Network Analysis of the 3D Genome

Gabriel Bratseth Stav

NCMM

Master Thesis
Cell Biology, Physiology and Neuroscience
60 credits

Department of Biosciences
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

June 2023

Abstract

The genome is not merely organized into a linear string spun around nucleosome but exhibits a complex hierarchical chromatin structure. These higher order chromatin structures take part in determining the 3D structure of the genome, taking part in the epigenetic regulatory landscape in the cell. Hi-C data captures this complex chromatin structure by capturing chromatin interactions genome-wide. An automated pipeline was made and used to process Hi-C data from five cell lines to networks. This Hi-C to network pipeline takes raw Hi-C data and returns significant chromatin interactions. Centrality metrics and community detection network theory approaches were then used to analyze and compare the chromatin networks between cell lines. Centrality distributions were calculated and contrasted between cell lines on a genome wide scale. Networks were determined to have a scale-free-like distribution, showing no difference in the genome-wide centrality distributions. The healthy tissue breast cell line MCF-10A and the cancer breast tissue cell line MCF-7 were analyzed further. Centrality metrics were determined in networks and compared across chromosomes, finding significant differences at specific chromosomes. Community detection algorithms were used to detect and compare the topology of the MCF-10A and MCF-7 networks. Large differences in topology were found across chromosomes, with smaller chromosomes being more similar. Chromosome compartments were determined and used to annotate the networks. Many nodes switch compartment status between MCF-10A and MCF-7, most of these nodes change from compartment B to compartment A. Taken together, these results capture the change in chromatin structure between healthy breast tissue cells (MCF-10A) and malignant breast tissue cells (MCF-7). This project can potentially shed light on how chromatin structure changes in breast cancer progression.

Acknowledgements

I would first and foremost like to thank my main supervisor, Marieke Kuijjer. Your insightful feedback and constructive critiques have always challenged my thinking and helped me underway. Your unwavering encouragement, even in the most trying times, was deeply appreciated. The skills and knowledge I have gained under your supervision will undoubtedly serve me well in my future career. I also wish to thank my co-supervisor Jonas Paulsen for the valuable insight provided at the finishing stages of the project.

To everyone at the Kuijjer and Mathelier groups at NCMM, I am grateful for the constructive feedback received during group meetings and presentations. Your collective knowledge and expertise enriched this project. I would also like to extend a special thanks to Roberto Rossini from the Paulsen group, who provided crucial assistance during the initial stage of my Hi-C analysis.

Lastly, I am deeply grateful to my family for their love and support during this challenging and rewarding journey.

Abbreviations

LCC	Largest connected component
NMI	Normalized mutual information
PIN	Protein interaction network
CIN	Chromatin interaction network
Mb	Mega base pairs
Kb	kilo base pairs
PCHi-C	promoter capture Hi-C
3C	chromosome conformation capture
NCHG	Statistical tool based on the non-central hypergeometric distribution
ICE	Iterative correction and eigenvector decomposition
TAD	Topologically association domain
LAD	Lamina associating domain
CTCF	CCCTC-binding factor
HMTase	Histone methyl transferase
LEF	Loop-extrusion factor
PTM	Post-translational modifications

Table of contents

1. Introduction

1.1 Chromatin

1.1.1 Organization of DNA

The total length of DNA in a human cell is around 2 meters in length, yet it must fit into the nucleus of the cell, which ranges in size from 5 – 20 micrometers in diameter in mammals (Lammerding, 2011). This severe compaction is achieved by histone proteins binding to DNA, which are positively charged proteins that have high affinity for the negatively charged phosphate backbone of DNA. The core histone proteins form octamers, consisting of two subunits of each histone protein (H2A, H2B, H3 and H4). DNA wraps around these octamers every 147 bp, forming the basic unit of chromatin, the nucleosome. Between each nucleosome is linker DNA, associated with H1 histones which stabilizes the linker DNA by binding both to DNA and to histones in the octamer. Chromatin can thus be defined as all DNA and RNA and their associated proteins, mainly histone proteins. The core repeating unit of chromatin is the nucleosomes, which stack together to form the helical 30 nm fiber. This fiber is further organized into larger structures, folding first into the 120 nm chromonema, then the 300 nm and 700 nm chromatid structures (Ou et al., 2017).

These larger chromatin structures form the structural basis of the chromosomal architecture, shown in Figure 1 on the proceeding page. This classical view on the nucleosome fibers and the secondary structure of chromatin has changed over the years as imaging technology and techniques have improved. One example of this is the finding that the 30-nm fiber largely does not occur *in vivo* in interphase chromosomes. Chromatin structures seem to be more irregular and dynamic, forming fractal-like structures instead the regular fiber architecture proposed in earlier studies (Nishino et al., 2012). The authors of the study concluded that no regular chromatin structures are found above the histone octamer of 11 nm, corroborated by other studies on chromatin secondary structures (Fussner et al., 2011) & (Maeshima et al., 2010). Although the classical view on the 30 nm fiber and higher order chromatin structure is challenged, what remains true is the formation of the nucleosome fiber to form chromatin, resulting in compaction of DNA.

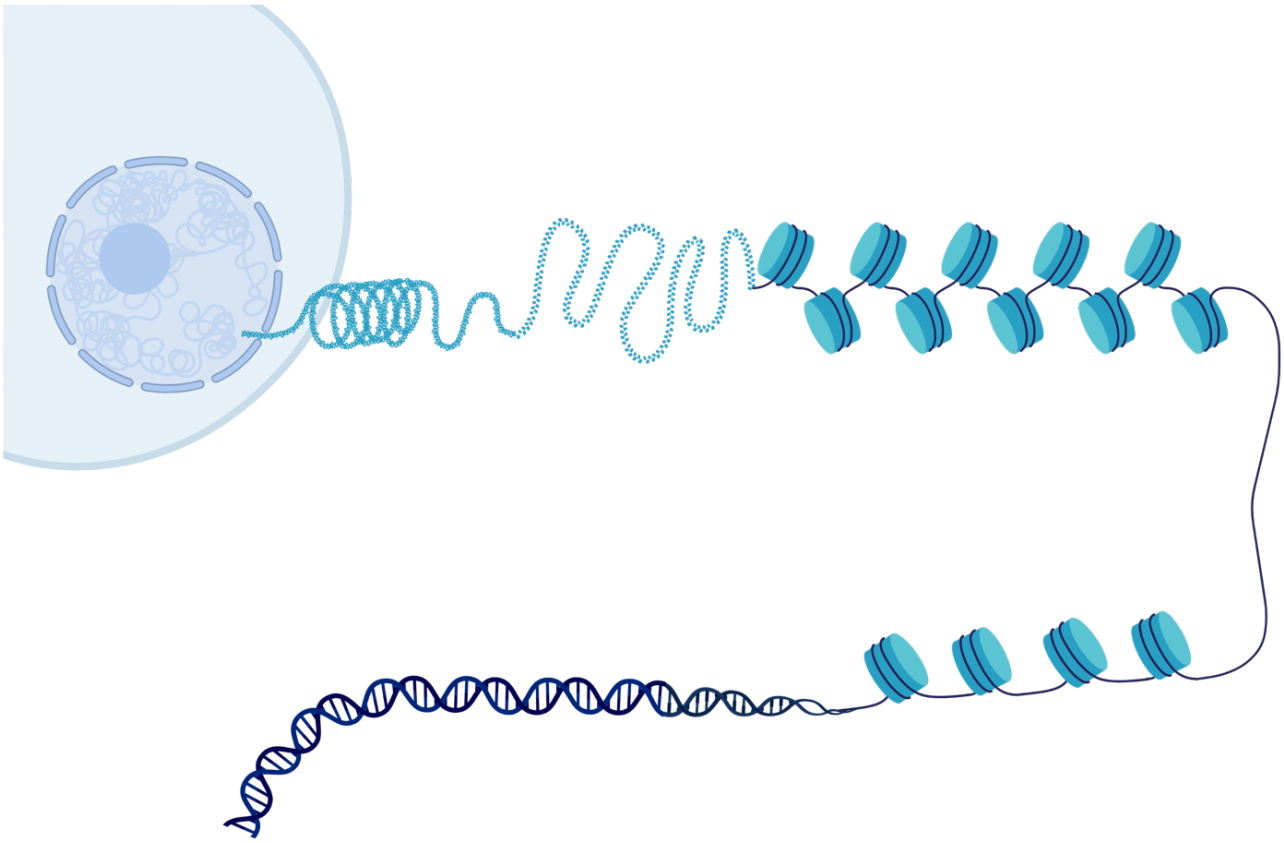


Figure 1*: Schematic view of the classical model of chromatin architecture – consisting of DNA wound around histone octamers (light blue) to form nucleosomes. Nucleosomes are wound together to form higher order structures of compacted chromatin, eventually forming the chromosome. Reprinted from Genomic Architecture, by BioRender.com (2023). Retrieved from: https://app.biorender.com/biorender-templates.*

1.1.2 Chromatin modifications and chromatin structure

In general, there are two chromatin states genome wide, which are heterochromatin and euchromatin. Heterochromatin is tightly packed and condensed, whereas euchromatin is loosely packed and decondensed. Chromatin state can mediate the ease of access to nucleic acids, whereby condensed chromatin prevents binding of proteins and thus can prevent regulatory proteins from binding. This chromatin state is controlled in part by covalent modifications to histone proteins. Each histone protein has “tails”, where parts of their amino acid structure protrude from the nucleosomes. A class of enzymes called histone modifying enzymes can covalently change specific residues on histone protein tails, affecting the secondary structure or the charge of residues. These post-translational modifications of histone proteins are reversible

and highly specific. One example illustrating this is the histone methyltransferase enzymes, which adds methyl groups to specific residues on histone protein tails, most often on lysine residues in histones H3 and H4 (Zhang et al., 2021). Methylation of H3K9 (histone H3, lysine residue K9) occurs by the HMTase SUV39H1 and is associated with the formation of heterochromatin and thus silencing of genes (Lomberk et al., 2006). These modifications can again be removed, and their effects reversed. This example illustrates the basis of the histone code hypothesis, which postulates that histone modifications are recognized by proteins which lead to downstream effects such as chromatin condensation and gene silencing (Strahl & Allis, 2000). Continuing the H3K9 methylation example with the histone code in mind, we find that such a methylation mark does indeed lead to downstream effects. Proteins that recognize histone modifications contain structural domains that bind to specific modified residues of histone tails, such as chromodomains and bromodomains. The HP1 protein is a chromatin reader and contains a chromodomain that recognizes H3K9me and upon binding to this residue, recruits more SUV39H1 proteins to extend the methylation pattern. This is the most common effect of HP1 and leads to the formation of heterochromatin and transcriptional silencing (Lomberk et al., 2006).

