential step in MDS-based methods of chromosome reconstruction is therefore the derivation of a complete set of distances from a contact map. The steps used by the ShRec3D algorithm are summarized in figure 3.11.

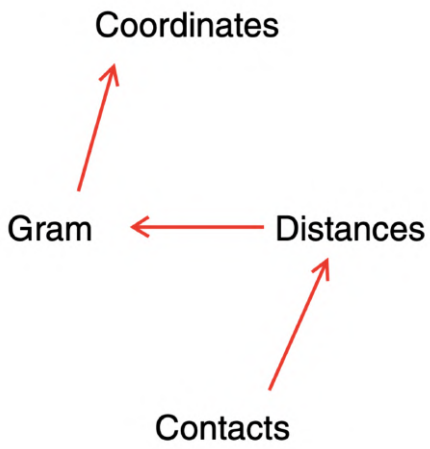


Figure 3.11: ShRec3D algorithm flowchart. From the contact map to the spatial coordinates. Adapted from [30].

Before to describe each single step of the ShRec3D algorithm it is necessary to define the framework in which it works. It deals with matrices which can be associated with a structure comprising N distinct points P_i $(i = 1, ..., N)$ in an n-dimensional space. In our case we consider experimental structures in the 3D space ($n = 3$). The origin O of the coordinate system describing the 3D points is taken to be their barycentre, because they are more suitable for structure visualization from a geometrical point of view.

3.7.1 Floyd-Warshall algorithm: shortest path analysis

The first step of the ShRec3D algorithm consists in deriving the full set of distances encoded in the distance matrix from the contact Hi-C map, using the concept of shortest path in graph theory. The distance matrix D is defined as an $N \times N$ matrix whose elements D_{ij} correspond to the Euclidean distance between the points P_i and P_j. In the case of Hi-C experiments, we considered the graph of nodes P_i $(i = 1, ..., N)$ defined by the Hi-C contact map seen as its weighted adjacency matrix M, where the link (i, j) between nodes i and j is endowed with a length equal to the inverse of the normalized contact frequency f_{ij}. The graph has to be connected, as it would not be possible to assign a distance between points belonging to two distinct components. The connectivity assumption means that for any pair of points P_i and P_j of the graph, it is always possible to find a path $(i_0, i_1, ..., i_k)$ with $i = i_0$ and $j = i_k$, such that $M_{ij}^k > 0$ for a strictly positive integer k. From a mathematical point of view, this condition coincides with the request for M to be irreducible, which is satisfied in our case of Hi-C experimental situation considered. Now, the ShRec3D algorithm defines the distance between two points by the length of the shortest path relating them. The latter is a path between the points P_i and P_j whose path length is minimal over all the paths relating them. Although the shortest path is not necessarily unique, its length takes a unique value. For computing the shortest paths and their length it is used the Floyd-Warshall algorithm. This algorithm is able to compare the shortest distances between every pair of vertices in the input graph. It does so by incrementally improving an estimate on the shortest path between two vertices, until the estimate is optimal. Consider the graph defined by the adjacency matrix M as described before. Further consider a function D_{ij}^k which returns the length of the shortest path from i to j such that the intermediate vertices are only from the set $V^k = \{1, 2, ..., k\}$. Our goal is to find the length of the shortest path from each i to each j considering any vertices from $V = \{1, 2, ..., N\}$. This is what we define as D_{ij}^N, which we will find recursively. Firstly, it is important to note that the inequality $D_{ij}^k \leq D_{ij}^{k-1}$ holds for any choice of k. In fact there is more chance of finding a shortest path if the graph allowed to use an additional vertex k, while the equality holds only when k is not an intermediate vertex on the shortest path $D_{ij}^{k-1} = D_{ij}^k$. However, if the vertex

k is such that $D_{ij}^{k} < D_{ij}^{k-1}$, then there must be a path from i to j using the vertices V^k that is shorter than any other path that does not make use of the vertex k. This path (see figure 3.12) can be break down into two different (shorter, otherwise we could further decrease the length) paths:

- a path D_{ik}^{k-1} from i to k using the vertices V^{k-1};
- a path D_{kj}^{k-1} from k to j using again the vertices V^{k-1}.

Interestingly, in either case, the subpaths contain merely nodes from V^{k-1}.

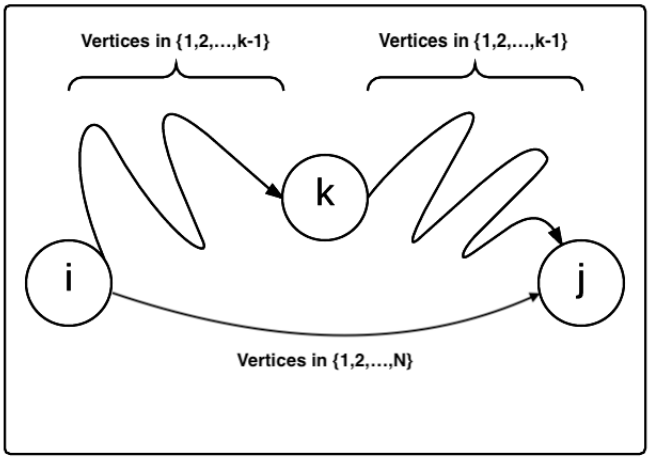


Figure 3.12: Graph showing the shortest path decomposition from node i to node j in $\{1, 2, ..., N\}$ into the sum of two paths passing through the node k with all the intermediate vertices in $\{1, 2, ..., k-1\}$ [14].

Therefore, from all of these observations we derive the following formula:

$$D_{ij}^{k} = min\left(D_{ij}^{k-1}\ ,\ D_{ik}^{k-1} + D_{kj}^{k-1}\right)$$

This equation represents the core of the Floyd-Warshall algorithm. It allows us to found the shortest path for all (i, j) pairs using any intermediate vertices. The algorithm proceeds by computing the shortest path D_{ij}^{k} for $k = 0$, which is basically $D_{ij}^{0} = w_{ij}$, where w_{ij} denotes the weight of the edge from i to j if exits or inf otherwise. Then the algorithm works for $k = 1$, $k = 2$ and so on by using the formula above. The process continues until $k = N$, where the shortest path D_{ij}^{N} is found as required. Therefore we have obtained the distance matrix D, whose entries are defined by $D_{ij} = D_{ij}^{N}$. Usefully, weak or vanishing contact

frequencies do not contribute to the distances, as the shortest paths will overstep the corresponding links (of large or infinite lengths) by construction. This method thus makes it possible to both reconstruct the whole set of distances and filter some of the experimental noise (low contact frequencies that may correspond to noise are rejected). Of great importance, this method defines a true distance. In fact, on one hand it is certainly symmetrical $D_{ij} = D_{ji}$ and vanishes only if the points are identical $D_{ij} = 0 \iff i = j$. On the other hand, by construction, the minimal path length from node i to node j is always smaller or equal to the sum of the minimal path length from node i to some node k with the minimal path length from such node k to node j. Thus, the shortest-path distance satisfies the triangular inequality $D_{ij} \leq D_{ik} + D_{kj}$ (with equality when a shortest path from i to j passes through k). Finally, the distance between neighboring nodes along the genome will be very small as desired (see figure 3.13), in fact the closer the loci are along the genome, the more they establish contacts; this distance is thus consistent with the polymer-like connectedness of each chromosome.

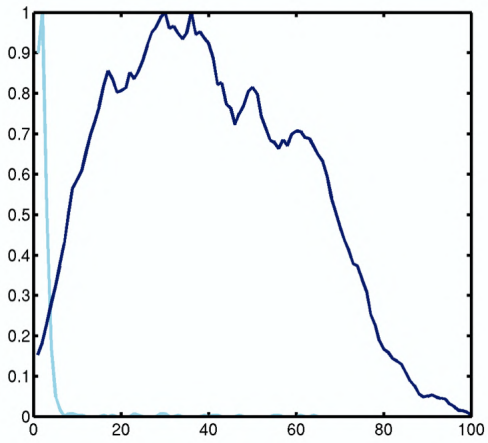


Figure 3.13: Polymer connectivity in ShRec3D reconstruction. Normalized histogram of the reconstructed distances $D_{i,i+1}$ between neighbors along the genome (light blue peaked curves), for genome-wide real Hi-C data, compared to the normalized histogram of all distances taken as a reference (dark blue broad curves) [30][27].

3.7.2 Gram matrix

In the first step we were therefore able to obtain the distance matrix D starting from the Hi-C contact matrix. Now we need to pass from the distances to the Gram matrix, following the flowchart in figure 3.11. The Gram matrix G is an $N \times N$ positive semidefinite matrix whose elements G_{ij} are defined as the scalar product of the coordinate vectors associated with points P_i and P_j. From a mathematical point of view it is possible to derive the Gram matrix from the knowledge of distances as demonstrated in Theorems 3.1 and 3.3 of [20] and a more tractable form for our purposes are shown in [30]. Firstly, we can express the distance d_{0i} between the barycenter O and the point P_i for any $i = 1, ...N$ as

$$d_{0i}{}^2 = \frac{1}{N}\sum_{j=1}^{N} D_{ij}{}^2 - \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k>j}^{N} D_{jk}{}^2$$

Moreover, we have use for an auxiliary matrix M, the metric matrix. It is defined by the elements

$$M_{ij} = \frac{1}{2}\left[d_{0i}{}^2 + d_{0j}{}^2 - D_{ij}{}^2\right]$$

In [41] it is shown that to ensure M to be positive semidefinite of rank n, the condition that D is a distance matrix associated with an Euclidean structure of N point in a space of dimension n has to be assured. Satisfactorily, this is exactly the definition of distance matrix, then M coincides with the Gram matrix G of the N points.

3.7.3 Coordinate matrix and 3D structure

The problem of reconstructing the coordinates from the sole knowledge of distances has been faced by the distance geometry. Multidimensional scaling (MDS) reconstruction, which brings in the notion of dimensional reduction, is able to find the n-dimensional structure ($n = 3$ in our case) to approximate as much as possible a given distance matrix. It is based on analytical formulas which can be found in [29] and [49], in which the purpose of MDS is extended beyond distance matrices with application to chromosome reconstruction. One of the most important theorems of distance geometry states that, given the Gram matrix to be positive semidefinite, the coordinates of N points P_i with

$i = 1, ..., N$ in a space of dimension n can be recovered from the first n eigenvectors E_α with $\alpha = 1, ..., n$ of the Gram matrix suitably normalized to 1 and rescaled by the square root of the corresponding eigenvalues λ_α, in formula

$$V_\alpha^{(i)} = E_\alpha^{(i)} \times \sqrt{\lambda_\alpha} \quad with \quad \sum_{i=1}^{N} {E_\alpha^{(i)}}^2 = 1$$

where $E_\alpha^{(i)}$ is the i-th component of the eigenvector E_α and $V_\alpha^{(i)}$ is the coordinate of the point P_i along the α-axis. This equation is presented and demonstrated in a part of the Theorem 3.1 of [41]. Finally, we are able to reconstruct the $n \times N$ coordinate matrix V comprising the Euclidean coordinates of the points $V_\alpha^{(i)}$. By supplementing the Floyd-Warshall algorithm to find the shortest path for the distance matrix, the Gram matrix and the coordinate matrix, we end up with the constructive algorithm ShRec3D. It allows us to display a 3D structure starting from any contact map, included the Hi-C contact maps.

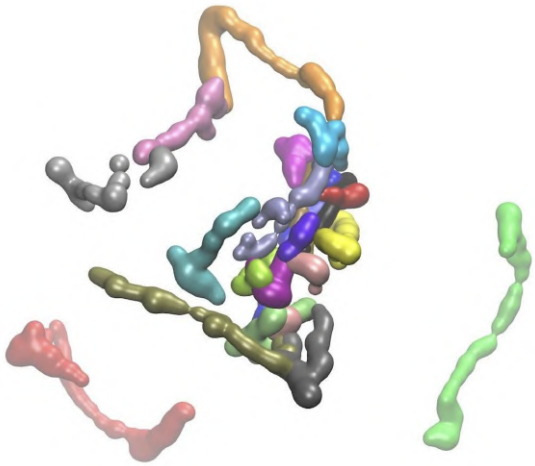


Figure 3.14: Visualization of human autosomal chromosomes using ShRec3D. Color labeling of the different chromosomes: 1: blue, 2: red, 3: grey, 4: orange, 5: yellow, 6: gold, 7: silver, 8: green, 9: pink, 10: cyan, 11: purple, 12: lime, 13: mauve, 14: ochre, 15: ice blue, 16: black, 17: light green,18: light cyan, 19: violet,20: magenta, 21: dark red, 22: light orange. (Hi-C data in lymphoblastoid cells) [30].

The reconstructed 3D coordinates are defined up to an arbitrary dilation, rotation, and possibly mirror symmetry and we will face this problem in the next chapter. Moreover, MDS truncates the metric matrix M into the posi-

tive semidefinite Gram matrix G of rank-3. Therefore the coordinates reconstruction takes into consideration only the dominant $n = 3$ (i.e., $\alpha = 1, 2, 3$) eigenvalues and associated eigenvectors of the metric matrix M, as if the other were vanishing. The validity of this dimensional reduction (from M to G) and subsequent step of coordinate reconstruction (from G to V) is assessed by examining the spectrum of M and ensuring that the largest three eigenvalues are separated by a large spectral gap from the remaining spectrum concentrated near 0, as seen in figure 3.15.

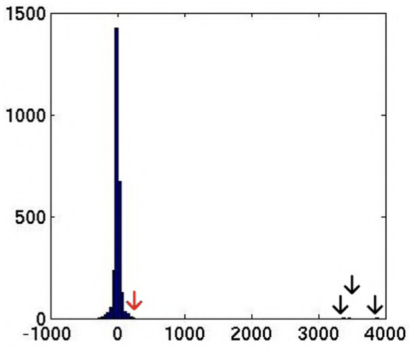


Figure 3.15: Spectrum (eigenvalue histogram) of the metric matrix derived from simulated contact maps using the shortest-path distance using the ShRec3D approach (both distances are dimensionless: the units on the abscissa axis depend on the chosen normalization for the contact frequencies). The three rightmost black arrows underline the first three eigenvalues and the leftmost red arrow underlines the fourth one, demonstrating the presence of a significant spectral gap [30].

All the steps carried out to get to the coordinates for the nodes of the chromosomes starting from a contact Hi-C map have been implemented in a Python code which, for convenience, is reported in the appendix C.

3.8 Procrustes Analysis

In the previous chapter we have seen how to generate a coordinate matrix which can be visualized as a series of points in the 3D space. A tool that will prove useful in the next analysis of the thesis is to be able to compare the shapes generated by the coordinate points provided by the ShRec3D algorithm in space. Firstly, is necessary to give some definitions regarding the concept of shape and landmark. With shape it is denoted all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. Thus, it can be considered as a member of an equivalence class formed by removing the translational, rotational and uniform scaling components, which are called Euclidean similarity transforms (see figure 3.16).

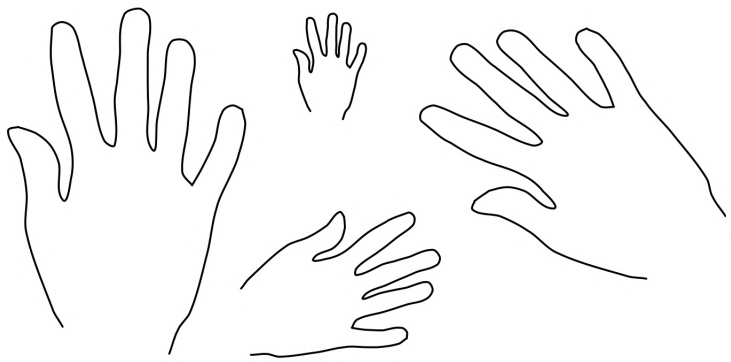


Figure 3.16: Four copies of the same shape, but under different Euclidean transformations [45].

One way to describe a shape is by locating a finite number of points along the outline of the shape. These points are called landmarks (see figure 3.17). They are divided into three groups (according to [23]): (1) anatomical landmarks, points assigned by an expert that corresponds between organisms in some biologically meaningful way; (2) mathematical landmarks, points located on an object according to some mathematical or geometrical property, i.e. high curvature or an extremum point and (3) pseudo-landmarks, constructed points on an object either on the outline or between landmarks.

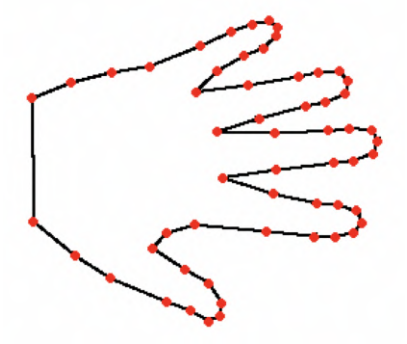


Figure 3.17: Example of how landmarks (red points) are used to represent a shape. Adapted from [1].

A mathematical representation of an N-point shape in n dimensions could be to concatenate each dimension into a $n \times N$-vector. The vector representation for three-dimensional shapes (i.e. $n = 3$) would then be

$$\mathbf{x} = (x, y, z)^T = (x_1, x_2, ..., x_N, y_1, y_2, ..., y_N, z_1, z_2, ..., z_N)^T$$

Now, according to the definition stated above, we need to perform the filtering of translation, rotation and uniform scaling transforms from the original shape, e.g. the one generated from ShRec3D coordinates, to obtain a true shape representation. This is carried out by introducing a coordinate reference to which all shapes are aligned, commonly known as pose. The procedure for finding such a coordinate reference is referred to as Procrustes analysis (from the Greek mythology, is the bandit who made his victims fit his bed either by stretching their limbs or cutting them off). It brings the shape set into shape space. The latter is defined as the set of all possible shapes of the object considered. Formally, the shape space σ_n^N of dimension K is the orbit shape of the non-coincident N point set configurations in the $\mathbb{R}^n$ under the action of the Euclidean similarity transformations [45]. In particular, the shape space is affected by a loss of dimensionality, due to the alignment procedure, with respect to that of a space of N points and dimension n. The latter is equal to nN, instead the shape space dimensionality is $K = nN - n - 1 - \frac{n(n-1)}{2}$. The peeling of dimensionality is easily explained by the translation, which removes n dimensions, the uniform scaling, which removes 1 dimension and the rotation, which removes $\frac{n(n-1)}{2}$ dimensions. In case it is possible to establish a relationship between the distance in this new shape space and the Euclidean distance in the original space, thus the set of shapes forms a Riemannian man-

ifold containing the object class in question, e.g. the hands or the ShRec3D reconstructions from Hi-C data. Specifically, this space is also denoted as Kendall shape space [12] and this relationship is known as shape metric. An example of shape metric is the Procrustes distance, which compare shapes with an equal amount of points N, as it will be useful in the further analysis. It corresponds to a least-squares type shape metric that requires two aligned shapes with one-to-one point correspondence. The alignment process involves three steps: translation, uniform scaling and rotation. The translational components can be removed by translating the object so that its centroid $(\bar{x}, \bar{y}, \bar{z})$ lies on the origin, formally

$$(x, y, z) \to \left(x - \frac{1}{N}\sum_{j=1}^{N} x_j,\ y - \frac{1}{N}\sum_{j=1}^{N} y_j,\ z - \frac{1}{N}\sum_{j=1}^{N} z_j \right)$$

Likewise, the scale component can be cut out by scaling the object so that the root mean square distance (RMSD) between the points to the translated origin is normalized to one. The RMSD is defined as

$$S = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \left[(x_j - \bar{x})^2 + (y_j - \bar{y})^2 + (z_j - \bar{z})^2 \right]}$$

where S is a statistical measure of the object's scale or size. It turns out to be 1 when the point coordinates are divided by the object's initial scale, so when the following coordinate transformation is applied

$$(x, y, z) \to \left(\frac{x - \bar{x}}{S}, \frac{y - \bar{y}}{S}, \frac{z - \bar{z}}{S} \right)$$

The last step consists in removing the rotation component, which is represented by a 3×3 rotation matrix R. To find the optimum value for R the singular value decomposition (SVD) can be used [12]. We will not analyze the SVD as it is beyond the scope of this thesis, but what is important to know is that one of the object is used to provide a reference orientation while the other is rotated around the origin until an optimum angle θ such that the sum of the squared distances (SSD) between the corresponding point is minimised is found. Finally, we are left with a new set of coordinates $(x, y, z) \to (u, v, w)$ which is optimally superimposed to the other one (see figure 3.18).

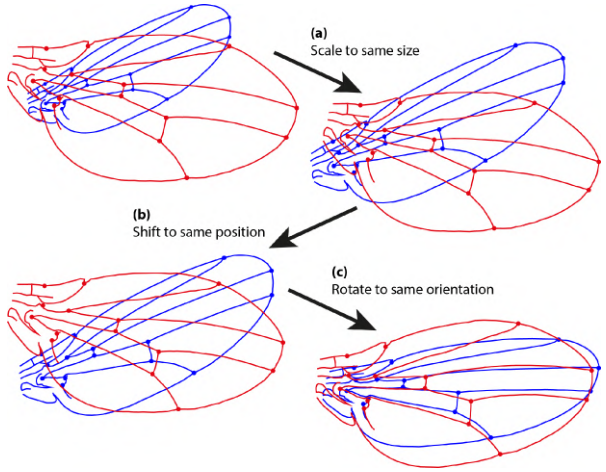


Figure 3.18: Example of Procrustes superimposition. Three transformation steps of an ordinary Procrustes fit for two configurations of landmarks. **(a)** Scaling of both configurations to the same size; **(b)** Transposition to the same centroid position; **(c)** Rotation to the orientation that provides the minimum sum of squared distances between corresponding landmarks [37].

Now, it is of relevant importance for us to be able to evaluate the difference between the shape of two objects by translating, scaling and optimally rotating them as illustrated above. As a statistical measure of the shape difference we take the square root of the aformentioned SSD between the corresponding points. Mathematically,

$$P_d = \sqrt{(u - x')^2 + (v - y')^2 + (w - z')^2}$$

where (x', y', z') is the vector representation for the 3D shape to which (u, v, w) is superimposed, as a result of the alignment procedure. This measure P_d is referred to as Procrustes distance.

Chapter 4

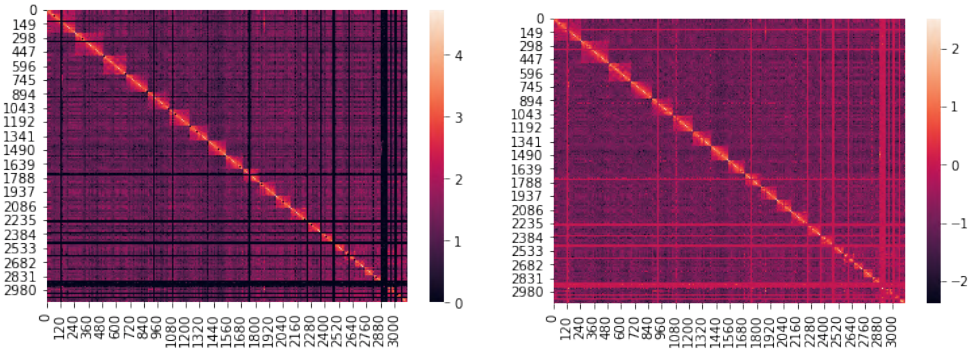
Results and discussion

In this chapter we present the results obtained from the spectral analysis starting from different types of Hi-C data at different resolutions. This analysis consists in identifying the eigenvalue spectrum of the considered Hi-C matrices and in extracting the essential matrices starting from the projectors related to the signal component. Ultimately we dealt with the production of synthetic images working instead with the noise component and leaving the signal component intact. All these Hi-C matrices mentioned above were then analyzed from different points of view and compared both with the aid of scatter plots and the ShRec3D algorithm to reconstruct their spatial coordinates.

4.1 Preliminary analysis: GM12878

Before going into the spectral analysis on single chromosomes, we want to verify and characterize the type of Hi-C data starting from one of the most common case studies in the literature: the GM12878 cell line. We therefore started from the complete cell line which includes all 23 chromosomes characterizing the human karyotype at 1Mb resolution. The data available to us on GM12878 were of two types: raw and ICE normalized Hi-C data. The former correspond to real contact counting of neighbouring loci matrices, while the latter are data processed through the normalization of the Iterative Correction and Eigenvectors decomposition (ICE) as described in paragraph 3.6.2. In order to be able to visualize the Hi-C maps, some preliminary operations on the data have been performed for the cell line considered. Firstly, the data have been imported in Python, then the logarithmic function to base 10 has been

applied to the whole data set. This has been necessary to visualize properly the contact frequencies and all the hidden patterns of the data due to the great difference between the main diagonal's values and those corresponding to the contact regions outside it. By exploiting *seaborn*, a Python data visualization library, we are able to visualize the heatmaps, i.e. color-encoded matrix plot, of the healthy cell line both for the raw and the ICE normalized data as shown in figure 4.1.



(a) *Raw Hi-C data* **(b)** *ICE normalized Hi-C data*

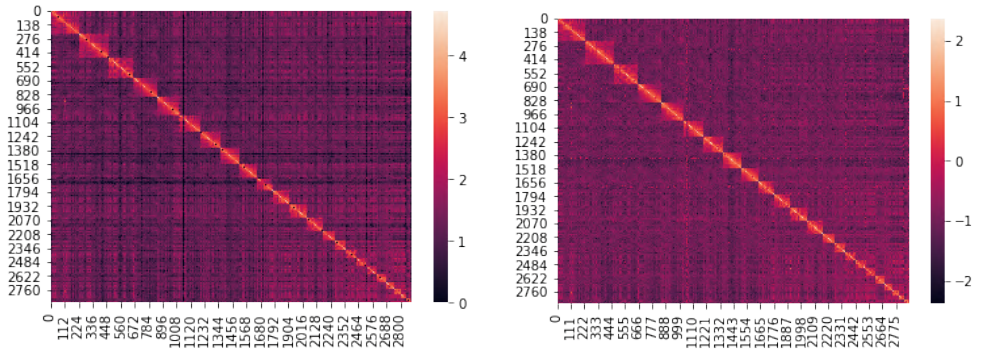
Figure 4.1: Raw and ICE normalized Hi-C maps for the whole GM12878 cell line matrices after taking the logarithmic function to base 10.

The metadata, in particular the intervals of coordinates within the matrices in figure 4.1, which indicate the beginning and the end of each chromosome, are the same both for the raw and ICE normalized Hi-C contact maps and they are listed in table 4.1.

Chromosome	Start	End	Chromosome	Start	End
1	1	250	12	2116	2249
2	251	494	13	2250	2365
3	495	693	14	2366	2473
4	694	885	15	2474	2576
5	886	1066	16	2577	2667
6	1067	1238	17	2668	2749
7	1239	1398	18	2750	2828
X	1399	1554	20	2829	2892
8	1555	1701	Y	2893	2952
9	1702	1843	19	2953	3012
10	1844	1979	22	3013	3064
11	1980	2115	21	3065	3113

Table 4.1: Metadata regarding the coordinate intervals for each chromosome of the GM12878 cell line.

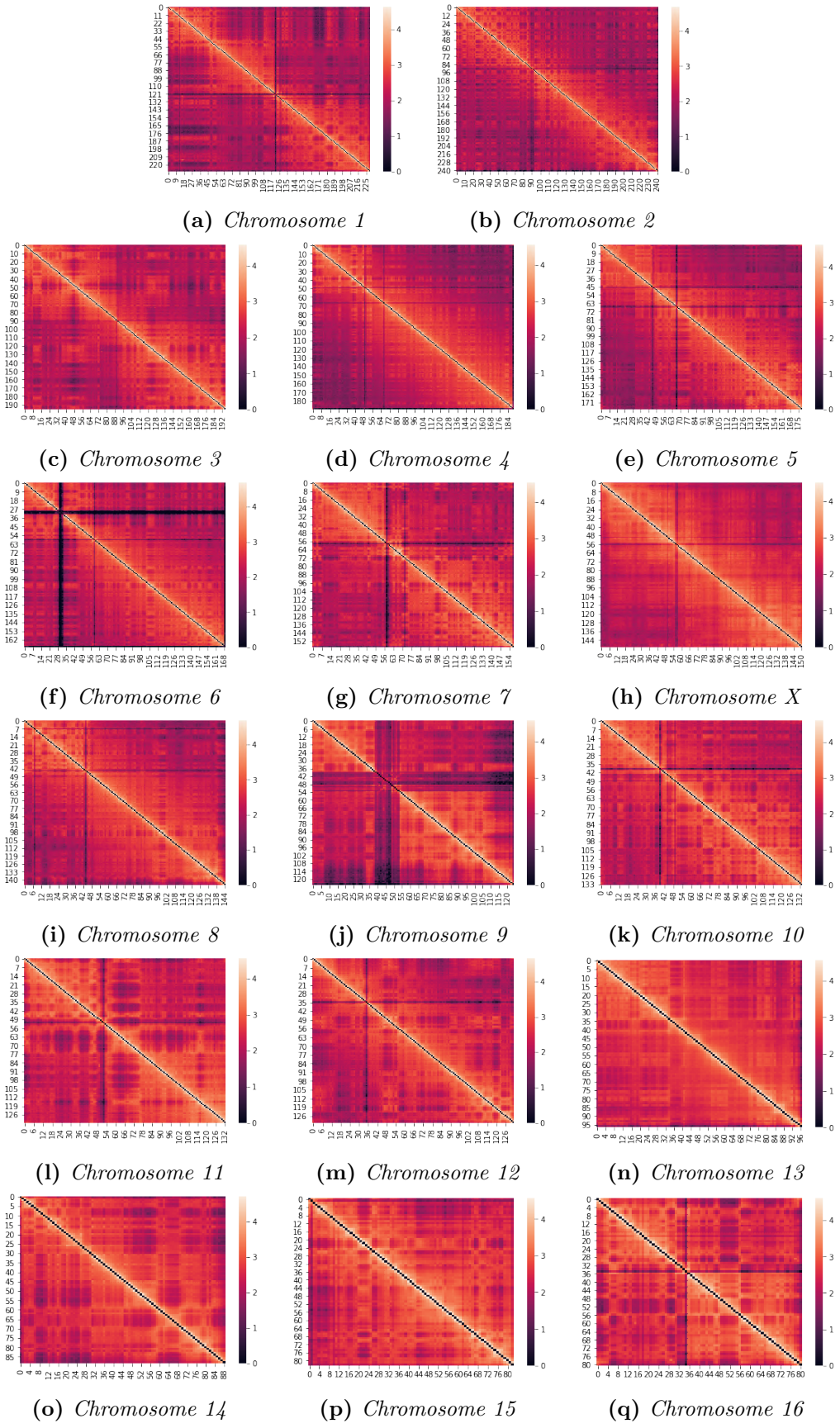
The black lines visible in figure 4.1 are indicators of no contact and/or very little contact among the different chromosomes. These lines are due to default holes (usually) present in the middle and end of the reference genome. To delete the superfluous black lines we used a python code capable of identifying the corresponding indexes of the rows and therefore of the columns of the (symmetric) matrix which add up to zero and removes them from the display. Moreover we have also removed the block and the corresponding coordinates of the Y chromosome since we are facing a female cell line, therefore the Y chromosome is not present, even if by default the Hi-C data taking procedure also includes its sequencing. After these preliminary procedures the Hi-C contact maps are shown as in figure 4.2).



(a) *Raw Hi-C data* **(b)** *ICE normalized Hi-C data*

Figure 4.2: Raw and ICE normalized whole Hi-C maps for the whole GM12878 cell line matrices after taking the logarithmic function to base 10, removing the Y chromosome and the rows and columns that sum up to zero.

Before moving on to the spectral analysis of the entire matrix, we want to make sure that the properties and characteristics of the Hi-C data can actually be traced back to those that have been observed so far by experts and that have been discussed in paragraph 2.3.2. Specifically, we want to see how the average value of the contacts within the individual chromosomes (contact probability) scales as a function of the genomic distance, i.e. distance in base pairs along the nucleotide sequence. To do this we first need to extract the submatrices relating to the single chromosomes from the Hi-C contact maps. In figures 4.3 and 4.4 we present all the chromosomes for the raw Hi-C contact maps, which are arranged in order of size, as in the original whole matrix in the arrangement as blocks along the main diagonal.



(a) *Chromosome 1* **(b)** *Chromosome 2*

(c) *Chromosome 3* **(d)** *Chromosome 4* **(e)** *Chromosome 5*

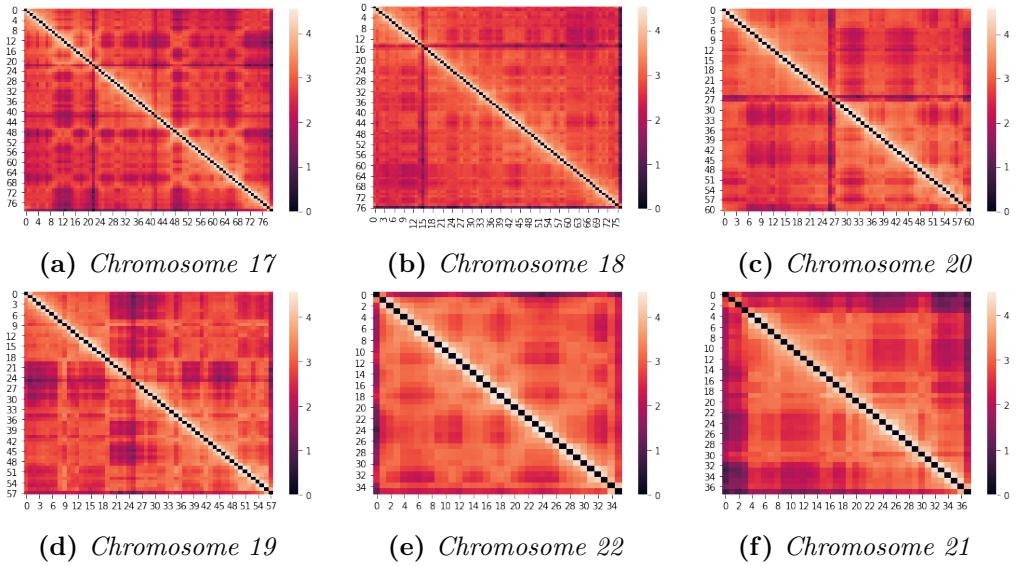
(f) *Chromosome 6* **(g)** *Chromosome 7* **(h)** *Chromosome X*

(i) *Chromosome 8* **(j)** *Chromosome 9* **(k)** *Chromosome 10*

(l) *Chromosome 11* **(m)** *Chromosome 12* **(n)** *Chromosome 13*

(o) *Chromosome 14* **(p)** *Chromosome 15* **(q)** *Chromosome 16*

Figure 4.3: Single chromosomes extracted from the whole raw Hi-C contact matrix.



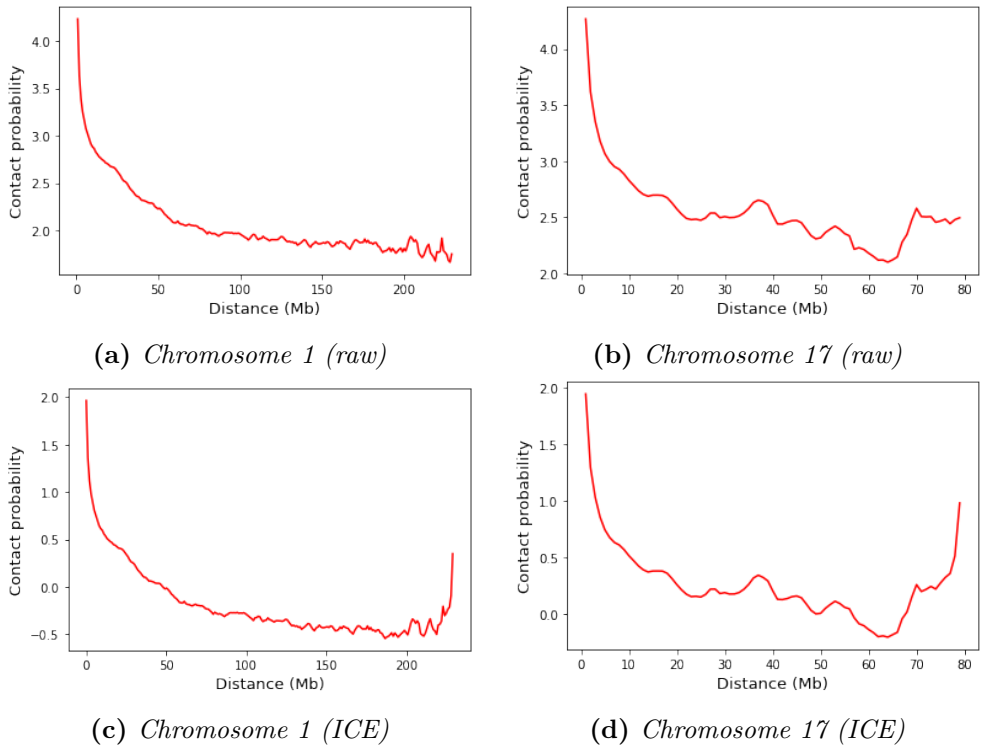
(a) *Chromosome 17* **(b)** *Chromosome 18* **(c)** *Chromosome 20*

(d) *Chromosome 19* **(e)** *Chromosome 22* **(f)** *Chromosome 21*

Figure 4.4: Single chromosomes extracted from the whole raw Hi-C contact matrix.

To test whether our data are consistent with known features of genome organization, such as chromosome territories (the tendency of distant loci on the same chromosome to be near one another in space), we compute the average intrachromosomal contact probability for pairs of loci separated by a certain genomic distance on that chromosome. We therefore calculated the average value of contact frequency starting from the matrices corresponding to each single chromosomes in figures 4.3 and 4.4 by varying the genomic distance. To achieve this goal, a Python code has been implemented to compute the average values of the entries for the single chromosome matrix considered and for each diagonal, which corresponds to a difference of 1 Mb in the genomic distance. The graphs of the contact probability as a function of the genomic distance for both the raw and the ICE normalized Hi-C maps are shown in figure 4.5, in which for convenience of displaying just the chromosome 1 and 17 are shown. The graphs related to the other chromosomes are presented in the appendix E.

From the plots we can notice that the contact frequency approximately decreases monotonically on every chromosome, suggesting polymer-like behavior in which the three-dimensional distance between loci increases with increasing genomic distance (diagonal in the Hi-C matrices). These findings are in agreement with the chromosome conformation capture results as described in paragraph 2.3.2. However in the final part of the graph it can be noticed that there are greater fluctuations. These are due to the fact that for diagonals

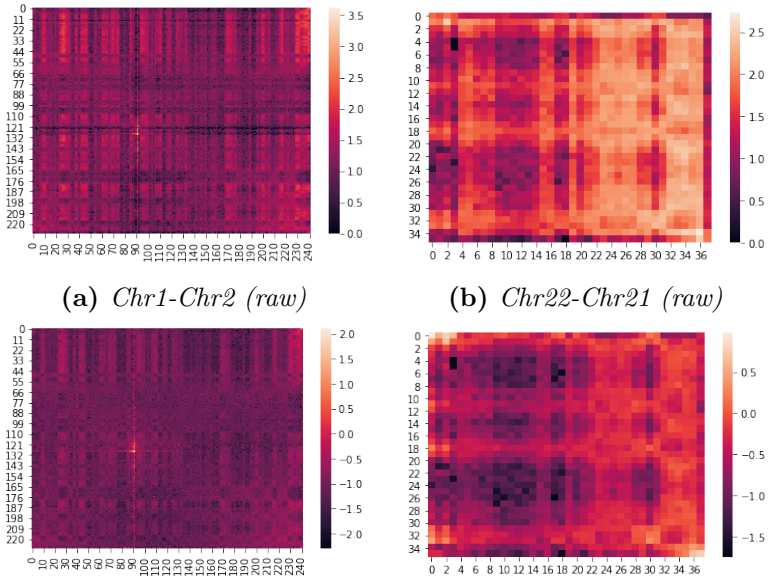


(a) *Chromosome 1 (raw)* **(b)** *Chromosome 17 (raw)*

(c) *Chromosome 1 (ICE)* **(d)** *Chromosome 17 (ICE)*

Figure 4.5: Contact probability graph as a function of the genomic distance (matrix diagonal) in Mb in case of single chromosomes 1 and 17 for both raw (top line) and ICE normalized (bottom line) Hi-C data.