This example illustrates a monocausal relationship between H3K9me and heterochromatin formation, but the full picture is not that simple. Histone modifications have different effects determined by the histone code surrounding them, for instance the combination of H3K9me and H3K4me is associated with active transcription (Berger, 2007).
PTMs of histones leads to the formation of euchromatin in similar manners, but with other histone modifying proteins, histone residues and potential downstream effects. In addition to histone modifications, DNA methylation and chromatin remodeling complexes also affect the chromatin state by inducing local changes to chromatin. DNA methylation is generally associated with the formation of heterochromatin (Buitrago et al., 2021) and chromatin remodeling complexes perform a wide range of functions by interacting with nucleosomes (Clapier et al., 2017). This paints a complex but partial picture of chromatin structure, where local changes like PTMs to histones or reshuffling of histones during transcription, affect the chromatin state.

1.2 Higher order chromatin structure

Chromatin structure is more complex than its organization around nucleosomes and subsequent packing that defines the chromatin state. Chromatin form self-interactions and 3D structures across multiple scales, defined as higher order chromatin structures. These higher order structures are not the result of stochastic processes, they form distinct structures and take part in a multitude of regulatory processes within the nucleus. Chromatin structure takes part in regulating gene expression, transcription timing and enhancer-promoter interactions. Many of these chromatin structures are conserved across tissues (McArthur & Capra, 2021). Disruption of the chromatin structure can lead to pathologies such as cancer and congenital disorder. It is thus important to study how the nature of these chromatin structures and how they interact, as they play important parts in the regulatory landscape of the nucleus.

1.2.1 Chromosome capture technologies

Several methods have been used throughout the years to capture higher order chromatin structures, like cryo-electron microscopy, light microscopy and by labeling chromosomes (Zakirov et al., 2022). An example of this is evidence found for chromosome territories by the use of fluorescence *in situ* hybridization (FISH) which labels DNA with fluorescent probes to evaluate the proximity between loci (Cremer & Cremer, 2010). Chromosome positioning in the nucleus was found to correlate with chromatin state, where gene rich chromosomes with more transcription was located closer to the nuclear center and chromosomes with less transcription and fewer genes located closer to the nuclear periphery (Jerkovic´ & Cavalli, 2021). It was also observed that chromosomes consisted of regions of "giant loops" around 1 Mb in size and small-scale loops around 100 kb in size (Cremer et al., 2006). The arrival of next-generation sequencing (NGS) technologies and chromatin conformation techniques provided new insights into higher order chromatin structures. Chromosome conformation capture, termed 3C, was the first of such chromatin conformation technologies and relied on capturing chromatin interactions between two loci and sequencing these regions by PCR. 3C experiments confirmed the existence of chromatin loops seen by FISH, but could only be used to detect the chromatin structure of known DNA sequences due to the PCR step of the protocol (Dekker et al., 2002). Many chromosome conformation methods are derived from 3C, enabling different analyses of

chromatin structure. 4C improves upon the original 3C protocol, improving the range limit from a few hundred kb to the whole genome, enabling capture of all chromatin contacts made from one locus. Both the 3C and 4C protocols examine the chromatin structure associated with specific loci. Other methods include 5C which capture chromatin interactions between multiple regions, and ChIA-PET which combines 3C with immunoprecipitation to find chromatin interactions associated with specific proteins (Han et al., 2018). Further improvements were made to the 3C protocol, combining it with NGS to enable capture of all chromatin contacts across the genome. This technology is termed high throughput chromosome conformation capture (Hi-C) and enables capture of chromatin interactions, and thereby structure, without relying on PCR primer design.

1.2.2 Hi-C data

Hi-C is a 3C derived technique and that captures chromatin interactions genome wide, which then can be used to infer structure. The Hi-C protocol (see Figure 2) starts with crosslinking of DNA with formaldehyde to link DNA with any proteins present to preserve the chromatin interactions. After the crosslinking of DNA, the cells are lysed, and chromatin is released from the cells. Next the crosslinked DNA is cut with a restriction enzyme which leaves 5' overhangs. These are then filled and marked with biotin. The two separate strands are ligated to create one chimeric DNA molecule derived from two separate loci, representing the chromatin interaction that took place at the time of crosslinking with formaldehyde. The DNA is then purified and can be sheared before biotin pull down occurs to preferentially select chimeric reads for sequencing. The end point of the protocol is two separate reads, each containing parts of the chromatin of the loci where the chromatin interactions took place (Lieberman-Aiden et al., 2009). The Hi-C protocol thus enables capture of loci that engage in chromatin interactions, capturing loci that are close together in space, but might be far away in linear sequence.

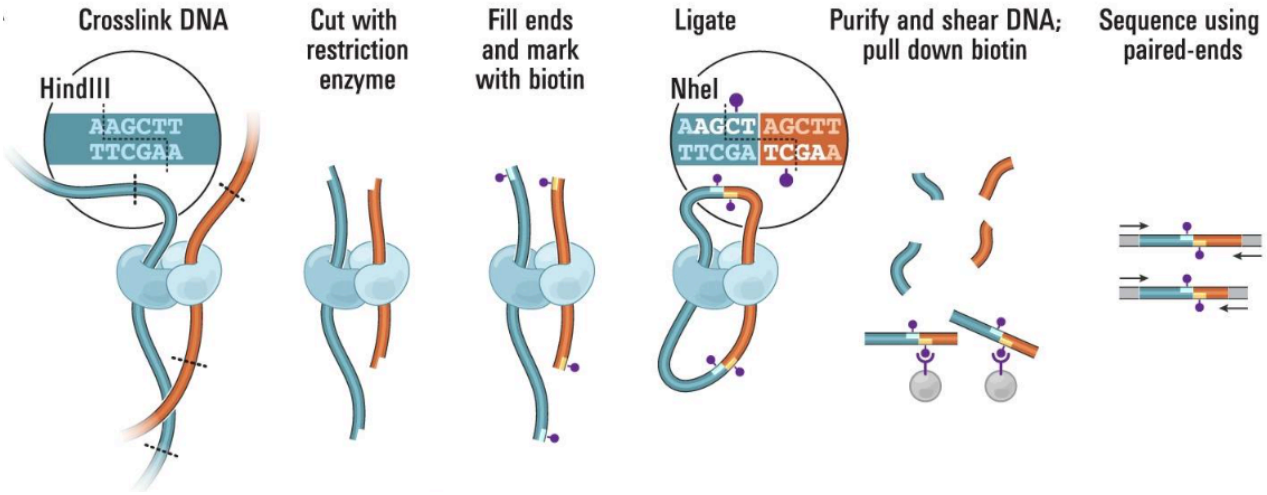


Figure 2: Overview of the Hi-C protocol steps focused on one chromatin interactions. The figure shows crosslinking of DNA and subsequent cutting by the restriction enzyme HindIII. The hanging ends are filled and marked with biotin, before the two loci are ligated by NheI. DNA is then purified, sheared and biotin enables the resulting ligation to be pulled down. The result of the protocol are two reads containing sequences from two separate loci involved in one chromatin interaction. Figure retrieved from figure 1A in Lieberman-Aiden, et al., (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, N.Y.), 326(5950), 289. https://doi.org/10.1126/science.1181369. Reprinted with permission from AAAS.

Since the reads contain parts of two separate loci, they contain information about a chromatin interaction. The output of a Hi-C experiment is raw data in the form of chimeric reads forming a Hi-C library, requiring further processing to align the reads to a reference genome and do quality filtering (see methods section). One important point to keep in mind with dilution Hi-C data is that each interaction frequency represents the population average, since Hi-C libraries are generated from a pool of cells. This means Hi-C data does not enable analyses of chromatin structures related to phenomena like cell cycle or specific conditions unless all cells in the population are synced in terms of cell cycle or treated with the same condition. If a specific chromatin interaction between two loci varies across two Hi-C libraries, it's not possible to infer if the chromatin structure is binary (i.e., it is there or not) or if there are several configurations of the chromatin structure present in the population. In one population the chromatin interaction might be transient and exist half of the time in all cells, but in the other population the chromatin interaction exists in half of the cells all the time. If these two populations of cells are compared, the interaction counts corresponding to the loci representing the chromatin interaction will be indistinguishable. This illustrates that Hi-C data describes the average chromatin structure from the population of cells that generated the Hi-C library (Lajoie et al., 2015). This population

average is also affecting individual populations, there might be a small proportion of cells that might always have a specific chromatin interaction between two loci, while in most cells this chromatin interaction does not take place. In a Hi-C library this phenomenon is indistinguishable from a weak population level interaction frequency between the two loci (Lajoie et al., 2015).

Hi-C data can be represented as a contact matrix spanning the entire genome, where each entry in the contact matrix represents the number of chromatin contacts between two regions of the genome. The upper limit on resolution of Hi-C data experiments is limited by the restriction fragment length and the sequencing depth of the data. The interaction frequencies in a Hi-C library are thus binned at non-overlapping segments of fixed length across the genome, and this is the Hi-C resolution of a processed Hi-C library and is determined computationally. A smaller bin size needs a higher sequencing depth to adequately represent the actual interaction frequencies of a chromatin region (Lajoie et al., 2015). Different resolutions will thus also represent chromatin interactions across different scales, selecting a bin size of 10 kb will enable different downstream analyses than a resolution of 1 Mb will. From the Hi-C contact matrix, a typical representation of interaction frequencies if the heatmap. These representations enable a look at varying structures of the genome and can be generated across all resolutions of Hi-C data and can be generated on a per chromosome basis or for the whole genome.

Figure 3 is an example of a typical heatmap and is generated from the IMR90 cell line using raw Hi-C data. The figure shows the entirety of chromosome 2, where darker colors indicate a higher interaction frequency, and a lighter color corresponds to a lower interaction frequency. From examining these heatmaps several genome wide patterns emerge, the strongest of which is the cis/trans interaction ratio. This is the ratio between the total number of intrachromosomal interactions versus the total number of interchromosomal interactions. Chromosome territories separate chromosomes into specific regions of the nucleus, which might explain the fact that inter-chromosomal (trans) interactions have a lower frequency than intra-chromosomal (cis) interactions. In general, cis interactions occur 40 – 60 times more frequently than trans interaction in Hi-C libraries. If there is noise in the dataset, which might be caused by fragments randomly ligating and interpreted as chromatin interactions, this affects both cis and trans

interactions by an equal amount. Therefore the trans interaction counts will be disproportionally affected compared to the cis interaction frequency counts (Lajoie et al., 2015).

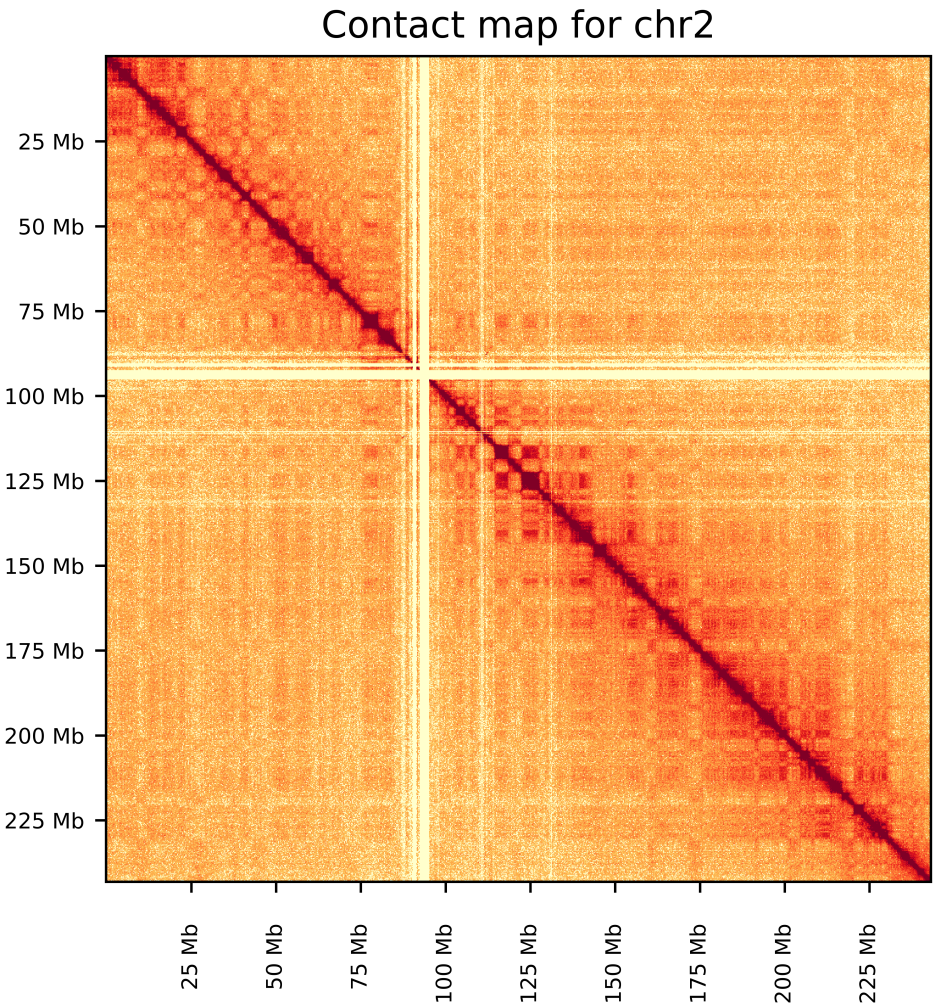


Figure 3: Heatmap generated from Hi-C data from the MCF-10A cell line. The chromosome 2 interaction matrix is plotted against itself, and the color scales from yellow to red as interaction frequency increases. The resolution of the data is 250 kb, each tick is 25 million base pairs (Mb).

Another genome-wide pattern observed from the Hi-C matrix is distance dependent decay, where the number of interactions decrease as the distance between two loci increases. They can be seen in the heatmap in Figure 3 above, where the highest interactions frequency regions are centered around the diagonal. Regions far away in linear sequence generally have a lower interaction frequency, but this pattern cannot be found in inter-chromosomal data. These trans interactions occur between chromosomes and are thus constrained by different factors than cis interactions. They are limited in their interaction counts by the folding of chromosomes and the territories

they occupy. These interactions patterns, the distance dependent decay and the cis/trans ratio, does not correspond to specific chromatin structures but are rather the results of the Hi-C protocol and the nature of Hi-C data.

1.2.3 Chromosome compartments

The largest intrachromosomal chromatin structures captured by Hi-C data are the chromosome compartments. These domains divide the chromosomes into two distinct regions, compartment A and compartment B (Lieberman-Aiden et al., 2009). Compartments preferentially interact with other compartments of the same type, meaning they interact with specific patterns in 3D enabling the capture of these chromatin interactions. This spatial segregation of compartments can thus be seen in the Hi-C matrix and visualized in Hi-C heatmaps, which are usually identified in Hi-C matrices at a 1 Mb resolution. Each compartment correlates strongly with the chromatin state, where compartment A is associated with active transcription and euchromatin. Compartment B on the other hand associates with heterochromatin and transcriptional repression. This correlation between compartment type and chromatin state is strong and remains true even for cell lines with potential genomic rearrangements like the lymphoblastoid cell line K562. Chromosome compartments functions as large domains to control transcription timing by modifying the chromatin state over large regions of the genome, which is seen in *Drosophila melanogaster* embryogenesis. During development an increasing amount of genes are switched to an active state, and this is captured by Hi-C data as compartments switch from compartment B to compartment A (Rowley & Corces, 2018). The maintenance of the chromosome compartments could be the result of proteins regulating transcription, rather than being directly affected by transcription itself. Inhibition of RNA polymerase II during embryogenesis in *D. melanogaster* led to reduced compartment formation, but not the total loss of chromosome compartments. This is also observed from the heat shock response in *D. melanogaster* where the degree of RNA Polymerase II loss correlates with loss of compartmentalization.

Identifying compartments be done using only the Hi-C matrix as input, as the compartments preferentially interact with themselves. This method relies eigenvector analysis, as the first eigenvector changes signs when compartments switch (Fortin & Hansen, 2015). Using this approach allows for identification of compartments A and B, which is done on a per-

chromosome basis. This method uses the first principal component, which is responsible for the largest variance in data, to denote chromosome compartments. In some cases the largest variance captured by the first principal component can capture other factors, instead capture the separation of chromosomes into chromosome arms instead of the compartment structure (Kalluchi et al., 2023). Chromosome compartments have also been found to contain different sub-compartments, indicating that the compartment structure is more complex than the A and B compartments. These sub compartments reflect the complex chromatin states present on the genome, where several histone modifications are associated with the higher order chromatin structure.

Compartments can also be differentially enriched differentially between cell lines, meaning if repressive histone marks such as H3K9me3 are associated with compartment B in one cell line, this might not be the case for all regions of compartment B in another cell line. In one study a clustering algorithm was used to define compartments, taking histone modifications into account, and compared with A and B compartments defined by principal component analysis. This approach resulted in the identification of four distinct compartments, suggesting that compartmentalization of chromosomes is more complex than previously thought (Nichols & Corces, 2021). With further increase in the Hi-C resolution to 1-kb the A and B compartments could be further subdivided into 6 distinct compartment types based on their histone modification and interaction profiles (Rao et al., 2014b). Compartments can now be called at sub-kilobase resolutions, finding that a larger proportion of compartment B than previously thought did not contain repressive histone marks which might suggest that compartment A drives compartmentalization (Zhou, 2022).

Comparing the A and B compartments between cell lines at lower resolutions can still be informative; when A and B compartments were compared between normal breast tissue (MCF-10A) and cancerous breast cell line (MCF-7), 12% of compartments switched between the tissues. Overall, there were more open compartments present in the cancer cell line, especially for the small chromosomes (chr16 – chr22). The altered compartment profile of the cancer cells reflected repression of Wnt signaling and correlated with compartment switching from A to B in the MCF-7 tissue (Barutcu et al., 2015). It is however unclear from this study if the compartment

switching in MCF-7 was due to deactivation of genes which influence the chromatin state, or if compartment switching occurred first which then led to changes in gene expression. Compartments can be visualized in Hi-C heatmaps due to the differential interaction patterns between the A and B compartments. To improve the visualization of the compartments in heatmaps, the normalized observed over expected interaction frequencies can be calculated. The expected interaction frequencies are calculated with regards to the distance decay of interactions in the Hi-C matrix. The “plaid” pattern seen in such heatmaps can be further enhanced by coloring the interactions based on the Pearson correlation between the normalized values in an interaction. An example of such a figure is Figure 4 on the next page, where this approach was used. There is a clear diagonal line of self-interacting domains, and large regions spanning multiple megabases belonging to different compartments, compartment A (red) and B (blue).

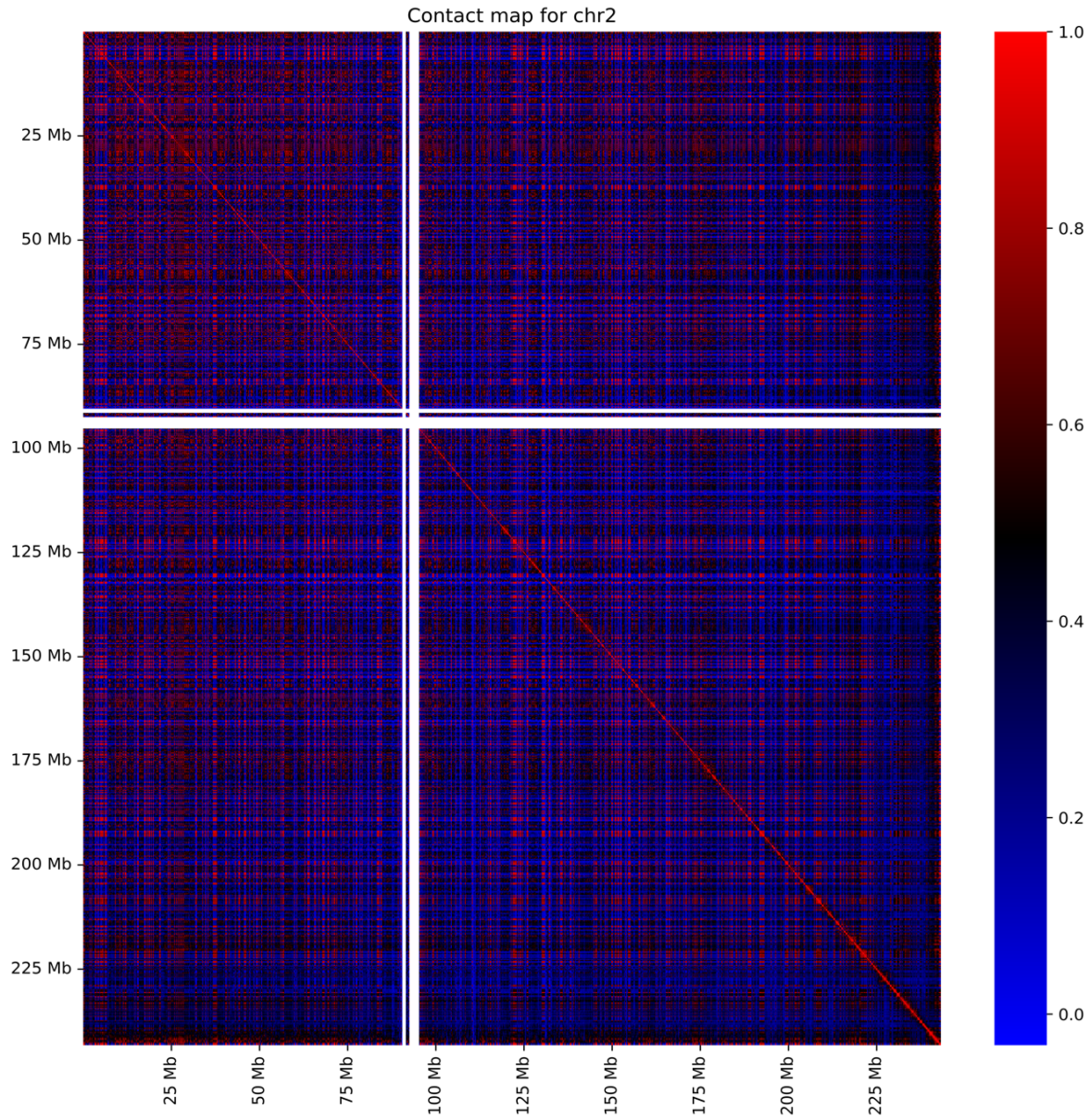


Figure 4: This figure shows the Hi-C heatmap for MCF10-A chromosome 2 at 250kb binned resolution. Compartments are visualized by doing iterative correction (matrix balancing – see methods) on the Hi-C matrix. Interactions are colored by the Spearman correlation between the observed over expected interaction frequency, ranging from 0 to 1. The colors show a "plaid" pattern, corresponding to compartment A (euchromatin) and compartment B (heterochromatin) present in the dataset.