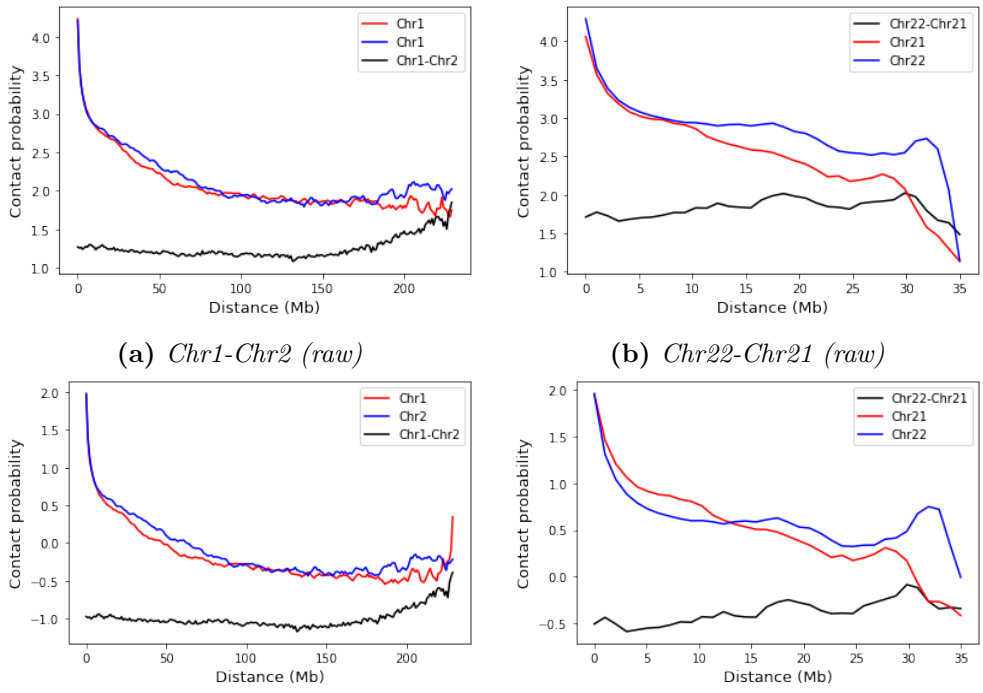
very far from the main one, there is a limited number of data from which we extract the average value, so the contact probability value may have significant differences, especially in case of smaller chromosomes such as 17, 21 and 22.
Once we have ascertained the correspondence between the intra-chromosomal interactions with the expectancy, we want to verify the trend for the inter-chromosomal ones. Just as an example we have taken the submatrices related to the interaction between chromosome 1 with chromosome 2 and chromosome 22 with chromosome 21. The choice has been made in order to compare both chromosomes with a large number of coordinates and those with smaller ones. The corresponding submatrices extracted from the whole Hi-C map both in case of raw and ICE normalized data are shown in figure 4.6.
As in the case of single chromosomes, we can make a graph showing the relation between contact probability and genomic distance. In figure 4.7 we can see a unique graph comprising both the single chromosomes and their corresponding interactions both for the raw and ICE normalized Hi-C data.



(a) *Chr1-Chr2 (raw)* **(b)** *Chr22-Chr21 (raw)*

(c) *Chr1-Chr2 (ICE)* **(d)** *Chr22-Chr21 (ICE)*

Figure 4.6: Inter-chromosomal interactions matrices extracted from the whole raw (top line) and ICE normalized (bottom line) Hi-C contact matrices for the interaction between chromosome 1 with chromosome 2 **(a,c)** and chromosome 22 with chromosome 21 **(b,d)**.



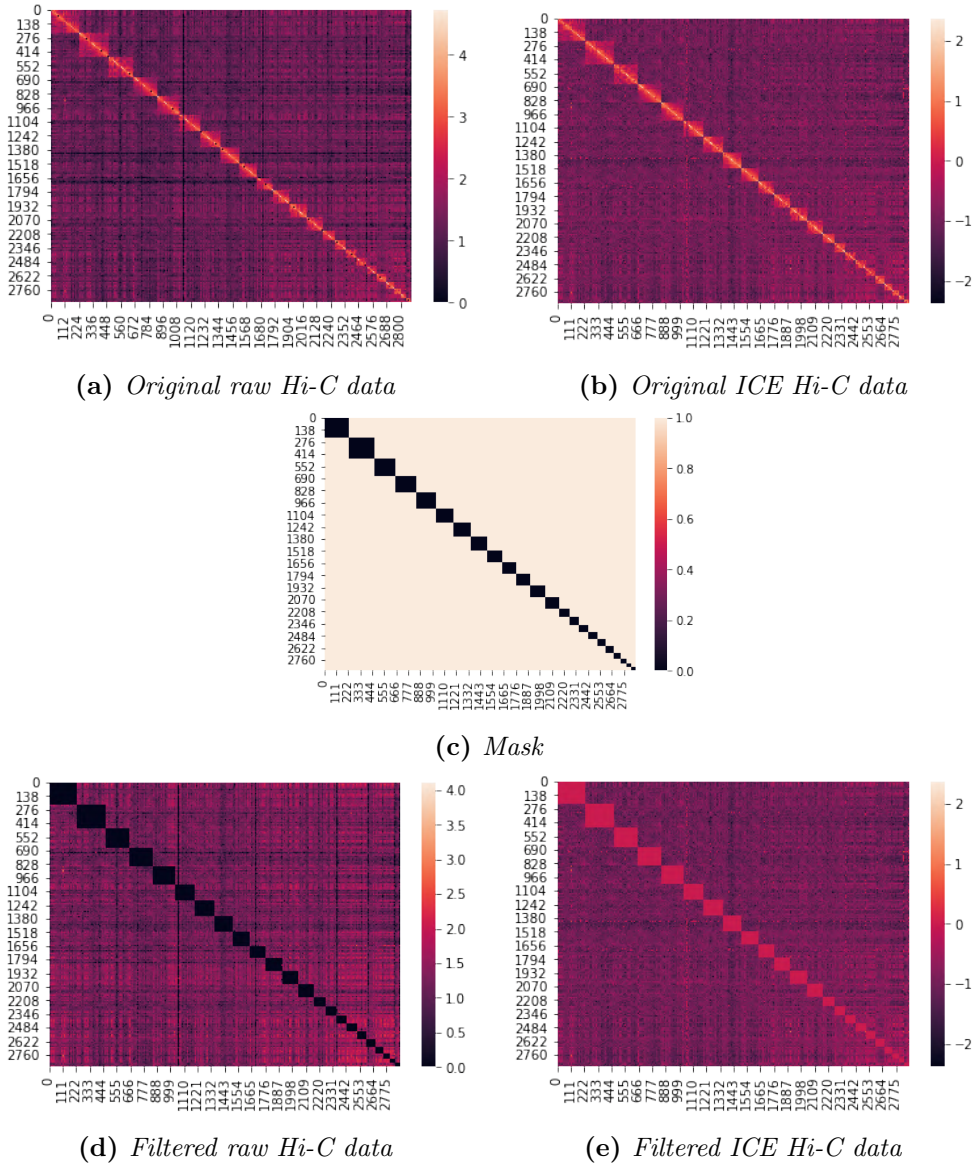
(a) *Chr1-Chr2 (raw)* **(b)** *Chr22-Chr21 (raw)*

(c) *Chr1-Chr2 (ICE)* **(d)** *Chr22-Chr21 (ICE)*

Figure 4.7: Contact probability graphs as a function of the genomic distance (diagonal) in Mb for both single chromosomes and their interaction for the whole raw (top line) and ICE normalized (bottom line) Hi-C contact matrices.

The graphs in figure 4.7 show a comparison between the contact probability trend for the individual chromosomes and the interaction between them. From the latter it can be seen that regarding the interactions between chromosomes, unlike those within the same chromosome in which we can observe the contact probability decay, the trend remains almost constant regardless of the genomic distance considered. Moreover, even at distances greater than 200 Mb, as seen for the longer chromosomes 1 and 2, the contact probability is always greater than the average contact probability between different chromosomes. This implies the existence of chromosome territories. Regarding the chromosomes 22 and 21, it can be seen that the fluctuations are more intensified than those of chromosomes 1 and 2. This is due to the fact that they are smaller and therefore, in particular for diagonals far from the main one, the contact probability values can undergo significant variations. Anyhow, it can be seen that, apart from values very far from the main diagonal in which fluctuations are noted, the values of contact probability relating to intrachromosomal interactions is always greater than that of interchromosomal interactions, regardless of the particular pair of chromosomes and the raw or ICE normalized Hi-C data considered.

Now we want to extend this procedure trying to visualize the trend of contact probability also for the whole matrix, both for ICE and raw Hi-C maps. In particular we are interested in characterizing interchromosomal interactions as a whole, as a function of their genomic distance and independently of the particular pair of chromosomes whose interaction is considered. To do this it is necessary to exclude from the analysis the single chromosomes organized in blocks along the diagonal. We have therefore implemented a Python code capable of creating a mask (see figure 4.8**(c)**) to select only the blocks along the diagonal, which are related to the single chromosomes. From the latter, boolean variables were then assigned to the single entries of the whole matrix in order to be able to select only the part of interaction external to the main blocks (see figure 4.8**(d,e)**). Once the desired matrix is obtained, it is therefore possible to calculate the average value of the contact probability of the k-th diagonal relating only to the interchromosomal interactions (see figure 4.9).



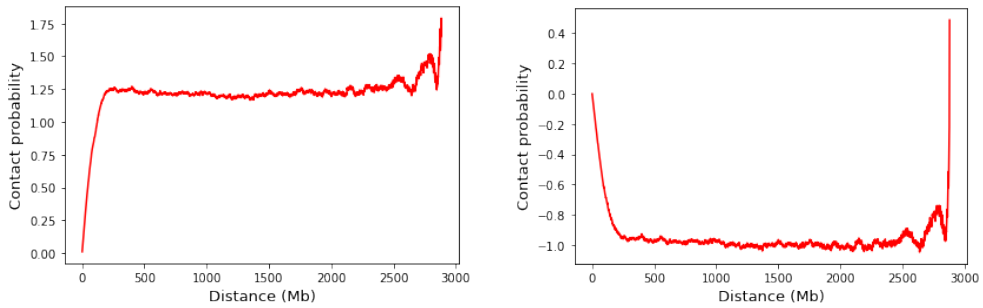
(a) *Original raw Hi-C data* **(b)** *Original ICE Hi-C data*

(c) *Mask*

(d) *Filtered raw Hi-C data* **(e)** *Filtered ICE Hi-C data*

Figure 4.8: Comparison among original raw and ICE normalized Hi-C data **(a,b)** with the corresponding filtered ones **(d,e)** by using the mask **(c)**.

As can be seen from the figure 4.9, the trend that follows the contact probability in case of whole matrix, both raw and ICE normalized, is similar to that relating to interchromosomal interactions, rather than intrachromosomal ones. This is an expected result, given that we are analyzing interchromosomal interactions, albeit related to different chromosomes pairs. A constant trend can therefore be seen for almost all genomic distances (main diagonals of the whole matrix). However, a significant difference can be noted in the contact probability values of the initial and terminal part of the graph, i.e. at very



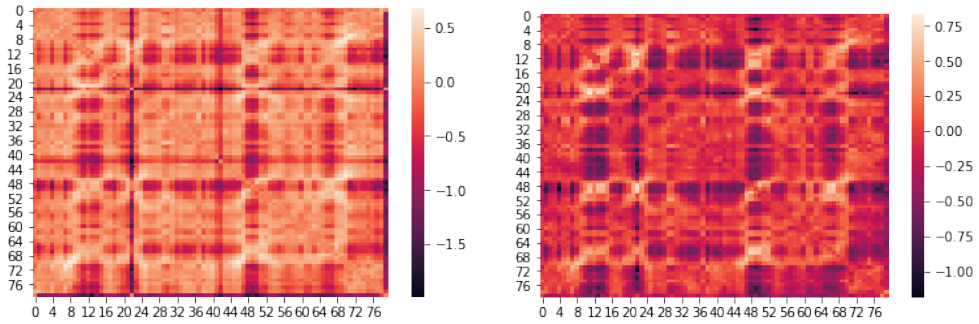
(a) *Raw Hi-C data* **(b)** *ICE normalized Hi-C data*

Figure 4.9: Contact probability graph as a function of the genomic distance (diagonal) in Mb for both the whole raw and ICE normalized filtered Hi-C contact matrices.

small or very large genomic distances. These large fluctuations, as in case of contact probability graphs observed so far (see figures 4.5 and 4.7), are due to the lack of data with which the average intensity of each entry is calculated along a given diagonal k of the matrix. The terminal part of the graph has fluctuations which can be explained (as in the case of individual chromosomes) by the fact that the whole Hi-C matrix, as the distance of the diagonal considered with respect to the main diagonal increases, contains less data. However, here we are dealing with the filtered total matrix, then it should be considered that even for diagonals very close to the main one, corresponding to small genomic distances, there is a very significant lack of data. This is due to the fact that in the filtered matrix all the blocks relating to the single chromosomes are excluded, which are placed along the main diagonal, as seen in figure 4.8. A further confirmation is given by the fact that just after a distance k in Mb equal to about the size of chromosome 2 (it is the largest, containing about 240 Mb as shown in table 4.1) the contact probability trend starts to become constant, meaning that from that genomic distance forward we are effectively taking all the average contact values between chromosomes. By adding up these two effects it is therefore possible to explain the trend displayed in the figure 4.9, which characterizes the interchromosomal interactions as a whole. We have then validated the fact that we are dealing with a Hi-C map that reflects the contact probability characteristics both in case of single chromosomes and their interactions, both for the raw and ICE normalized matrices, which so far do not exhibit any significant differences.

4.1.1 Spectral analysis for single chromosomes

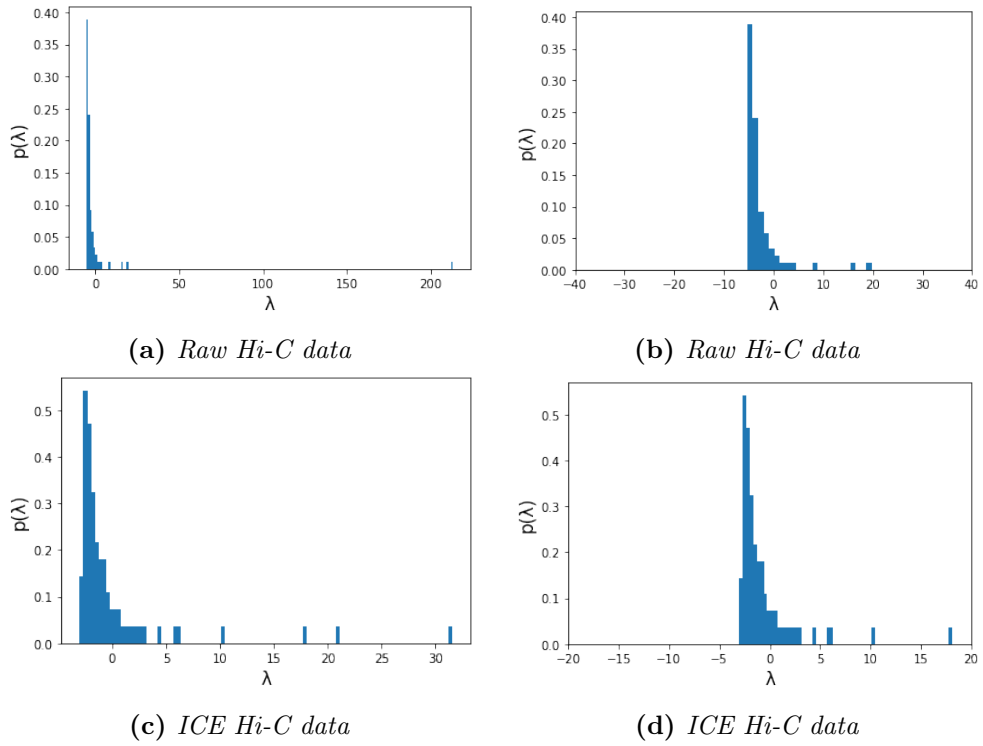
In the first place, it is important to check whether our spectral analysis on the Hi-C data related to the cell line GM12878 at 1 Mb resolution both in case of raw and ICE normalized data provide compatible results. Currently, for the sake of simplicity we restrict the analysis to the chromosome 17 (for the analysis of the whole matrix see the appendix A). After extracting the submatrix related to the chromosome 17 from the entire Hi-C matrix containing all the human genome, we take the Observed-over-Expected normalization, which discounts genomic-distance biases and thus puts on equal footing interactions at different sequence separations, for both the raw (see figure 4.10**(a)**) and ICE normalized data (see figure 4.10**(b)**), after taking the base 10 logarithm in order to compare them in the following spectral analysis.



(a) *Raw Hi-C data* **(b)** *ICE Hi-C data*

Figure 4.10: Raw **(a)** and ICE normalized **(b)** Hi-C data matrices of the chromosome 17 after the OoE normalization and taking also the base 10 logarithm.

Now we make the spectral decomposition of the OoE normalized matrices (after taking the base 10 logarithm) by computing the eigenvectors and the corresponding eigenvalues. In figure 4.11 histograms of the eigenvalues spectrum are shown for each Hi-C data type (raw and ICE normalized). The latters are accompanied by the relative zoom around zero to highlight the trend of the peak. In this way it is possible to better distinguish the shape of the histogram trend in the region where we expect, based on what we have seen in paragraph 3.2, a symmetric function that follows that of Wigner's semicircle. As regards instead the eigenvalues outside the peak around zero, we expect them to be distributed in an isolated and sparse manner. From the eigenvalue distributions in figure 4.11 we can recognise that all of them shared a spectrum different from a symmetric distribution around zero as we expected for a

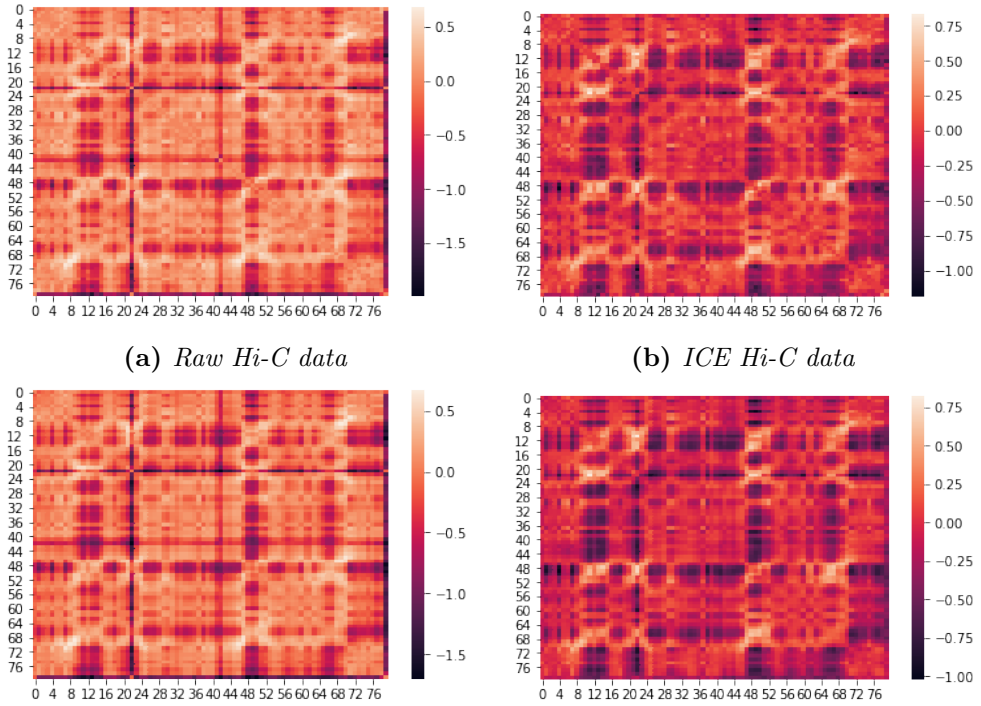


(a) *Raw Hi-C data* **(b)** *Raw Hi-C data*

(c) *ICE Hi-C data* **(d)** *ICE Hi-C data*

Figure 4.11: Histograms of the eigenvalue probability distribution $p(\lambda)$ obtained from the raw **(a)** and ICE normalized **(c)** Hi-C data matrices of the chromosome 17 after the OoE normalization and taking the base 10 logarithm with the corresponding zoom around zero **(b,d)**.

symmetric random matrix. It can still be appreciated that there is a distribution of eigenvalues concentrated in the proximity of zero, while the others eigenvalues are dispersed far from it. In the first place we follow the approach seen in [17], whereby we can determine the essential Hi-C matrix (essHi-C), from the spectral summation of the $n^* = 10$ projectors formed by the eigenvectors related to the eigenvalues of greater absolute value (see figure 4.12**(c,d)**). It must be said that from the spectrum of the eigenvalues we do not notice the distribution given by the Wigner's semicircle law and just by seeing at the histograms in figure 4.11 it is visible that part of the first ten largest (in module) eigenvalues fall into the peak, then the essential matrix will plausibly also contain some projectors relating to the noise and not only to the signal component. However for the purposes of this analysis we will follow the same procedure. Furthermore, in this preliminary analysis it is also useful to verify if the matrix reconstructed starting from all the projectors associated to the all set of eigenvectors gives back the original matrix as expected (see figure 4.12**(a,b)**).

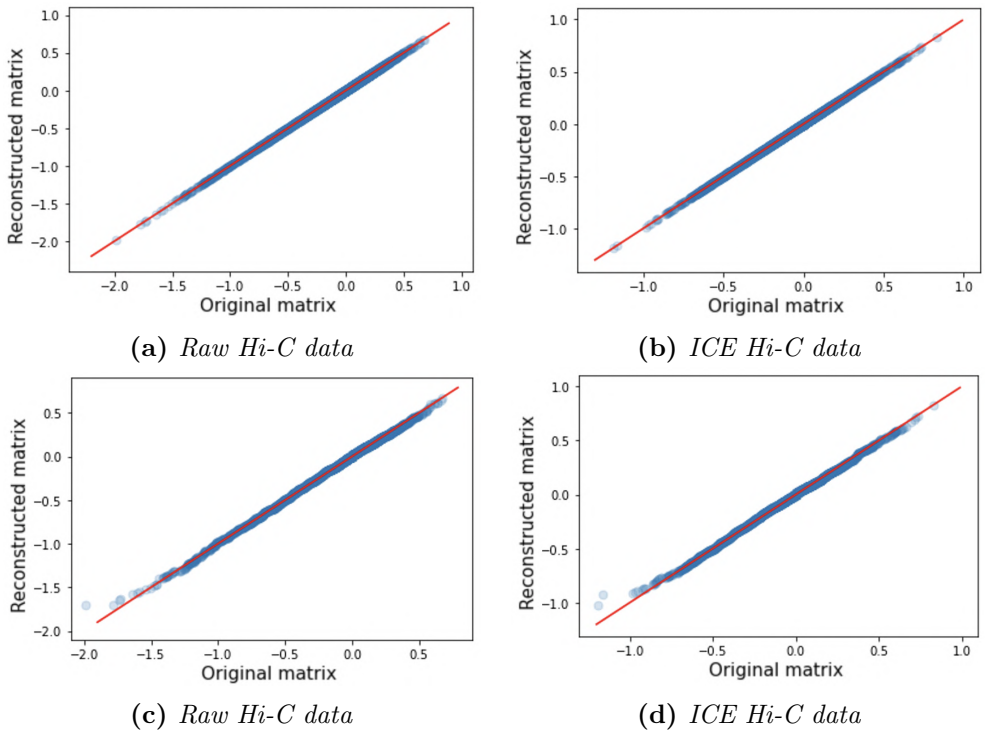


(a) *Raw Hi-C data* **(b)** *ICE Hi-C data*

(c) *Raw Hi-C data* **(d)** *ICE Hi-C data*

Figure 4.12: Raw and ICE normalized Hi-C data matrices of the chromosome 17 reconstructed starting from all **(a,b)** and the only first 10 **(c,d)** projectors associated to the corresponding eigenvectors after the OoE normalization and taking the base 10 logarithm.

Looking at the essential Hi-C matrices reconstructed using only the first 10 projectors in figures 4.12**(c,d)**, it is possible to notice how all the main patterns that characterize the interchromosomal interactions of chromosome 17 are preserved and enhanced, for both raw and ICE normalized Hi-C maps. In fact we can see how the interaction patterns are more marked than in the original matrices, this is due to the fact that many of the projectors that refer to the noise component have been eliminated from the reconstruction. To check whether the reconstructed matrices are really compatible with that of the original matrices, we make the scatter plots and compute the Pearson correlation coefficient ρ. Since the Hi-C matrices are symmetrical, to avoid double counting of the points within the graph we extracted the upper triangular matrix and compared their values with a scatter plot. The graphs are shown in figure 4.13 with the regression line in red and the Pearson correlation coefficients listed in the table 4.2.



(a) *Raw Hi-C data* **(b)** *ICE Hi-C data*

(c) *Raw Hi-C data* **(d)** *ICE Hi-C data*

Figure 4.13: Scatter plots for raw and ICE normalized Hi-C data matrices of chromosome 17 reconstructed starting from all **(a,b)** and the only first 10 **(c,d)** projectors associated to the corresponding eigenvectors after the OoE normalization and taking the base 10 logarithm.

Figure 4.13	ρ
(a)	0.9999999999999998
(b)	0.9999999999999998
(c)	0.9995
(d)	0.9994

Table 4.2: Pearson correlation coefficient values related to each scatter plot in figure 4.13.

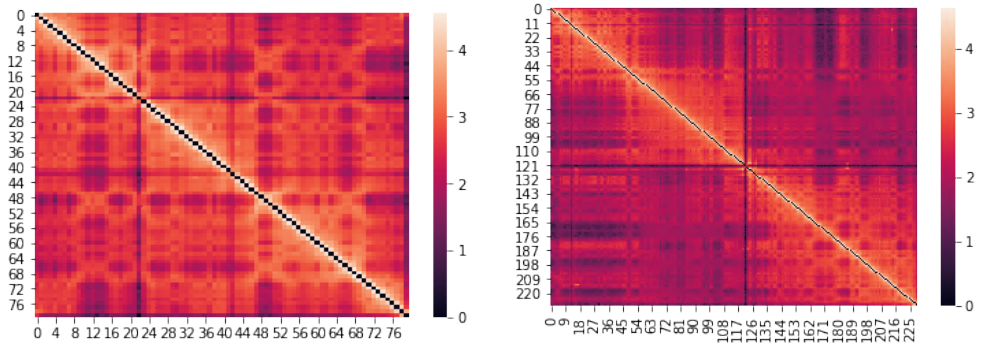
As we can see from the Pearson correlation coefficient values listed in table 4.2, regarding the matrices reconstructed by using all the projectors, related to the entire set of eigenvectors, they are almost one as expected for a linear correlation. However for the matrices reconstructed from the projectors associated to the corresponding eigenvectors related only to the greater 10 eigenvalues, the values of correlation coefficients decrease, even though they remain highly close to one. This is caused by the fact that the latters are reconstructed from not all the projectors or components, so they should not be identical to the original

matrices, but still close. All the considerations done so far remain valid both for the raw and ICE normalized Hi-C matrices. Therefore this analysis shows how both the matrices are equivalent in relation to the spectral properties of the reconstructed matrices. In this way in the following analysis we can choose between raw and ICE normalized matrices indifferently. Furthermore, in this paragraph we have performed the Observed-over-Expected normalization on the contact Hi-C matrix of the chromosome, which tries to dampen the dependence on the genomic distance of the contact value between the chromosome loci by normalizing the intensity with the average value of all the intensities found at that specific genomic distance. The latter therefore attempts to get rid of the decay of the contact value as the genomic distance increases, trying to uniform the entire matrix by bringing out the interchromosomal patterns regardless of whether the pairs of loci are far or close in the genomic sequence. This normalization is particularly useful when one wishes to observe the chromosome compartmentalization in the outermost corners of the Hi-C matrix which are otherwise challenging to bring out clearly. However, for the purposes of the present thesis work, whose ultimate goal is to produce synthetic data, it is necessary to start from a matrix that is as faithful as possible to the real data, trying to maintain all of its properties, including the characteristic decay which is intrinsic to the physics of chromatin folding within the nucleus. Moreover the Hi-C matrices with and without OoE normalization are strongly correlated ($\rho = 0.817$). For these reason, both statistical and biological, from now on we will deal with Hi-C matrices without an Observed-over-Expected normalization.

4.1.2 Chromosomes 1 and 17 inspection

Now we present a more in-depth spectral analysis on single chromosomes 1 and 17 taking the base 10 logarithm of the corresponding raw Hi-C maps. The choice fell on these two chromosomes since they present a variability both in biological terms and in terms simply of size. In fact, looking at the figures 4.14 of the single chromosomes, it can be seen that the chromosome 1, one of the largest, has different properties from chromosome 17. For example, within the Hi-C matrix relating to chromosome 1, we can distinctly identify the two blocks (square submatrices) along the diagonal which are related to the arms of the chromosome itself and the related interactions within the antidiagonal

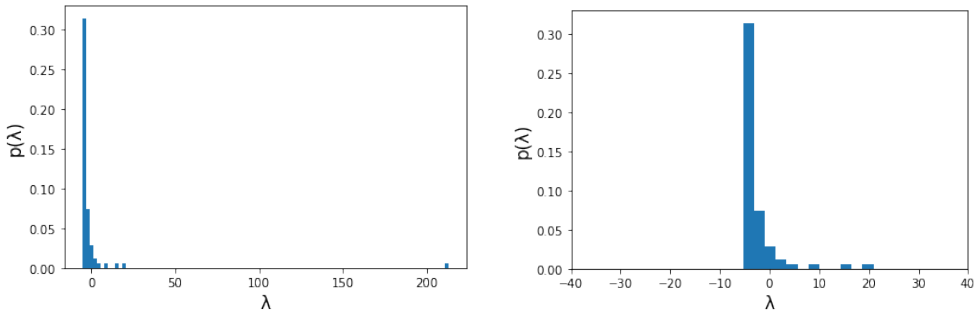
ones. All these properties are not found instead in chromosome 17, which is one of the smallest and which appears in its characteristic pattern structure of interaction between the loci. It was therefore avoided to consider even smaller chromosomes with respect to 17 since in that case it would not have been possible to capture well all those properties of the Hi-C matrix as such. The latter was also taken as a model in [17]. In this way we can have two proxies for the two extreme situations that can occur within the variability of chromosomes.



(a) *Chromosome 17* **(b)** *Chromosome 1*

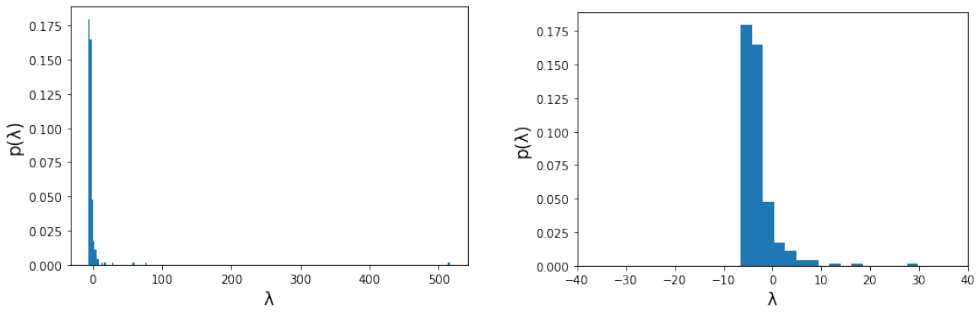
Figure 4.14: Chromosomes 17 **(a)** and 1 **(b)** extracted from the whole raw Hi-C matrix.

In particular, from the analysis of the eigenvalues spectrum, we should try to establish a threshold for which it is possible to separate the random part from the signal one which characterizes the Hi-C matrix spectrum as in figures 4.15 and 4.16.



(a) *Chromosome 17* **(b)** *Chromosome 17*

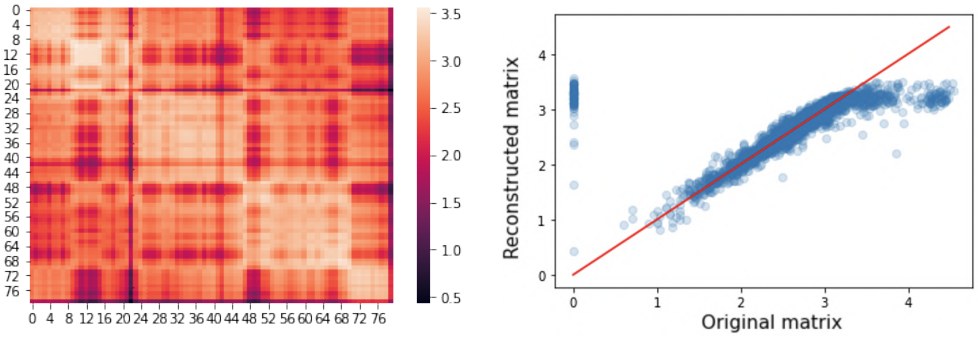
Figure 4.15: Histograms of the eigenvalues probability distribution $p(\lambda)$ for the chromosome 17 **(a)** with a zoom on the corresponding random component **(b)**.



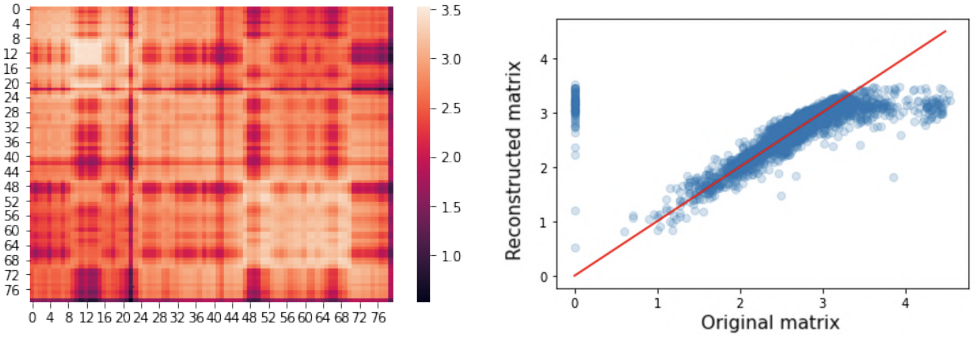
(a) *Chromosome 1* **(b)** *Chromosome 1*

Figure 4.16: Histograms of the eigenvalues probability distribution $p(\lambda)$ for the chromosome 1 **(a)** with a zoom on the corresponding random component **(b)**.

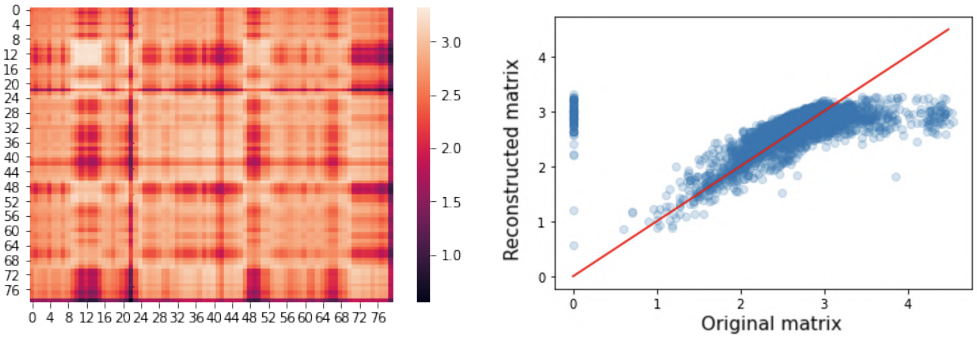
As can be seen from the histograms in figures 4.15 and 4.16, they are not at all symmetric as it should be for a random matrix, both for chromosome 1 and 17. This is probably due to the lower fluctuation of the data which are therefore more precise making the distribution asymmetric with respect to zero. Despite this we have selected the threshold taking as reference the value of the negative eigenvalue at the end of the peak of the distribution close to zero and we have chosen the modulus of that value as threshold. It follows that for chromosome 17 we selected as eigenvalues relating to the essential part of the matrix those such that $|\lambda| > 5$, while for chromosome 1 we chose $|\lambda| > 7$. We have therefore reconstructed the essential matrices starting from the maximum number of eigenvalues that are outside what we currently define as random component of the spectrum and gradually lowering their number $n^* = 4$ for chromosome 17 and $n^* = 8$ for chromosome 1. In fact, in this case study analysis we would try to establish how the essential Hi-C matrices change as the number of projectors used to reconstruct them decreases. The resulting EssHi-C matrices with the corresponding scatter plots with respect to the original ones for both the chromosomes 17 and 1 are shown in figures 4.17 and 4.18 respectively. In the next analysis, since there is not much difference in terms of reconstruction of chromosome 1 between the cases $n^* = 8$ and $n^* = 4$, we will use the latter to compare the two chromosomes.



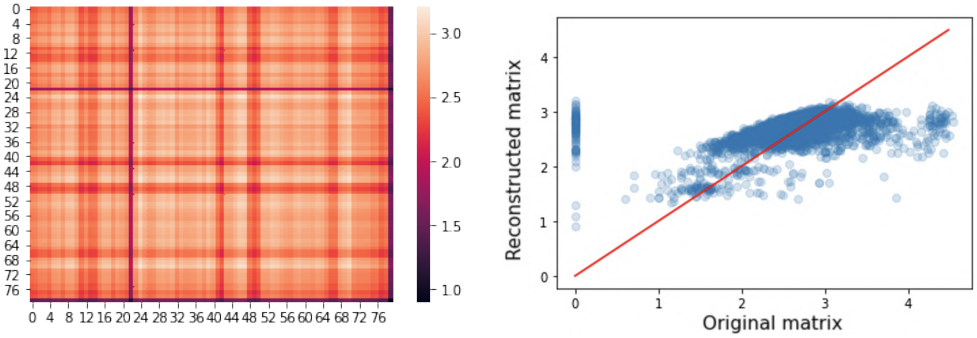
(a) *EssHi-C with* $n^* = 4$ **(b)** *Scatter plot* ($\rho = 0.575$)



(c) *EssHi-C with* $n^* = 3$ **(d)** *Scatter plot* ($\rho = 0.567$)

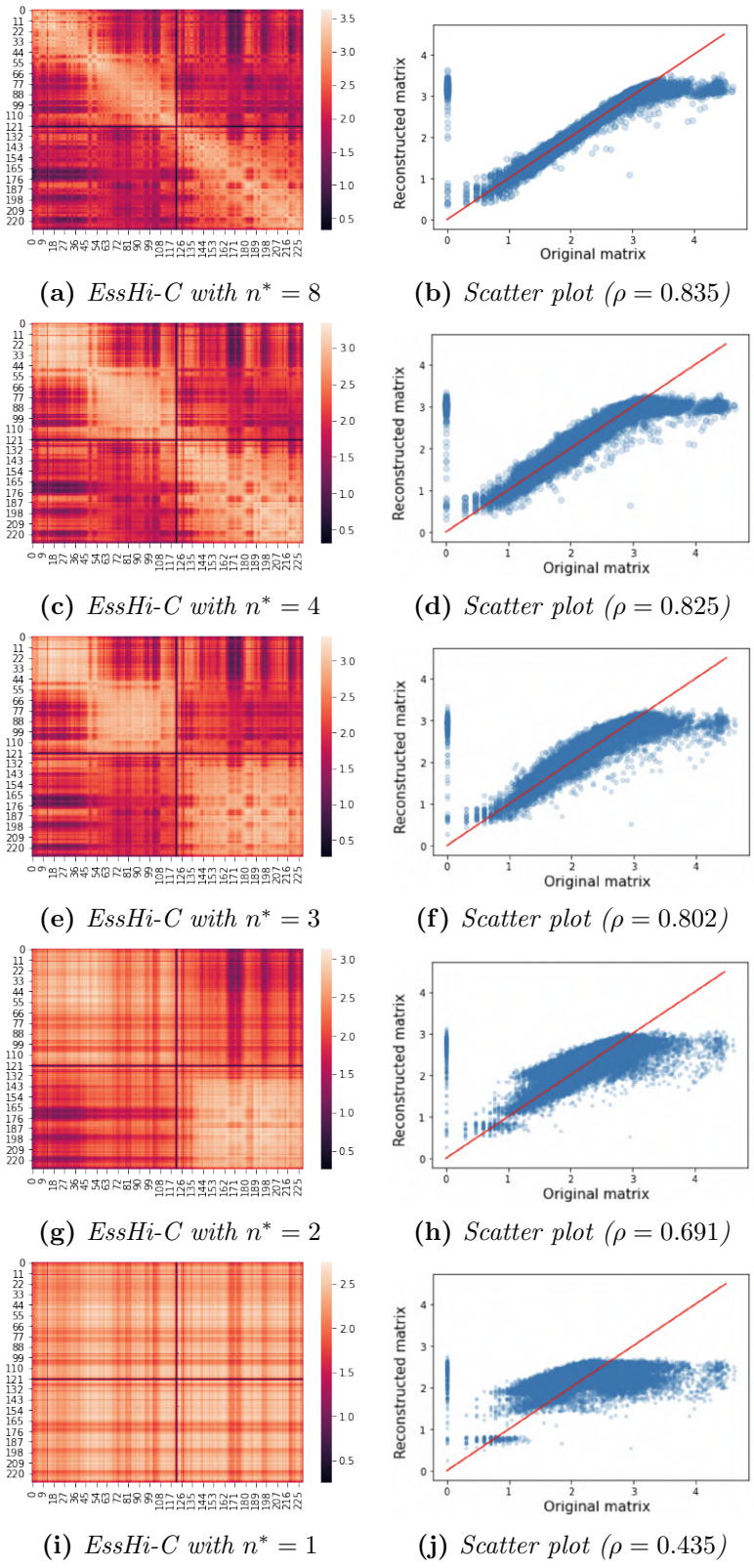


(e) *EssHi-C with* $n^* = 2$ **(f)** *Scatter plot* ($\rho = 0.516$)



(g) *EssHi-C with* $n^* = 1$ **(h)** *Scatter plot* ($\rho = 0.418$)

Figure 4.17: First column: essential matrices from the Hi-C data of the chromosome 17 reconstructed starting from different n^* highest-ranking projectors. **Second column:** the corresponding scatter plots between the original raw Hi-C map and the reconstructed one with the Pearson correlation coefficient ρ in brackets.