1.2.4 Topologically associating domains

Topologically associating domains (TADs) are another chromatin structure seen in intra-chromosomal Hi-C matrices. These domains are smaller in size than chromosome compartments, first thought to have an median size of 880kb (Dixon et al., 2012), but more recent evidence

suggests this number is closer to 185 kb (Rao et al., 2014b). TADs are chromatin domains increased self-interactions, in which interactions inside the domain occur much more frequently than between different domains. Around 91% of the genome is organized into TADs; they play important parts in regulating chromatin interactions and controlling the transcriptional process. Boundary regions in TADs are enriched for the transcriptional repressor CCCTC-binding factor (CTCF), which acts to stop the spread of heterochromatin, forming barriers between the regions inside the TAD from the chromatin state outside of it (Dixon et al., 2012). Factors associated with active gene expression are also enriched at TAD boundaries, indicating the TADs form as regulatory regions which limit the external chromatin state from interfering with gene regulation within the domain.

The formation of TADs can be explained by the loop extrusion model, which states that chromatin is extruded by a loop extrusion factor (LEF) until it encounters a boundary element (BE). The LEF binds to DNA and moves along the genome in both directions, forming a loop of chromatin behind it, when it reaches a BE, it stops, and a loop of chromatin is formed. This hypothesis was tested using polymer simulations which supported the loop extrusion model and its role in TAD formation (Fudenberg et al., 2016), and further strengthened when loop extrusion by yeast condensin was observed ex vivo using real-time imaging (Ganji et al., 2018).
It is now thought that structural maintenance of chromosome (SMC) protein complexes, cohesin and condensin, are responsible for the loop extrusion process that generates TADs. They form circular complexes that bind to DNA and extrudes the chromatin in an ATP-dependent manner, until they reach boundary domains such as CTCFs. CTCF binds to DNA in a directional manner and most long-range cis interactions are enriched for CTCF at their anchors. These interactions form when CTCF sites converge directionally, meaning that the LEF only stops at CTCF domains when they point in the same direction – when they converge (Davidson & Peters, 2021). TADs might then be formed by the extrusion of multiple such loops that stop when they meet convergent CTCF domains, forming regions of increased interaction within the TAD borders. TADs can be seen from the unprocessed Hi-C heatmap as square blocks of increased interaction frequency along the diagonal, although visual inspection cannot be used to quantify TADs or define them accurately – as seen in Figure 5.

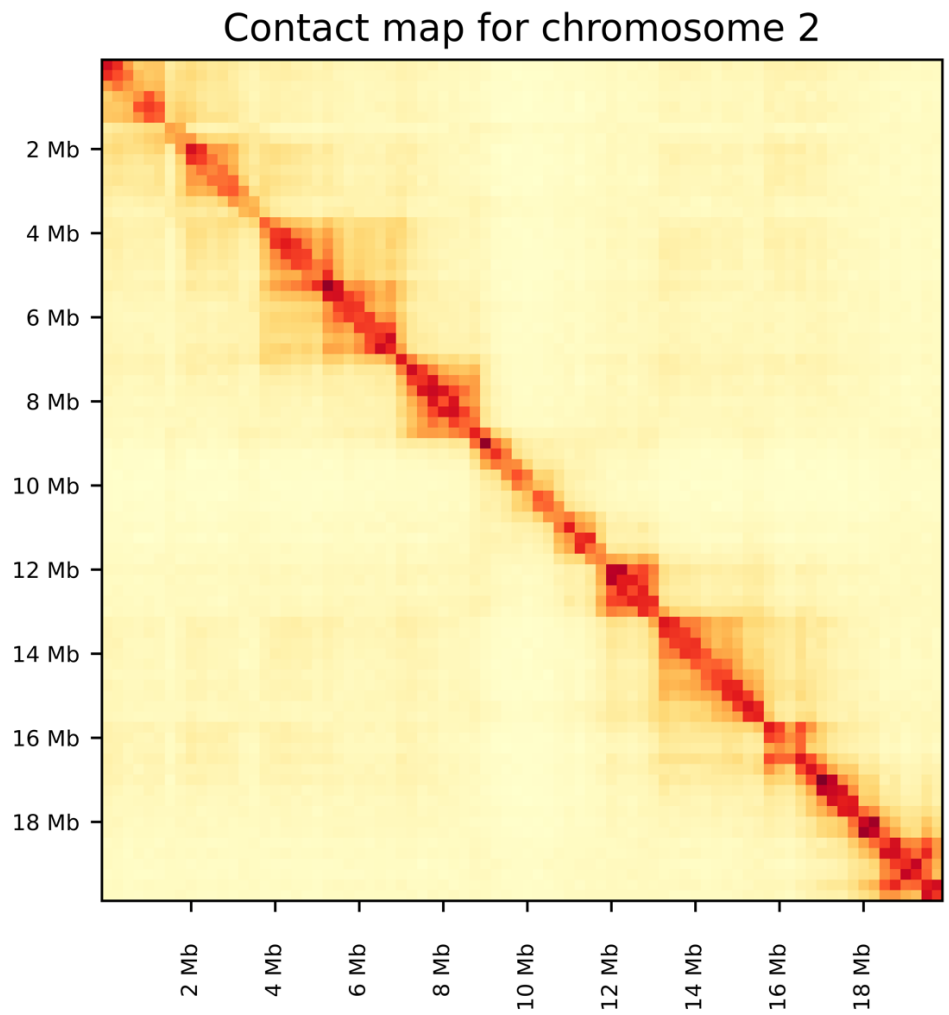


Figure 5: This figure is a heatmap of Hi-C data from MCF-10A cell line, showing the first 20 Mb of chromosome 2. The resolution of the data is 250 kb, which is too coarse to accurately depict TADs, and no TAD caller algorithm has been used. Domains can be seen along the diagonal, where several domains overlap and form nested structures, several of which might be defined as TADs.

TADs are usually defined computationally from the Hi-C matrix alone, through a wide range of algorithms that leads to big discrepancy in results between tools (Dali & Blanchette, 2017). These algorithms use the interaction frequencies between regions in the Hi-C data to define TADs and their boundaries genome wide. TADs can be nested, which several of these algorithms do not account for, meaning called TADs miss about 25% of TADs defined through manual annotation. Still, this approach of calling TADs genome wide has enabled a deeper look into TADs and their effect on the regulatory process in the cell. The boundaries between TADs enriched for CTCF and the physical separation of TADs in space helps prevent interactions

forming between different TADs. Disruptions of the TAD boundaries can thus lead to aberrant enhancer promoter interactions forming. The deletion of boundaries surrounding the Epha4a locus led the formation of *de novo* enhancer-promoter interactions and the introduction of limb malformations in mice. Analysis of families with a history of limb malformation found that this TAD deletion also occurs in humans and leads to similar limb malformations (Lupiáñez et al., 2015). Disruption of TADs and TAD borders have similarly been found to play important role in tumorigenesis and cancer progress in a wide variety of cancer subtypes (S. Deng et al., 2022).

MCF-7 breast cancer cells have increased proliferation, caused in part by increased upregulation of the ESR1 gene. This gene is in a TAD thought to be regulated by the non-coding RNA (ncRNA) Eleanor, which is located at the ESR1 TAD boundaries. The ESR1 TAD is in a transcriptionally active state, and forms interactions with other TADs in compartment A. One of these interactions is with a TAD housing the FOXO3 gene, encoding a transcription factor involved in apoptosis. Upon Eleanor inhibition, the interaction between these two TADs is lost, causing silencing of genes within the ESR1 TAD while not affecting the FOXO3 expression. This suggests that the ncRNA Eleanor enables upregulation of ESR1 mediated by chromatin interactions to the FOXO3 TAD. Upon loss of Eleanor, the interaction between the TADs is lost. This leads to silencing of the ESR1 gene while not affecting FOXO3 expression, enabling the cancer cells to undergo apoptosis while having decreased proliferation. Disruption of chromatin interactions, in this case by loss of the ncRNA Eleanor, provides potential therapeutic targets in cancer medicine (Abdalla et al., 2019). TADs can also associate with the B compartment (heterochromatin) and these TADs are often associated with the nuclear lamina and termed lamina associated domains (LADs). These LADs are enriched for H3K9me2/me3 histone modifications, meaning they contain heterochromatin and thus their gene expression levels are low. Insertion of genes into LADs lead to less transcripts than when the gene was inserted into inter-TAD regions (Briand & Collas, 2020). The compartmentalization of the genome into TADs and LADs thus provides the cell with additional layers of regulatory control. The positioning of these domains and the chromatin interactions within and between them is vital for the regulation of gene expression.

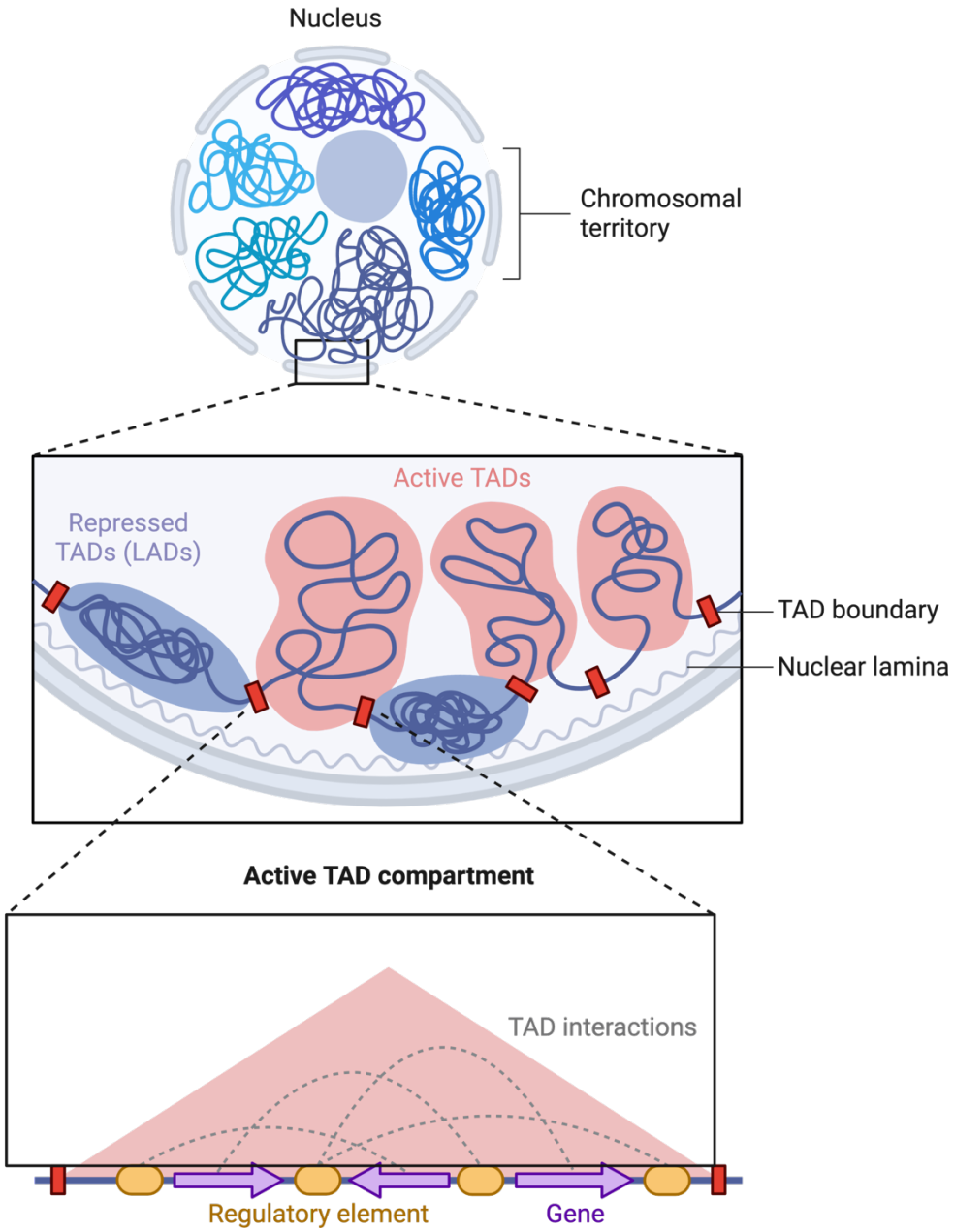


Figure 6: Schematic overview of higher order chromatin structures found in the nucleus. Chromosomes are organized into chromosome compartments, within chromosomes are the larger structured A and B compartments. Within these compartments we find TADs and LADs, denoting regions of increased self-interaction mediated by chromatin loop formation between regulatory elements and genes. Figure made with Biorender.com, adapted from "Chromosome Organization in the Nucleus: TADs" template.

1.2.5 Loops

Loops are chromatin domains thought to be formed by loop extrusion, as explained earlier. The main function of these loops is to promote enhancer promoter interactions though the looping of chromatin. Several loops can form within a TAD, but both TADs and loops are demarcated by regions such as CTCF. It is thus possible that TADs are population average structures seen in Hi-C data, emerging from the overall average loop interactions across many different cells (Hansen et al., 2018). This perspective also explains why loop formation is dynamic while TADs are more conserved across time. Loops can be seen in high resolution Hi-C heat maps, as small points of increased interaction frequencies perpendicular to the diagonal. These points in the Hi-C heatmap, found at the corners of the squares denoting TADs, are interactions between anchors in the loop, which are the regions brought close together in space by the loop formation: the enhancer-promoter interaction.

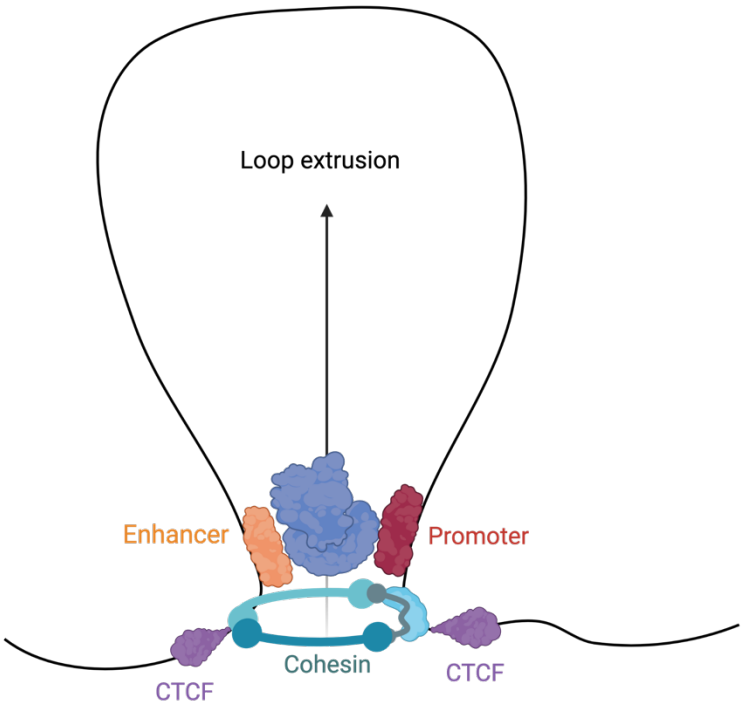


Figure 7: Schematic view of the loop extrusion mechanisms mediated by the SMC complex cohesin. The loop extrusion process continues until the cohesin complex reaches converging CTCF domains (Purple). The formation of a chromatin loop brings enhancers and promoter close together in physical space, mediated by activator proteins (orange), mediator proteins (blue) and transcription factors (red). Figure made with Biorender.com.

Disruption to chromatin loop structure can alter gene expression directly. This is demonstrated by the globin genes; the transcription of different globin genes determines the hemoglobin type. In adults, hemoglobin A is the predominant type, while in fetal development hemoglobin F is the predominant type. Distal enhancers in the locus control region (LCR) control the expression of the globin genes, mediation by the chromatin loop formation. By the engineering a novel protein, GG1-SA, the LCR chromatin loops can be changed. This GG1-SA protein contains a self-associating subunit (Ldb1SA) involved in chromatin loop formation and the GG1 subunit, which binds to the γ-globin promoter. The effect of this protein is thus binding to the γ-globin promoter and subsequent loop formation to the LCR. GG1-SA leads to altered expression of globin genes, mediated by changes in chromatin loop structure, and expression of fetal hemoglobin. This finding might be important in patients with mutations in the β-globin gene, whereby expressing fetal hemoglobin by changing the enhancer-promoter interaction mediated by GG1-SA, could be a therapeutic strategy in sickle cell anemia (W. Deng et al., 2014).

1.3 Network theory

A network can be defined as a data structure or a method to represent data, in which each object of the network is a node, and the interaction between these objects are edges. A network, also defined as a graph, thus represents objects and their relation. Networks are often used to represent and model data which has interactions between the components, and have been widely used in social sciences, economics, informatics, machine learning, and in system biology. Biological systems are often complex and contain layers of interactions acting in multidirectional manners. Representing biological systems as networks is not only useful for visualization purposes, but adds structure to the data, where each object in the network is viewed in the context of the system it takes part in. Graph theory is the mathematical field in which networks are studied and analyzed, methods from this field can thus be applied to biological data when represented as networks. Networks have been used to represent and model protein interactions, gene co-expression, metabolic pathways and 3C data as chromatin networks.

1.3.1 Networks and Hi-C data: The basics

Hi-C data can be represented as networks, where the interaction matrix is used to construct a network. In such networks, the nodes are regions of the genome, and the edges are the chromatin interactions between regions of the genome. Such networks can be made when converting the interaction matrix to formats capable of generating networks. Converting the interaction matrix to an adjacency matrix is often done, although many formats can be used to generate networks, as the data structure only needs to denote every connection between all nodes in the network. An adjacency matrix in its simplest form is defined as a square matrix, similarly to the interaction matrix in Hi-C datasets. Interactions between two nodes are denoted as a 1, and 0 for no interactions (Koutrouli et al., 2020).

Thus, the adjacency matrix captures every node and their edges. Representing Hi-C data as networks instead of 2D matrices or heatmaps can then be done after conversion to a format like the adjacency matrix. In such networks the chromatin interactions in Hi-C data are proxies for the physical chromatin structure in a cell. These interactions are close in space but can be far away in sequence. In network representations of Hi-C data the linear genomic distance between nodes is not what determines the distance between the nodes in a layout, but their interactions.

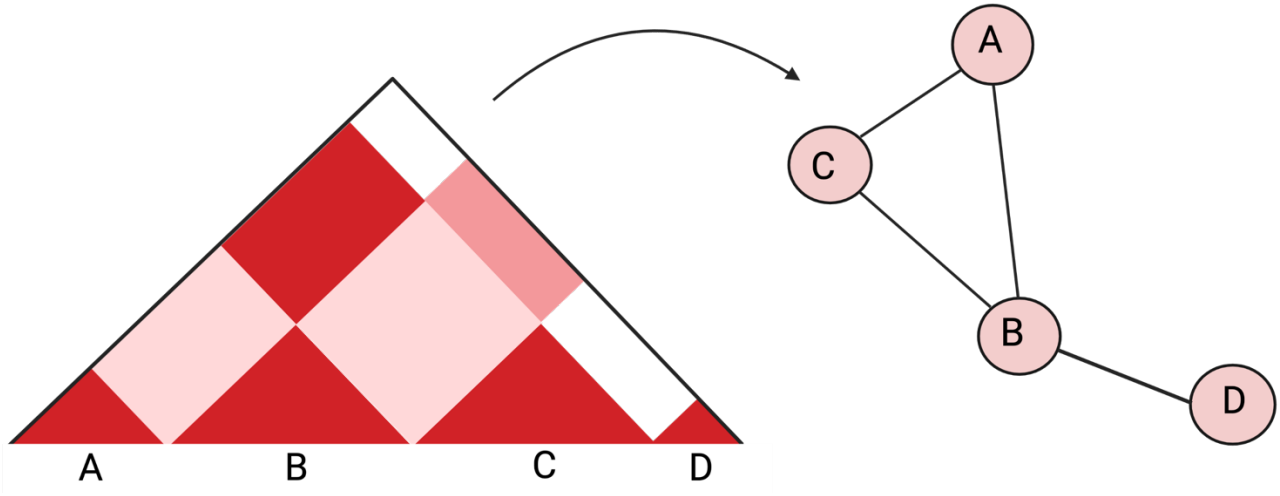


Figure 8: This figure depicts a simplistic schematic of a TAD and the corresponding chromatin network generated from it. The TAD has sub-regions which interact (denoted A-D) with varying intensities. Both representations of the Hi-C data depict the same interactions, but the network model is better fit to visualize the interactions between chromatin regions. In the network representation the sub-TADs are nodes (A-D) corresponding to the sub-TADs and their interaction patterns. Figure made with Biorender.com.