(a) *EssHi-C with* $n^* = 8$ **(b)** *Scatter plot* ($\rho = 0.835$)

(c) *EssHi-C with* $n^* = 4$ **(d)** *Scatter plot* ($\rho = 0.825$)

(e) *EssHi-C with* $n^* = 3$ **(f)** *Scatter plot* ($\rho = 0.802$)

(g) *EssHi-C with* $n^* = 2$ **(h)** *Scatter plot* ($\rho = 0.691$)

(i) *EssHi-C with* $n^* = 1$ **(j)** *Scatter plot* ($\rho = 0.435$)

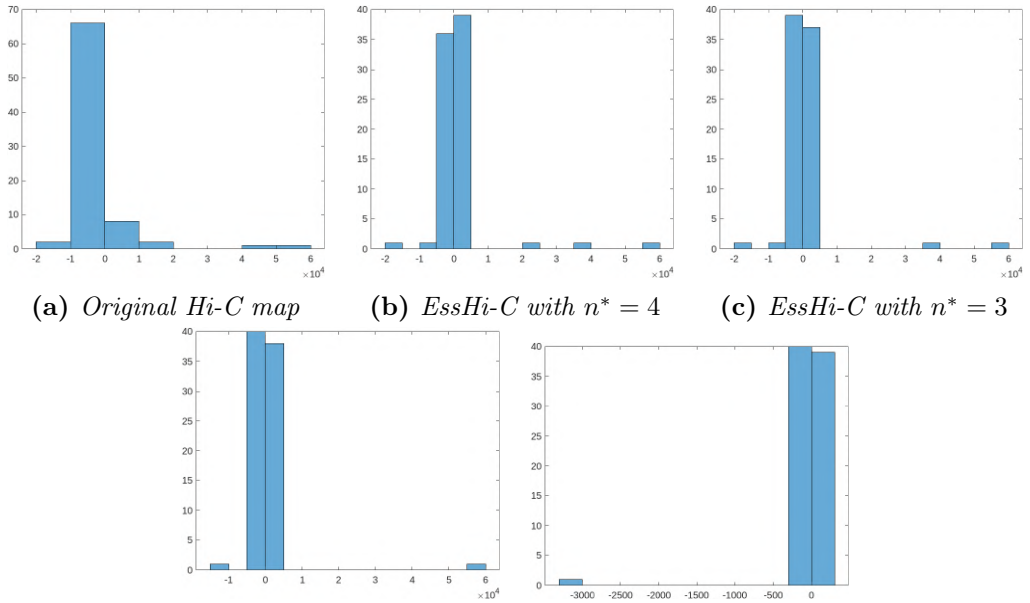
Figure 4.18: First column: essential matrices from the Hi-C data of the chromosome 1 reconstructed starting from different n^* highest-ranking projectors. **Second column:** the corresponding scatter plots between the original raw Hi-C map and the reconstructed one with the Pearson correlation coefficient ρ in brackets.

From the essential Hi-C matrices both in case of chromosome 1 and 17 we can notice that as the number of projectors used to reconstruct the matrix decreases, the distinctive contact patterns fade out, up to the case in which only one or two projectors are exerted, which are informative just for the chromatin compartmentalization. This is demonstrated by the fact that the corresponding scatter plots values worsen in correlation, deviating more from the straight line $y = x$, as shown by the decrease in the Pearson correlation coefficient's values. In fact when the number of projectors decrease and consequently the number of eigenvectors and related eigenvalues n^* used for the Hi-C maps reconstruction, the difference in the reconstructed data increases, up to the case of $n^* = 1$ in which the scatter plot is distributed almost along a straight line with null angular coefficient, index of highly uncorrelated values ($\rho = 0.435$). In addition, by looking at the shape of the scatter plots, it can be seen that the values are on average well reconstructed in the central area, but tend to deviate from the bisector for higher values. The latter correspond precisely to the values along the main diagonal of the Hi-C map, typically more highlighted, which are reconstructed worse than the external ones, enhancing the characteristic patterns of normally concealed interactions among distant loci.

4.1.3 3D chromosome reconstruction

The last step of the analysis regarding the GM12878 consists in visualizing the chromosomes 1 and 17 by means of the ShRec3D algorithm as described in 3.7. Using this algorithm it is possible to obtain the 3D coordinates for the single chromosomes, starting from appropriately normalized matrices according to the SCN normalization described in 3.6.2. Furthermore it is necessary, as for the Hi-C matrices, to remove all rows and columns that add up to zero and, in this case, to replace each isolated zero in the matrix with a non-zero constant suitably adjusted according to each SCN normalized Hi-C matrix. This procedure was done by using a MATLAB code able to identify each zero and replace it with different non-zero numbers equal to half of the minimum non-zero value of each Hi-C map. In this way we ascertain that the Hi-C map, seen as adjacency matrix, described a graph with no isolated nodes, which in the MATLAB code, when we take the inverse of the entries to determine the inverse contact frequencies, corresponding to the edge weights between any pair of nodes, causes the value to diverge infinitely. Once the preprocessing

was performed, it was therefore possible to execute the ShRec3D algorithm to obtain the spatial coordinates for each single locus of the chromosomes. However, for the assumptions on which the algorithm is based, it is necessary to ensure that the spectrum of the eigenvalues of the Gram matrix obtained after the preprocessing is characterized by a peak around zero with positive eigenvalues at a large distance from it. Histograms of eigenvalues for single chromosomes 17 and 1 are shown respectively in figures 4.19 and 4.20.

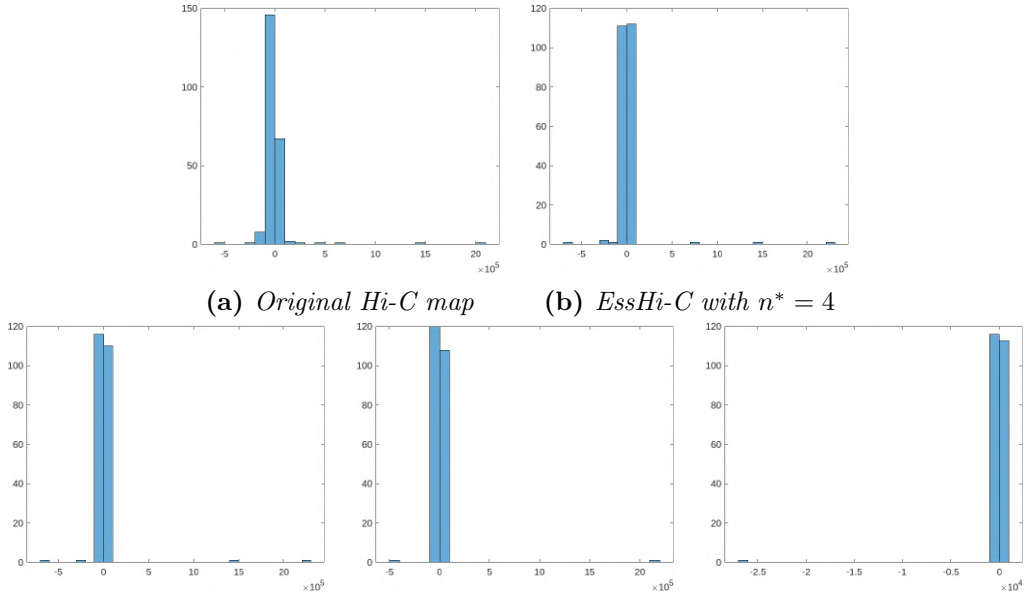


(a) *Original Hi-C map* **(b)** *EssHi-C with* $n^* = 4$ **(c)** *EssHi-C with* $n^* = 3$

(d) *EssHi-C with* $n^* = 2$ **(e)** *EssHi-C with* $n^* = 1$

Figure 4.19: Histograms of the eigenvalues from the essential matrices for the chromosome 17 reconstructed starting from different n^* highest-ranking projectors.

As it can be seen from the histograms in figures 4.19**(a)** and 4.20**(a)**, both chromosomes 17 and 1 have the characteristics required by the SchRec3D algorithm. In particular, in the distributions it is possible to notice how the spectrum of the eigenvalues has a peak around zero and that a certain number of isolated (positive) eigenvalues are present at a great distance from it. This number should be equal to 3, like the number of spatial coordinates to be obtained from the rank-3 Gram matrix. However we can notice that in the case of essential matrices reconstructed starting from a low number of projectors, there are fewer (see figures 4.19**(c-d)** and 4.20**(c-d)**) or even none (see figures 4.19**(e)** and 4.20**(e)**) of the (positive) eigenvalues far from the main peak, in which the 3 largest eigenvalues are all concentrated within the peak around zero. Starting from the spatial coordinates generated by the ShRec3D



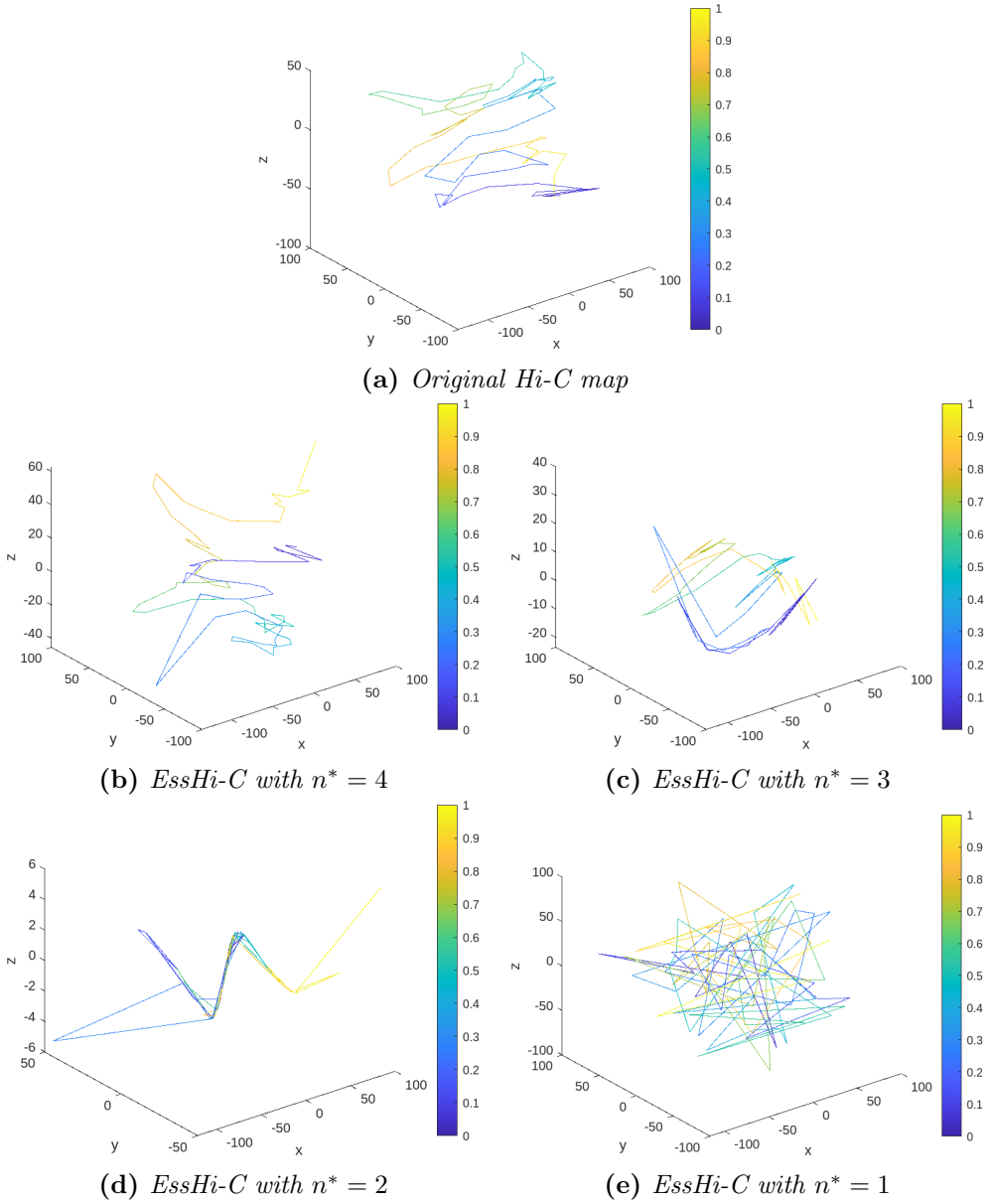
(a) *Original Hi-C map* **(b)** *EssHi-C with* $n^* = 4$

(c) *EssHi-C with* $n^* = 3$ **(d)** *EssHi-C with* $n^* = 2$ **(e)** *EssHi-C with* $n^* = 1$

Figure 4.20: Histograms of the eigenvalues from the essential matrices for the chromosome 1 reconstructed starting from different n^* highest-ranking projectors.

algorithm we then used a MATLAB code to visualize the individual nodes and join them in pairs via a color-coded line. The different color for each pair of points identifies the position along the genome where the nodes are located, so that regions that correspond to nearby points along the genome can be better viewed with similar colors. The 3D images obtained from the spatial coordinates provided by the algorithm are shown in figures 4.21 and 4.22. From the images reconstructed with a different number of projectors it is easy to distinguish the different levels of organization of the chromosome in space. Specifically, it can be noted that the essential matrices reconstructed starting from a large number of projectors ($n^* = 4$) thoroughly reproduce the spatial configuration of the corresponding original chromosome, while as the employed projectors decrease only certain shapes are enhanced. For example, in the case of chromosome 1, the image reconstructed with $n^* = 3$ projectors evidently shows a division of the chromosome into two arms, which correspond to the two arcs marked with different colors (see figure 4.22**(c)**). In this case it is evident how the use of a colorbar, which refers to the different points in space according to the location of the loci along the chromosome, is extremely useful in characterizing the various elements that are present in the reconstruction from the point of view of the correspondence between network, Hi-C

matrix and spatial configuration. As regards the reconstructions starting from a low number of projectors, it can be seen how the spatial configuration differs greatly from the original one, up to the extreme case of a single projector in which it is not possible to distinguish any chromosomal structure that can provide significant biological information (see figures 4.21**(c)** and 4.22**(c)**).



(a) *Original Hi-C map*

(b) *EssHi-C with* $n^* = 4$

(c) *EssHi-C with* $n^* = 3$

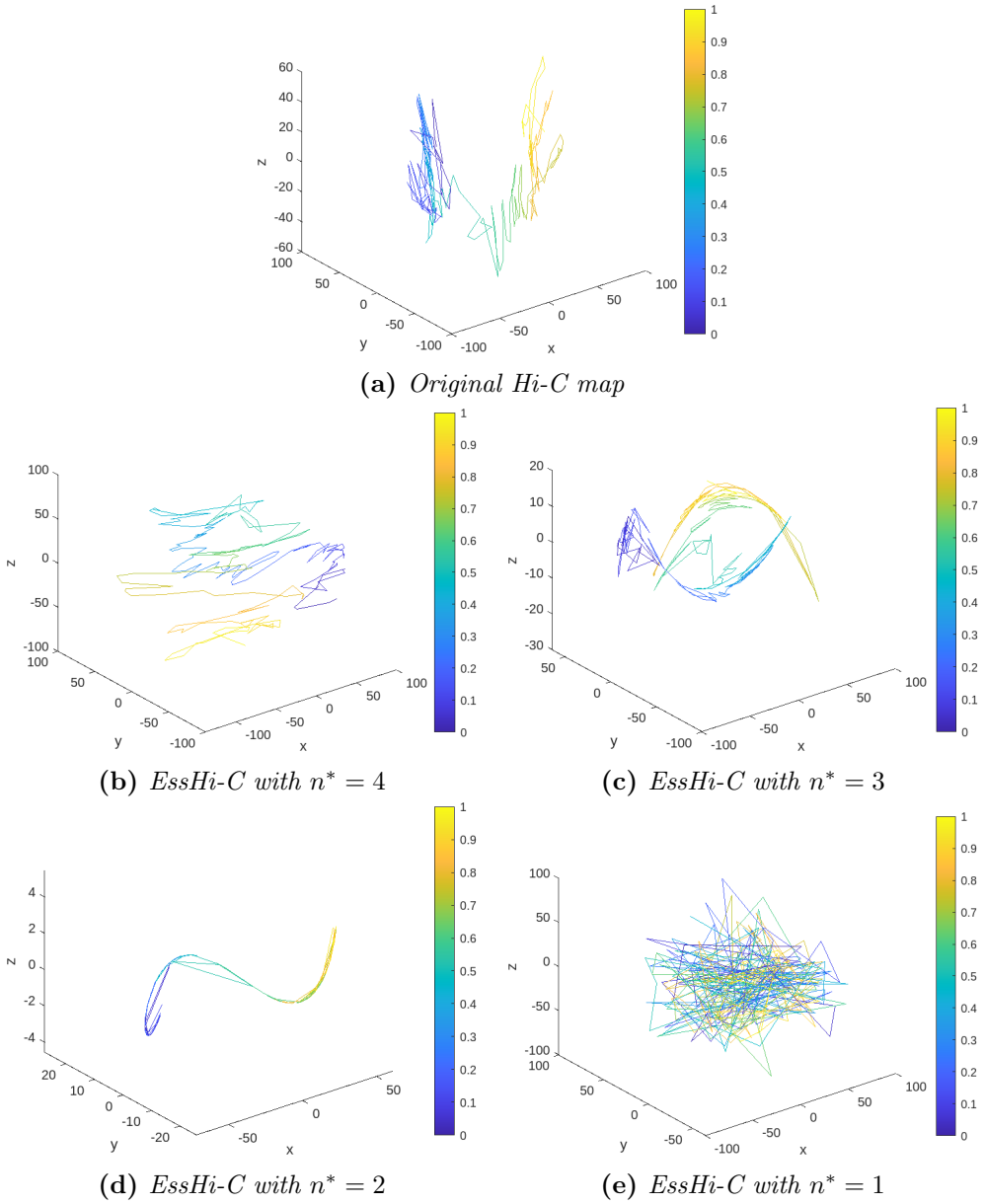
(d) *EssHi-C with* $n^* = 2$

(e) *EssHi-C with* $n^* = 1$

Figure 4.21: 3D images reconstructed by using the SchRec3D algorithm starting from the original and the essential Hi-C matrices for the chromosome 17 with different levels of reconstructions by varying the number of n^* highest-ranking projectors. The colorbar indicates the coordinates position along the genome.

Therefore, we can appreciate that the ShRec3D algorithm provides with even

more evidence how the Hi-C maps that should be produced at different levels of reconstruction are different from each other. In fact, the success of the reconstruction of the essential matrix can be clearly seen, just as the differences can be clearly appreciated when the number of projectors is reduced, unlike what can be appreciated with a scatter plot or with a simple value of data comparison.



(a) *Original Hi-C map*

(b) *EssHi-C with* $n^* = 4$

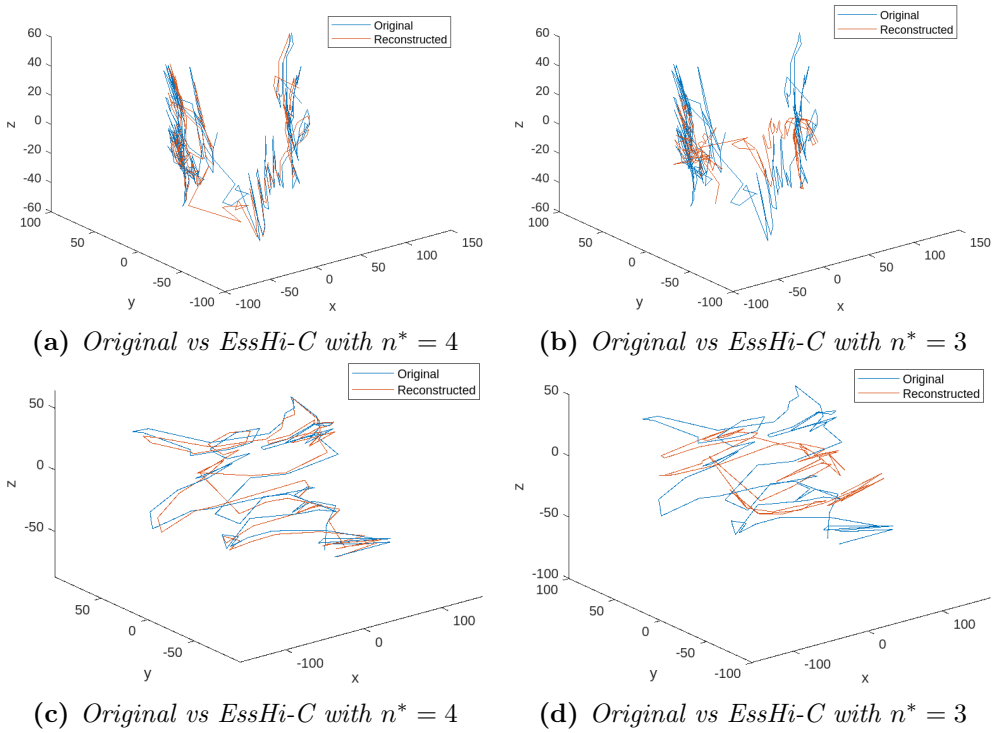
(c) *EssHi-C with* $n^* = 3$

(d) *EssHi-C with* $n^* = 2$

(e) *EssHi-C with* $n^* = 1$

Figure 4.22: 3D images reconstructed by using the SchRec3D algorithm starting from the original and the essential Hi-C matrices for the chromosome 1 with different levels of reconstructions by varying the number of n^* highest-ranking projectors. The colorbar indicates the coordinates position along the genome.

To clearly visualize the goodness of the reconstruction we performed a procrustes analysis (following the steps described in paragraph 3.8) between the reconstruction of the original chromosome using ShRec3D and the one made starting from the essential matrices reconstructed starting from a number of projectors $n^* = 4$ (figure 4.23**(a,c)**) and $n^* = 3$ (figure 4.23**(a,c)**). The cases with a lower number of projectors were rejected from this analysis because it is evident how the reconstruction did not work well and therefore returns a very different spatial conformation from the original one. It is also interesting to note how the superimposition highlights the goodness of the reconstruction starting from a different number of projectors. In particular, for both the chromosomes it can be seen that the reconstructions are particularly well superimposed for $n^* = 4$, while for $n^* = 3$ it gets worse.



(a) *Original vs EssHi-C with* $n^* = 4$ **(b)** *Original vs EssHi-C with* $n^* = 3$

(c) *Original vs EssHi-C with* $n^* = 4$ **(d)** *Original vs EssHi-C with* $n^* = 3$

Figure 4.23: 3D images reconstructed by using the SchRec3D algorithm from the essential matrices of the chromosome 1 (top line) and 17 (bottom line) starting from $n^* = 4$ **(a,c)** and $n^* = 3$ **(b,d)** highest-ranking projectors (red line) together with the corresponding original chromosome reconstruction (blue line).

The procrustes distance is used to evaluate the similarity between the different structures reconstructed using ShRec3D and superimposed using procrustes analysis both for the original and essential Hi-C maps of the chromosomes 1 and 17. The procrustes distance P_d is a measure of dissimilarity between two

shapes by computing the point-wise distance as the sum of squared differences between the corresponding points in the two shapes. It is then normalized to the scale S of the original reconstruction which is not affected by procrustes analysis. The procrustes distance then returns a numerical scalar in the range $[0, 1]$, where 0 indicates a perfect alignment between the two shapes and 1 means that the dissimilarity is maximum. The values of procrustes distance for each reconstruction, even for the ones related to a lower number of projectors, are listed in table 4.3.

Chromosome	n^*	P_d
1	4	0.012
1	3	0.162
1	2	0.419
1	1	0.995
17	4	0.011
17	3	0.155
17	2	0.472
17	1	0.965

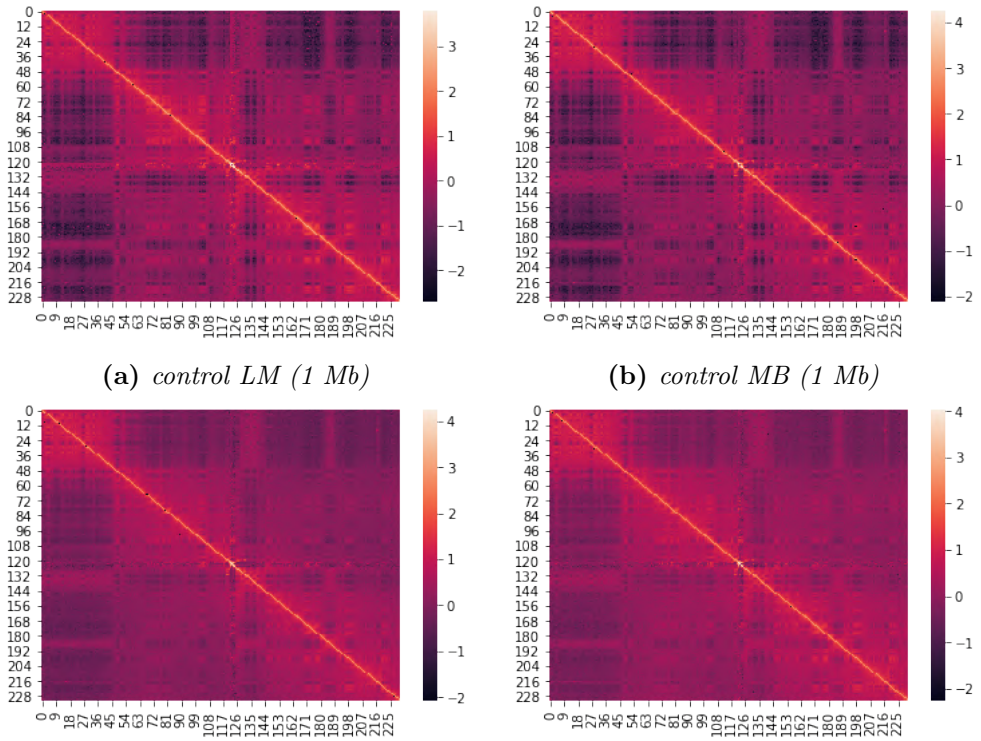
Table 4.3: Procrustes distance values for the 3D shapes reconstructed by using the ShRec3D algorithm starting from the original Hi-C matrices and the essential ones at different reconstruction levels both for chromosome 1 and chromosome 17.

From the procrustes distance values in table 4.3, we have a corroboration of the excellent goodness of reconstruction in the case of chromosome 17 starting from $n^* = 4$ ($P_d = 0.011$) projectors and a worsening at $n^* = 3$ ($P_d = 0.155$). Both for chromosome 1 and for chromosome 17 it can also be seen that for reconstructions starting from a lower number of projectors, there is a notable worsening of the reconstruction, which if it is still good for $n^* = 3$ ($P_d = 0.155 - 0.162$), it gets considerably worse for $n^* = 2$ ($P_d = 0.419 - 0.472$) and even almost collapses to 1 for $n^* = 1$ ($P_d = 0.995 - 0.965$), indicating an absolute dissimilarity between the recostructions starting from the original Hi-C maps and the essential ones.

4.2 Case study: case vs control

So far we have done training on the method we want to use to verify the goodness of the reconstruction of the Hi-C matrices, i.e. through the use of scatter plots and correlation coefficients, up to the more refined method which

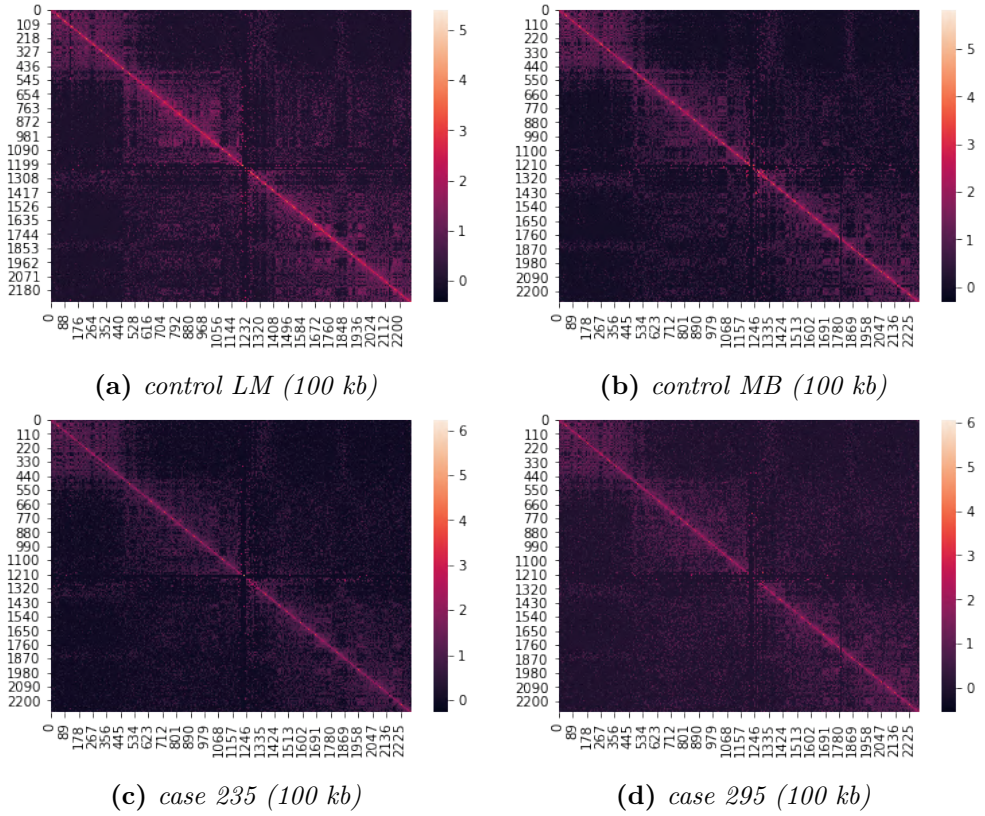
is provided by using the ShRec3D algorithm, which combines the properties of networks and the three-dimensional structure of chromosomes. We have therefore so far verified the whole methodological framework, the intrinsic properties of the Hi-C data and the ability to reconstruct the signal and distinguish the noise component by performing a spectral analysis. Now we want to apply this strategy to a medical application case study for which we have two cases (235 and 295) and two controls (LM and MB), for a total of four different samples of which we have four Hi-C maps of the chromosome 1. The Hi-C matrices that we are going to use at this time are both at 1 Mb resolution, as in the previous case of GM12878, but also at an higher resolution of 100 kb. The request for an higher Hi-C map resolution is suitable for our purpose, because it will help us to understand if the trend of the noise component, which differs from the expected Wigner's semicircle law, is actually due to the fact that with a too low resolution the data might suffers from a loss of noise which changes the spectral characteristics.



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

(c) *case 235 (1 Mb)* **(d)** *case 295(1 Mb)*

Figure 4.24: Heamaps of the Hi-C matrices for the chromosome 1 of the two controls: LM and MB (top line) and the two cases: 235 and 295 (bottom line) at 1 Mb resolution and after taking the base 10 logarithm.



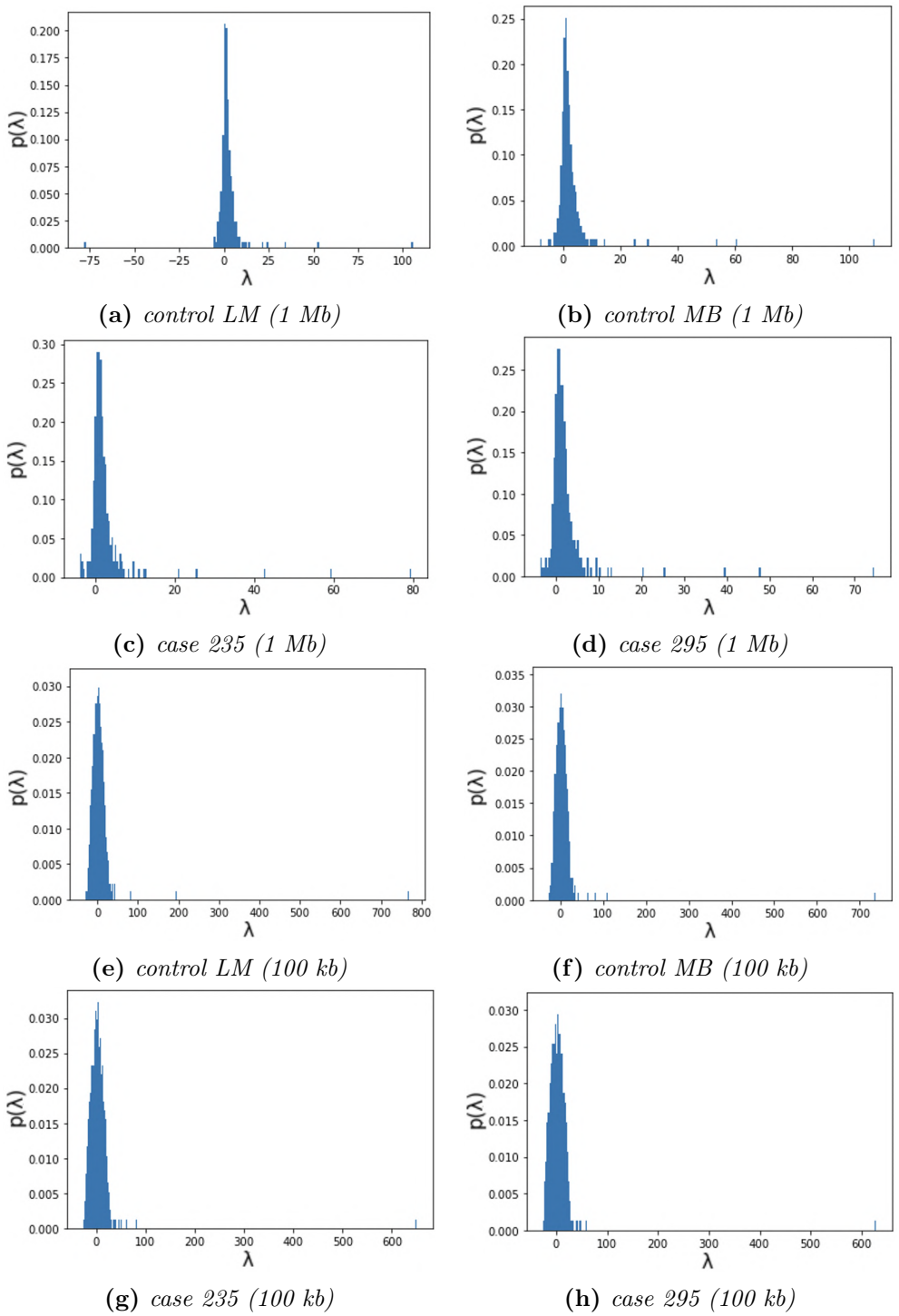
(a) *control LM (100 kb)* **(b)** *control MB (100 kb)*

(c) *case 235 (100 kb)* **(d)** *case 295 (100 kb)*

Figure 4.25: Heamaps of the Hi-C matrices for the chromosome 1 of the two controls: LM and MB (top line) and the two cases: 235 and 295 (bottom line) at 100 kb resolution and after taking the base 10 logarithm.

The Vanilla-Coverage normalization was applied to the Hi-C data and the lines which add up to zero has been removed and the base 10 logarithm has been taken as well. The results, both at 1 Mb and 100 kb resolution are shown as heatmaps respectively in figures 4.24 and 4.25. Already starting from a visualization as a heatmap we can see some, albeit minimal, pattern differences between cases and controls, however we do not expect large differences in terms of 3D rearrangement of the genome for the type of rare disease from which the cases are affected. The latter will be explored with the ShRec3D analysis in the next paragraphs. Subsequently we produced the eigenvalue spectrum relating to both the 1 Mb resolution Hi-C maps and those with a higher resolution of 100 kb. They are shown in figure 4.26. As can be seen, the histograms show how, both at 1 Mb and at 100 kb resolution, the spectrum is composed of a main peak around the origin and of a series of dispersed eigenvalues, even at large distances from the peak. However, already at this level we note an asymmetry of the peak around zero which is more evident in

the histograms that refer to lower resolution Hi-C matrices (1 Mb) than in those with a resolution of 100 kb, which instead appear much more symmetrical.



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

(e) *control LM (100 kb)* **(f)** *control MB (100 kb)*

(g) *case 235 (100 kb)* **(h)** *case 295 (100 kb)*

Figure 4.26: Spectra of the eigenvalues probability distribution $p(\lambda)$ related to the Hi-C matrices for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at different resolutions (1 Mb and 100 kb).

Subsequently, we want to verify that the trend related to the noise component follows at least approximately the one expected, given by Wigner's semicircle function. The latter, as described by the theory of random matrices and discussed in the paragraph 3.2, corresponds to the distribution of the modulus of the eigenvalues extracted starting from a symmetric random matrix. For this purpose, we decided to generate a symmetric random matrix, for each of the four Hi-C matrices examined, with entries distributed according to a Gaussian distribution with mean equal to that of the corresponding original data and standard deviation suitably adjusted. The latter is tuned in such a way as to maximize the correlation value that resulted by comparing the spectrum of the original data with the simulated ones. Specifically, we calculated the Pearson correlation coefficient, which is less sensible to the punctual differences among the simulated data, between the original data and the simulated ones, by varying a multiplicative constant k to be added to the standard deviation relating to the original data. Furthermore, having noticed a parabolic trend of the correlation coefficient as a function of k, we decided to proceed with a parabolic fit of the data. Finally the constant k was estimated as the value corresponding to the abscissa of the vertex of the fitting parabola. As an example of interpolation, the trend of the correlation coefficient is shown in figure 4.27 as a function of the multiplicative constant k in the case of control MB at 1 Mb resolution. The spectra of the eigenvalues for the original Hi-C matrices and those of the corresponding simulated matrices are shown in figure 4.28.

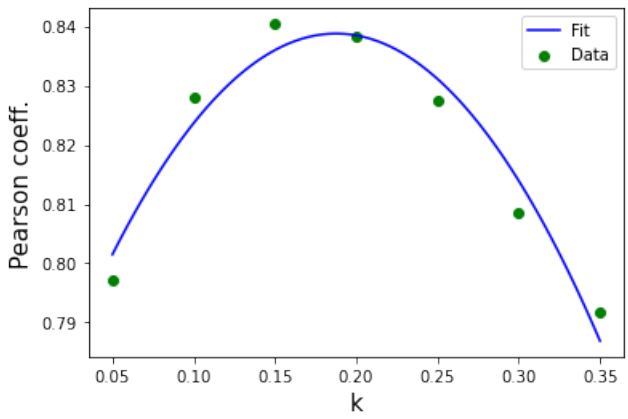
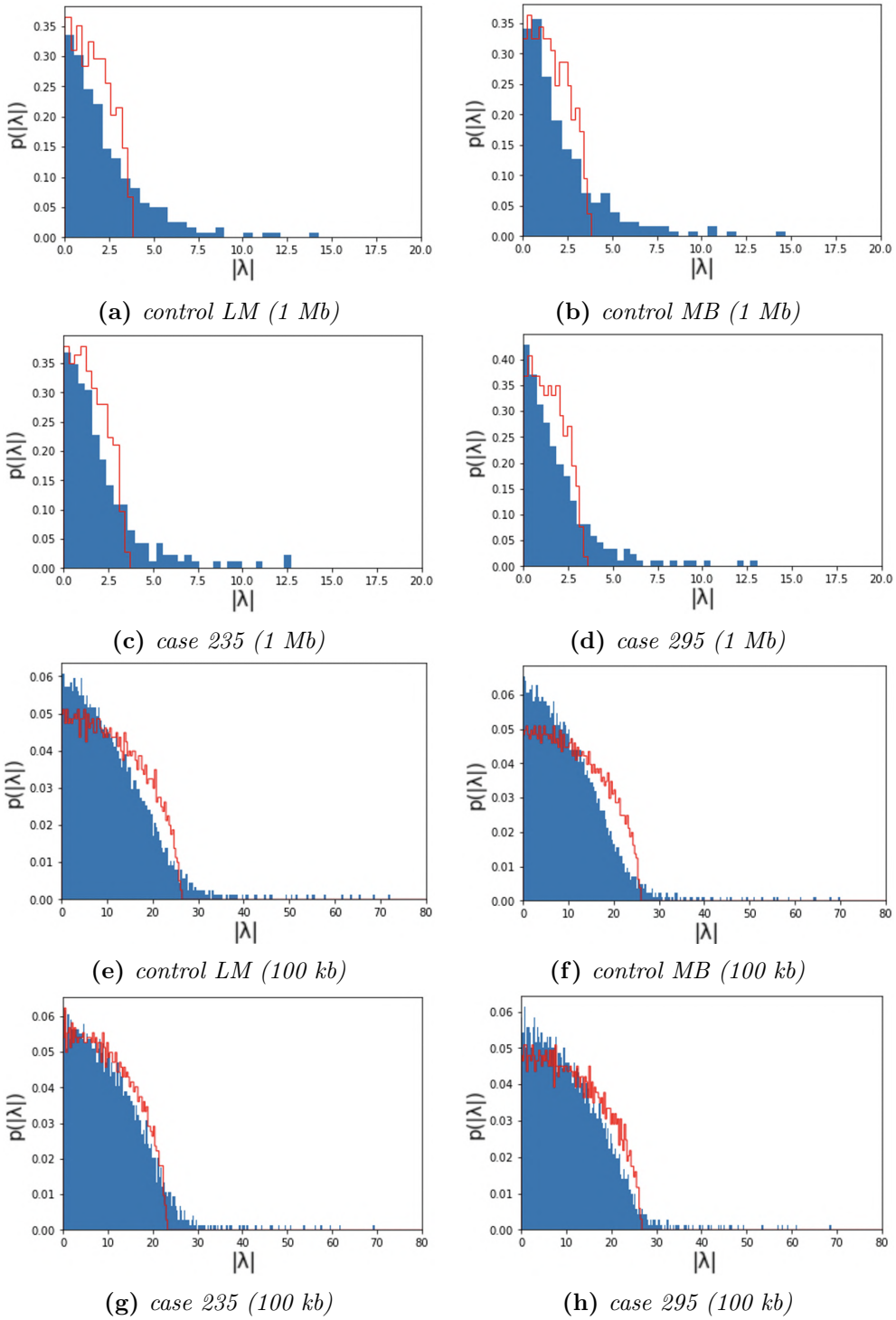


Figure 4.27: Pearson correlation coefficient data (green dots) as a function of the constant k for adjusting the standard deviation of the simulated random matrix with a parabolic fit (blue line) for the control MB at 1 Mb resolution Hi-C map. The correlation between the data and the parabolic fit is $\rho = 0.973$ (p-value $< 10^{-3}$).



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

(e) *control LM (100 kb)* **(f)** *control MB (100 kb)*

(g) *case 235 (100 kb)* **(h)** *case 295 (100 kb)*

Figure 4.28: Zoom for the spectra of the module of the eigenvalues $p\,(|\lambda|)$ (blue) in the range $0 < |\lambda| < 80$ related to the Hi-C matrices for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at different resolutions (1 Mb and 100 kb) with the superimposition of the corresponding symmetric random matrix distribution with Gaussian distributed entries with mean equal to the original one and suitably adjusted standard deviation (red).

The overlapping of the spectrum of the simulated random matrices with the original ones, seen in figure 4.28, confirms the fact that the 100 kb Hi-C matrices actually have a spectrum of the eigenvalue module of the random component more similar to that of a real random matrix than it happens for Hi-C matrices with lower resolution of 1 Mb. In the latter case, in fact, the trend is much peaked around zero with a large tail-like drop. The latter are instead much less marked at resolutions of 100 kb, both in controls (LM and MB) and cases (235 and 295), and the shape better reflects the expected Wigner's semicircle function.

4.2.1 Essential Hi-C maps: a novel approach

Now we want to reconstruct the essential matrices for the four original Hi-C matrices, both for the controls and for the cases at the two different resolutions. To extract these matrices we need to distinguish the signal component from the noise component, thus setting a threshold that varies in order to satisfy this request. It follows that the eigenvalues with modulus greater than this threshold have been counted in the reconstruction of the essential matrix, while the others have been discarded since they are labeled as belonging to the random component of the spectrum. The signal part consists of the eigenvalues scattered along the entire spectrum, while those relating to the random component are distributed according to Wigner's semicircle law. However, from the spectra of the eigenvalue modulus there is no clear separation between the two components, so we tested different methods for estimating the threshold and therefore the number of projectors n^* to be used for the reconstruction of the essential Hi-C matrices. In the first place we considered the spectra of the eigenvalues and estimated the absolute value of the minimum eigenvalue of the distribution as the threshold value. In this case, however, it was necessary to look at each spectrum to see if there were any outlier eigenvalues of particularly negative value and therefore detached from the semicircle distribution centered around zero. So as a second alternative we have considered the eigenvalue distributions generated by the random symmetric matrices as seen in the previous paragraph. In this case the threshold has been estimated as the absolute value of the minimum eigenvalue this time of the simulated distribution. In this way it was possible to implement a recursive Python code for estimating the threshold, since the random distribution follows that

of Wigner's semicircle function, as per the theory of random matrices, without the necessity to discount outliers or to make customized modifications on the choice of the threshold. Both of these methods provide a compatible threshold value result and consequently a compatible number of projectors n^* employed for the reconstruction of each essential matrix. Particularly, passing from the first method to the second one results in a percentage change of the number of projectors n^* related to the signal component $\Delta_{\%} < 0.5\%$ for 100 kb resolution and $\Delta_{\%} < 1.5\%$ for 1 Mb resolution for any Hi-C matrices. The greater, albeit slight, difference for the lower resolution of 1 Mb is due to the marked asymmetry present in the spectrum of the eigenvalues around zero. This makes it more difficult for both methods to correctly measure the separation threshold between the peak around zero, which has large tails whose slight threshold variation involves a variation, albeit small, in the number of eigenvalues and therefore of projectors to be associated with the signal component. However, these subtle differences between the two methods do not correspond to any substantial deviation in the essential matrices reconstructed starting from a number of projectors that differ by such a small percentage. This is confirmed by the graph shown in figure 4.29, which shows the Pearson correlation coefficients between the entries of the reconstructed essential matrix and the corresponding original Hi-C matrices for the control MB (1 Mb resolution) as the number n^* of projectors used to reconstruct the essHi-C increases.

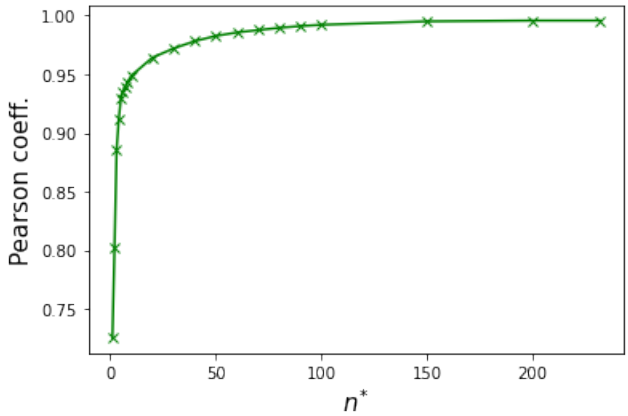
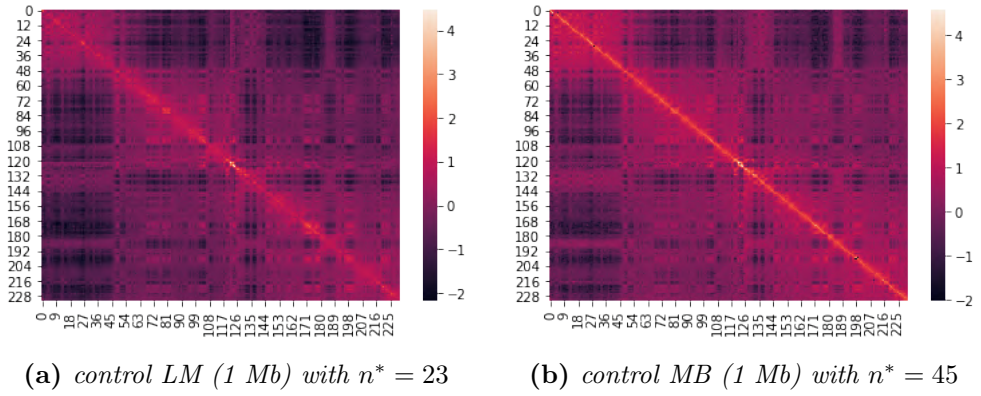


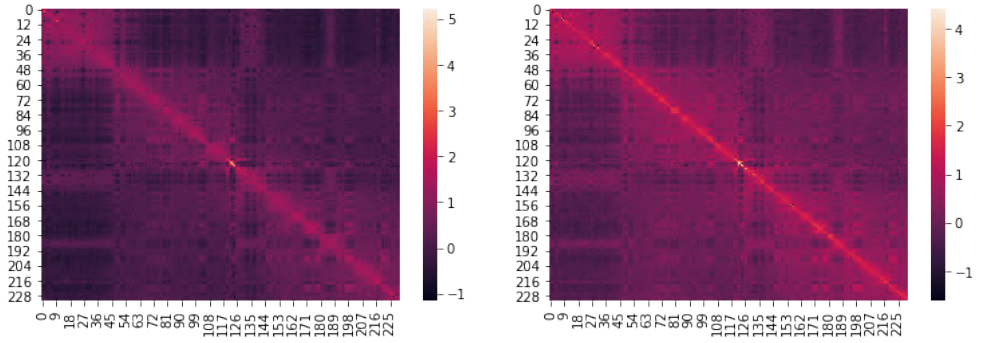
Figure 4.29: Pearson correlation coefficient as a function of the number of projectors n^* used to reconstruct the essential Hi-C matrix control MB with a 1 Mb resolution.

From the graph in figure 4.29, it can be seen that as the number n^* of projectors used to reconstruct the essential matrix varies, the Pearson correlation

coefficient calculated between the upper triangular matrices of the essHi-C and the corresponding original matrix has a large variation up to certain n^* value, which in general depends on the particular Hi-C matrix considered, and then remains almost constant, until it reaches a value close to 1 when the number of projectors n^* coincides with the linear size of the matrix considered. This trend is found for all the four samples analyzed and confirms how, once a certain minimum threshold has been verified to be exceeded, small percentage variations of n^* do not significantly affect the reconstruction's goodness of the essential matrix. Having verified the equivalence of the methods, we chose to use the second one, which uses the distribution of eigenvalues starting from the simulated matrix, given its ability to be used recursively and to be as conservative as possible on the number n^* of projectors to use for the reconstruction of the essHi-C. The latters, obtained starting from the four samples of cases and controls, are shown in figure 4.30 for those at 1 Mb resolution and in figure 4.31 for those at 100 kb resolution.

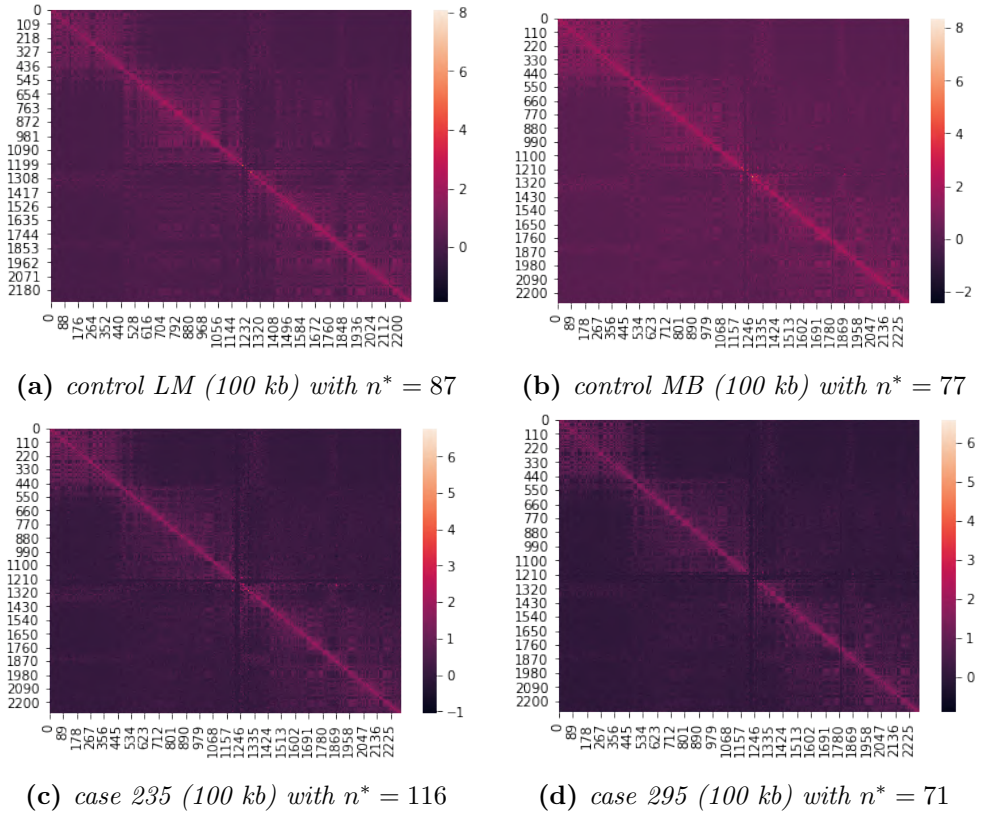


(a) *control LM (1 Mb) with* $n^* = 23$ **(b)** *control MB (1 Mb) with* $n^* = 45$



(c) *case 235 (1 Mb) with* $n^* = 25$ **(d)** *case 295 (1 Mb) with* $n^* = 43$

Figure 4.30: Essential Hi-C matrices from the corresponding original Hi-C maps of the chromosome 1 for the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution.



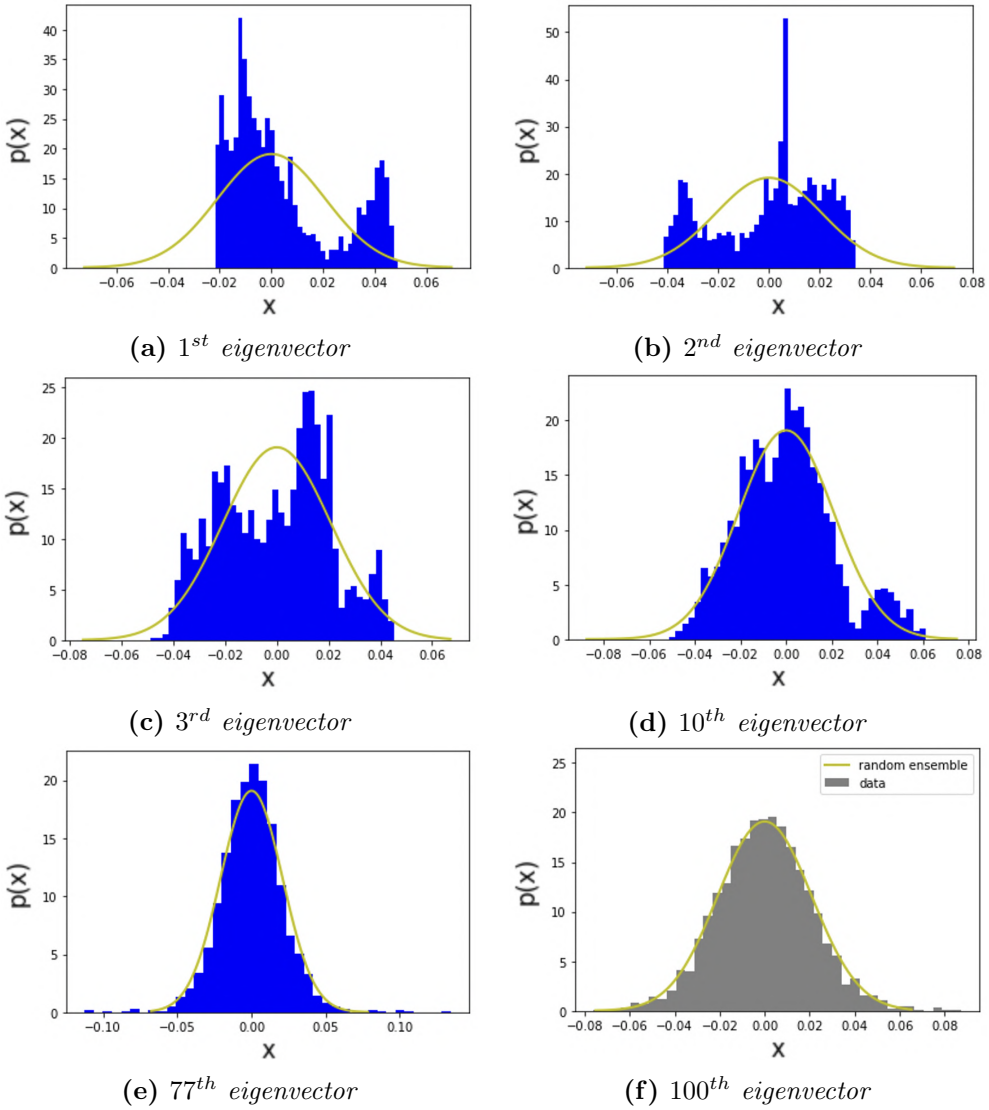
(a) *control LM (100 kb) with* $n^* = 87$ **(b)** *control MB (100 kb) with* $n^* = 77$

(c) *case 235 (100 kb) with* $n^* = 116$ **(d)** *case 295 (100 kb) with* $n^* = 71$

Figure 4.31: Essential Hi-C matrices from the corresponding original Hi-C maps of the chromosome 1 for the two controls (LM and MB) and cases (235 and 295) at 100 kb resolution.

Even just looking at the images as heatmaps in figures 4.30 and 4.31 we can see how they are very similar to the original ones. It is therefore apparent that the reconstruction was successful. Furthermore, it can be seen how the intrachromosomal interaction patterns are enhanced and the images therefore appear smoother, i.e. without that characteristic noise of the original Hi-C maps. This change, although slight if you look just at the heatmaps, is due precisely to the way the essential matrices are reconstructed using an alternative and novel method based on finding a threshold capable of distinguishing the signal region and isolating it from the noise component. Furthermore, with this method all those eigenvalues and therefore the corresponding eigenvectors that make up the n^* projectors used for the reconstruction are considered, no longer making an arbitrary choice on their number. This assures that none of the intrinsic properties and characteristics of the original Hi-C data are lost in the transition to its essential form. A corroboration of the valid choice for the threshold and therefore for the number of projectors n^* to be used in the

reconstruction of each essential matrix can also be found in the form of the distribution for the components of the n^*-th eigenvector. Specifically, from random matrix theory we have seen that the distribution of the eigenvector components is a Gaussian with zero mean and variance equal to $\frac{N}{2}$, where N is the linear size of the matrix considered. Therefore we want to verify the Gaussianity condition for the components of the eigenvectors related to the eigenvalues in decreasing module. The images of probability distributions for the n-th highest-ranking eigenvector components are shown in figure 4.32.



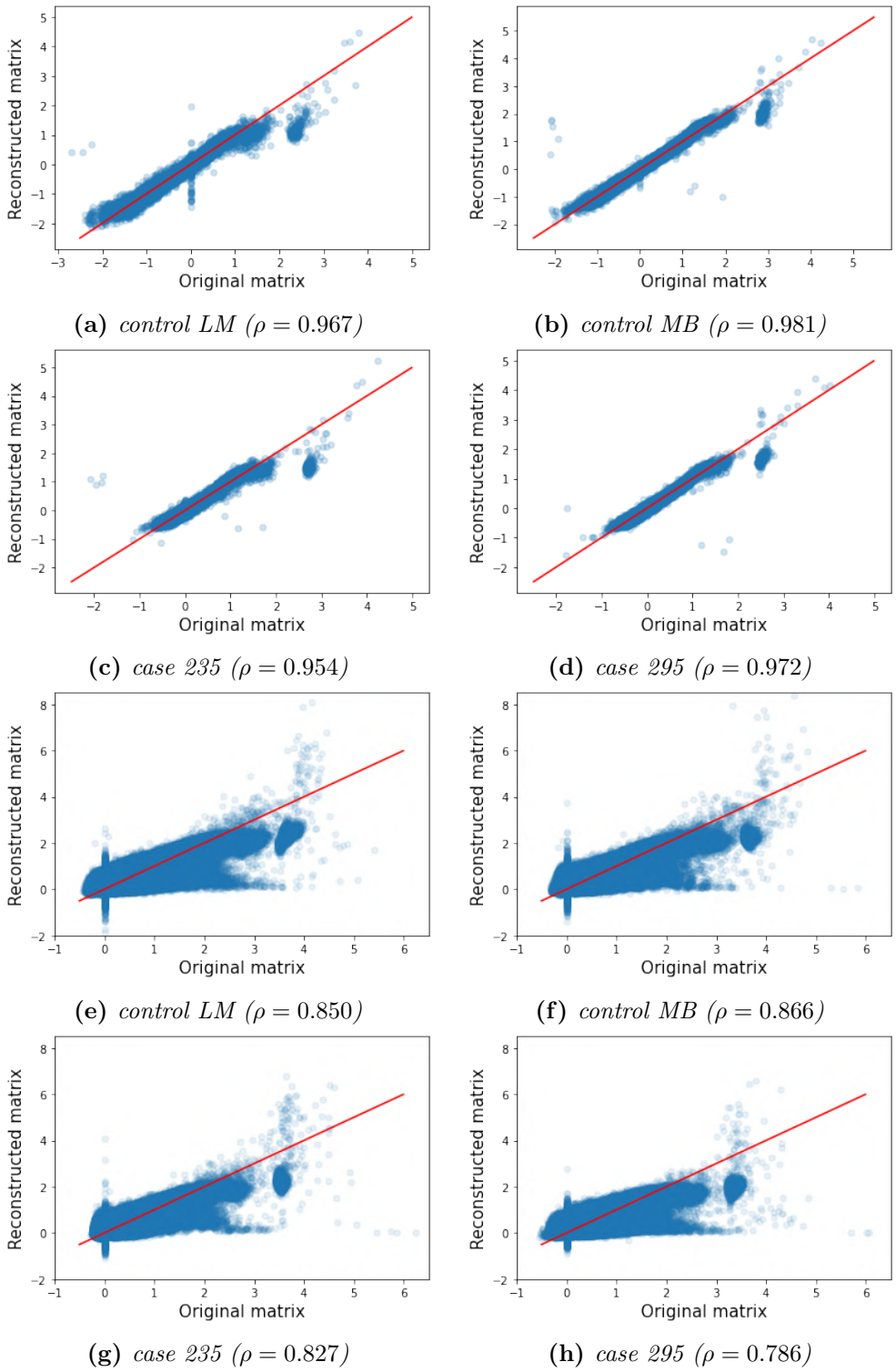
(a) 1^{st} *eigenvector* **(b)** 2^{nd} *eigenvector*

(c) 3^{rd} *eigenvector* **(d)** 10^{th} *eigenvector*

(e) 77^{th} *eigenvector* **(f)** 100^{th} *eigenvector*

Figure 4.32: Probability distributions $p(x)$ for the components of Hi-C matrix eigenvectors of different ranks (1, 2, 3, 10, 77, 100) for the control MB at 100 kb resolution; in green the same distribution from the simulated random matrix.

What can be seen in the example in figure 4.32 for the control MB at 100 kb resolution is that in the first highest-ranking eigenvectors 4.32**(a)** the trend of the probability distribution of its components differs completely from the Gaussian one of a random matrix. As the ranking of the eigenvectors decreases, and therefore as the modulus of the associated eigenvalue decreases too, we see how the trend tends to become Gaussian 4.32**(b-e)**, up to the 100-th 4.32**(f)** highest-ranking eigenvector in which the Gaussianity is much more accurate. The result is that starting from the n^*-th eigenvector we begin to see the typical Gaussian of the simulated random matrix. Furthermore the choice of the threshold placed in correspondence with a number of projectors n^*, and therefore of eigenvectors associated with the eigenvalues of greater modulus, turns out to be a good choice also from the point of view of the expected trend of the probability distribution of the components for the n^*-th eigenvector. In fact, this confirms that up to a certain highest-ranking eigenvector n^*, compatible with that obtained from the threshold using the methods described previously, the probability distribution begins to assume an almost Gaussian behaviour. This implies that we are passing from the signal component, for which we do not expect a Gaussian trend, to the noise component, for which we expect the Gaussian trend of the random matrix. Now that we have validated through various methods the goodness of the choice of the threshold and therefore of the number of projectors n^* to reconstruct the essential matrix, we want to verify if the essHi-C maps are compatible and in what manner with the original matrices from which they are obtained. To make these checks two different tools will be used: the scatter plots and the ShRec3D algorithm. In this paragraph we deal with the first method, while for the second we refer to the next paragraph. As already seen in the case of the preliminary operations carried out on the GM12878 in paragraph 4.1.2, it is possible to compare the individual values for each entry of the original Hi-C matrices and of the corresponding essHi-C by constructing scatter plots. In addition, by using the scatter plots it is possible to certify the correlation between the two matrices, and through the correlation value it is also possible to quantify it. The scatter plots obtained in this way for the four samples at the different resolutions are shown in figure 4.33.



(a) *control LM* ($\rho = 0.967$) **(b)** *control MB* ($\rho = 0.981$)

(c) *case 235* ($\rho = 0.954$) **(d)** *case 295* ($\rho = 0.972$)

(e) *control LM* ($\rho = 0.850$) **(f)** *control MB* ($\rho = 0.866$)

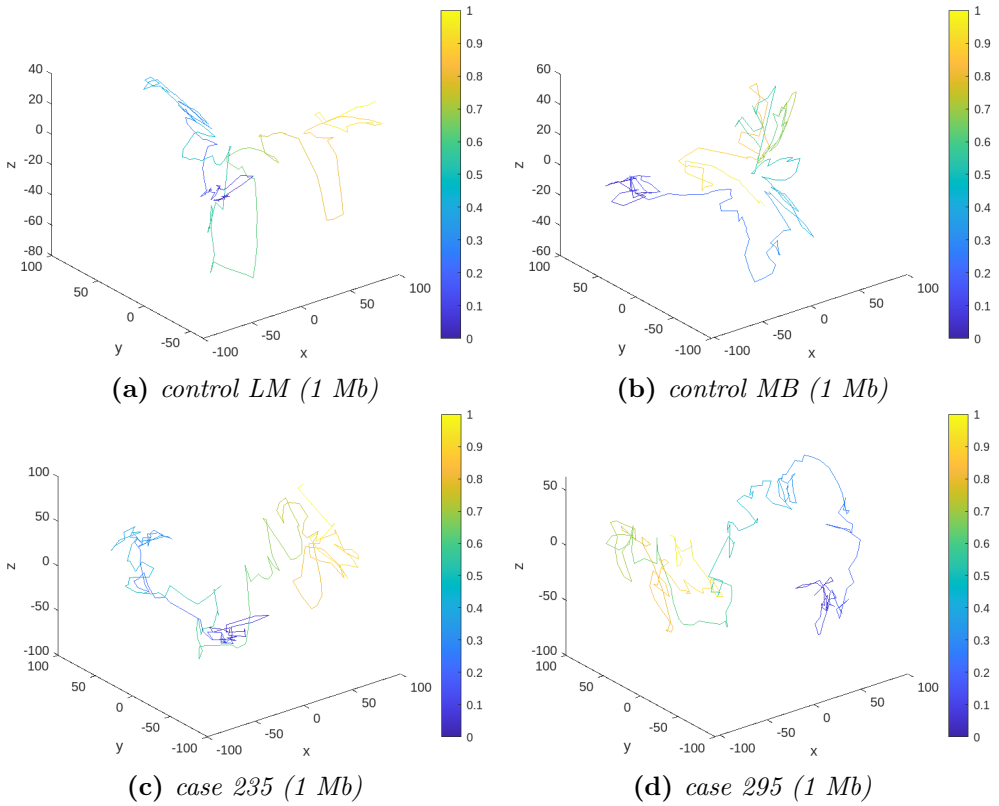
(g) *case 235* ($\rho = 0.827$) **(h)** *case 295* ($\rho = 0.786$)

Figure 4.33: Scatter plots between the original Hi-C matrices for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at different resolutions of 1 Mb **(a-d)** and 100 kb **(e-h)** with the corresponding reconstructed Hi-C matrices; in red the regression line. The Pearson correlation coefficient value ρ is listed for each plot.

From the scatter plots in figure 4.33 it can be seen how the essential matrices are significantly well reconstructed in comparison with the corresponding original ones, confirmed by with very high correlation values both in the case of Hi-C matrices with high resolution of 100 kb and even more for low resolution of 1 Mb. The latter difference is due in part to the use of a higher percentage number of projectors in the case of low resolution Hi-C data, since the spectrum of the eigenvalues is more asymmetric, thus making it more difficult to estimate the noise component and to separate it from that of signal. However this is not the only reason, in fact the difference in the correlation between the original and reconstructed Hi-C data is also due to the intrinsic properties of the higher resolution Hi-C maps, which turn out to be much noisier than the lower resolution ones, returning hence a worse point-wise correlation. Furthermore, what can be seen from the use of scatter plots is precisely the possibility of looking at the different points and at the way in which they are distributed in space with respect to the linear trend expected for a maximum correlation equal to 1. In particular, the values of the reconstructed matrices are linearly distributed along the expected bisector $y = x$, apart from the highest values in which a cloud of points can clearly be seen detaching from the linear trend. This means that the points with higher contact values are the worst reconstructed from the essential matrix, which correspond to the values close to the main diagonal in the original Hi-C matrices displayed as heatmaps. Therefore, in these last points, the reconstruction of the essential matrix is no longer reliable, resulting in an objective limit to the reconstruction. Despite that, it must be emphasized that the overall reconstruction is excellent, and in particular it helps us to highlight the characteristics of the Hi-C matrices otherwise hidden by the fluctuations due to the intrinsic noise. This will be even more evident from the next analysis.

4.2.2 ShRec3D reconstruction

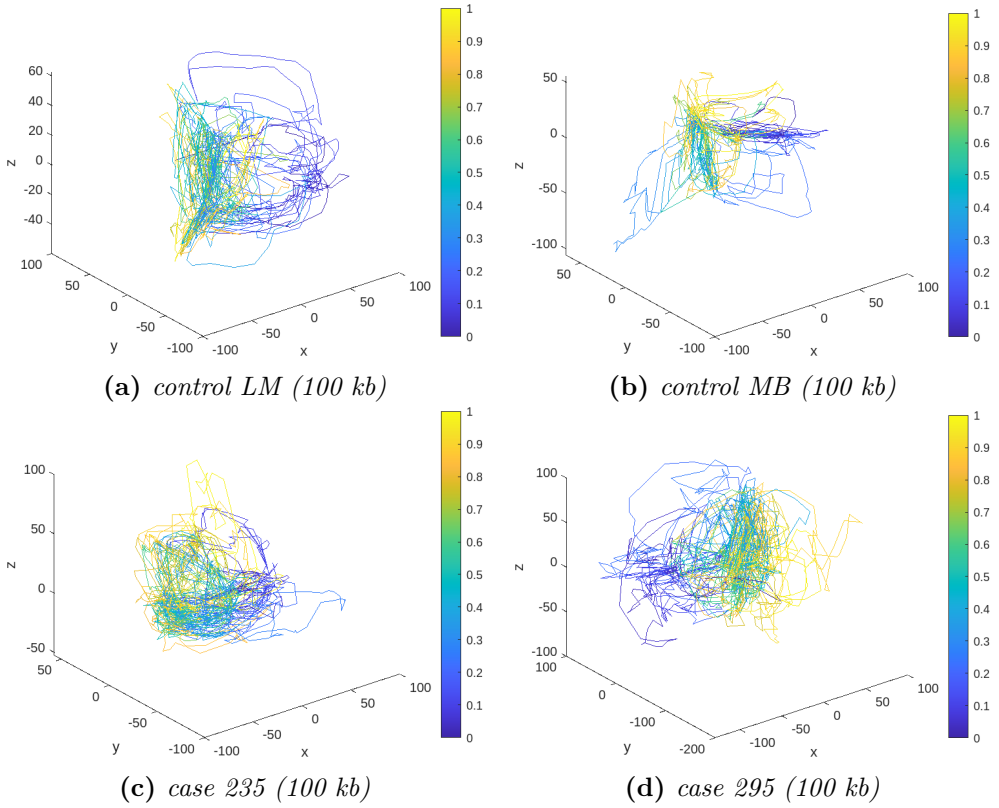
In this section we use another approach to study the goodness of reconstruction of the essential matrices from the point of view of the chromosome conformation in space. This was done taking into consideration the ShRec3D algorithm for the description of which we refer to the section 3.7. At first, the algorithm has been applied to both the original Hi-C matrix samples of 1 Mb and 100 kb resolution and the 3D plots of the obtained coordinates are reported in figures 4.34 and 4.35 respectively. Each graph is accompanied by a colorbar which indicates different points in space with different colors, following the linearity of the genome. It is important to note that the Hi-C matrices used by the algorithm have previously been normalized via the SCN and processed by removing the rows that add up to zero and replacing the isolated zeros with a value equal to half of the non-zero minimum of the matrix entries, in accordance with what was done for GM12878 in paragraph 4.1.3.



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

Figure 4.34: Three-dimensional images reconstructed starting from the spatial coordinates for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The colorbar indicates the coordinates position along the genome.



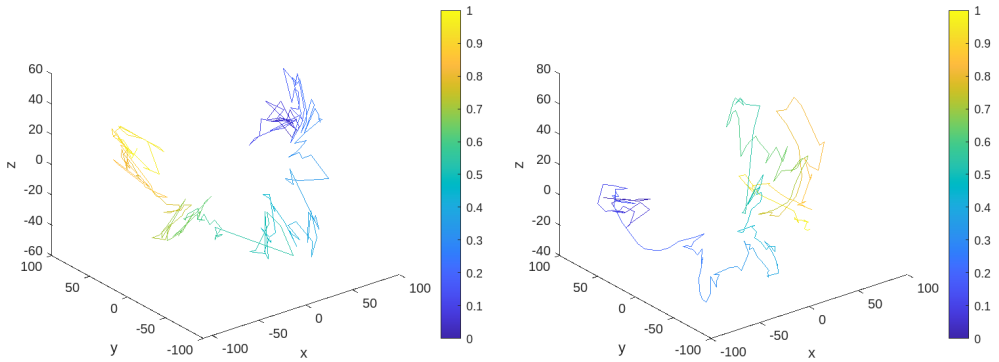
(a) *control LM (100 kb)* **(b)** *control MB (100 kb)*

(c) *case 235 (100 kb)* **(d)** *case 295 (100 kb)*

Figure 4.35: Three-dimensional images reconstructed starting from the spatial coordinates for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 100 kb resolution obtained from the ShRec3D algorithm. The colorbar indicates the coordinates position along the genome.

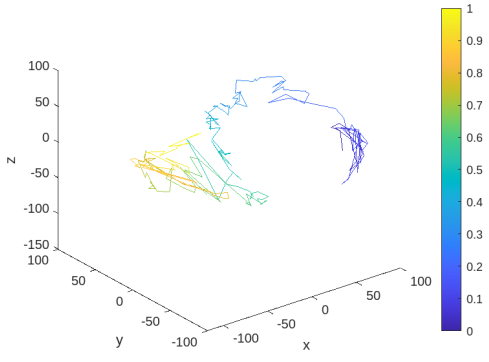
We also recall that, as for GM12878, also in this case the hypotheses of a large gap between the three largest positive eigenvalues with respect to the peak concentrated close to zero were verified for the distribution of the eigenvalues of the matrices after preprocessing. To make the reading easier, we put the histograms of the eigenvalues for each Hi-C matrix that we will analyze with ShRec3D from now on in the appendix D. From the figures 4.34 and 4.35, we can notice a significant difference in terms of Hi-C data noise, in fact while the chromosomes for the cases and controls at 1 Mb of resolution present a DNA filament that can be distinguished very clearly, in the case at 100 kb of resolution the chromatin presents a very pronounced tangling, in which characteristic patterns cannot be clearly distinguished. In particular, as regards the 100 kb cases, a strong noise is well notable which makes it almost impossible to identify any internal pattern, while for the controls, although still noisy, the structure appears more defined and patterns emerge in which the

dot density is greater than in other regions of space. This appearance could be a symptom of the disease to which the cases are linked. Without going into details, the disease from which the cases are affected is a prion disease, a neurological disease caused by prions, proteins that break down and accumulate and which can cause brain disorder [22]. However the samples analyzed are taken from peripheral blood, in which one would not expect to see huge differences between cases and controls. With a reconstruction analysis of ShRec3D, a difference emerges in the compactness of the spatial conformation of the chromatin. Through the ShRec3D algorithm we have also reconstructed the three-dimensional coordinates starting from the essential matrices, which are listed in figures 4.36 and 4.37.

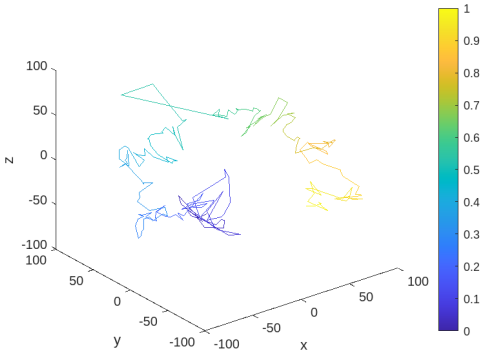


(a) *control LM (1 Mb)*

(b) *control MB (1 Mb)*

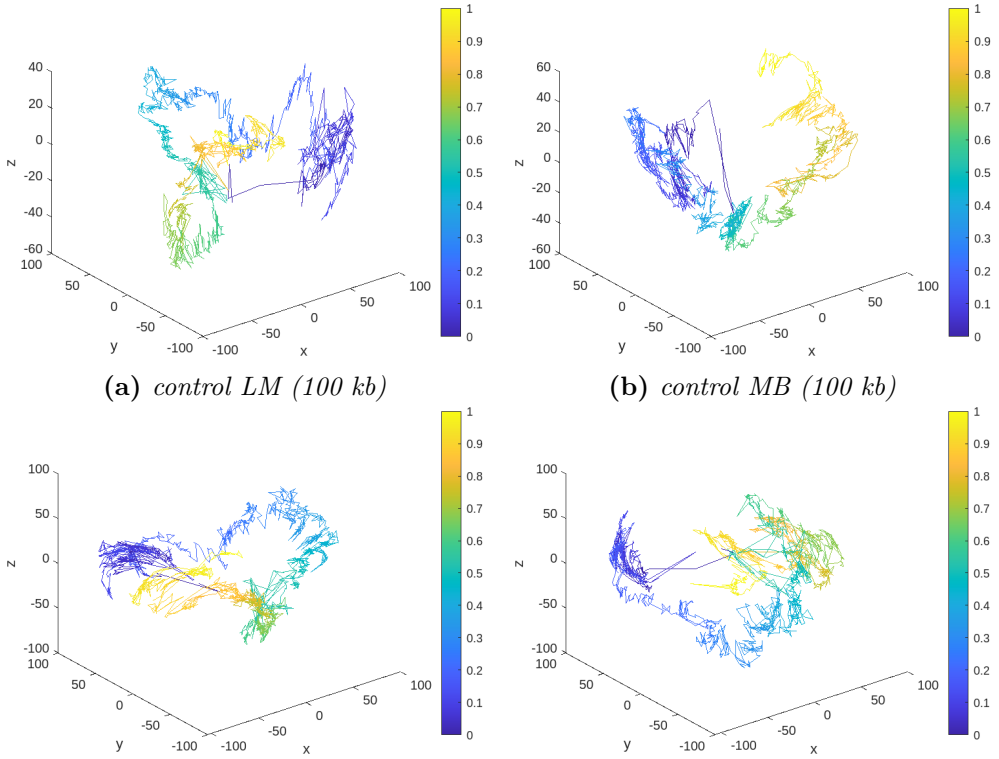


(c) *case 235 (1 Mb)*



(d) *case 295 (1 Mb)*

Figure 4.36: Three-dimensional images reconstructed starting from the spatial coordinates for the essential matrices for chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The colorbar indicates the coordinates position along the genome.