There are many types of network representations, the one in Figure 8 is an unweighted and undirected graph. This means that the edges between the nodes are not ranked by some metric, which is often the case in biological networks. Weighted edges can denote interaction strength between loci, in chromatin networks this weight can be derived from the raw interaction counts between nodes (Koutrouli et al., 2020). Chromatin networks are undirected, because one region cannot unidirectionally engage in a chromatin interaction with another chromatin region. In other words, all chromatin interactions are bidirectional, and thus chromatin networks representing chromatin interactions are undirected. Traversal of graphs is an important concept for many network algorithms and metrics. A walk is a set of nodes and edges in a network, denoting a specific traversal from one node to another. The two possible walks in the graph in figure 8 from node A to node D is A-B-D and A-C-B-D. Paths are like walks, but every edge must be unique. A valid walk in the graph in Figure 8, is A-B-C-B, while an example of a path is A-B-D.
When walks are cyclic they end up on the same node as they started at, and are aptly termed cycles, an example of a walk in Figure 8 graph is A-B-C-A (Pavlopoulos et al., 2011).
The distance between two nodes in a network is the shortest path in between them, represented as number of edges.

Subgraphs are smaller sections of the larger graph (networks). Some subgraphs form cliques, when every node is connected to every other node – a complete subgraph. The triangle formation in Figure 8 formed between node A-B-C is an example of a clique. Similarly, clusters are subgraphs of the larger networks which form more tightly connected regions, but nodes in clusters are not required to be complete. A graph is termed connected if there is a path from any node to any other node, as in there are no nodes disconnected from the main graph. Many graphs form from several components, which are smaller regions of the larger graphs where all nodes are connected. With these definitions in mind, the graph in Figure 8 is an undirected, unweighted complete graph.
Several metrics can be calculated for nodes in a graph, one important metric is the degree, which is how many edges a node has associated with it. In the example graph in Figure 8, node D has a degree of one and node B a degree of 3. Figure 9 is an example of a heatmap, and a network generated from real data, but these concepts apply to all undirected, unweighted chromatin

networks. The network in figure 9 is a processed network, but the heatmap represents normalized data that has not been fully processed.

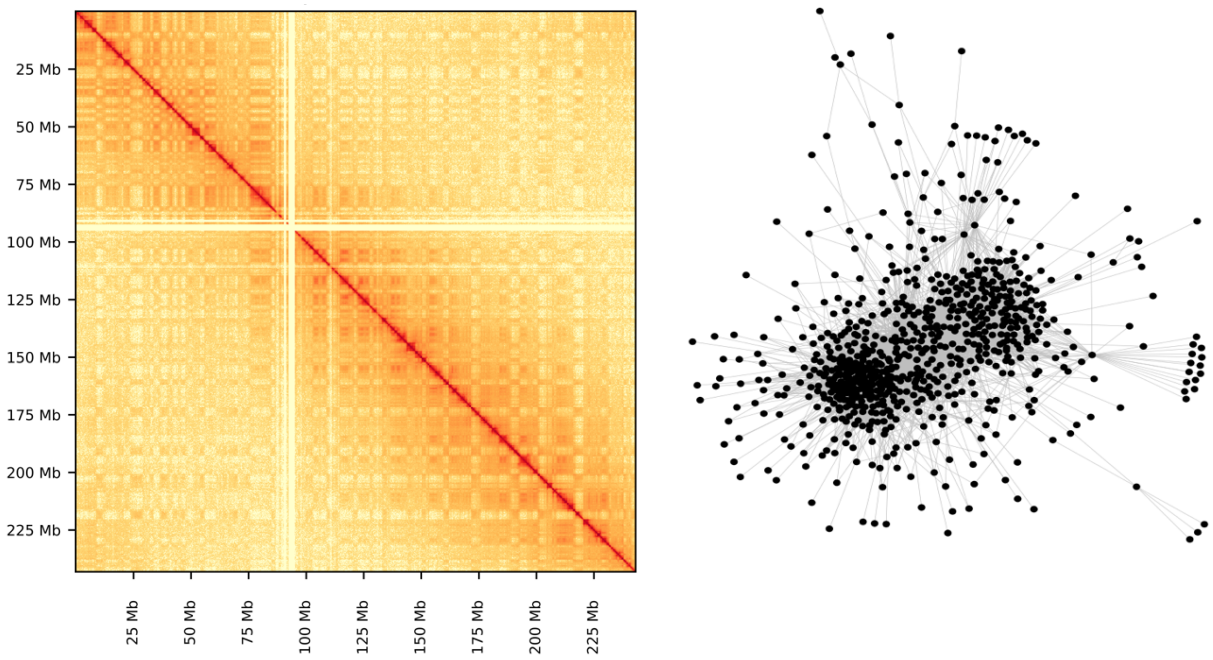


Figure 9: Heatmap and corresponding network generated from chromosome 2 from the IMR-90 cell line at 250kb resolution. The network represents processed data while the heatmap does not, thus the number of interactions in the heatmap does directly not correspond to the number of nodes and edges seen in the network.

There are different approaches to creating chromatin networks. The first being derived from spatial data gained from microscopy and labeling to describe direct spatial location of chromatin. The second approach relies on modeling chromatin interactions to create spatial models of chromatin networks at varying scales. These approaches are used to create 3D models of chromatin, based on polymer physics or constraint modeling. The final method of representation in chromatin networks are the abstract topological representations. In this approach chromatin interactions are represented as networks directly from Hi-C datasets, without embedding the nodes spatially or modeling the physical conformation of chromatin. Graph theory approaches are used to describe these networks and their structure, which might then be used to describe the phenotype (Pancaldi, 2023). Chromatin networks considered in this thesis is of the latter type, not spatially embedded nor temporally annotated, but abstract representations of chromatin interactions. All networks discussed in this work are undirected, unweighted networks unless stated otherwise.

1.3.2 Centrality

The centrality of a network is a collection of summary statistical metrics used to describe individual nodes in a network. Such centrality metrics are important tools that allow for identification of important nodes in a network. Many centrality metrics exist (Das et al., 2018), but some of the most common ones are degree centrality, closeness centrality and betweenness centrality. These metrics inform us of a node's importance in the network and the nodes influence over other nodes. The degree centrality is a measure of every node's degree in a network. In an undirected network the degree is simply the number of edges associated with each node. Closeness centrality is a measure of how close nodes are to other nodes in the network. It is defined in undirected networks as the reciprocal sum of shortest paths between a specific node and all other nodes in the network. Closeness centrality is defined mathematically for one node i as:

$$C_{closeness}\,(i) = \frac{1}{\Sigma\, d_{i,j}}\,, \qquad for\ all\ j \neq i$$

Where i is a specific node, j are all other nodes and the $\Sigma\, d_{i,j}$ term denotes all shortest path length from i to j.

Betweenness centrality is a measure of how important a node is for communication between different parts of a network. Betweenness centrality is calculated in undirected networks as the shortest path between every pair of nodes in the network, for each node summing the number of shortest paths pass through that specific node. Repeating this for all nodes in the network results in the raw betweenness count of every node in the network. Betweenness centrality is thus defined for one node i as:

$$\mathrm{C}_{betweennes}\,(i) = \frac{\sigma_{x,y}i}{\sigma_{x,y}}, \qquad for\ all\ x \neq i \neq y$$

Where $\sigma_{x,y}$ is the sum of shortest paths between nodes x, y and $\sigma_{x,y}i$ is the sum of shortest paths between nodes x, y passing through node i (Koutrouli et al., 2020). All these centrality metrics can be normalized in respect to the size of the network.

1.3.3 Community detection

Centrality metrics are used to describe singular nodes of interest and how they relate to the networks overall structure, the network topology. To characterize larger structures relating to network topology it is also important to consider groups of nodes. These groups of related nodes can be termed clusters or communities – and there is a wide array of algorithms to define such groups. The goal of defining communities in a network is to group nodes based on their structural similarity. There are differences in the clustering between networks and it is not obvious which level of clustering best defines communities within a network. A network could for instance be partitioned into two parts – but this is not very informative (Newman, 2006). One approach to solve this problem and define communities is by using modularity maximizing algorithms. Modularity, as defined by Newman & Girvan in 2004, is a measure of density of edges within a community compared to outside a community. It can be defined as:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - k_i k_j)\, \delta(c_i, c_j)$$

Here m is defined as the number of edges in the network, and A is the adjacency matrix representing the network while A_{ij} represents one element in this adjacency matrix for nodes i, j. The product of the degree between nodes i, j is represented as $k_i k_j$. The communities that the nodes i, j belong to are defined as c_i, c_j. The Kronecker delta function is δ, which returns 0 for edges between nodes in separate communities, only returning 1 for edges where nodes are assigned to the same community. This function allows for distinguishing between nodes assigned to different communities. Modularity values for a network where communities are defines ranges from -1 to 1. The closer the modularity is to 1 the better the community structure is, for real world networks the values of a network with good community structure does not approach 1 but ranges from 0.3 to 0.7, a higher modularity in a network means nodes within a community have denser connections then outside of communities. The result of this when implemented in community detection algorithms is the optimization for modularity, where this function is applied to each set of nodes in a network.

There are several such implementations of modularity-based community detection algorithms, each with differing implementation and optimization strategies. Some common modularity-based algorithms are the fast-greedy algorithm and the Louvain algorithm (Lancichinetti & Fortunato, 2009). There are many other approaches for defining communities which do not rely on modularity optimizations. In general, one can categorize community detection algorithms as modularity based approaches, traditional algorithms and dynamic algorithms (Javed et al., 2018). These categories are not clear cut, for instance fast greedy optimizes for the higher modularity score but is also a hierarchical clustering algorithm, which is defined as a traditional approach. What they have in common is that they define nodes as part of only one community, but this might not reflect the reality of biological networks, as they are dynamic. This means that optimal approach might be to assign nodes are members of several communities to better reflect the reality of biological networks (Javed et al., 2018).