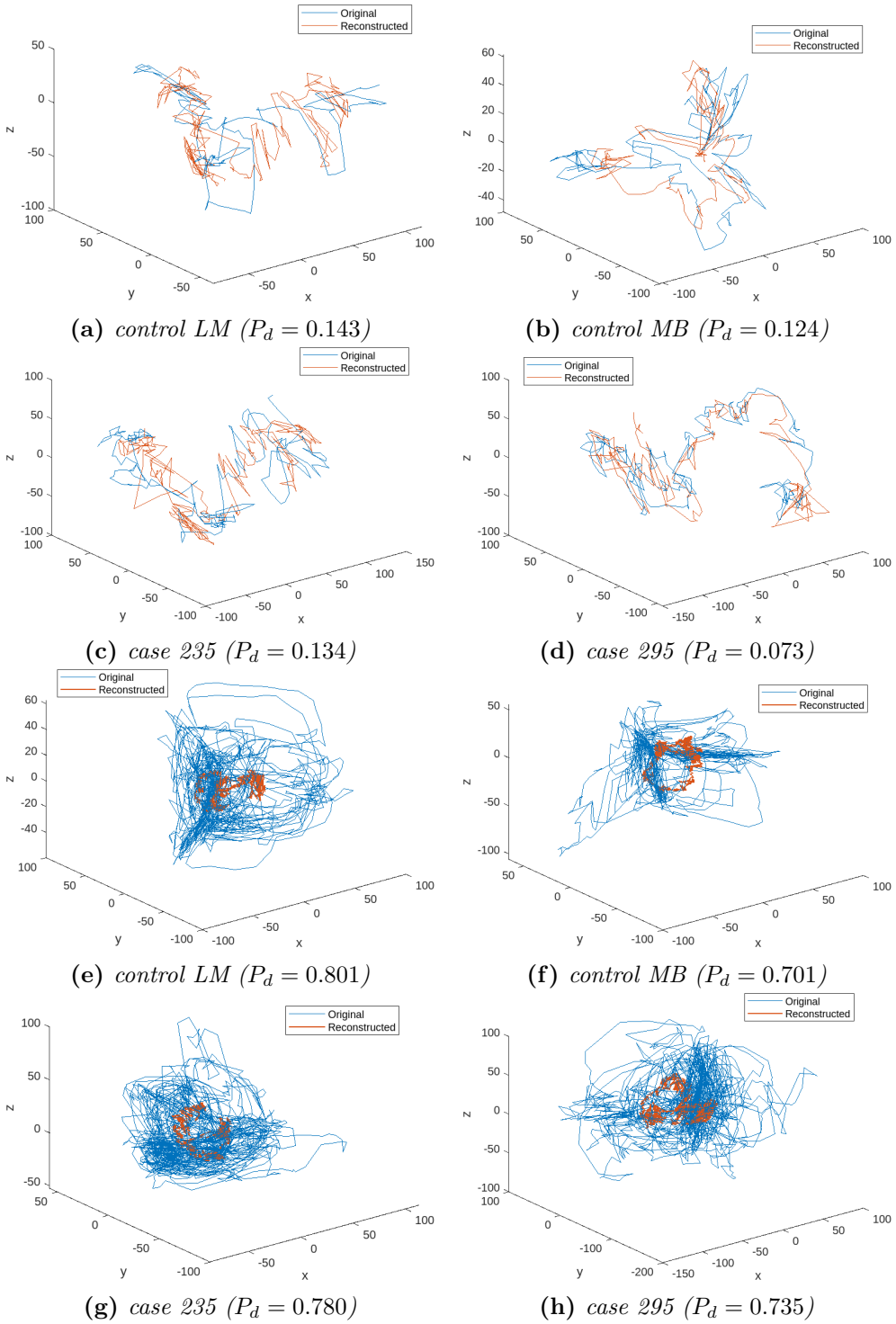


(a) *control LM (100 kb)* **(b)** *control MB (100 kb)*

(c) *case 235 (100 kb)* **(d)** *case 295 (100 kb)*

Figure 4.37: Three-dimensional images reconstructed starting from the spatial coordinates for the essential matrices for chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 100 kb resolution obtained from the ShRec3D algorithm. The colorbar indicates the coordinates position along the genome.

In the original matrices 3D reconstruction in figures 4.34 and 4.35, it is immediately striking that while at 1 Mb the orange and blue lines are more or less close in space, at 100 kb the orange lines are all concentrated in the centre and the blue lines have a much wider range, which in all probability distinguish the noise component of the Hi-C data. However in the reconstruction made starting from the essential matrices this is no longer true, in fact in both resolutions we can see how the filaments are very distinct and colored without particular regions easily attributable to noise. This further confirms the goodness of the reconstruction of the essential matrices as essHi-C and can be seen even better by superimposing the two reconstructions, original and essential, after having performed an appropriate procrustes analysis as described in the paragraph 3.8. The images relating to the superimposition between the different reconstructions for original matrices with the respective reconstructed essential matrices are shown in figure 4.38, in which are listed also the procrustes distance P_d values for each reconstructed pair.



(a) *control LM* ($P_d = 0.143$) **(b)** *control MB* ($P_d = 0.124$)

(c) *case 235* ($P_d = 0.134$) **(d)** *case 295* ($P_d = 0.073$)

(e) *control LM* ($P_d = 0.801$) **(f)** *control MB* ($P_d = 0.701$)

(g) *case 235* ($P_d = 0.780$) **(h)** *case 295* ($P_d = 0.735$)

Figure 4.38: Three-dimensional plots starting from the spatial coordinates for the chromosome 1 both of the original (blue line) and reconstructed (red line) essHi-C maps of the two controls (LM and MB) and cases (235 and 295) at 1 Mb **(a-d)** and 100 kb **(e-h)** resolution obtained from the ShRec3D algorithm. The procrustes distance value is listed for each reconstructed pair.

From the superimposition between the original and the essential Hi-C maps three-dimensional reconstruction it can be seen how in case of the Hi-C matrices at 1 Mb resolution it results to be excellent, confirmed by values of procrustes distance $P_d < 0.15$ for any sample. On the contrary, regarding the 100 kb resolution matrices, the overlap doesn't match at all, in fact the procrustes distance values are $P_d > 0.7$ for any of them. This analysis confirms how the noise level of the high resolution data is predominant in comparison with the essential matrix with which it is superimposed.

4.2.3 Case-control comparative study

In this section we carry out a comparative study between the two cases (235 and 295) and the two controls (LM and MB) to highlight differences in both the original matrix and the corresponding reconstructed essential matrix. In particular, we create scatter plots between each pair of original matrices, for a total of six combinations including the pairs of cases, controls and the four possible mixes between them (see appendix B). For each scatter plot, the correlation between the entries of the matrix pairs was then calculated.

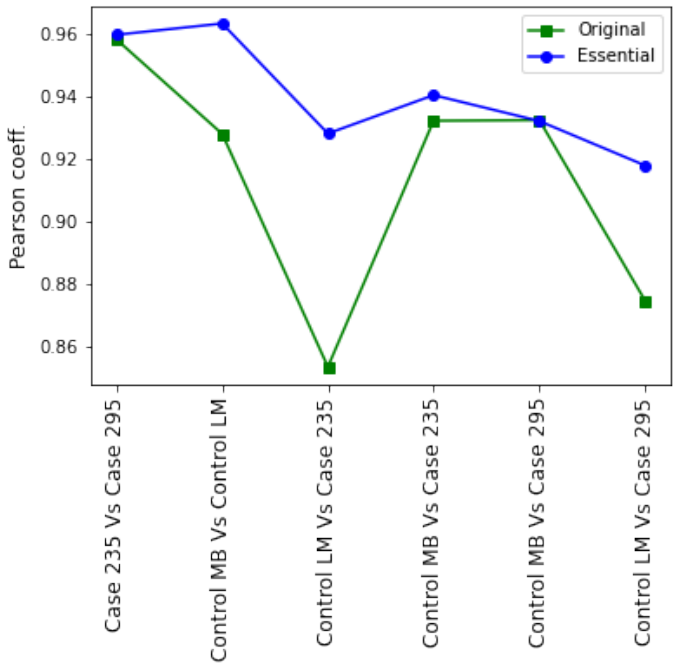


Figure 4.39: Pearson correlation coefficient values for the different combinations of pairs between the two cases (235 and 295) with the two controls (LM and MB) for the original Hi-C matrices (green line) and the corresponding essential Hi-C matrices (blue line) at 1 Mb resolution.

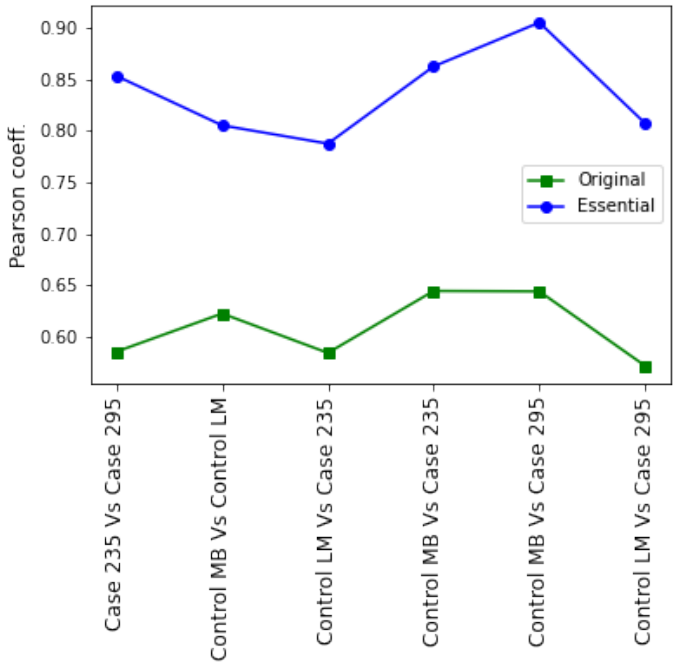
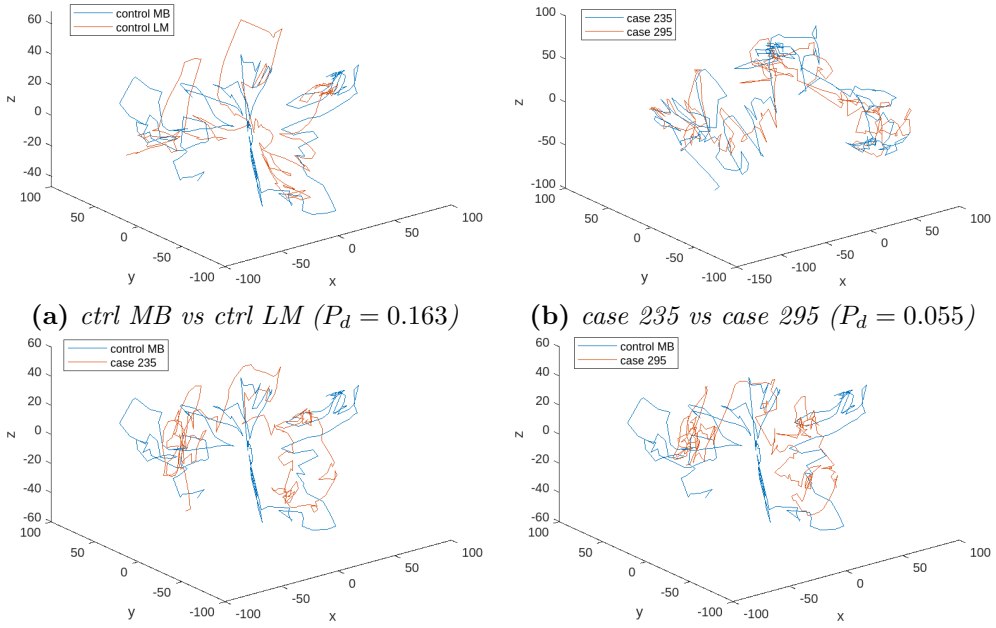


Figure 4.40: Pearson correlation coefficient values for the different combinations of pairs between the two cases (235 and 295) with the two controls (LM and MB) for the original Hi-C matrices (green line) and the corresponding essential Hi-C matrices (blue line) at 100 kb resolution.

What we would expect to find is that the pairs formed by the two cases and the two controls are more correlated than the mixes. However, as shown in the green line of figure 4.40 which refers to the original Hi-C maps for the samples with high resolution of 100 kb, this is not the case at all. We can in fact see that the first two points of the green line, referring to the original matrices, are not the highest at all, but that these are lower than some of the mix of cases and controls. Even more surprisingly, this is also the case when considering the corresponding pairs of essential matrices. In fact, even in case of the blue line in figure 4.40, it is once again noticeable that the expectation of having cases and controls with higher correlation values between the mixes is not fulfilled. This is surely due to an intrinsic noisiness of the Hi-C data at a high resolution of 100 kb, which therefore does not meet those conditions that we would normally expect in a case-control comparative study. However, these are fulfilled when switching to higher resolutions of 1 Mb. In fact, in figure 4.39, the correlation coefficients between the two cases (235 and 295) and between the two controls (LM and MB) are indeed the highest, both when comparing the original matrices (green line), but even more so when

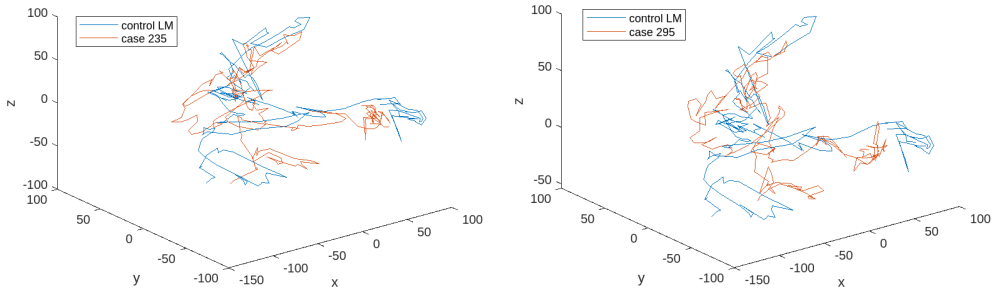
considering the respective essential matrices (blue line). In the latter case, we can see how the essential matrices not only lower the noise but also capture those properties characteristic of each matrix, thus enhancing the correlation between pairs where there is a profound similarity. The high noise present in the case of 100 kb resolution is also reflected in the large difference in correlation values between the original ($0.572 < \rho < 0.644$) and essential matrix ($0.787 < \rho < 0.905$) pairs, which tend to eliminate noise. In the 1Mb case, this difference is not so great precisely because the correlation is already very high for the original data pairs ($0.854 < \rho < 0.958$), but it does improve in case of the essential matrices ($0.918 < \rho < 0.963$) in particular in detecting those expected differences between pairs formed by only cases or only controls and the mixed ones. Ultimately, to confirm what the correlation indicated, we used the ShRec3D algorithm to obtain the coordinates of each sample at 1 Mb resolution. These were then paired and, by means of a procrustes analysis, appropriately compared. The results obtained are shown in figures 4.41 and 4.42.



(a) *ctrl MB vs ctrl LM* ($P_d = 0.163$) **(b)** *case 235 vs case 295* ($P_d = 0.055$)

(c) *ctrl MB vs case 235* ($P_d = 0.217$) **(d)** *ctrl MB vs case 295* ($P_d = 0.254$)

Figure 4.41: Three-dimensional plots starting from the spatial coordinates for the chromosome 1 of the pairs of original Hi-C maps of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The procrustes distance value is listed for each reconstructed pair.



(a) *ctrl LM vs case 235 ($P_d = 0.158$)* **(b)** *ctrl LM vs case 295 ($P_d = 0.188$)*

Figure 4.42: Three-dimensional plots starting from the spatial coordinates for the chromosome 1 of the pairs of original Hi-C maps of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The procrustes distance value is listed for each reconstructed pair.

In the 3D reconstructions of the pairs of the different samples of chromosome 1, it is confirmed that at 1 Mb the pairs of the same group (controls and cases) are more similar to each other than the mixed ones. This is indicated by the procrustes distance value, which is strongly close to zero for the pairs of controls ($P_d = 0.163$) and cases ($P_d = 0.055$), while it deviates more from zero in almost all the mixes ($\bar{P}_d = 0.204$). This analysis, together with those of correlation between the original matrices and the essential ones, confirms a greater reliability of the data at a resolution of 1 Mb than those at a higher resolution of 100 kb.

4.2.4 SynHi-C: synthetic component analysis

In the analysis carried out so far we have verified the goodness of the essential matrices reconstructed through a number of projectors estimated starting from the eigenvalues related to the signal component, and thus eliminating those relating to the noise component characterized by Wigner's semicircle curve. The inspections were carried out on different levels, both by considering the correlation from the scatter plots between the original Hi-C maps and the essential ones, and by using the ShRec3D algorithm which allowed us to represent in the three-dimensional space the arrangement and folding of the chromatin. Finally, we tested the quality of the Hi-C data at the two different resolutions and we realized that those with an higher resolution of 100 kb are not very reliable in the light of the case-control comparative study. This therefore leads us to use for the following analysis the lower resolution Hi-C data equal to 1 Mb in which, although at the beginning there were uncertainties related to the number of projectors estimation for reconstructing the essential matrices, they proved to be less noisy and more reliable for the purposes of this thesis work. Now we want to illustrate an innovative method to generate synthetic Hi-C data starting from the instruments obtained so far. As described in section 3.5, the idea to build synthetic matrices starts from the spectral analysis of Hi-C matrices, considering both signal and noise components. In particular, we identify the specific characteristics of each single original Hi-C matrix with those of the corresponding essential matrix, which is built starting from the projectors formed by the eigenvectors and the corresponding eigenvalues which constitute the signal component. We have already discussed how to derive these essential matrices in section 4.2.1. Now, however, we intend to deal with the noise part, which up to now has always been discarded, but which is instead important for obtaining the variability in the synthetic Hi-C data. Here we have to reconstruct a Hi-C matrix starting from the projectors relating to those eigenvectors whose eigenvalues are part of the noise component. In this case, to create a variability of the Hi-C data, we made a permutation of the eigenvalues related to the noise component and then coupled them to the eigenvectors after the reshuffling to obtain the projectors with which an error matrix was reconstructed. The latter is then added to the essential matrix to obtain the new synthetic Hi-C matrix or synHi-C matrix. The scheme of the synthetic Hi-C maps production is depict in figure 4.43.

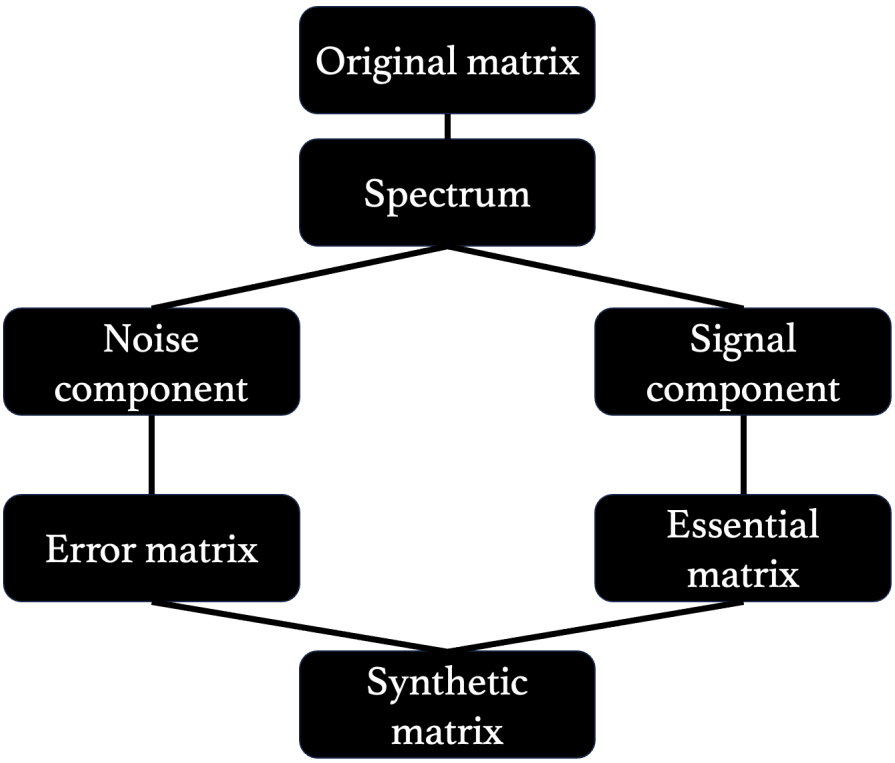
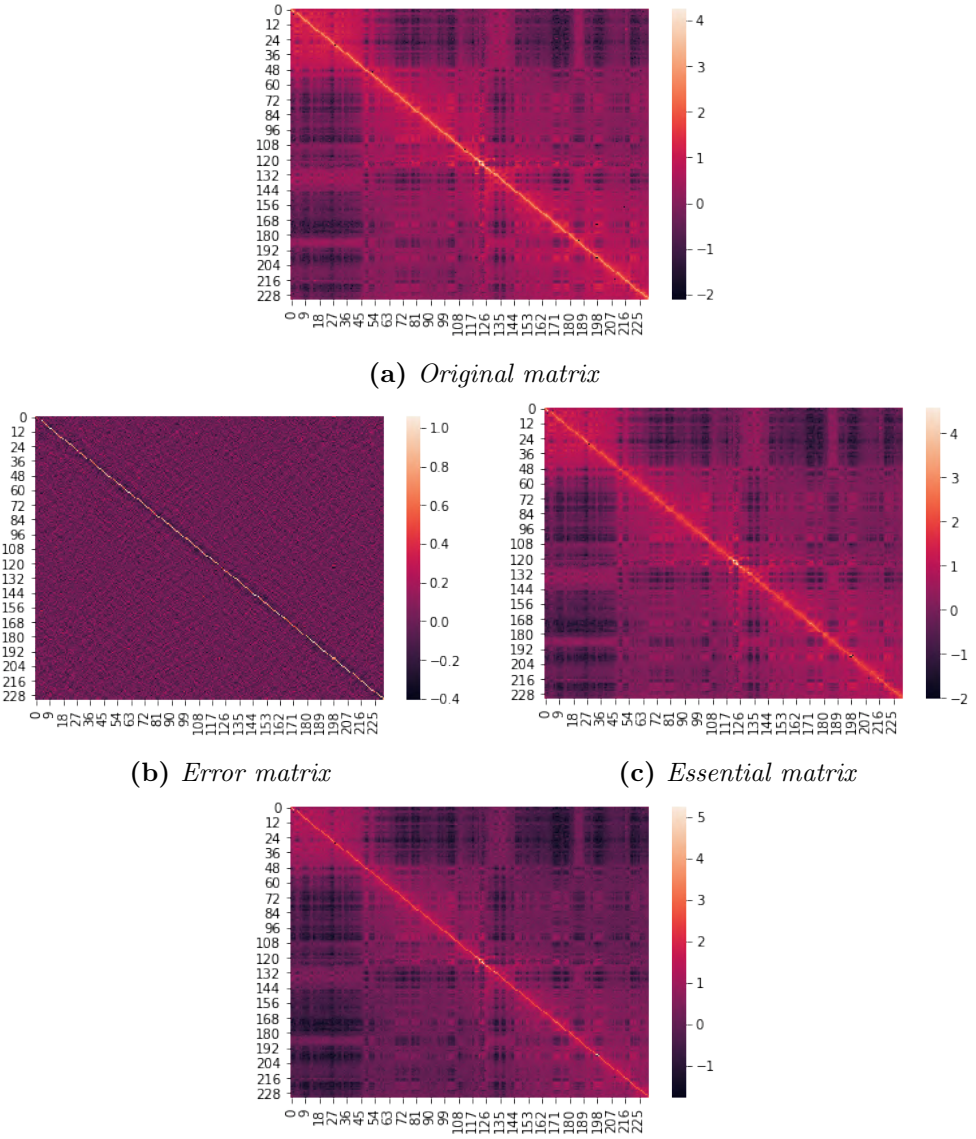


Figure 4.43: Synthetic Hi-C maps production flowchart.

Basically, since each eigenvector corresponds to a weight given by the corresponding eigenvalue, permuting the eigenvalues while keeping the eigenvectors fixed leads to a reshuffling of the weights given to each projector which will reconstruct the error matrix. This reshuffling process leads each time to a different result in terms of weight to be given to the individual projectors, thus creating a variability in the error matrix and therefore in the synthetic matrix. Now we want to apply the procedure described to generate synthetic matrices starting from the Hi-C maps of the two controls (LM and MB) and the two cases (235 and 295). Firsly, we generated a synthetic matrix for each of the aformentioned matrices. The figure 4.44 shows the heatmaps relating both to one of the error matrices obtained after the eigenvector reshuffling and to the essential matrix reconstructed starting from the different components of the spectrum, in addition to their sum which constitutes the synthetic matrix itself. A feature that can already be seen in figure 4.44 is that the error matrix has a marked line along the main diagonal in which the values are particularly high. This confirms what was seen in the scatter plot between the essential and original matrix (see figure 4.33) in which the worst reconstructed values were precisely those along the tail, i.e. those along the main diagonal.



(a) *Original matrix*

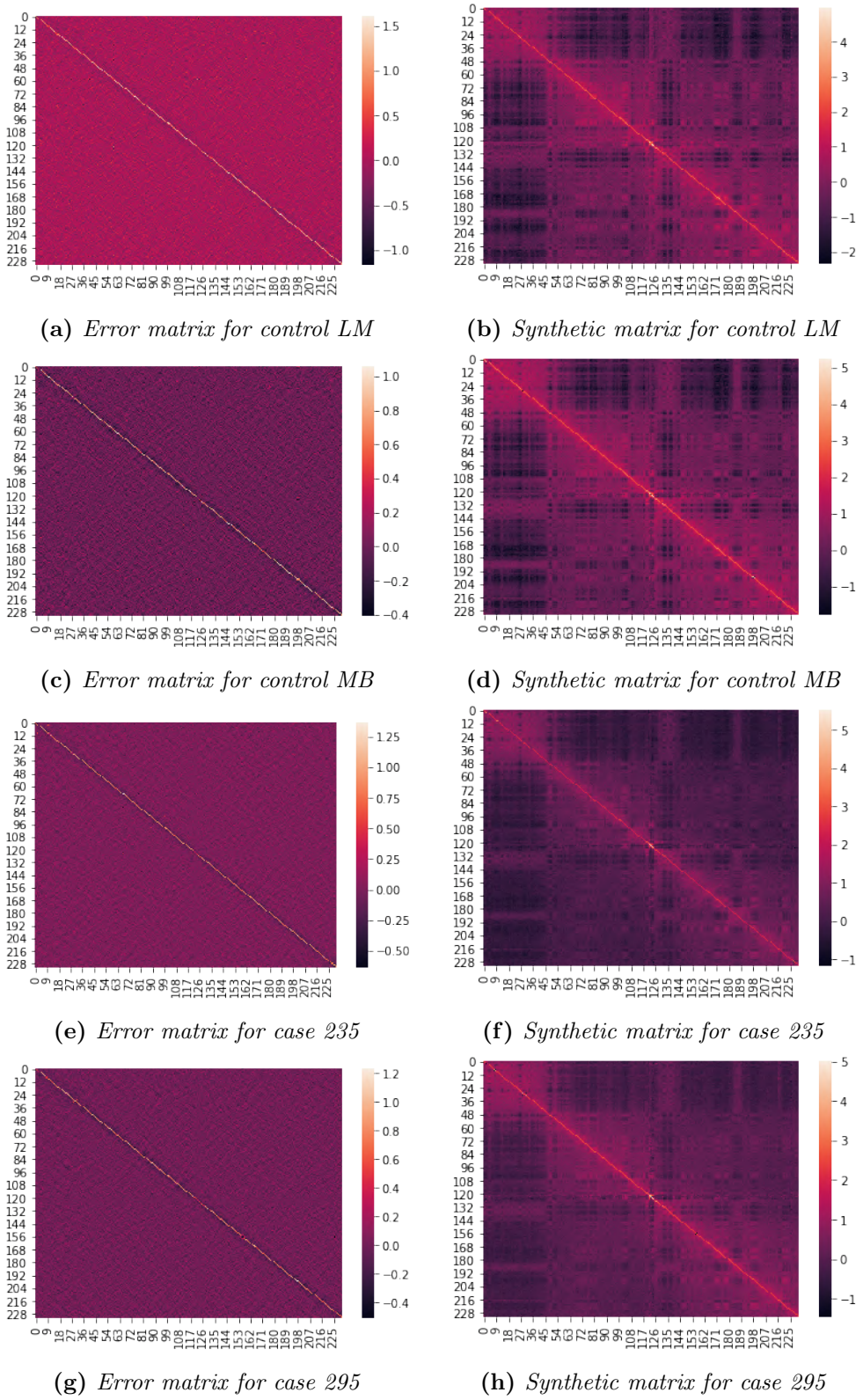
(b) *Error matrix*

(c) *Essential matrix*

(d) *Synthetic matrix*

Figure 4.44: Heatmaps of the synthetic matrix construction **(d)** starting from the sum between the error matrix **(b)** with the essential matrix **(c)** of the Hi-C original matrix **a** of the control MB at 1 Mb resolution.

Now, as expected, these values instead correspond to those reconstructed better by the error matrix which is complementary to the essential one, acting on the noise component and not the signal one. As can be seen in figure 4.45, the error matrices all have the same specific main diagonal highlighted, while the synthetic matrices inherit their variability from the latter.



(a) *Error matrix for control LM*

(b) *Synthetic matrix for control LM*

(c) *Error matrix for control MB*

(d) *Synthetic matrix for control MB*

(e) *Error matrix for case 235*

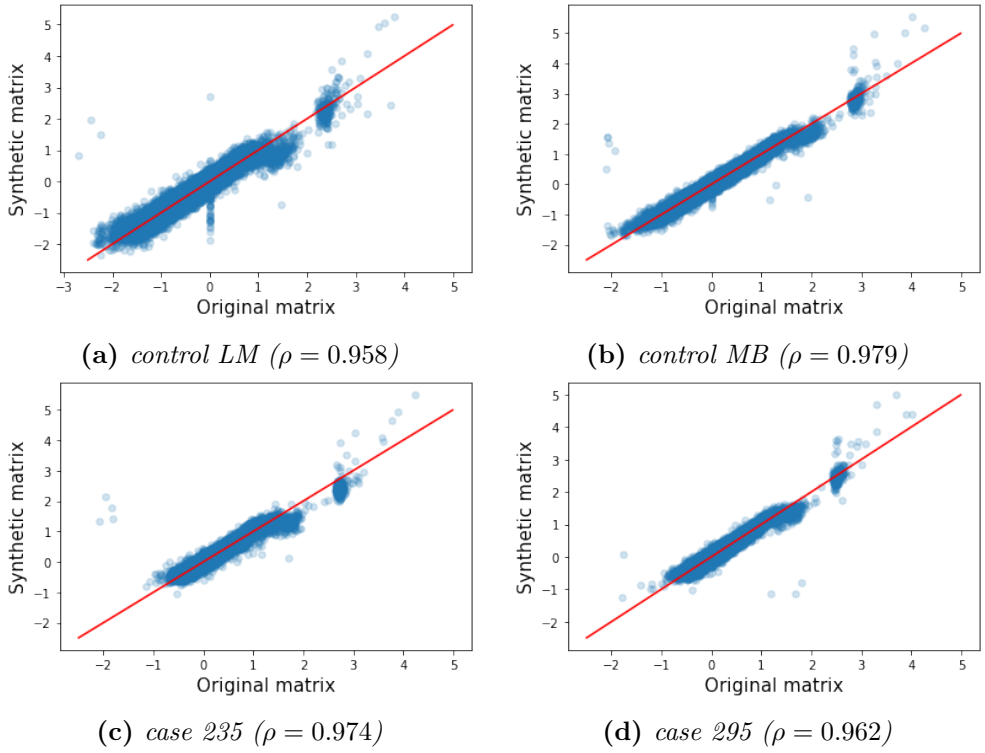
(f) *Synthetic matrix for case 235*

(g) *Error matrix for case 295*

(h) *Synthetic matrix for case 295*

Figure 4.45: Heatmaps for the chromosome 1 both of the error matrix (first column) and the synHi-C matrix (second column) for the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution.

To compare the synthetic matrices with the original ones and therefore understand the level of similarity of the first ones, we first made the scatter plots between the values of the entries extracted from the upper semidiagonal (remember that the matrices are symmetrical, so we avoid useless double counting) for both matrices. They are shown in figure 4.46.

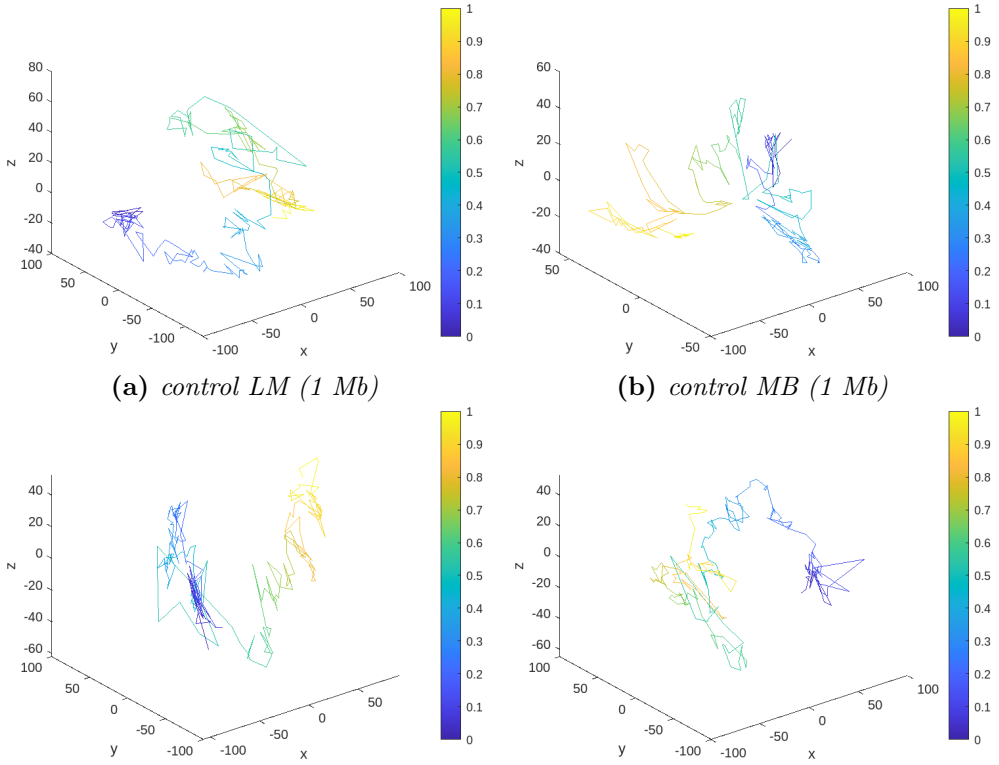


(a) *control LM* ($\rho = 0.958$) **(b)** *control MB* ($\rho = 0.979$)

(c) *case 235* ($\rho = 0.974$) **(d)** *case 295* ($\rho = 0.962$)

Figure 4.46: Scatter plots between the original Hi-C matrices for the chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution with the corresponding synthetic Hi-C matrices; in red the regression line. The Pearson correlation coefficient value ρ is listed for each plot.

As previously observed by looking at the error matrices, also the scatter plots confirm a very strong goodness of reconstruction and similarity of the synthetic matrices, where thw Pearson correlation values are $\rho > 0.958$ for any sample. In fact we can see how even that cloud of points detached from the rest of the points, which in case of the essential matrices was underestimated, is now in line with the best correlation trend $y = x$. Moreover, using the ShRec3D algorithm we reconstructed a three-dimensional graph of the chromosome 1 starting from the synHi-C maps of the four different samples (for the eigenvalue histograms see the appendix D). The latters are shown in figure 4.47, where the colorbar indicates the coordinates position along the genome.

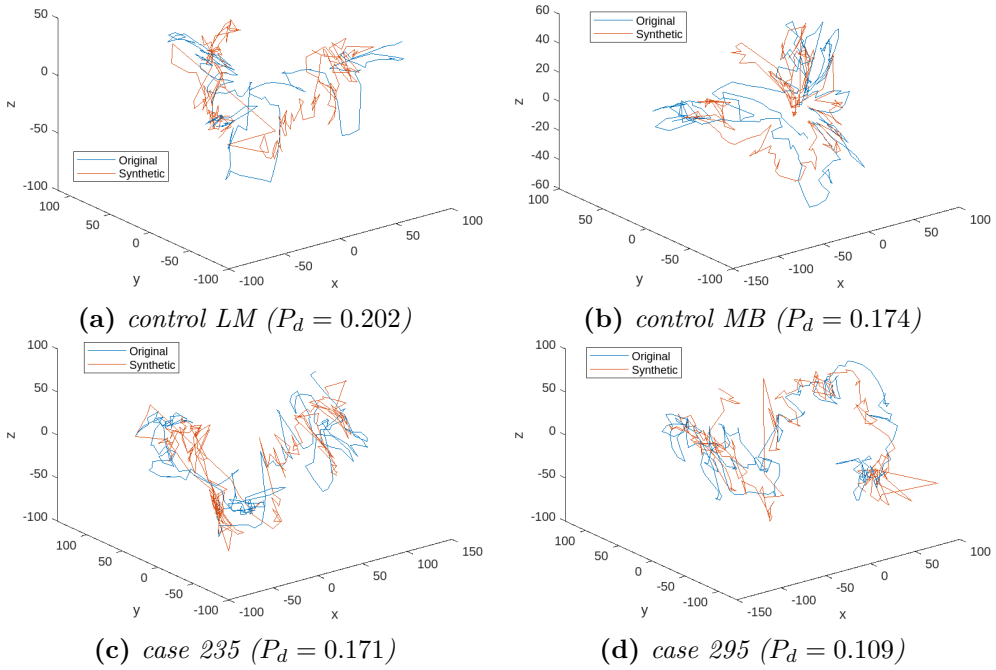


(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

Figure 4.47: Three-dimensional images reconstructed starting from the spatial coordinates for the synthetic matrices for chromosome 1 of the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The colorbar indicates the coordinates position along the genome.

Now we want to verify the similarity between the reconstructions, in particular between the original ones presented in figure 4.34, with the synthetic ones that we have generated and which are presented in figure 4.47. Also in this case, to quantify the distance between the two forms we use the procrustes distance, not before having carried out a procrustes analysis on the synthetic matrix, so as to align it as much as possible with the original reference one. The pairs of chromosomes are shown in figure 4.48 and marked with different colors, with the value of procrustes distance P_d listed for each pair of original chromosomes and the ones reconstructed starting from the corresponding synHi-C. Thanks to the comparison between the two reconstructed Hi-C matrices, it is possible to better delineate the points of contact from those more distant from each other. Furthermore, the procrustes distance parameter is very close to 1 for any couple ($\bar{P}_d = 0.164$) confirming an excellent similarity in terms of 3D configuration between the synthetically reconstructed chromosomes and the



(a) *control LM* ($P_d = 0.202$) **(b)** *control MB* ($P_d = 0.174$)

(c) *case 235* ($P_d = 0.171$) **(d)** *case 295* ($P_d = 0.109$)

Figure 4.48: Three-dimensional plots reconstructed starting from the spatial coordinates for the chromosome 1 of the pairs formed of the original (blue line) and the corresponding synthetic (red line) Hi-C maps for the two controls (LM and MB) and cases (235 and 295) at 1 Mb resolution obtained from the ShRec3D algorithm. The procrustes distance value is listed for each reconstructed pair.

original ones. The values of procrustes distance are less close to 1 than the ones obtained from the comparison with the essential matrices ($\bar{P}_d = 0.119$) as seen in figure 4.38, due to the re-introduction of noise which, on the other hand, is not present in the essential matrices. This could cause a misleading in thinking that essential matrices are more suitable as synthetic matrices, effectively resulting in a chromosome configuration more similar to the original one. However, it is important to note that in the essential matrices one of the fundamental and characteristic components of the Hi-C matrices is not present, i.e. the typical noise component, which is due to the reshuffling of the eigenvalues and therefore of the weights to be assigned to each reconstructing projector. To verify if the properties of the original Hi-C data are preserved also for the synthetic matrices, we compared the pairs of samples both between cases and controls and the mixes of both. The comparison, as in the case seen in figure 4.39, was made by building scatter plots between the input values of each pair of Hi-C matrices and from there we then extracted the Pearson correlation value, as an indicator of the similarity between the two matrices. The graph showing the correlation values for all six possible pairs of samples

both for the original matrices and the corresponding synthetc ones is shown in figure 4.49.