Community structures can be compared across networks, but many algorithmic implementations of such methods assume that the communities in each network contain the same number of nodes. This is often not the case in chromatin networks, as which might not be the case for chromatin networks. Methods developed to compare networks with differing number of nodes and node sets in community structures are termed unknown node-correspondence methods (Tantardini et al., 2019). Comparing community clusters is a difficult problem because communities can be nested, contain different nodes and have different sizes. Normalized mutual information (NMI) is a metric that scores the similarity between two clusters – or community structures. Algorithmic implementations of NMI that allow for differences in community sizes can be used to compare the community structures between two graphs (Lancichinetti et al., 2009). An improvement on the normalization method for this algorithm was made to better handle different overlaps of communities (McDaid et al., 2011). These approaches were originally aimed at comparing community structures determined by different community detection algorithms. But NMI can also be used to calculate the similarity between communities of two different networks, using an implementation of the NMI metric which allows for differing number of nodes between communities. The NMI is a normalized score between 0 and 1, a higher score means the community structure is more similar between the compared networks. NMI works by comparing the mutual information between two variables, when comparing two

networks with differing communities, it measures how much information is gained about the community structure in one network by examining the other networks community structure. There are several other approaches for comparing community structures that allow for different node sets, community structures and variables nestedness. These NMI methods are general mathematical approaches and can be used to find the similarities between communities. In addition to these methods, a myriad of biological network approaches have been developed that seek to define differences between different networks (Ideker & Krogan, 2012). Many of these approaches do not consider communities structures directly, but instead evaluate the differences between communities on an edge-to-edge basis.

One recent approach that considers differences in community structure between two biological networks is the Alpaca algorithm (Padi & Quackenbush, 2018). In this approach, one network with communities in considered the ground truth, and compared to another network's community structure. This model considers differential community structures, where the expected number of edges in the perturbed network is calculated based on the community structure found in the baseline network (the ground truth). The modularity scores are then compared between the networks and uses a Louvain algorithm to maximize this differential modularity score to determine the differential communities between two networks. Most approaches used in biological networks assume the networks are weighted networks and that there is a known node-correspondence between the networks, as is often the case when working with protein interaction networks (PINs) or gene co-expression networks. Besides these approaches, there are also global summary approaches that can be used to compare the community structure between two networks, for instance the modularity score. General approaches, NMI, and modularity, measure the overall structural differences between communities in two networks. Unlike the more specialized approaches like the Alpaca algorithm, the generalized metrics do not inform on specific community differences in networks. Instead, they give an overall similarity of the partition differences between the two networks, reflecting the overall community structure similarities.

1.3.4 Chromatin networks

Biological networks often form structural architectures which are seen in other real-world networks, namely that they are scale-free and form core-periphery structure. A scale-free network has an asymmetry in the distribution of degree centrality, meaning that most nodes have low degree while a few nodes have a very high degree. Networks are deemed scale-free if they follow such a power-law. Scale-free networks are less prone to structural changes when removing individual nodes, if random nodes are removed from a scale free network it is less likely to affect the overall topology of the network. This is because there is a large difference in degree distribution between nodes in such networks, where most nodes have a low degree and are less important for the networks overall structure (Albert et al., 2000). In biological terms, this organizational feature might allow higher tolerance in general to alteration of components in the network. If a disease affects a hub node it is then expected to perturb the network, and the underlying phenotype, more than a peripheral node (Furlong, 2013). One example of this is in the HIST1 gene cluster, which form several long-range (<200kb) and super-long-range (<500kb) interactions. If these interactions are lost, the cluster would break into several smaller components. It was further found that strong promoters and enhancer have are hub nodes and occupy more central positions in CINs, while weak enhancers and promoters are located in peripheral regions of the network (Sandhu et al., 2012). This further illustrates that network structures reflects the underlying biological function of these systems.

Core-periphery structures are observed in a wide variety of systems pertaining to information flow, like the brain and social groups. The core contains nodes which are central to the network and affect many other nodes. This network structure was observed in a subset of hub TADs in cortical neurons, where disease causing single-nucleotide polymorphisms (SNPs) in neurological disease are enriched in these core TADs. This indicates that disruption to core TADs lead to disruption to the network topology, while a change in the periphery does not (Huang et al., 2019). Both the scale-free and core-periphery structure of networks serve to make networks more robust against perturbance and channel communication through specific paths. A node with high centrality, e.g., degree or betweenness, will thus have a larger effect on the topology of the network. This indicates that centrality metrics reflect the underlying biology in the system.

Networks also allow for annotation of datasets describing the linear genome, which can be integrated into the networks. An example of this is annotating nodes in the network with histone modification data from ChIP-Seq. Several datasets can be integrated at the same time to provide a clearer view of the genome in 3D and how it relates to function. This approach allows for one-dimensional genomic data to be related to the 3D structure of chromatin. Chromatin assortativity is an example of this, it is a measure of the correlation between features of nodes in networks, ranging from -1 to 1. This metric can be used to determine if similar nodes preferentially interact with each other. A feature can be histone modifications or promoters, and this approach can be used to define the interaction patterns in chromatin networks. Histone modifications correlate with chromatin state and was determined to be assortative in promoter-capture chromatin networks of neutrophils. Some variability was found between individuals, for instance in the H3K27Ac histone modification which is associated with active transcription (Madrid-Mencía et al., 2020). These findings suggest that chromatin with similar features preferentially interact, which is known to be the case for large scale chromatin structures like the A and B compartments. Using network approaches to find assortativity can thus shed light on chromatin structures and how they interact.

Networks also allow for multiscale hierarchical representations of biological data, where nodes can represent different structures at differing scales. Such representations of biological networks might reflect the complexity of biological networks more accurately since they capture the complex nestedness found in biological systems. A hierarchical network was created by combining Hi-C, promoter capture Hi-C (PCHi-C) and compartment profiling data to create a single multiscale network. Using this approach allows for exploration of chromatin interactions across multiple scales, with integration of multiple annotations like enhancers and promoters being possible. Comparing different murine cell lines using multiscale networks annotated with transcription factor sites and methylation profiles revealed changes across all scales. Primary and naïve stem cells were compared across multiple Hi-C resolutions and with respect to the promoter sites. The network was then annotated with methylation states and large differences between the cell lines were found. For instance, the cadherin gene is controlled by enhancer-promoter binding mediated by CTCF loops to form an interaction hub. This hub seems to form during development, which might be priming this region for transcription when the naïve

pluripotent stem cells (PSC) transition to primed PSC (Chovanec et al., 2021). This approach underlines the flexibility of networks, they can contain multitudes of annotated data and represent hierarchical structures, such as chromatin organization. Network theory approaches have also been used to call TADs allowing for capture of their nested structures, using modularity maximizing community algorithms to detect nested TADs (Norton et al., 2018) across multiple scales (Yan et al., 2017).

Nodes in chromatin networks are organized into components of varying size. Around 40% of nodes in ChIA-PET networks formed one giant connected component, termed the largest connected component (Sandhu et al., 2012). This was also found to be the case in promoter-capture Hi-C networks (Lace et al., 2020). This implies that a large proportion of the genome has widespread potential for co-regulation, although this represents a population average, meaning in single cells this largest connected component (LCC) might differ significantly from average LCC in the population-level network. ChIA-PET chromatin interaction networks from K562 and MCF-7 cell lines were found to form scale-free networks (Sandhu et al., 2012).
Local changes do occur between cancer cell lines and healthy cell lines. By creating CINs annotated with H3K4me and H3K27ac marks, it was wound that super-enhancers and broad domain promoters interact more frequently. Using network approaches to examine these interactions revealed that the connectivity between these elements were conserved across cell lines. Decomposing chromatin networks into graphlets (topologically unique subgraphs) revealed that different chromatin features exhibit unique topological patterns. For instance, the super-enhancers occupy more central locations and are more likely to form cliques compared to enhancers. The EMP2 oncogene in the MCF-7 breast cancer cell line had a different network structure than the K562 and GSM12878 cell lines, being connected to more enhancers. (Thibodeau et al., 2017). To determine whether chromatin network patterns reflect the underlying biology, machine learning models can be trained on CINs and used to predict cell-activity. Such a model was trained and could distinguish between super-enhancers and enhancer, as well as broad domain- and regular promoters with an AUC of <0.7 for all cell lines tested (Thibodeau et al., 2017). While this model is not the most accurate, it does indicate that networks approaches are valuable tools for understanding the underlying biology. Similar approaches using graphlets and network centralities were used to characterize chromatin networks from

healthy and cancer-derived human blood cell lines. There were differences between healthy and cancer cells, indicating that structural changes in the chromatin network reflect the underlying phenotype of the genome. Leukemia genomes had a lower modularity score, indicating poorer community structure. In concordance with this overall structural change, driver genes of chronic lymphocytic leukemia adopt more central positions in the CIN to become hub nodes (Malod-Dognin et al., 2020). Machine learning models have also been used to show that Hi-C interaction models as networks can predict gene co-expression, without input from external data (AUC <0.77). When the model trained on topological networks representing Hi-C interactions in murine cortical cells it could accurately predict gene co-expression. This might be an effect of chromatin assortativity; similar chromatin features co-localize spatially and gene-pairs that co-express interact with these assortative features.

Co-expression of genes was not found to have a one-to-one relationship with chromatin interactions, but was rather the result of indirect interactions between several chromatin regions (Babaei et al., 2015). This means that representing and modeling chromatin interactions as networks allows for better representation of the complex underlying biology.

2. Aims of study

The overarching goal of this master project is twofold: (1) Develop an automated processing pipeline that generated Hi-C networks from raw Hi-C data, and (2) use network theory approaches such as degree centrality and community detection to characterize and compare these networks.

3. Methods

3.1 Datasets

Hi-C data from five different cell lines was acquired and processed to networks. Two breast tissue-derived cell lines were used, MCF-10A and MCF-7. The MCF-10A cell line originates from healthy mammary epithelial tissue and is widely used as a model for healthy breast tissue (Qu et al., 2015). MCF-7 is a commonly used model for cancer research as they are transformed mammary epithelial tissue that is estrogen responsive (Vantangoli et al., 2015).
Three additional cell lines from healthy tissue was processed and analyzed. HUVEC is derived from the vein of the umbilical cord and are primary tissue cells (Baudin et al., 2007).
The IMR-90 cell line is derived from myofibroblasts isolated from lung tissue (Ehler et al., 1996). HA-c are astrocytes isolated from the cerebellum (Dekker, 2017), and is referred to in the results by its NCBI sample accession number (gsm2824367). All datasets come from cell lines except for HA-c and HUVEC, which are primary cells. The Hi-C libraries were all generated using the HindIII restriction enzyme, except for IMR-90 which used MboI. All Hi-C libraries was aligned to the hg19 reference genome. All datasets were generated from dilution Hi-C protocols, except for the HA-c dataset, which was generated from an in-situ Hi-C protocol.

The MCF-10A and MCF-7 Hi-C data (Barutcu et al., 2015) used in this study can be accessed from the European Nucleotide Archive (ENA) at EMBL-EBI with project accession number PRJNA277846. The individual run accessions are found at SRR1909069 (MCF-10A) and SRR1909070 (MCF-7), located at: https://www.ebi.ac.uk/ena/browser/view/PRJNA277846.