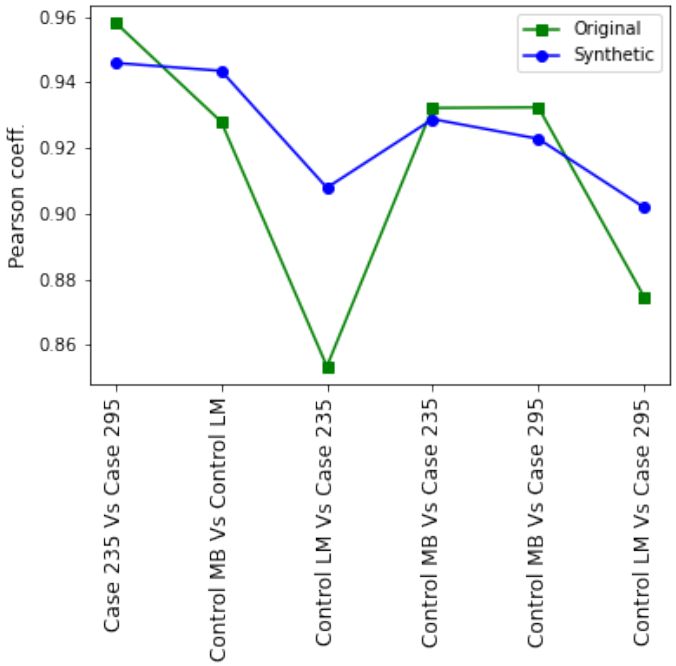


Figure 4.49: Pearson correlation coefficient values for the different combinations of pairs between the two cases (235 and 295) with the two controls (LM and MB) for the original Hi-C matrices (green line) and the corresponding synthetic Hi-C matrices (blue line) at 1 Mb resolution.

In this case it can be seen that the pairs consisting of only cases (235 and 295) or only controls (LM and MB) have correlation coefficients with a larger gap with the mixed pairs than it happens for the corresponding original Hi-C matrices. Furthermore, on average, the correlation is higher in the case of synthetic matrices than in the originals. Both these characteristics are due to the presence of the signal component given by the essential Hi-C matrix. This can be seen even better in figure 4.50, which includes the correlation trends between the possible pairs in the three cases of original, essential and synthetic matrix. In fact, in this case it is distinctly perceived how, compared to the pairs of original samples, the essentials and the synthetics are both better from the point of view of highlighting the expected characteristics between pairs of cases and controls compared to the mixed ones, having higher correlation between samples of the same type than mixed pairs. However, above all it can be seen that between the pairs of essential and synthetic matrices there is an agreement in terms of correlation trend and this is due precisely to the way of

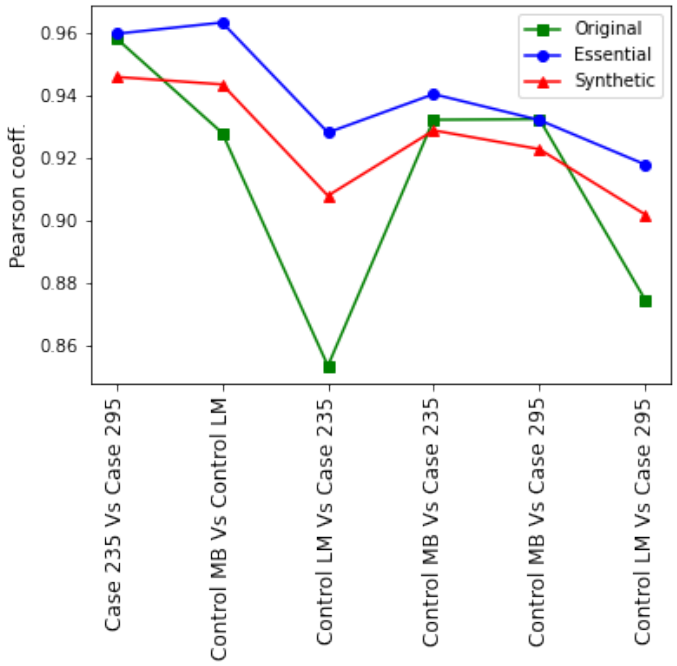


Figure 4.50: Pearson correlation coefficient values for the different combinations of pairs between the two cases (235 and 295) with the two controls (LM and MB) for the original Hi-C matrices (green line) and the corresponding essential Hi-C matrices (blue line) and the synthetic ones (red line) at 1 Mb resolution.

constructing the synthetic matrices starting from the essential ones. The addition of noise in the synthetic matrices causes the Pearson correlation coefficients to be on average lower than those relating to the essential matrices, but in any case higher than the original correspondents. Ultimately, instead of concentrating as we have done so far on verifying that the synthetic matrices actually reproduce those characteristic properties of the Hi-C matrices of the starting samples, now we want to focus our attention on a specific matrix, e.g. the control MB at 1 Mb resolution, and produce a large quantity of synthetic Hi-C matrices. Remarkably, the latter corresponds precisely the main purpose of using synthetic matrices, and that is to be able to produce large quantities of a Hi-C matrices. For our samples it is of crucial importance given that they derive from rare diseases which are already hardly studied at the Hi-C level and for which very few samples are available. For these reasons, it is particularly useful to increase the number of samples and so to have a wider statistics. We have applied the scheme already shown in figure 4.43 to generate a total of 100 synthetic matrices, each time recombining the weights, given by the eigenvalues, of the projectors used for the reconstruction process.

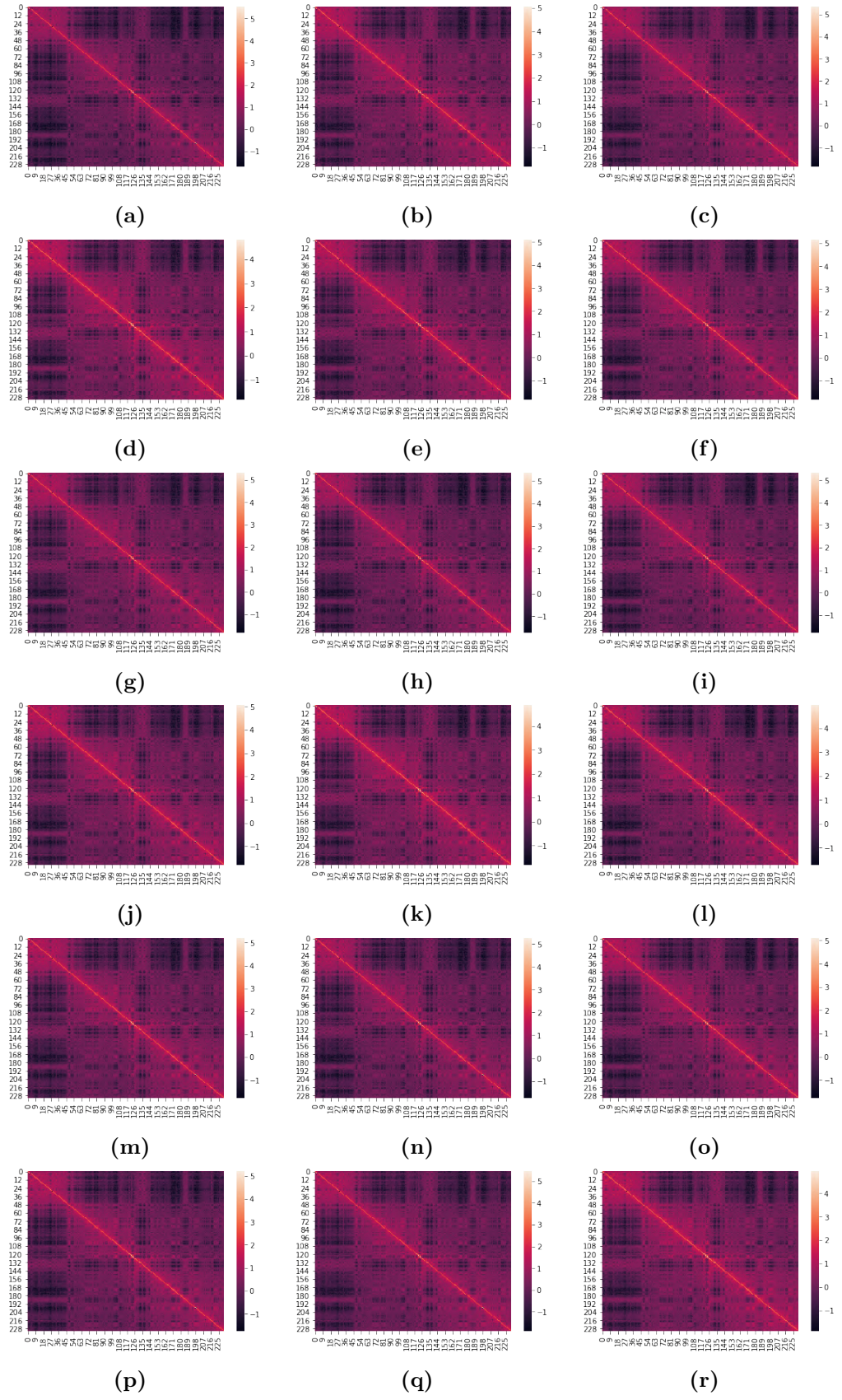


Figure 4.51: Examples of 18 different synthetic Hi-C matrices generated starting from the original Hi-C matrix related to the control MB at 1 Mb resolution.

In figure 4.51 we have reported 18 examples of synHi-C maps among the 100 synthetically generated. They are very similar, at least from the point of view of visualization via heatmap, even if the differences between the values of the entries can be appreciated if you look at the colorbar in which the values corresponding to each color are marked. Obviously the differences are given by the noise component which is introduced according to the technique illustrated previously in figure 4.43. Furthermore, to explore the generated synthetic data, we computed for each of the 100 synthetic matrices the Pearson correlation coefficient between the entries of the synthetically reconstructed matrix with the original ones. These correlation values are then reported in the histogram in figure 4.52, in such a way as to understand the distribution followed by the correlation for a large number of synthetic matrices.

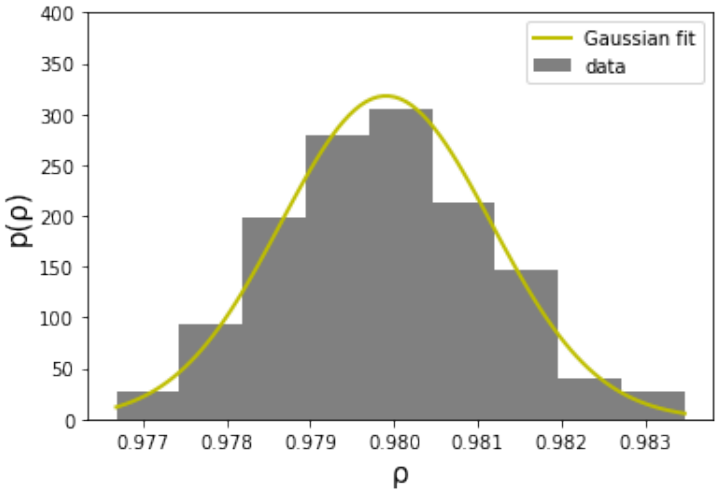


Figure 4.52: Histogram for the Pearson correlation coefficient ρ (grey) for 100 pairs of synthetic Hi-C matrices with the original control MB at 1 Mb resolution. In green the Gaussian best fit (p-value $< 10^{-3}$).

As can be seen from the histogram in figure 4.52 there is excellent compatibility between the Gaussian and the histogram of the correlation coefficients (p-value $< 10^{-3}$). This confirms that indeed the correlation coefficients of each synthetic matrices are distributed as a random variable with real values which tend to concentrate around a single average value. Therefore, the fluctuations introduced starting from the expected configuration are not unbalanced but truly random, confirming the unbiased synthetic matrix generation process. The best fit parameters of mean with its standard deviation turns out to be equal to $\rho = (0.980 \pm 0.001)$.

4.3 Blender and Virtual Reality

In this paragraph we present the results obtained from the three-dimensional visualization both starting from the total matrix GM12878 and from the pairs of samples of cases and controls. To obtain these graphic visualizations, we started from the same coordinate files as obtained through the use of the ShRec3D algorithm and we imported and processed them through the Blender software. Blender is the free and open source 3D creation suite. It supports the entirety of the 3D pipeline: modeling, rigging, animation, simulation, rendering, compositing and motion tracking, even video editing and game creation. It has a flexible Python controlled interface (API). Advanced users employ Blender's API for Python scripting to customize the application and write specialized tools; often these are included in Blender's future releases. Blender has a wide variety of tools making it suitable for almost any sort of informatics and media production. As its key features it enumerates:

- Being a fully integrated 3D content creation suite, offering a broad range of essential tools, including Modeling, Rendering, Animation and Rigging, Video Editing, VFX, Compositing, Texturing, and many types of Simulations.
- It is a cross platform, with an OpenGL GUI that is uniform on all major platforms (and customizable with Python scripts).
- It has a high-quality 3D architecture, enabling fast and efficient creation workflow.
- It boasts active community support.
- It has a small executable, which is optionally portable.

Blender has several workspaces, the default startup shows the "Layout" workspace in the main area which is where our results are graphically visualized. Among other workspaces, for our purposes we exploited the scripting one (for interacting with Blender's Python API and writing scripts). Once the ShRec3D algorithm has produced the spatial coordinates of the chromosomes in the cell line GM12878 and of the chromosome 1 of both the two cases (235 and 295) and the two controls (LM and MB), we proceeded with writing a python code exploiting Blender API to visualize edges connecting each couple of coordinated consecutively. Our first attempt aimed at creating Blender

edges connecting coordinates interpreted as vertices of a mesh. In Blender a *mesh* is a collection of vertices, edges, and faces that describe the shape of a 3D object: a vertex is a single point; an edge is a straight line segment connecting two vertices; a face is a flat surface enclosed by edges. However, after successfully creating desired edges we discarded this approach since it did not allow to color edges, a necessary step to better do visual comparisons between different cases and to better recognize each single chromosome's location in space in the cell line GM12878 analysis. Therefore, we proceeded with the following method: instead of creating edges to connect consecutively all vertices we define a function able to create cylinders between each couple of vertices. The key of this method is the size of the cylinder: as the radius of the cylinder gets smaller, the cylinder's dimension resembles a simple line (i.e. edges). Also, cylinders (as well as other mesh with surfaces) can be assigned with a material and, as a consequence, with a color. The function able to do what above mentioned is called cylinder_between and is shown in figure 4.53. The function takes as input parameters six spatial co-

```
import bpy
import math
import re

#function creating cylinder between input coordinates
def cylinder_between(x1, y1, z1, x2, y2, z2, r):

  dx = x2 - x1
  dy = y2 - y1
  dz = z2 - z1
  dist = math.sqrt(dx**2 + dy**2 + dz**2)

  bpy.ops.mesh.primitive_cylinder_add(
      radius = r,
      depth = dist,
      location = (dx/2 + x1, dy/2 + y1, dz/2 + z1)
  )

  phi = math.atan2(dy, dx)
  theta = math.acos(dz/dist)

  bpy.context.object.rotation_euler[1] = theta
  bpy.context.object.rotation_euler[2] = phi



#filename defined giving its path in the computer
filename = '/Users/noemisgambelluri/Desktop/Matrice_tot_ess.csv'

with open(filename, 'r') as fp:
    lines = fp.read().splitlines()
    lines = [x.strip() for x in lines]
    lines = [re.split(r'\s+',x) for x in lines]
    lines = [list(map(float,x)) for x in lines]

#lines is a list of tuples containing floats
#access each element of each tuple as required by the function
for i in range(len(lines)-1):
    cylinder_between(lines[i][0],lines[i][1],lines[i][2], lines[i+1][0], lines[i+1][1], lines[i+1][2], 0.5)
```

Figure 4.53: Function definition of cylinder_between code.

ordinates corresponding to the couple of vertices that need to be connected and the cylinder's radius dimension. First, the distance between the pair of

coordinates under consideration is computed, the obtained value represents the depth of the built cylinder. The cylinder is built using the pyhton library bpy, specifically the function bpy.ops.mesh.primitive_cylinder_add(). Bpy.ops provides python access to calling operators, this includes operators written in C, Python or macros. In this function a mesh operator is used in order to construct a cylinder ("primitive_cylinder_add") giving as parameters the radius, the depth ("dist" computed before) and location (corresponding to the centre location) of the cylinder. After that, taking into considerations all possible configurations of the cylinders in the 3D space, the azimuthal and polar angle have been defined through the function bpy.context.object.rotation_euler. Bpy.context.object function has access to the information regarding the currently active object, in our case the cylinder, setting rotations to be equal to the azimuthal and polar angle previously defined. Once the definition of the cylinder_between function has been completed, we proceeded with importing the cell line GM12878 and the chromosomes 1 of the two cases and of the two controls. Each filename is defined storing its path in the variable "filename" and is then opened and organized in list of tuples of floats, corresponding to the vertices coordinates, in the variable "lines". Finally, to call the cylinder_between function, we access the six coordinates corresponding to the couple of implied vertices with a for loop working on each of the three elements of tuples' list stored in "lines". Thereafter we continued with the second section of the code shows in figures 4.54 and 4.55. The second part of the Blender code aims at specifically selecting only a portion of all the constructed cylinders in order to assign different materials hence different colors to each portion of cylinders. It uses bpy.context.view_layer.objects to have access to all active objects of the type "mesh" (such as all the previously constructed cylinders with the function) and the operator select_all to deselect all the automatically selected cylinder. This is done to enable a preferred manual selection of portions of cylinders. Computationally, this specific selection is performed by a for loop on the cylinders' names: in obj.active all objects with a specific name (e.g., Cylinder.001 and Cylinder.002) chosen by the user are selected. At this point, selected portions of active objects of the type "mesh", in our case cylinders, are assigned with a previously created material. Cell line GM12878 contains information about 23 chromosome, thus a material for each of them needs to be created. To do so, the function bpy.data.materials.new() has been used. It is also possible to assign a color to each material through diffuse_color by

```
import bpy
import math
import re


objs = bpy.context.view_layer.objects
names = [o.name for o in objs if o.type == 'MESH']

bpy.ops.object.select_all(action='DESELECT')

for i in range(0, 230):
    objs.active = objs.get(names[i]).select_set(True)


matg = bpy.data.materials.new('Green') #chr1
matg.diffuse_color = (0, 1, 0, 0.8)
matg.specular_intensity = 0.3

matr = bpy.data.materials.new('Red') #chr3
matr.diffuse_color = (1, 0, 0, 0.8)
matr.specular_intensity = 0.3

matb = bpy.data.materials.new('Blue') #chr2
matb.diffuse_color = (0, 0, 1, 0.8)
matb.specular_intensity = 0.3

matp = bpy.data.materials.new('Purple') #chr4
matp.diffuse_color = (0.627,0.125 ,0.941 , 0.8)
matp.specular_intensity = 0.3

mata = bpy.data.materials.new('azzurro') #chr5
mata.diffuse_color = (0, 0.6 ,0.8 , 0.8)
mata.specular_intensity = 0.3

matgiallo = bpy.data.materials.new('giallo') #chr6
matgiallo.diffuse_color = (1, 1 ,0 , 0.8)
matgiallo.specular_intensity = 0.3

matvs = bpy.data.materials.new('verde scuro') #chr7
matvs.diffuse_color = (0.13, 0.37, 0.13 , 0.8)
matvs.specular_intensity = 0.3

matocra = bpy.data.materials.new('ocra') #chr9
matocra.diffuse_color = (0.8, 0.5 ,0.2 , 0.8)
matocra.specular_intensity = 0.3

matmar = bpy.data.materials.new('marrone') #chrX
matmar.diffuse_color = (0.55, 0.28 ,0.15 , 0.8)
matmar.specular_intensity = 0.3
```

Figure 4.54: First part of the code: materials creation.

setting the RGB code and the desired intensity as parameters, as it can be seen in figure 4.55.

```
matstep = bpy.data.materials.new('steppa')#chr20
matstep.diffuse_color = (0.9, 0.74 ,0.23 , 0.8)
matstep.specular_intensity = 0.3

matbl = bpy.data.materials.new('bluino')#chr19
matbl.diffuse_color = (0.23, 0.4 ,0.65 , 0.8)
matbl.specular_intensity = 0.3

matgr = bpy.data.materials.new('grigio')#chr22
matgr.diffuse_color = (0.23, 0.23 ,0.23 , 0.8)
matgr.specular_intensity = 0.3


matrss = bpy.data.materials.new('rosso scuro')#chr21
matrss.diffuse_color = (0.89, 0.09 ,0.05 , 0.8)
matrss.specular_intensity = 0.3


for o in bpy.context.selected_objects:
    if o.type == 'MESH':
        o.active_material = matrss
```

Figure 4.55: Second part of the code: assigning materials to each cylinder.

4.3.1 GM12878 cell line 3D visualization results

The following materials have been created for each chromosome of the whole GM12878 cell line: matg (green) for Chromosome 1, matb (blue) for chromosome 2, matr (red) for chromosome 3, matp (orchid) for chromosome 4, mata (light blue) for chromosome 5, matgiallo (yellow) for chromosome 6, matvs (viridian) for chromosome 7, matfux (fuxia) for chromosome 8, matocra (ochre) for chromosome 9, matmar (brown) for chromosome X, matrosa (pink) for chromosome 10, mataran (orange) for chromosome 11, matvvs (forest green) for chromosome 12, matva (cyan) for chromosome 13, matprug (mauve) for chromosome 14, matviola (purple) for chromosome 15, matind (indigo) for chromosome 16, matcil (cherry) for chromosome 17, matfr (strawberry) for chromosome 18, matbl (frost blue) for chromosome 19, matstep (golden brown) for chromosome 20, matrss (dark red) for chromosome 21, matgr (grey) for chromosome 22. In figures 4.56 and 4.57 we reported the images for the raw Hi-C map generated with Blender which have been also visualized by using the Virtual Reality (VR). To visualize the single chromosomes, all the cylinders connecting two nodes belonging to different chromosomes have been eliminated. The same has been done for the essential matrix, varying the cylinder radius, whose images are presented in figures 4.59, 4.58 and 4.60.

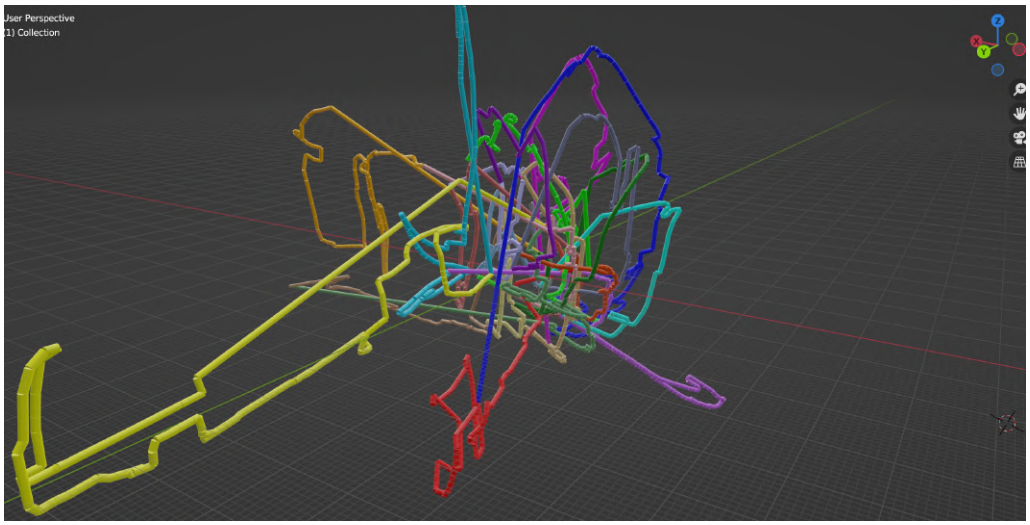


Figure 4.56: ShRec3D reconstruction starting from the original matrix of the whole raw Hi-C data by using Blender. Each material (and the corresponding colour) is referred to a different chromosome.

In this case, by using Blender and even more by observing the 3D image using the Virtual Reality, it is possible to better notice the individual chromosomes and their folding within the genome.

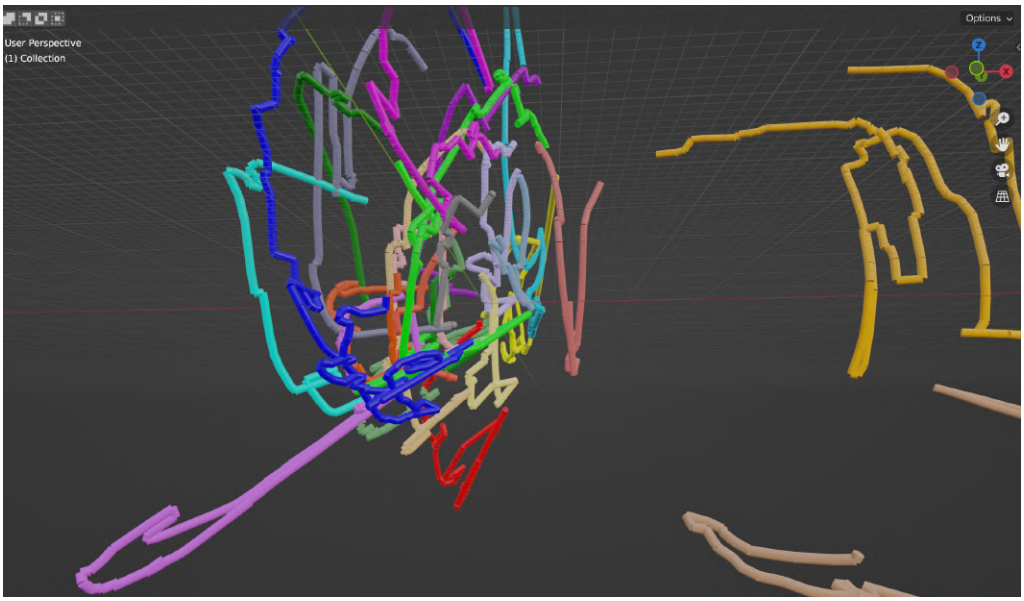


Figure 4.57: Different perspective for the ShRec3D reconstruction starting from the original matrix of the whole raw Hi-C data by using Blender. Each material (and the corresponding colour) is referred to a different chromosome.

Unlike the case of the original matrix, in the synthetic one the divisions between the chromosomes are even better appreciated, in fact the noise has been completely removed from the latter, giving a clearer view of the individual chromosomes. In particular, it can be noticed that the larger chromosomes tend to position themselves in the external regions, such as chromosomes 1 and 2, while the smaller ones tend to compact themselves in the central regions.

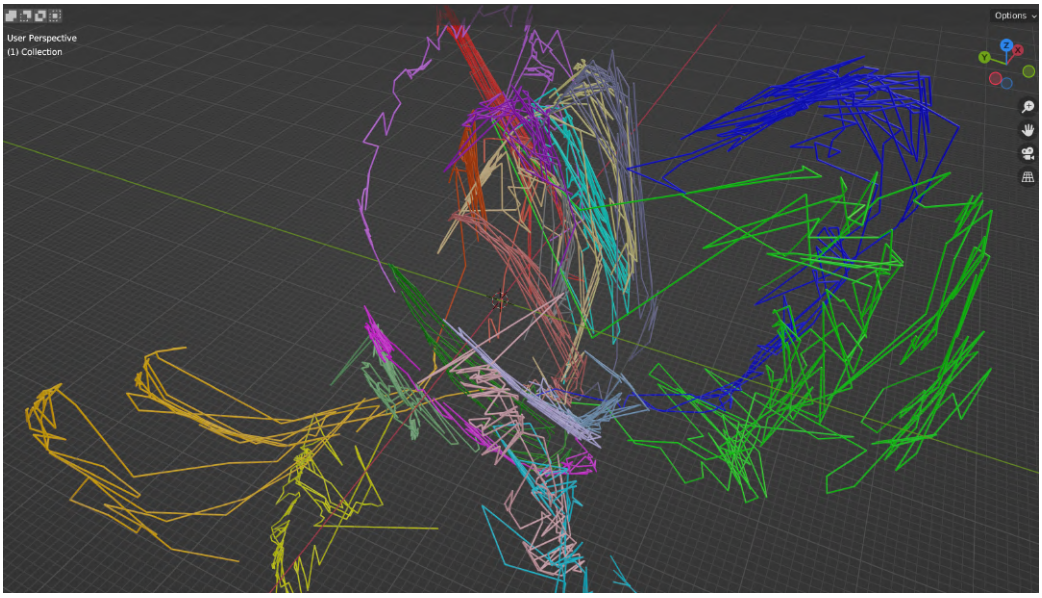


Figure 4.58: ShRec3D reconstruction starting from the essential matrix of the whole raw Hi-C data by using Blender. Each material (and the corresponding colour) is referred to a different chromosome.

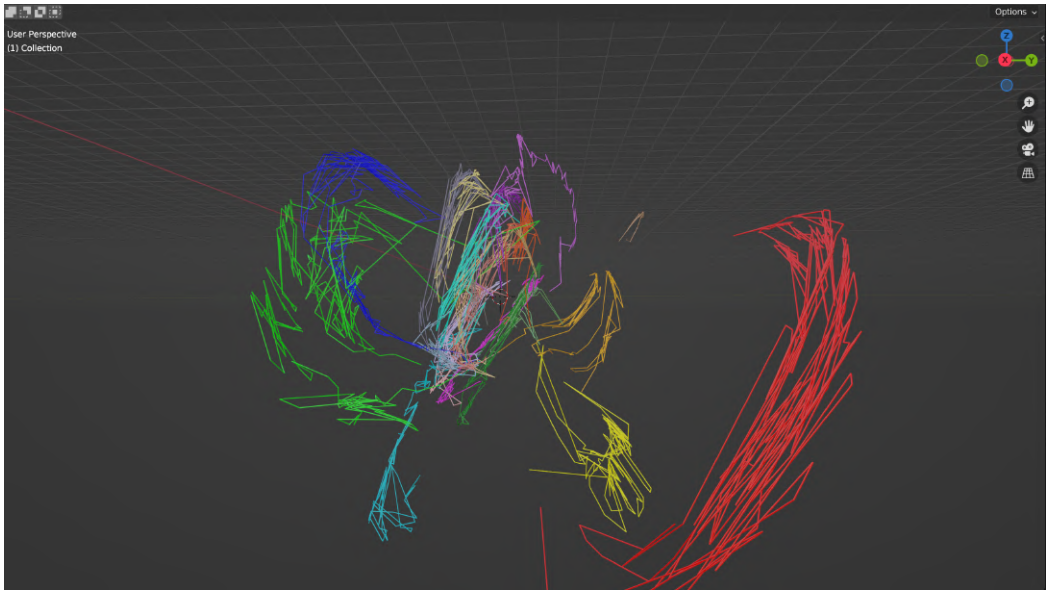


Figure 4.59: Different perspective for the ShRec3D reconstruction starting from the essential matrix of the whole raw Hi-C data by using Blender. Each material (and the corresponding colour) is referred to a different chromosome.

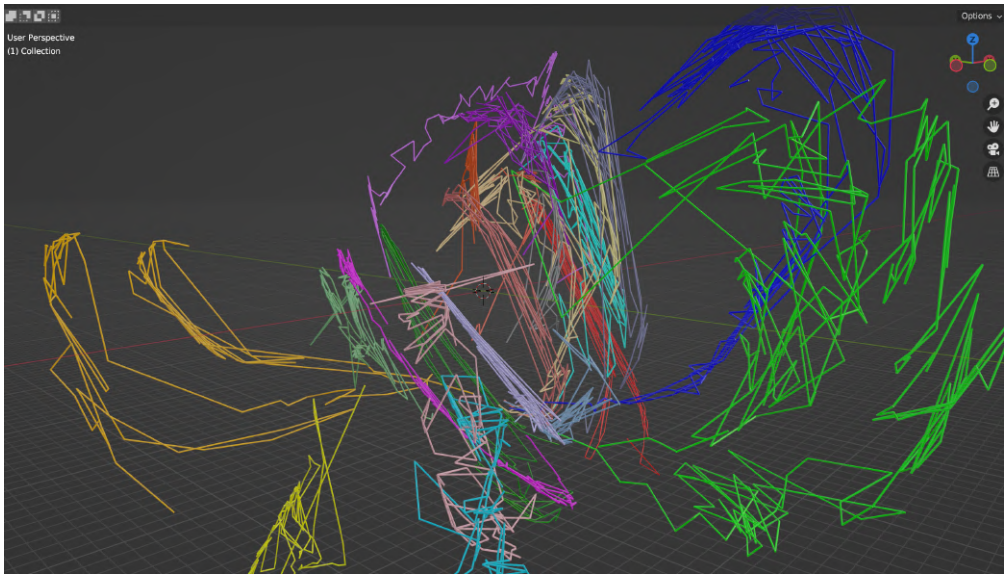
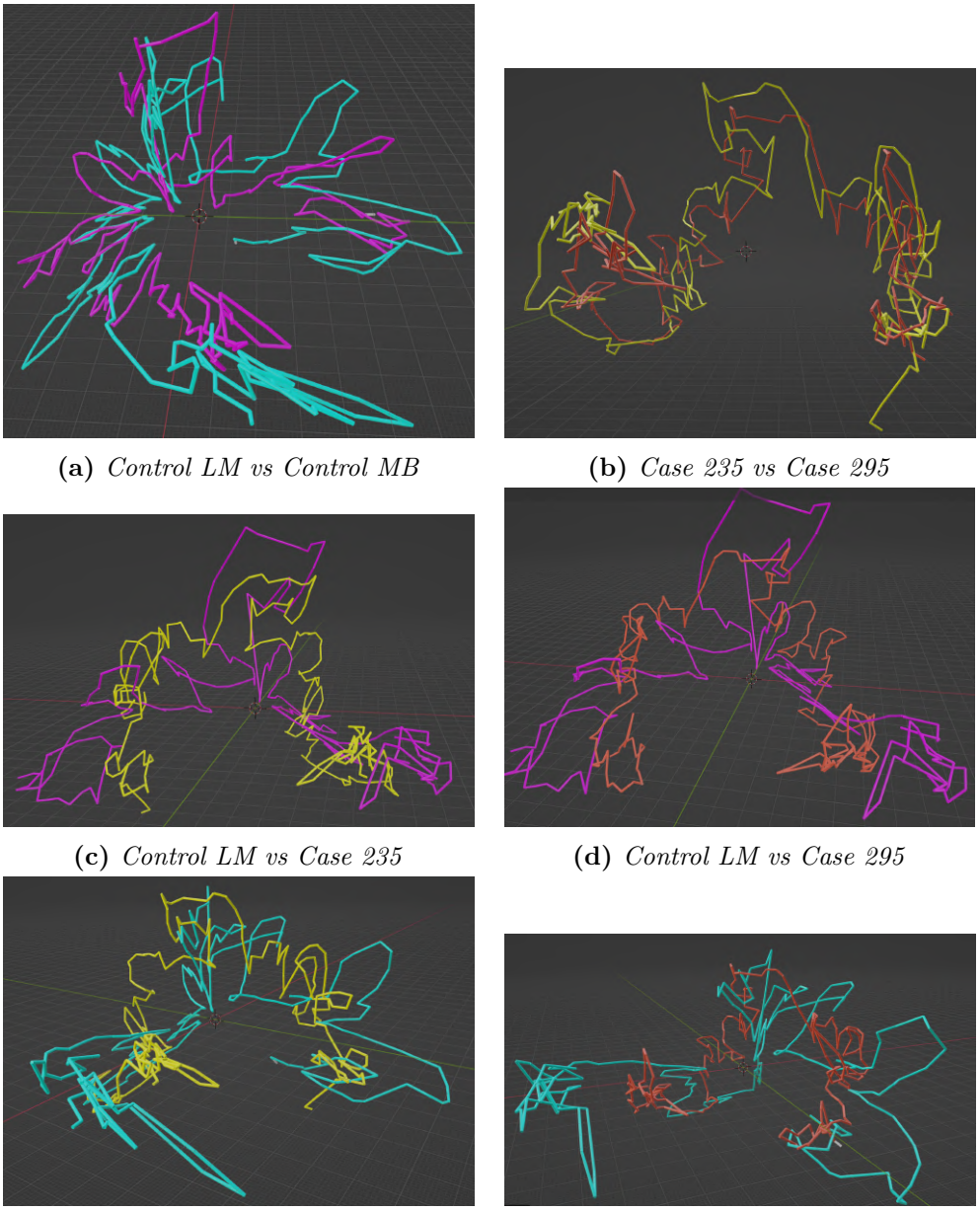


Figure 4.60: Different perspective for the ShRec3D reconstruction starting from the essential matrix of the whole raw Hi-C data by using Blender. Each material (and the corresponding colour) is referred to a different chromosome.

4.3.2 Case study 3D visualization results

Regarding the medical application case study, we worked with the chromosome 1 of two cases (case 235 and case 295) and two controls (LM and MB). To make better comparisons, we assigned to each a different material: matva (cyan) to MB, matfux (fuxia) to LM, matrss (dark red) to case 295 and matgiallo (yellow) to case 235. The six possible pairs between cases and controls built in Blender are shown in figure 4.61. Also in this case it is possible to better distinguish the points of contact and those of detachment between the individual nodes of each pair of chromosomes.



(a) *Control LM vs Control MB*

(b) *Case 235 vs Case 295*

(c) *Control LM vs Case 235*

(d) *Control LM vs Case 295*

(e) *Control MB vs Case 235*

(f) *Control MB vs Case 295*

Figure 4.61: ShRec3D reconstruction starting from the original Hi-C matrices of the four samples (two cases: 235 and 295; and two controls: LM and MB) by using Blender. Each sample has a different colour. Control MB: cyan; Control LM: fuxia; Case 235: yellow; Case 295: dark red.

Chapter 5

Conclusions

Hi-C matrices are milestones for the qualitative and at the same time quantitative study of genome folding, its organization into chromosomal territories, compartments and topological domains. In this type of data we have seen how it was possible to characterize the signal-to-noise ratio starting from a spectral analysis on different types of Hi-C data at different resolutions (1 Mb and 100 kb). Through the spectral analysis of the Hi-C matrices, under an appropriate preprocessing, one of the main characteristics of these data has been highlighted: the distinction between the signal component, linked to the intrinsic biological properties, and the noise component, linked to the statistical fluctuations. We have seen how the latter follow the distribution known as Wigner's semicircle function, which describes the distribution of eigenvalues for a symmetric random matrix. On the other hand, the signal part is characterized by a small percentage of isolated eigenvalues in the spectrum, even at large distances from the noise semicircle centered at zero. Starting from these considerations, the part of the signal spectrum was therefore isolated in order to be able to reconstruct, through the projectors corresponding to the signal component, the essential matrix (essHi-C), which contains all the significant biological information on the chromatin structure. To compute the number of signal projectors, simulated random matrices were generated, whose distribution of eigenvalues coincides precisely with Wigner's semicircle function, in such a way as to estimate the demarcation threshold between the two distinct components. The possibility of splitting the two components that characterize the Hi-C matrices was first tested on a standard case, that of cell line GM12878, and was then applied to a case study consisting of four samples,

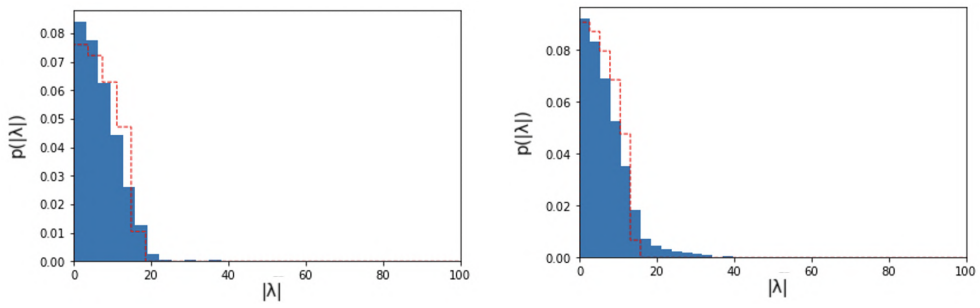
two controls (LM and MB) and two cases (235 and 295), the latter related to a rare prion disease. To validate the goodness of reconstruction of the essential matrices with respect to the original ones, various tools were used including scatter plots, ShRec3D algorithm and virtual reality (VR). The former made it possible to directly verify the correlation between the essential and the original matrices, providing positive results for the analyzed case study, whose samples resulted to be reconstructed significantly well, both for controls and for cases at 1 Mb ($\bar{\rho} = 0.969$) and 100 kb resolution ($\bar{\rho} = 0.832$). With particular reference to the case study, a further check was performed concerning the comparison between the different cases and controls, both for the original and for the essential matrices, which highlighted how, in particular for the less noisy matrices at lower resolution of 1 Mb and even more for the essential matrices, as expected there is a greater correlation between the pairs consisting of only cases or only controls ($\bar{\rho} = 0.962$) compared to the mixed pairs which contain both ($\bar{\rho} = 0.930$). Both from the analysis of the case study and the preliminary analysis on the GM12878 cell line, further confirmation of this is found by obtaining the three-dimensional conformations of the individual chromosomes. In fact we have seen how, after having performed an appropriate prorustes analysis, the similarity between the three-dimensional shapes of chromatin folding are very similar both between essential and original matrices of GM12878 ($\bar{P}_d = 0.012$) and the cases ($\bar{P}_d = 0.104$) and controls ($\bar{P}_d = 0.134$) at 1 Mb resolution. Spatial visualization was further deepened by a Virtual Reality analysis, able to show even more in detail those regions of space in which the chromosomes are more similar in terms of spatial configuration from those that are more distant. Finally, having verified the robustness of the experimental design used to extract the signal and noise components, we focused on the generation of synthetic Hi-C (synHi-C) matrices. The innovative method developed in this thesis consists in considering not only the signal component, given by the essential matrix, but also the noise component, through an appropriate reshuffling of the eigenvalues with respect to the eigenvectors related to the Wigner's semicircle. By adding the two matrices reconstructed starting from the signal projectors with the noise-related ones, a synHi-C is obtained with the same average properties of the original matrix, but with an intrinsic variability that respects the noise pattern of the Hi-C data type. To test the good reconstruction of synthetic matrices, the tools described above were used, which once again confirm that not only the spatial reconstruction using

ShRec3D and VR is faithful to a large extent to the original one ($\bar{P}_d = 0.164$), but that it is even more significant from a biological point of view, the cases and controls being more similar to each other ($\bar{\rho} = 0.945$) than to the mixed pairs ($\bar{\rho} = 0.916$). Indeed the case study analysis shows how the expected properties of greater correlation between only cases and only controls is highlighted even more when the essential matrices are considered, and since the synthetic matrices are built starting from the latter, it is also reflected in the synHi-C artificially generated. Finally, the statistical properties were also tested, generating a significant number of synthetic matrices of a given sample and verifying how the distribution of the Pearson correlation coefficient, calculated for each pair of synthetic matrix with the original one, follows a Gaussian distribution (p-value $< 10^{-3}$). This confirms that the synthetic matrix generation process is unbiased. The generation of synthetic data is useful due to the fact that it allows us to overcome the laboratory experimental phase, which is often difficult and expensive, but in particular in the case of Hi-C data it is even more fundamental as the data analyzed in the case study are little studied at the Hi-C level and there are just few samples to work with. Moreover, with the production of synHi-C it is therefore possible to have greater control over the type of data being studied, to understand its properties related to the different signal and noise components and to characterize all those methodological and variability issues of the experiment, tuning the parameters. Starting from this new way of generating synthetic data of Hi-C maps, a huge amount of future scenarios open up. Among these certainly there is the possibility of carrying out unsupervised clustering of synthetically generated cases and controls or in different cell lines. Even from the biological point of view many future directions open up, in fact, given the possibility of directly controlling the essential matrices and the number of associated projectors, it is possible to satisfy requests for the study of customized biological characteristics. Especially, the possibility of studying the intrinsic variability of a sample at different levels of organization, for example at the level of topological domains, which are configured starting from a different number of projectors used to reconstruct the essential matrix and therefore the corresponding synthetic matrix.

Appendix A

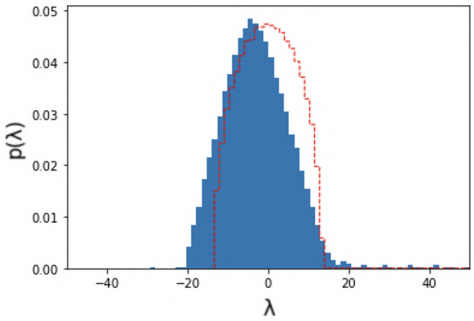
Whole GM12878 spectral analysis

In this appendix we want to probe the spectral properties of the entire healthy cell line GM12878. In particular we computed the probability distribution of the eigenvalues (just for the raw Hi-C data) and of their modulus (for both raw and ICE normalized Hi-C data) of the Hi-C matrix related to the cell line just mentioned. Figures A.1 and A.2 show this distribution as an histogram in which in dashed line it is superimposed the same distribution but calculated for a symmetric random matrix of equal linear size with entries distributed as a Gaussian with mean equal to that of the Hi-C matrix and a suitably adjusted variance.



(a) *Raw Hi-C data* **(b)** *ICE normalized Hi-C data*

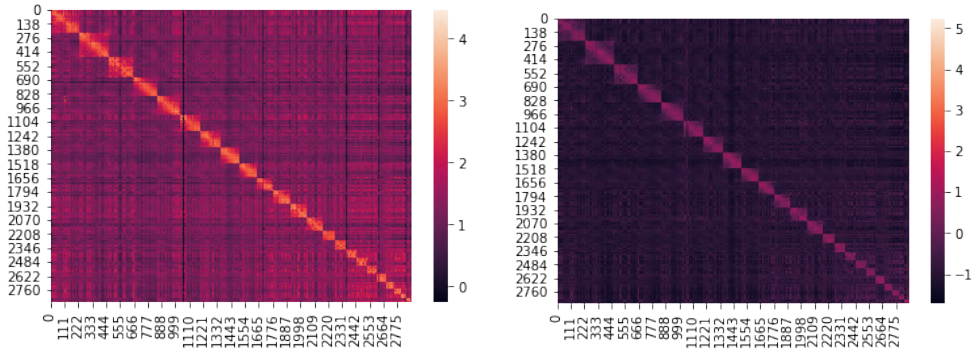
Figure A.1: Histograms of the eigenvalues' absolute values distribution for $|\lambda| < 100$ obtained from the whole raw **(a)** and ICE normalized **(b)** Hi-C data matrices (blue) and from the corresponding random matrices with the same mean and suitably adjusted standard deviation (dashed red line).



(a) *Raw Hi-C data*

Figure A.2: Histogram of the eigenvalues distribution for $-50 < \lambda < 50$ (blue) obtained from the whole raw Hi-C data matrix with the corresponding spectrum from a random matrix with same mean and suitably adjusted standard deviation (dashed red line).

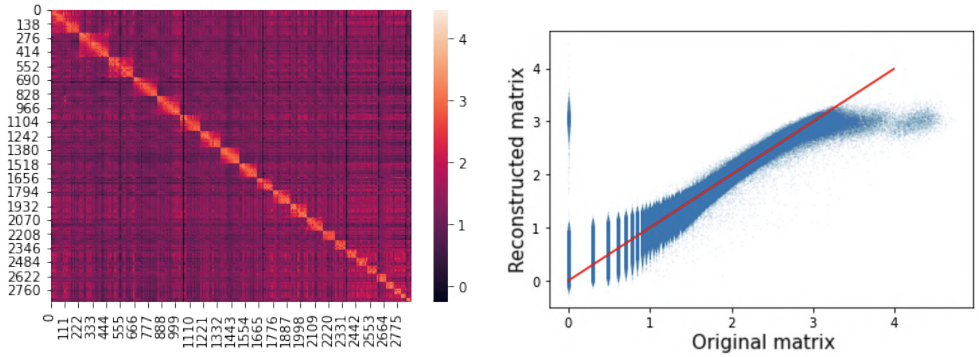
From the eigenvalue distributions in figure A.2 we can recognise that both for the raw and the ICE normalized data the spectrum that extends up to $|\lambda| \approx 20$ is largely consistent with that of random matrices, even though the spectrum is not absolutely symmetric with respect to zero, and except for a set of eigenvectors with atypically large eigenvalues in modulus. For ease of display, the histogram is limited to the first $|\lambda| < 100$ eigenvalues, even though it extends with isolated eigenvalues up to $|\lambda| \approx 3809$ for the raw Hi-C data and up to $|\lambda| \approx 2647$ for the ICE nomalized one. We discounted the non-specific random component from the matrices so to isolate their essential component. The essHi-C matrices are obtained from the spectral summation of the projectors formed by the eigenvectors related to the higher eigenvalues' modulus. The essential matrix has been reconstructed starting from all the projectors, with the exception of those relating to the eigenvectors whose modulus of the corresponding eigenvectors is lower than a certain threshold. The latter is chosen considering the spectrum of the modulus of the eigenvalues that form the random component, which coincides with the semicircle identified by those same eigenvalues generated starting from the random matrix. We have therefore reconstructed the essential matrices, both in the case of raw and ICE normalized matrices, starting from a number of projectors $n^* = 46$, which is equal to the number of eigenvalues outside the random component identified by $|\lambda| < 20$. The essHi-C matrices are shown in figure A.3.



(a) *Raw Hi-C data* **(b)** *ICE normalized Hi-C data*

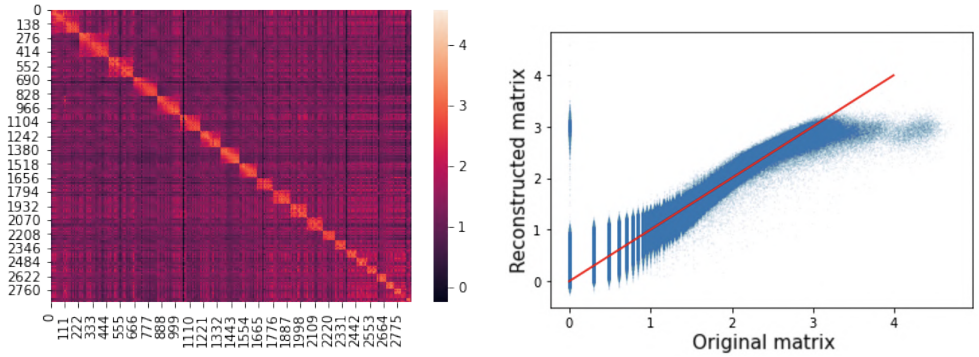
Figure A.3: Essential raw and ICE normalized Hi-C matrices.

From the essential matrices obtained in figure A.3 it can be seen how they are much less noisy and more clearly defined both as regards the blocks of the single chromosomes and for the interaction domain in which distinctive contact patterns emerge. We have therefore reconstructed the essential matrices gradually lowering the number of projectors n^*. The results are shown in figures A.4 and A.5, which are placed side by side with those related to the scatter plots between the entries of the upper triangular matrix (so to avoid double counting, since the matrix is symmetric) of the original Hi-C maps and the corresponding essential ones.



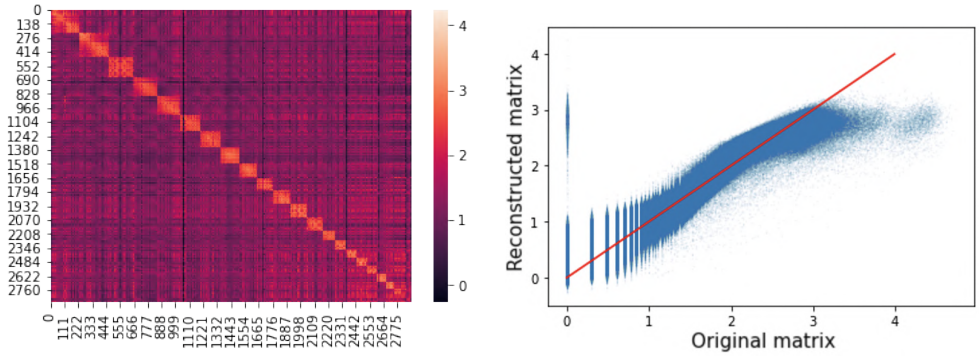
(a) *EssHi-C with* $n^* = 46$ **(b)** *Scatter plot* ($\rho = 0.935$)

Figure A.4: **(a)**: essential matrix from the whole Hi-C data reconstructed starting from $n^* = 46$ highest-ranking projectors. **(b)**: the corresponding scatter plot and Pearson correlation coefficient ρ.



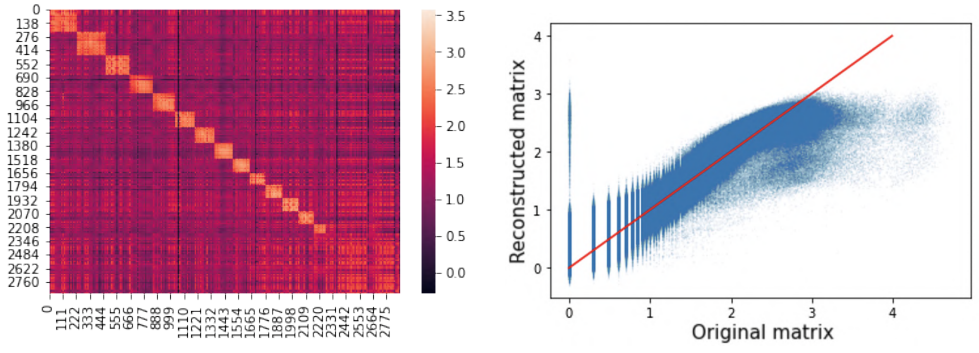
(a) *EssHi-C with* $n^* = 36$

(b) *Scatter plot* ($\rho = 0.932$)



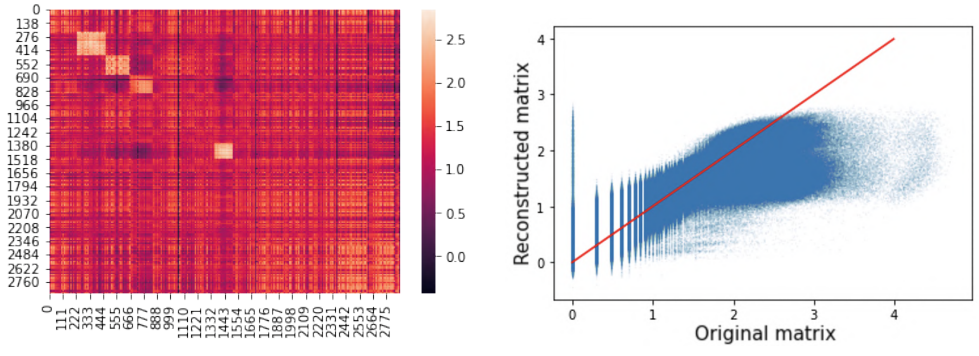
(c) *EssHi-C with* $n^* = 26$

(d) *Scatter plot* ($\rho = 0.925$)



(e) *EssHi-C with* $n^* = 16$

(f) *Scatter plot* ($\rho = 0.899$)



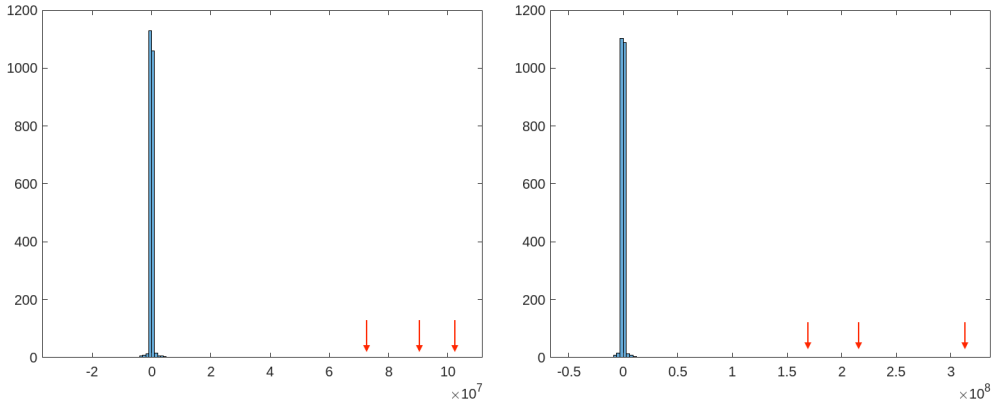
(g) *EssHi-C with* $n^* = 6$

(h) *Scatter plot* ($\rho = 0.794$)

Figure A.5: First column: essential matrices from the whole Hi-C data reconstructed starting from different n^* highest-ranking projectors. **Second column**: the corresponding scatter plots and Pearson correlation coefficient ρ.

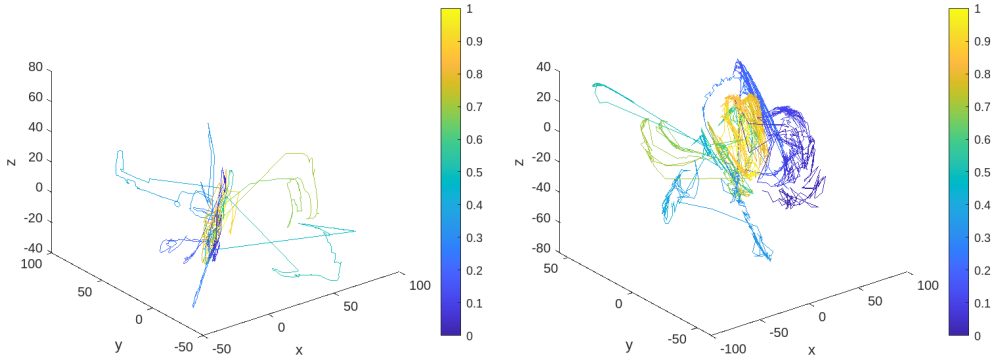
In case of the essential matrices of the whole matrix seen in figures A.4 and A.5, the correlation of the reconstructed images with respect to the original one remains almost constant as the number of projectors decreases, up to a certain value of $n^* \approx 6$ in which the correlation coefficient varies significantly until it is no longer even possible to distinguish the blocks relating to the single chromosomes, except in case of the larger ones such as chromosome 2, chromosome 3 and the chromosome X. What is striking is precisely the presence of the X chromosome, which is not among the largest chromosomes present, and the absence of other larger chromosomes such as 1 and 4. In fact, one would expect to observe that the smaller blocks disappear as the number of projectors used for the reconstruction is reduced. This is due to the fact that the smaller blocks are considered as noise and therefore they no longer appear along the main diagonal. However, in this case the X chromosome stands out among all the others, both larger and smaller, remaining clearly visible together with two other chromosomes (2 and 3) and part of chromosome 4. This could be the sign of some similarity in terms of structure between the aforementioned chromosomes. Finally, it can be seen that with a number of projectors equal to $n^* = 16$ it returns to the situation in which all the largest chromosomes are present except the smallest ones which blend into the noise and by further increasing n^* we get to reconstruct all the different chromosomes, including the smaller ones. Regarding the scatter plots, it can be seen that as the number of projectors decreases, they are increasingly scattered from the correlation line $y = x$, meaning that the values are worse reconstructed. In particular, a characteristic tail is also noted in all the scatter plots, meaning that for larger values the reconstruction works worse and in particular there is an underestimation of the latters. They correspond to the values along the main diagonal, i.e. those around the blocks of the single chromosomes.

Also in the case of the whole matrix we applied the ShRec3D algorithm to obtain images relating to the matrix as a whole of chromosomes. The spectra of the corresponding Gram matrices, both for the raw and the essential cases, are reported in figure A.6, where the arrows indicate the three greater eigenvalues, which as expected are separated from the main peak around zero. The three-dimensional plots obtained starting from the original Hi-C matrix and the essential one are displayed in figure A.7.



(a) *Original Hi-C matrix* **(b)** *Essential Hi-C matrix* ($n^* = 46$)

Figure A.6: **(a)**: original matrix of the whole raw Hi-C data. **(b)**: the corresponding essential matrix reconstructed starting from $n^* = 46$ highest-ranking projectors.



(a) *Original Hi-C matrix* **(b)** *Essential Hi-C matrix* ($n^* = 46$)

Figure A.7: **(a)**: ShRec3D reconstruction starting from the original matrix of the whole raw Hi-C data. **(b)**: ShRec3D reconstruction starting from the corresponding essential matrix reconstructed by using $n^* = 46$ highest-ranking projectors.

In the figure A.7 a colorbar has been used to indicate with different color gradations the sequence of points that follow one another along the linear chain of the entire genome. In this way it is possible to notice that there are denser regions of space characterized by very similar colors. These correspond precisely to the chromosome territories and therefore to the blocks of the Hi-C whole matrix. To compare the two 3D configurations, original and essential, we superimposed them by applying the procrustes analysis and then calculated the procrustes distance P_d. The two superimposed images are shown in figure A.8.

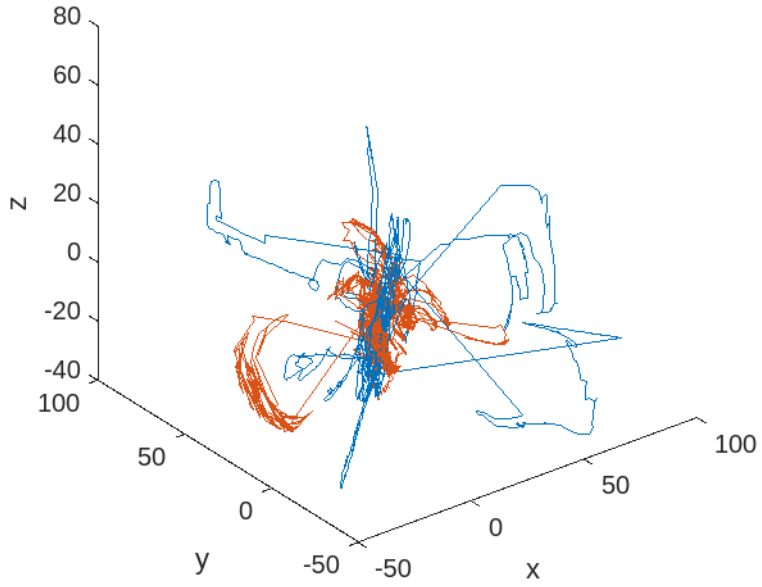


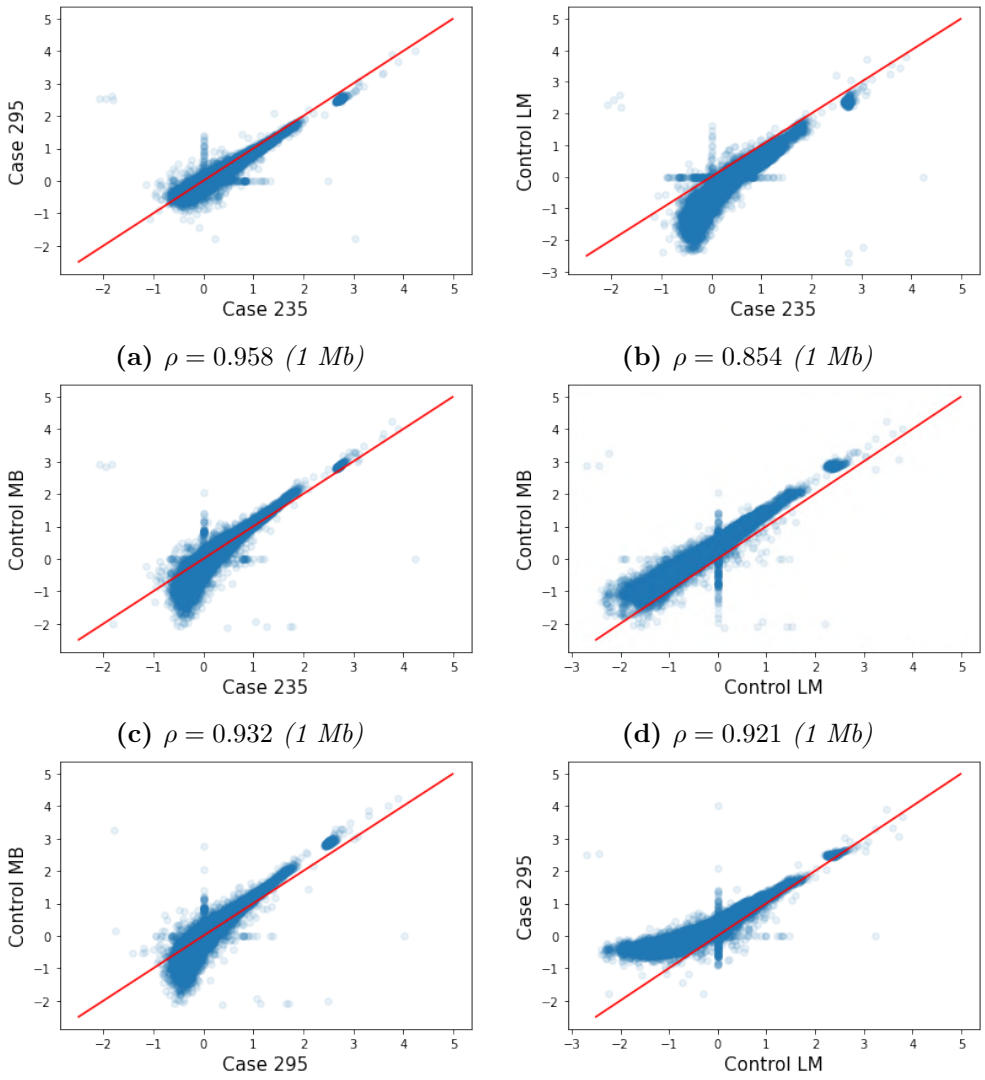
Figure A.8: ShRec3D reconstructions starting from the original matrix of the whole raw Hi-C data (blue line) together with the corresponding essential matrix (red line) reconstructed by using $n^* = 46$ highest-ranking projectors ($P_d = 0.561$).

From the figure representing the overlap between the two 3D configurations it is not immediate to extract detailed information on single chromosomes. This is due to the fact that most of the chromosomes are concentrated and tangled in the center, overlapping each other. At least at this level, we can limit ourselves to considering the value of procrustes distance, which turns out to be equal to $P_d = 0.561$. This value tells us that the overlap is not optimal, but not completely fallacious either. Surely, at least judging by the external chromosomes, it can be seen that they are not properly superimposed. This is probably due to noise present in the original Hi-C matrix. For a more detailed analysis, see the paragraph 4.3.

Appendix B

Scatter plots

Here we show the scatter plots between each pair of samples for the two cases (235 and 295) and the two controls (LM and MB) for the original Hi-C matrices at 1 Mb (see figure B.1) and 100 kb (see figure B.2), the corresponding essential matrices (1 Mb: see figure B.3 and 100 kb: see figure B.4) and the synthetic Hi-C matrices generated starting from the original matrices at 1 Mb resolution (see figure B.5). The corresponding Pearson correlation coefficient ρ is reported for each graph.

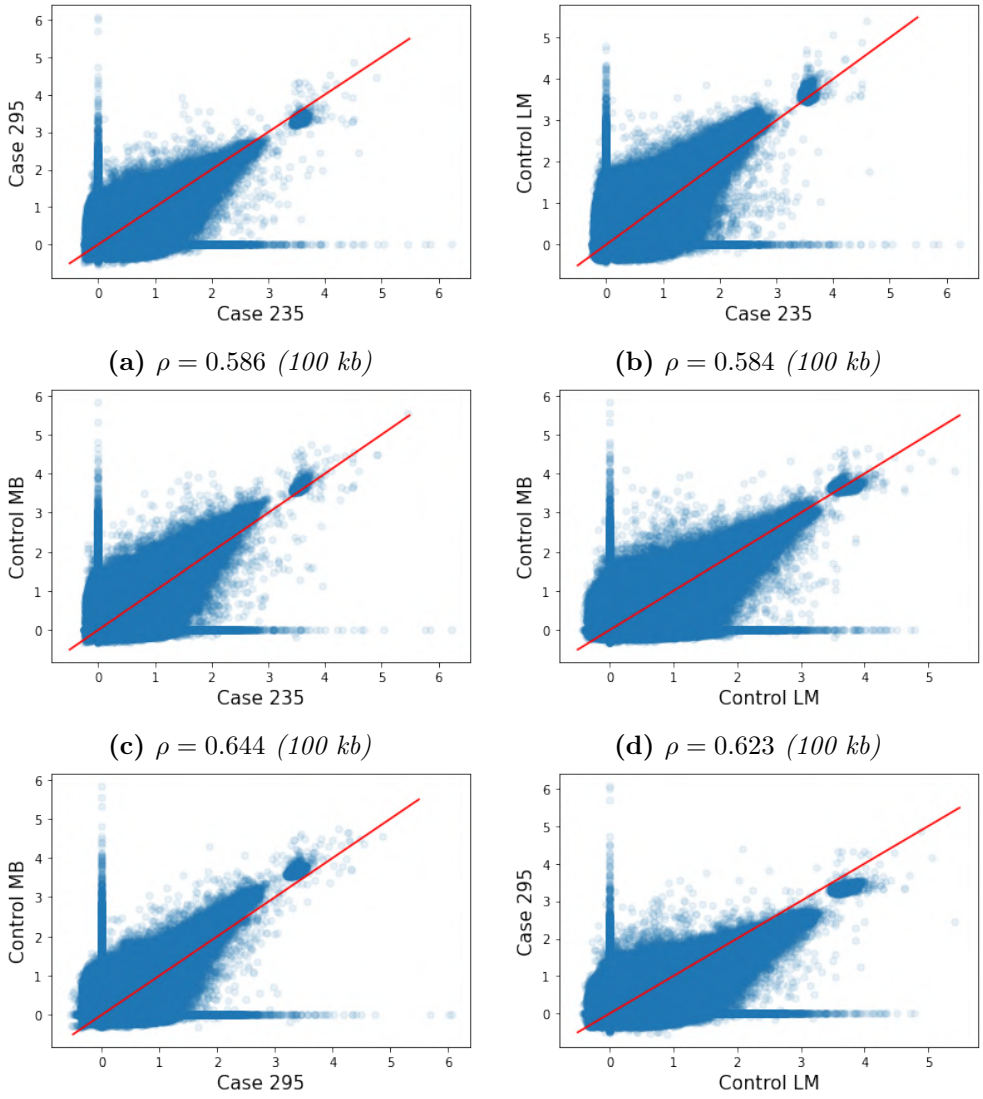


(a) $\rho = 0.958$ *(1 Mb)* **(b)** $\rho = 0.854$ *(1 Mb)*

(c) $\rho = 0.932$ *(1 Mb)* **(d)** $\rho = 0.921$ *(1 Mb)*

(e) $\rho = 0.932$ *(1 Mb)* **(f)** $\rho = 0.875$ *(1 Mb)*

Figure B.1: Scatter plots from the original matrices after the preprocessing step for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb resolution. The Pearson correlation coefficient ρ is listed for each plot.

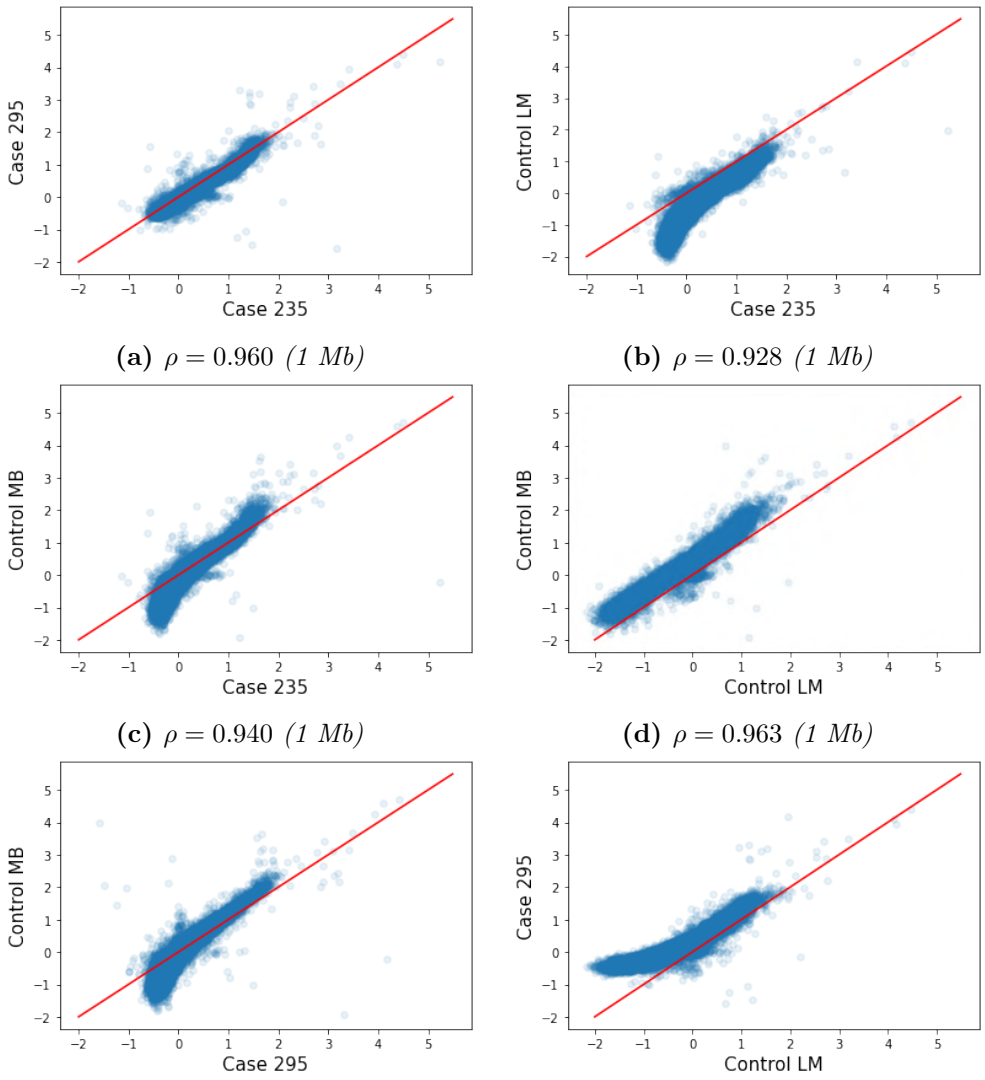


(a) $\rho = 0.586$ *(100 kb)* **(b)** $\rho = 0.584$ *(100 kb)*

(c) $\rho = 0.644$ *(100 kb)* **(d)** $\rho = 0.623$ *(100 kb)*

(e) $\rho = 0.644$ *(100 kb)* **(f)** $\rho = 0.572$ *(100 kb)*

Figure B.2: Scatter plots from the original matrices after the preprocessing step for the two cases (235 and 295) and the two controls (LM and MB) at 100 kb resolution. The Pearson correlation coefficient ρ is listed for each plot.

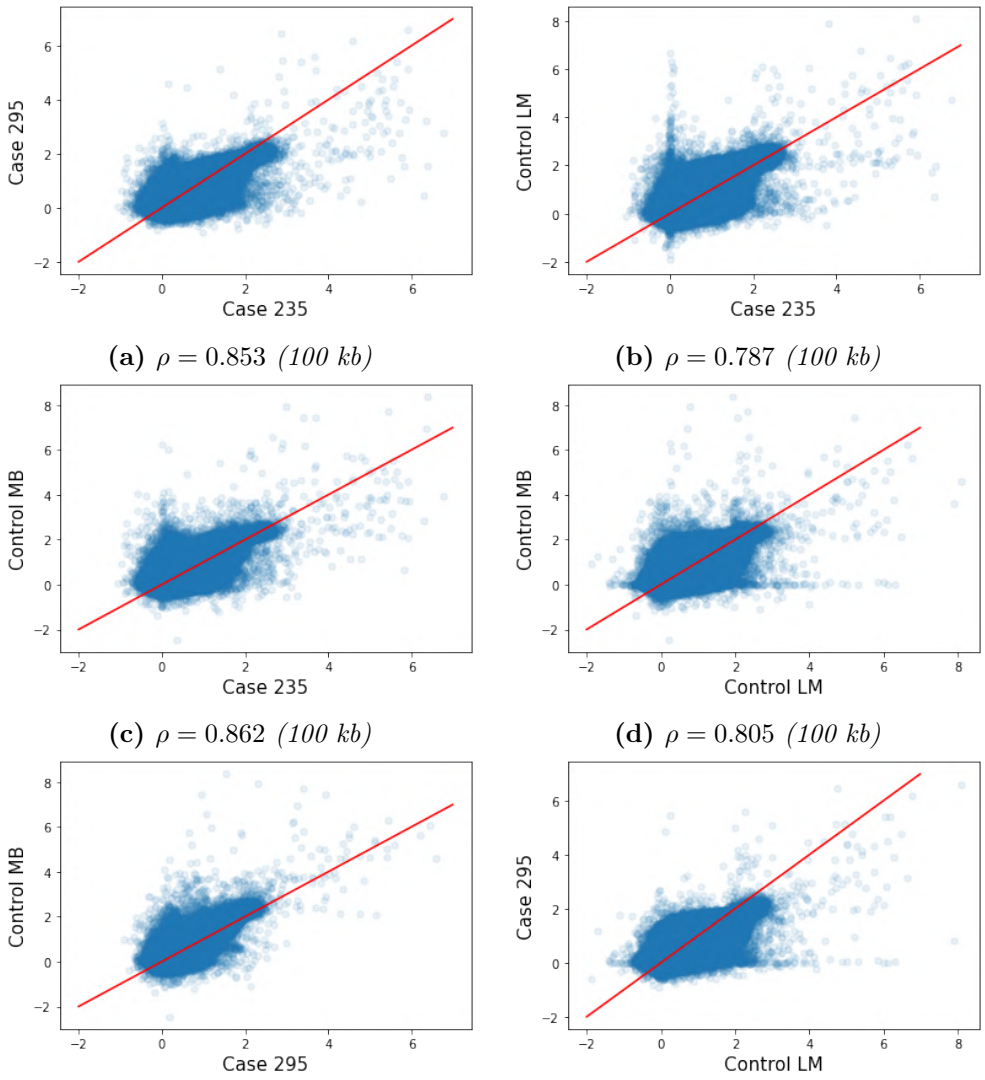


(a) $\rho = 0.960$ *(1 Mb)* **(b)** $\rho = 0.928$ *(1 Mb)*

(c) $\rho = 0.940$ *(1 Mb)* **(d)** $\rho = 0.963$ *(1 Mb)*

(e) $\rho = 0.932$ *(1 Mb)* **(f)** $\rho = 0.918$ *(1 Mb)*

Figure B.3: Scatter plots from the essential matrices after the preprocessing step for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb resolution. The Pearson correlation coefficient ρ is listed for each plot.

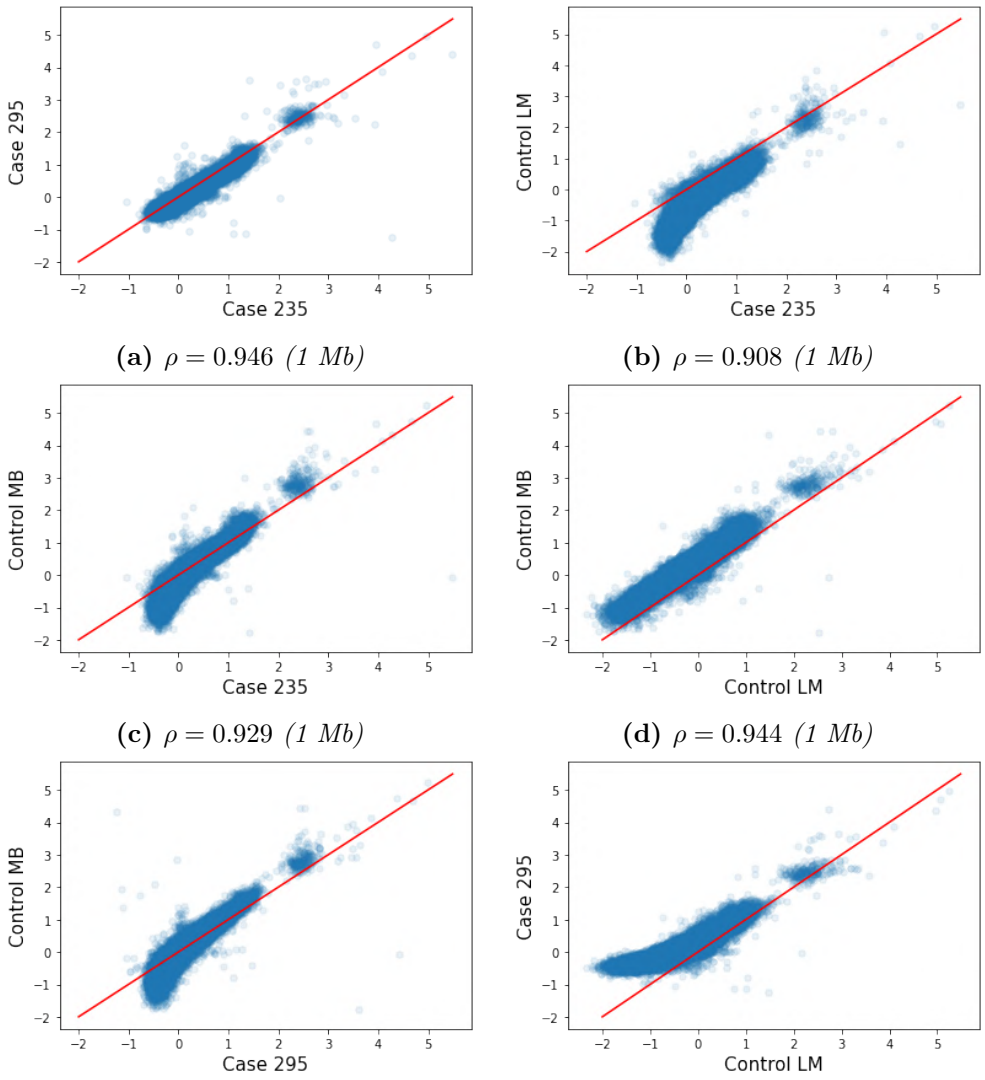


(a) $\rho = 0.853$ *(100 kb)* **(b)** $\rho = 0.787$ *(100 kb)*

(c) $\rho = 0.862$ *(100 kb)* **(d)** $\rho = 0.805$ *(100 kb)*

(e) $\rho = 0.905$ *(100 kb)* **(f)** $\rho = 0.807$ *(100 kb)*

Figure B.4: Scatter plots from the essential matrices after the preprocessing step for the two cases (235 and 295) and the two controls (LM and MB) at 100 kb resolution. The Pearson correlation coefficient ρ is listed for each plot.



(a) $\rho = 0.946$ *(1 Mb)* **(b)** $\rho = 0.908$ *(1 Mb)*

(c) $\rho = 0.929$ *(1 Mb)* **(d)** $\rho = 0.944$ *(1 Mb)*

(e) $\rho = 0.923$ *(1 Mb)* **(f)** $\rho = 0.902$ *(1 Mb)*

Figure B.5: Scatter plots from the synthetic matrices after the preprocessing step for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb resolution. The Pearson correlation coefficient ρ is listed for each plot.

Appendix C

ShRec3D Python code

To perform the ShRec3D algorithm as described in the paragraph 3.7 from a computational point of view we wrote a code using the Python programming language shown in figure C.1.

```
from scipy.sparse.linalg import eigsh
import numpy.matlib
def SchRec3D(A):
    Data = 1./A;
    n = len(A[0])
    for k in range(0,n):
        i2k = np.matlib.repmat(Data[:,k], n, 1).T
        k2j = np.matlib.repmat(Data[k,:], n, 1)
        Data = np.minimum(Data, i2k+k2j)
    n = len(Data)
    center=np.zeros((n,1))
    for i in range(0,n):
        for j in range(0,n):
            center[i] = center[i] + (Data[i,j])**2 - 1/n*np.dot(Data[j,j:n],Data[j:n,j])
    center[center<0] = 0
    center = (center/n)**(1/2)
    distmat=(1/2)*(np.matlib.repmat((center*center),1,n)+np.matlib.repmat((center*center).T,n,1)-(Data*Data))
    V,D=eigsh(distmat, k=3)
    XYZ=np.array([D[:,0]*(V[0])**(1/2), D[:,1]*(V[1])**(1/2), D[:,2]*(V[2])**(1/2)])
    scale=100/np.max(np.real(XYZ))
    XYZ=XYZ*scale

    return XYZ
```

Figure C.1: ShRec3D Python code.

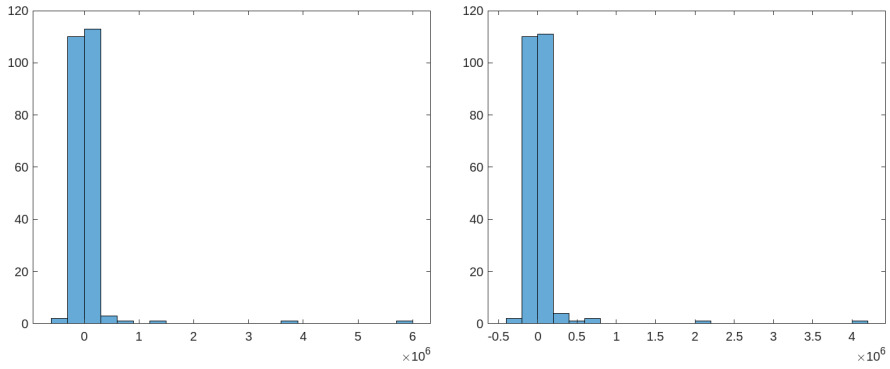
As a first step, we need to apply the Floyd-Warshall algorithm. To do this we first modified the Hi-C matrix by replacing each input value with its inverse, thus assigning a distance between the nodes of the graph described by the Hi-C map seen as an adjacency matrix, equal to the inverse of the respective frequency of normalized contact. We then wrote a for loop that allowed us to compute the minimum distance between each pair of nodes i and j. The latter was first implemented to calculate the single possible paths between node i and node j in order for them to pass through a generic node k among all those present. To calculate the distance between node i and the generic node k we used the repmat function of the numpy matlib library. It allows to repeat a given matrix (or array in our case) $M \times N$ times. In our case this distance

has been calculated considering the k-th data vector of the transformed Hi-C matrix, repeated n times along the columns (to go from i to k) or rows (to go from k to j), where n is the linear size of the Hi-C matrix. The sum of these two matrices, which represents the distance sought between nodes i and j passing through a generic node k that was run in the for loop along n, was then compared point-wise with the original one. The minimum of the two was therefore taken to obtain, once the cycle was completed, the shortest path matrix containing in each entry D_{ij} the minimum distance between the two nodes. At this point we have used the formulas seen in the paragraph 3.7.2 to first calculate the distance between each point i with the center of gravity, which for convenience has been abbreviated as center. Starting from these distance values calculated through two for loops as i and j vary over the entire length n of the matrix, we have calculated the metric matrix by reconstructing the distance matrix once again using the repmat function. Finally, having obtained in this way the Gram matrix ("distmat" in the code), we have extracted through the function of the scipy library eigsh, the eigenvalues and eigenvectors of the found symmetric real square matrix. Finally, the latter were used to obtain the XYZ coordinates, assigning to each of them the product between the eigenvector and the root of the corresponding eigenvalue. The coordinates have also been rescaled to get a better view that doesn't affect the newly calculated distances.

Appendix D

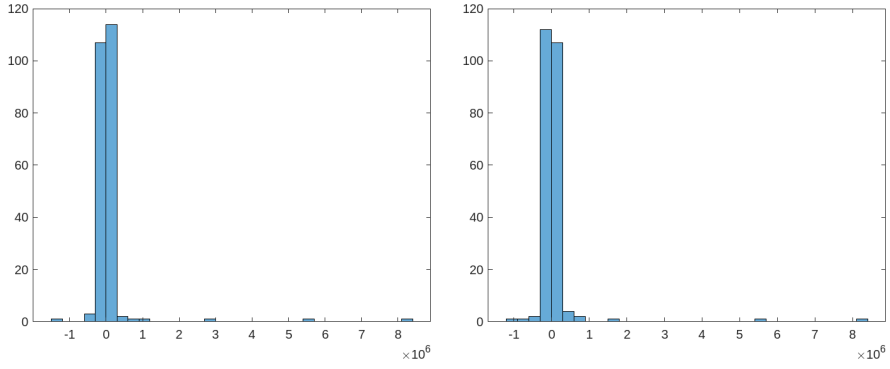
Gram matrix eigenvalue spectra

In this appendix we report the histograms of the spectrum of the eigenvalues both for the original Gram matrices at 1Mb (figure D.1) and 100 kb (figure D.2) resolution, and the essential ones (figure D.3) after carrying out the preprocessing of removing the rows and columns that add to zero, the replacem-



(a) *control LM (1 Mb)*

(b) *control MB (1 Mb)*

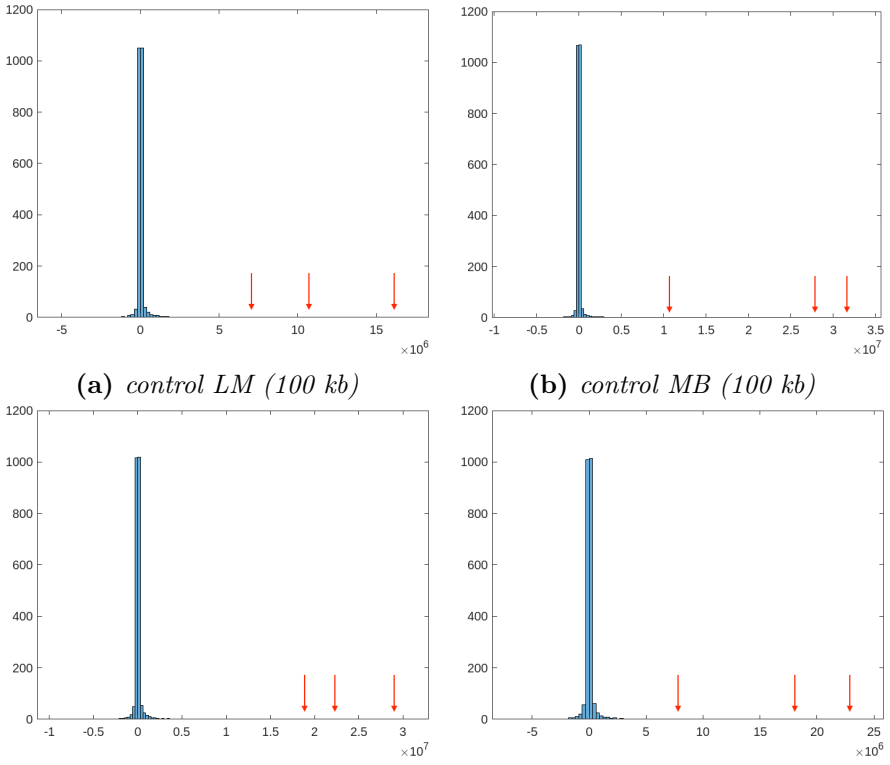


(c) *case 235 (1 Mb)*

(d) *case 295 (1 Mb)*

Figure D.1: Histograms of the eigenvalues from the original matrices after the preprocessing step and the SCN normalization for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb resolution.

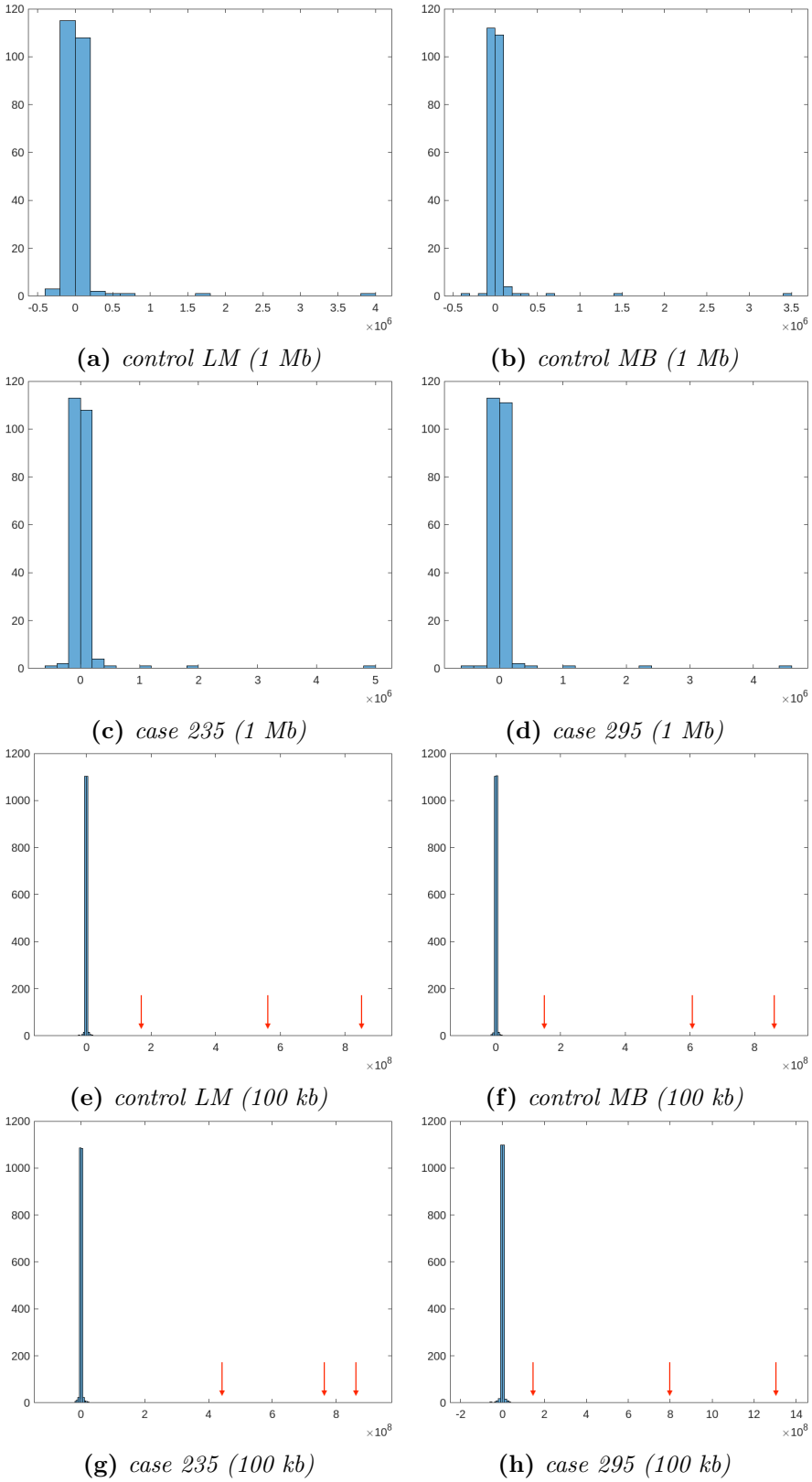
ent of the isolated zeros with the half of the minimum non-zero value between the matrix entries and the SCN normalization. These histograms actually prove to meet the requirements for using the spectral gap ShRec3D algorithm between the three largest eigenvalues and the central peak around zero.



(a) *control LM (100 kb)* **(b)** *control MB (100 kb)*

(c) *case 235 (100 kb)* **(d)** *case 295 (100 kb)*

Figure D.2: Histograms of the eigenvalues from the original matrices after the preprocessing step and the SCN normalization for the two cases (235 and 295) and the two controls (LM and MB) at 100 kb resolution.



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

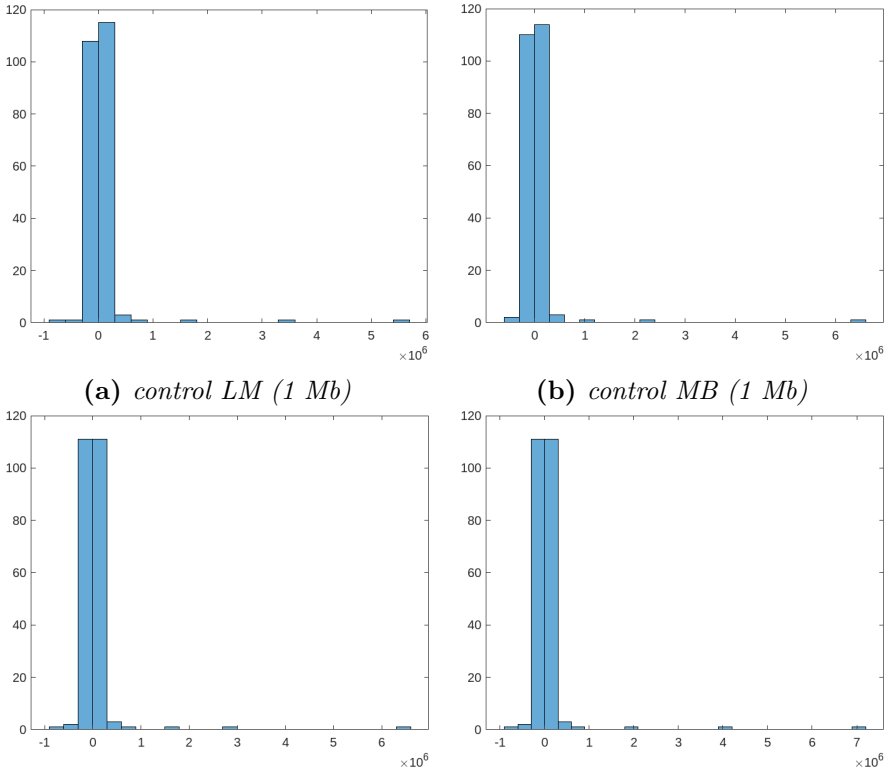
(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

(e) *control LM (100 kb)* **(f)** *control MB (100 kb)*

(g) *case 235 (100 kb)* **(h)** *case 295 (100 kb)*

Figure D.3: Histograms of the eigenvalues from the essential matrices after the preprocessing step and the SCN normalization for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb **(a-d)** and 100 kb **(e-h)** resolution.

The same histograms were also created for the synthetic matrices and are reported in figure D.4. Also in this case we can appreciate that the three largest values have a quite large spectral gap with respect to the main peak around zero.



(a) *control LM (1 Mb)* **(b)** *control MB (1 Mb)*

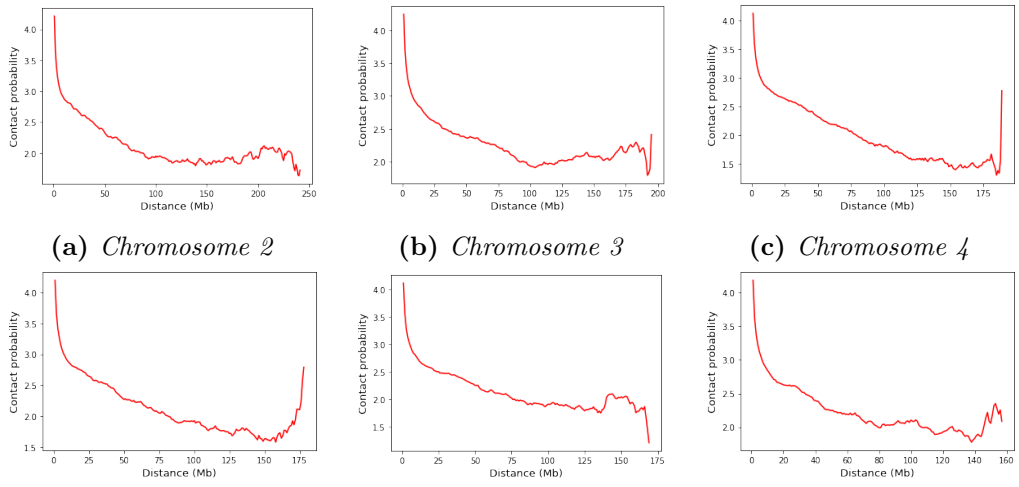
(c) *case 235 (1 Mb)* **(d)** *case 295 (1 Mb)*

Figure D.4: Histograms of the eigenvalues from the synthetic matrices after the preprocessing step and the SCN normalization for the two cases (235 and 295) and the two controls (LM and MB) at 1 Mb resolution.

Appendix E

Contact probability graphs

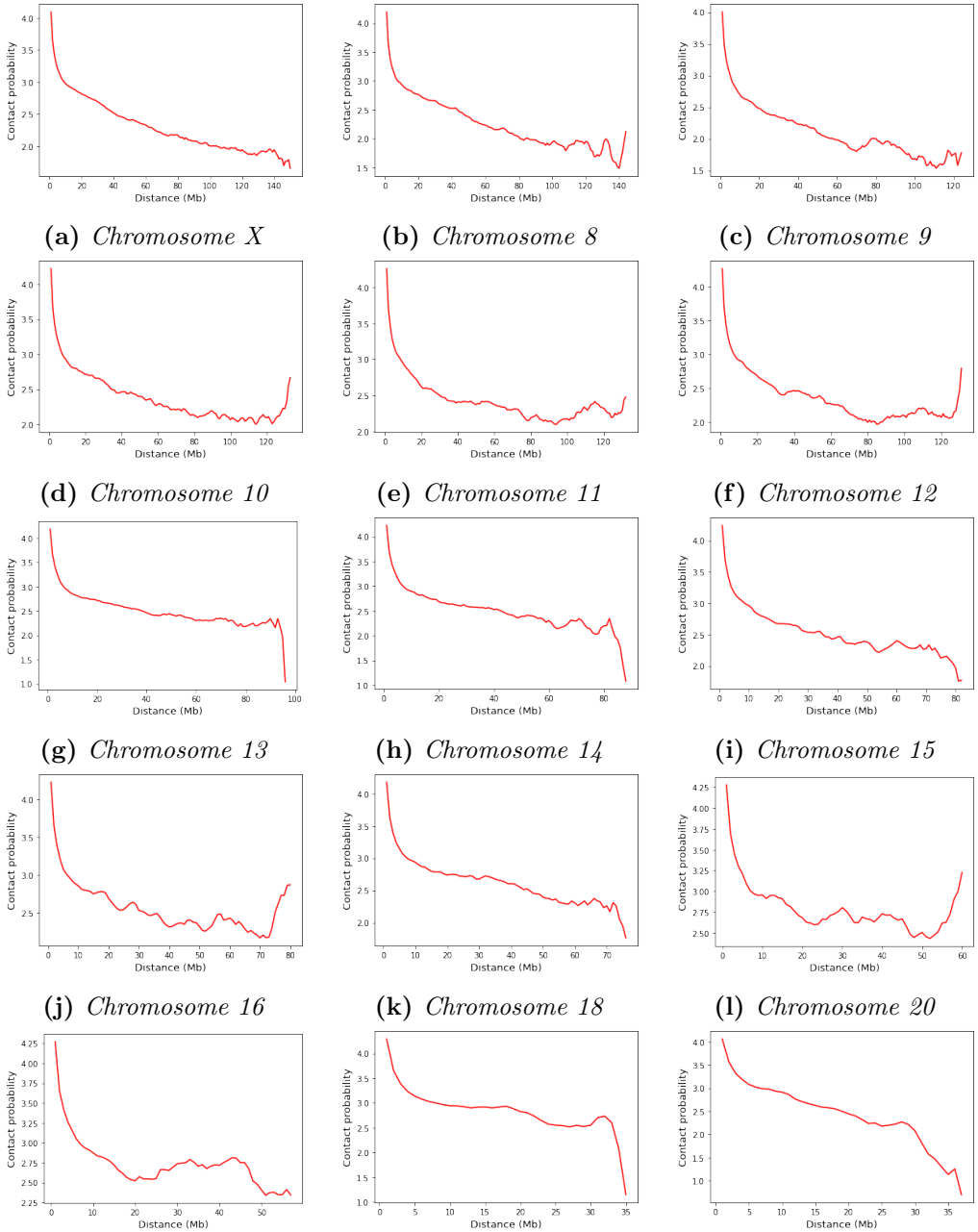
In this appendix we listed the graphs of contact probability as a function of the genomic distance for both the raw and the ICE normalized Hi-C maps related to all the other chromosomes excluding the chromosome 1 and the chromosome 17 already shown in figure 4.5. For doing that we therefore computed the average value of contact frequency starting from the matrices corresponding to each single chromosomes in figures 4.3 and 4.4 by varying the genomic distance. They are shown in figures E.1 and E.2 for the raw Hi-C matrix and in figures E.3 and E.2 with regard to the ICE normalized ones.



(a) *Chromosome 2* **(b)** *Chromosome 3* **(c)** *Chromosome 4*

(d) *Chromosome 5* **(e)** *Chromosome 6* **(f)** *Chromosome 7*

Figure E.1: Contact probability graph as a function of the genomic distance (diagonal) in Mb in case of single chromosomes extracted from the whole raw GM12878 Hi-C matrix.



(a) *Chromosome X* **(b)** *Chromosome 8* **(c)** *Chromosome 9*

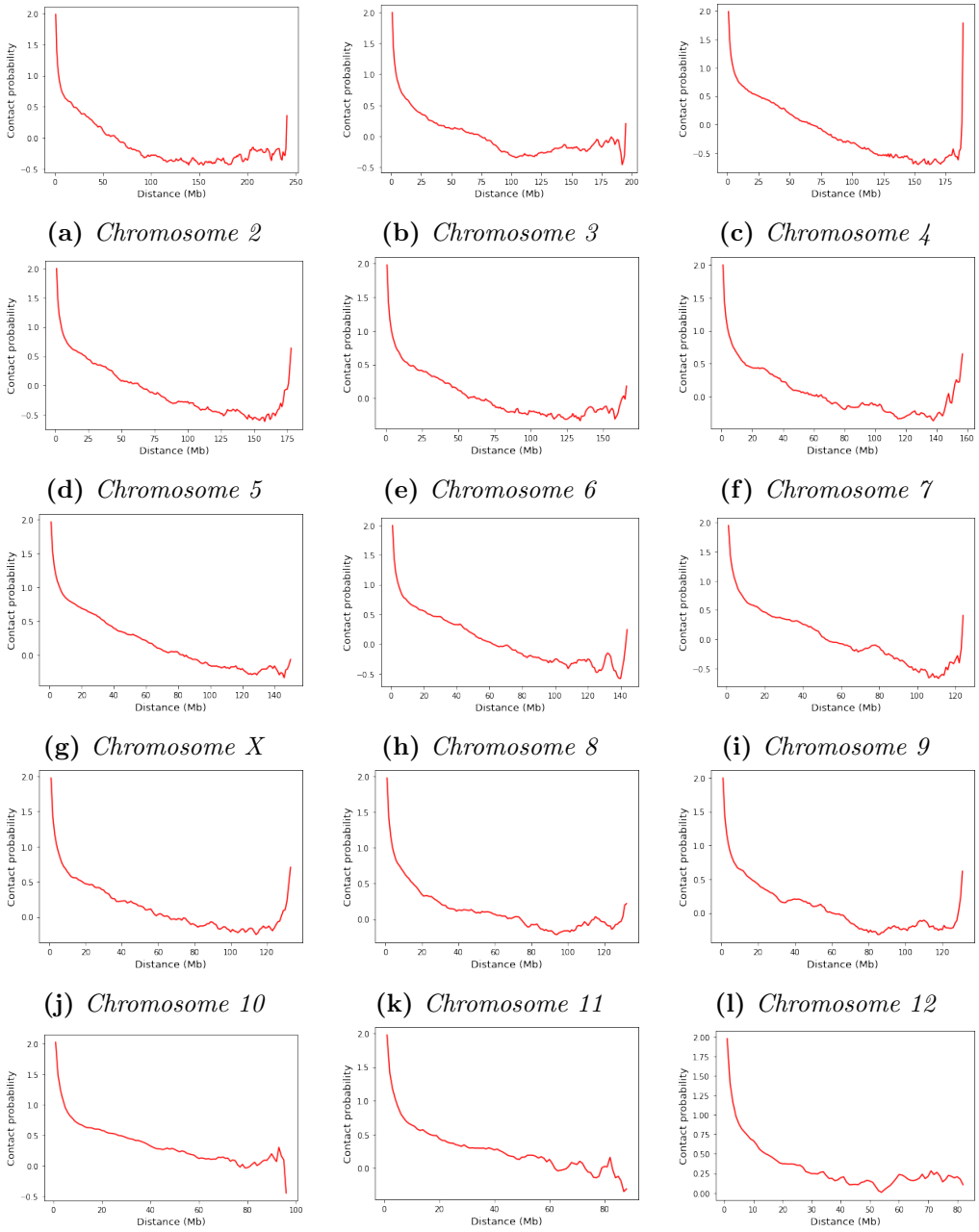
(d) *Chromosome 10* **(e)** *Chromosome 11* **(f)** *Chromosome 12*

(g) *Chromosome 13* **(h)** *Chromosome 14* **(i)** *Chromosome 15*

(j) *Chromosome 16* **(k)** *Chromosome 18* **(l)** *Chromosome 20*

(m) *Chromosome 19* **(n)** *Chromosome 22* **(o)** *Chromosome 21*

Figure E.2: Contact probability graph as a function of the genomic distance (diagonal) in Mb in case of single chromosomes extracted from the whole raw GM12878 Hi-C matrix.



(a) *Chromosome 2* **(b)** *Chromosome 3* **(c)** *Chromosome 4*

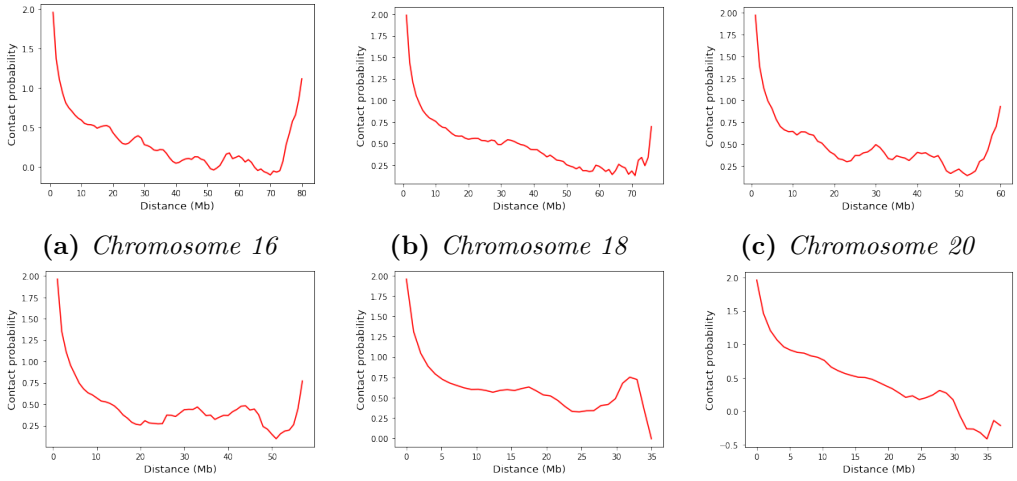
(d) *Chromosome 5* **(e)** *Chromosome 6* **(f)** *Chromosome 7*

(g) *Chromosome X* **(h)** *Chromosome 8* **(i)** *Chromosome 9*

(j) *Chromosome 10* **(k)** *Chromosome 11* **(l)** *Chromosome 12*

(m) *Chromosome 13* **(n)** *Chromosome 14* **(o)** *Chromosome 15*

Figure E.3: Contact probability graph as a function of the genomic distance (diagonal) in Mb in case of single chromosomes extracted from the whole ICE normalized GM12878 Hi-C matrix.



(a) *Chromosome 16* **(b)** *Chromosome 18* **(c)** *Chromosome 20*

(d) *Chromosome 19* **(e)** *Chromosome 22* **(f)** *Chromosome 21*

Figure E.4: Contact probability graph as a function of the genomic distance (diagonal) in Mb in case of single chromosomes extracted from the whole ICE normalized GM12878 Hi-C matrix.

From the plots we can again appreciate that the contact frequency approximately decreases monotonically on every chromosome, regardless the Hi-C data type (raw or ICE normalized) suggesting polymer-like behavior in which the three-dimensional distance between loci increases with increasing genomic distance (diagonal in the Hi-C matrices). These findings are in agreement with the chromosome conformation capture results as described in paragraph 2.3.2. However, in the final part of the graph it can be notice that there are greater fluctuations. These are due to the fact that for diagonals very far from the main one, there is a limited number of data from which we extract the average value, so the contact probability value may have significant differences, especially in case of smaller chromosomes such as 16, 22 and 21.

Bibliography

[1] Amy Ross, Procrustes Analysis, Department of Computer Science and Engineering University of South Carolina, SC 29208;

[2] Chen J, Hero AO 3rd, Rajapakse I. Spectral identification of topological domains. Bioinformatics. 2016 Jul 15;32(14):2151-8. doi: 10.1093/bioinformatics/btw221. Epub 2016 May 5. PMID: 27153657; PMCID: PMC4937202;

[3] CMSE 890-001: Spectral Graph Theory and Related Topics, MSU, Spring 2021, Lecture 01: Introduction to Spectral Graph Theory January 19, 2021, Lecturer: Matthew Hirn;

[4] Cournac, A., Marie-Nelly, H., Marbouty, M. et al. Normalization of a chromosomal contact map. BMC Genomics 13, 436 (2012). https://doi.org/10.1186/1471-2164-13-436;

[5] Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2001 Apr;2(4):292-301. doi: 10.1038/35066075. PMID: 11283701;

[6] Dankar, F.K.; Ibrahim, M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Appl. Sci. 2021, 11, 2158. https://doi.org/10.3390/app11052158;

[7] Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013 Jun;14(6):390-403. doi: 10.1038/nrg3454. Epub 2013 May 9. PMID: 23657480; PMCID: PMC3874835;

[8] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002 Feb 15;295(5558):1306-11. doi: 10.1126/science.1067799. PMID: 11847345;

[9] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 Apr 11;485(7398):376-80. doi: 10.1038/nature11082. PMID: 22495300; PMCID: PMC3356448;

[10] Durbin, R. M. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010);

[11] EreborMountain/Shutterstock.com;

[12] F. L. Bookstein. Landmark methods for forms without landmarks: localizing group differences in outline shape. Medical Image Analysis, 1(3):225–244, 1997;

[13] Fan R. K. Chung, Lectures on Spectral Graph Theory, University of Pennsylvania, Philadelphia, Pennsylvania 19104, chung@math.upenn.edu;

[14] Floyd-Warshall Algorithm, Brilliant.org, from https://brilliant.org/wiki/floyd-warshall-algorithm/;

[15] Fortin, JP., Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol 16, 180 (2015). https://doi.org/10.1186/s13059-015-0741-y;

[16] Frank Firk, & Steven Miller (2009). Nuclei, Primes and the Random Matrix Connection. Symmetry, 1(1), 64–105;

[17] Franzini S, Di Stefano M, Micheletti C. essHi-C: essential component analysis of Hi-C matrices. Bioinformatics. 2021 Aug 9;37(15):2088-2094. doi: 10.1093/bioinformatics/btab062. PMID: 33523102;

[18] Gartner, "Maverick Research: Forget About Your Real Data – Synthetic Data Is the Future of AI," Leinar Ramos, Jitendra Subramanyam, 24 June 2021;

[19] Gower, John C. and Dijksterhuis, Garmt B.: Procrustes Problems, Oxford University Press, 2004;

[20] Havel, T. F., Kuntz, I. & Crippen, G. M. (1983). The theory and practice of distance geometry. Bulletin of Mathematical Biology, 45(5), 665–720. doi:10.1007/bf02460044;

[21] https://hyperskill.org/learn/step/5645;

[22] https://www.niaid.nih.gov/diseases-conditions/prion-diseases.

[23] I.L.Dryden, K.V. Mardia, Statistical Shape Analysis, Wiley, Chichester, (1998);

[24] Imakaev, M., Fudenberg, G., McCord, R. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9, 999–1003 (2012). https://doi.org/10.1038/nmeth.2148;

[25] Jaadi, Zakaria, A step-by-step explanation of principal component analysis (PCA), Retrieved June (2021);

[26] Jean-Francois Rajotte, Robert Bergen, David L. Buckeridge, Khaled El Emam, Raymond Ng, Elissa Strome, Synthetic data as an enabler for machine learning applications in medicine, iScience, Volume 25, Issue 11, 2022, 105331, 2589-0042, https://doi.org/10.1016/j.isci.2022.105331;

[27] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2011 Dec 25;30(1):90-8. doi: 10.1038/nbt.2057. PMID: 22198700; PMCID: PMC3782096;

[28] Kendall, D. G. (1989). A Survey of the Statistical Theory of Shape. Statistical Science, 4(2), 87–99. http://www.jstor.org/stable/2245331;

[29] Kruskal, J.B. and Wish, M. (1978) Multidimensional Scaling. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-011, Sage Publications, Newbury Park. http://dx.doi.org/10.4135/9781412985130;

[30] Lesne, A., Riposo, J., Roger, P. et al. 3D genome reconstruction from chromosomal contacts. Nat Methods 11, 1141–1143 (2014). https://doi.org/10.1038/nmeth.3104;

[31] Li, G., Cai, L., Chang, H. et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 15 (Suppl 12), S11 (2014). https://doi.org/10.1186/1471-2164-15-S12-S11;

[32] Lieberman-Aiden, Erez and Van Berkum, Nynke L and Williams, Louise and Imakaev, Maxim and Ragoczy, Tobias and Telling, Agnes and Amit, Ido and Lajoie, Bryan R and Sabo, Peter J and Dorschner, Michael O and others, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, science, American Association for the Advancement of Science (2009);

[33] Lyu H, Liu E, Wu Z. Comparison of normalization methods for Hi-C data. Biotechniques. 2020 Feb;68(2):56-64. doi: 10.2144/btn-2019-0105. Epub 2019 Oct 7. PMID: 31588782;

[34] Nambiar, Mridula and Kari, Vijayalakshmi and Raghavan, Sathees C., Chromosomal translocations in cancer, Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, Elsevier (2008);

[35] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012 Apr 11;485(7398):381-5. doi: 10.1038/nature11049. PMID: 22495304; PMCID: PMC3555144;

[36] Osorio, D., Yu, X., Yu, P. et al. Single-cell RNA sequencing of a European and an African lymphoblastoid cell line. Sci Data 6, 112 (2019). https://doi.org/10.1038/s41597-019-0116-4;

[37] Procrustes analysis, https://en.wikipedia.org/w/index.php?title=Procrustes_analysis&oldid=1151503166;

[38] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Epub 2014 Dec 11. Erratum in: Cell. 2015 Jul 30;162(3):687-8. PMID: 25497547; PMCID: PMC5635824;

[39] Roger Van Peski, Random matrix theory and the semicircle law;

[40] Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918 (2007);

[41] Schoenberg, I. J. (1937). On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space. Annals of Mathematics, 38(4), 787–793. https://doi.org/10.2307/1968835;

[42] Servant, N., Varoquaux, N., Heard, E. et al. Effective normalization for copy number variation in Hi-C data. BMC Bioinformatics 19, 313 (2018). https://doi.org/10.1186/s12859-018-2256-5;

[43] Servant N., Varoquaux N., Lajoie BR., Viara E., Chen CJ., Vert JP., Dekker J., Heard E., Barillot E. HiC-Pro: An optimized and flexible pipeline for Hi-C processing. Genome Biology 2015, 16:259 doi:10.1186/s13059-015-0831-x http://www.genomebiology.com/2015/16/1/259;

[44] Stansfield, J.C., Cresswell, K.G., Vladimirov, V.I. et al. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. BMC Bioinformatics 19, 279 (2018). https://doi.org/10.1186/s12859-018-2288-x;

[45] Stegmann, Mikkel B., Gomez , David Delgado: A Brief Introduction to Statistical Shape Analysis, Technical University of Denmark, Lyngby, 2002;

[46] Synthetic data: Unlocking the power of data and skills for machine learning – Data in government. Data in government, 20 August 2020, https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/;

[47] Szabo, Quentin & Bantignies, Frédéric & Cavalli, Giacomo. (2019). Principles of genome folding into topologically associating domains. Science Advances. 5. 10.1126/sciadv.aaw1668;

[48] The, E. P. C. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012);

[49] Varoquaux N, Ay F, Noble WS, Vert JP. A statistical approach for inferring the 3D structure of the genome. Bioinformatics. 2014 Jun 15;30(12):i26-33. doi: 10.1093/bioinformatics/btu268. PMID: 24931992; PMCID: PMC4229903;

[50] Wigner semicircle distribution, https://en.wikipedia.org/w/index.php?title=Wigner_semicircle_distribution&oldid=1150268003;

[51] Y. Lu, H. Wang, and W. Wei, Machine Learning for Synthetic Data Generation: A Review. 2023, arXiv:2302.04062;

[52] Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017 Nov;27(11):1939-1949. doi: 10.1101/gr.220640.117. Epub 2017 Aug 30. PMID: 28855260; PMCID: PMC5668950;

[53] Yardımcı, G.G., Ozadam, H., Sauria, M.E.G. et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol 20, 57 (2019). https://doi.org/10.1186/s13059-019-1658-7;

[54] Ye, C., Paccanaro, A., Gerstein, M. et al. The corrected gene proximity map for analyzing the 3D genome organization using Hi-C data. BMC Bioinformatics 21, 222 (2020). https://doi.org/10.1186/s12859-020-03545-y;

[55] Zhao, B. et al. The NF-kB Genomic Landscape in Lymphoblastoid B Cells. Cell Reports 8, 1595–1606 (2014);

[56] Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, Sejnowski TJ, Dixon JR, Ecker JR. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. Proc Natl Acad Sci U S A. 2019 Jul 9;116(28):14011-14018. doi: 10.1073/pnas.1901423116. Epub 2019 Jun 24. PMID: 31235599; PMCID: PMC6628819;

Acknowledgments

I would like to thank my supervisor Prof. Daniel Remondini for his support in my thesis work. I would also like to thank my co-supervisor Dr. Alessandra Merlotti, for guiding me in this exciting journey and for her always invaluable advice. A special thanks to my mum Palmina and my dad Giuseppe, together with my brother Davide without whom I would not have been who I am today. Thanks also to my wonderful uncles Sandra and Giovanni. Finally, I would like to thank my friends Flavio, Leonardo and Alessia for all the affection they have shown me, especially when I needed it most; together with a memory for dear Angela and Alfonso.