The IMR-90 Hi-C data (Rao et al., 2014a) was retrieved from the ENCODE portal with experiment accession number ENCSR645ZPH located at:
https://www.encodeproject.org/experiments/ENCSR645ZPH/.
The individual reads have accession numbers ENCFF238TYE and ENCFF361VGE.

HUVEC Hi-C data was also retrieved from the ENCODE portal, with the following experiment accession number ENCSR008UIB located at:
https://www.encodeproject.org/experiments/ENCSR008UIB/

Individual reads have accession numbers: ENCFF468XVT and ENCFF395VSI.

The HA-c Hi-C dataset has the GEO accession number GSM2824367, the ENCODE experiment accession number ENCSR011GNI and was retrieved from: https://www.encodeproject.org/experiments/ENCSR011GNI/. The individual reads have accession numbers: ENCFF198DDF and ENCFF901VRD.

3.2 Pipeline

A pipeline was made to pre-process raw Hi-C data into edge-lists to generate networks, based on the Chrom3D pipeline (Paulsen et al., 2018). The pipeline automatically processes Hi-C data from several cell lines across many resolutions, in parallel. The pipeline takes input files from Hi-C Pro output, formats them correctly, finds significant interactions, adjusts for false discovery rate, and returns data in a network-compatible format, the default being edgelists. This allows for automatic generation of networks in python that are compatible with iGraph, NetworkX and Cdlib, common network processing and visualization tools. An overview of the pipeline is shown in figure 10 below, highlighting the most important steps from raw Hi-C data to networks representing the chromatin interactions present in the Hi-C datasets.

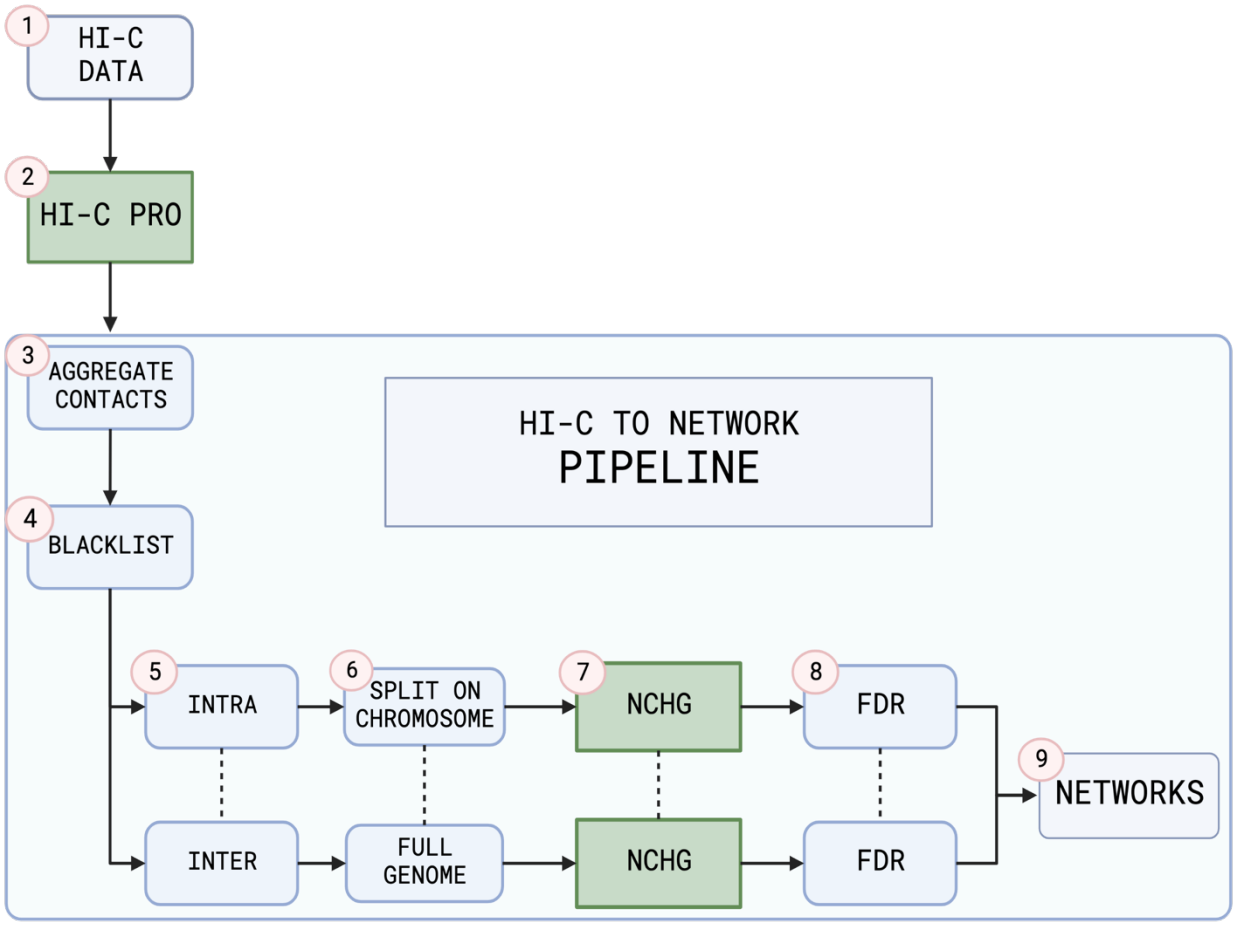


Figure 10: Schematic overview of pre-processing pipeline used to generate networks from raw Hi-C data. 1) Acquiring raw Hi-C data in paired-end FASTQ format is the first step, this is the input to Hi-C Pro. 2) Hi-C Pro aligns the reads to the reference genome of choice and filters low quality data. Hi-C Pro returns a matrix file with the number of interactions per loci, which maps to a BED file denoting the genome segments. 3) This is the first input to the network pipeline, which maps the interactions to the binned genome, creating a BEDPE file. 4) This BEDPE file is run through a series of steps which removes "blacklisted" regions like centromeres and low mappability regions. 5) The full genome file is split into intra- and interchromosomal datasets. 6) The intrachromosomal files is further split into each chromosome, while the interchromosomal file is kept in its full genome state. 7) These files are fed in parallel to the NCHG tool, which finds significant interactions both for intra- and interchromosomal data. 8) The intrachromosomal files are concatenated, both intra- and interchromosomal data is controlled for false discovery rate (FDR) using Benjamini-Hochberg procedure. 9) The intra- and interchromosomal files are formatted to edgelist files and imported to Python to generate iGraph- and NetworkX-compatible networks. Green square boxes denote external tools used. Figure made with Biorender.com.

3.2.1 Hi-C Pro

Hi-C Pro is a tool used to generate Hi-C datasets from raw Hi-C libraries. It takes paired sequenced FASTQ files as input and generates full-genome contact maps. It is a powerful tool that aligns the reads to a reference genome, filters out low quality data and allows for correction of bias, all while running efficiently in parallel (Servant et al., 2015). Reads are first aligned to their respective locations on the reference genome and interactions are classified. The first step when running Hi-C Pro is mapping, this is the process of aligning reads individually to a reference genome. Some reads will map to several regions of the genome and will be unable to be aligned using the Bowtie2 aligner. This is handled by Hi-C Pro aligning the 5' end of the read to the genome by finding the ligation sites, so that unmapped chimeric reads can also be aligned. For a read to be considered valid it must map to two different restriction fragments, while also representing valid reads as generated from the Hi-C protocol. Some reads to not form valid read pairs and Hi-C Pro has extensive quality controls in place to filter these reads out. Some examples of this are singletons (reads from one loci), form dangling ends (larger than expected reads), self-circle reads (reads form circles) and relegation (two restriction sites ligate). Reads that do not fulfill these quality criteria are discarded, while the valid reads are aligned to restriction fragments. After alignment of valid reads to restriction fragments, the genome is then divided into equal size bins of the pre-determined sizes, designating the resolution of the dataset. All valid reads are then aligned to its corresponding bin to generate a full-genome dataset containing all Hi-C interactions (Servant et al., 2015).

The reference genome used in this pipeline was hg19 (Kent et al., 2002). Hi-C Pro uses the Bowtie2 aligner tool, which requires indexed versions of the reference genome to enable more efficient mapping of reads. These indexed reference genome files were generated using Bowtie2's indexing functionality (Langmead & Salzberg, 2012). In addition to the Hi-C Data and the reference genome, Hi-C Pro requires two reference files: One file containing the sizes of each chromosome and one file containing the sizes of all restriction fragments. The restriction fragment file should correspond to the restriction fragment used in the generation of the Hi-C library. These reference files for hg19 are part of the Hi-C Pro repository and enable correct mapping of Hi-C reads to the reference genome.

3.2.2 Aggregation, blacklist, and centromeres

The next step of the pipeline is taking the output from Hi-C Pro which is cell-line and resolution specific. This means Hi-C Pro generates two files for each Hi-C library per resolution, the matrix file contains each interaction and interaction frequency, while the BED file contains the genomic region (the bins). There is a mapping between the interaction counts and the bins, allowing for the creation of a single file, aggregating the frequencies and the bins into a single file. This file is in the BEDPE format, which contains each interaction as regions of the genome and the interaction frequency between each region. This process is done for each resolution, and each cell line. These BEDPE files are then checked for overlap with blacklisted regions (Amemiya et al., 2019). The ENCODE blacklist dataset contains regions of high-signal, irregular and unstructured regions, such as high-repeat regions resulting from e.g., transposable elements. These regions are removed from the datasets to ensure higher quality of downstream data analysis. The next step in the pipeline is similar, centromeres are removed by checking for overlap with a BED file containing centromeric regions for hg19 generated from USCS genome browser. Both filtering for blacklisted regions and centromeric regions is done with the python package Pybedtools (Dale et al., 2011).

3.2.3 Statistical significance

Hi-C data is important to test for significance. During the Hi-C protocol, loci in spatial proximity will crosslink and be captured as chromatin interactions. There is inherent noise in Hi-C datasets resulting from the random movement of chromatin, resulting in a background noise in the data. In addition to this, there are also biases present ranging from batch effects to sequence specific bias such GC content. Several different approaches have been developed to deal with the noise and bias in Hi-C data to determine significant chromatin interactions (Ay & Noble, 2015). Matrix balancing is a commonly used normalization strategy, aiming to reduce inherent bias in Hi-C contact maps. The method operates under the assumption that if there were no bias present, all loci would produce equal amounts of Hi-C reads in the experiment, and the number of reads in a Hi-C experiment would be proportional to the number of interactions. The Hi-C matrix is decomposed such that the observed interactions sum to equal values for each locus and the decomposition term captures the bias. In matrix balancing approaches distance dependent decay

is not accounted for, which can be problematic for intrachromosomal data. Furthermore, no confidence or estimate of significance is attached to interactions.

To account for distance dependent decay and capture significant Hi-C interactions, the Hi-C to network pipeline uses the NCHG statistical tool (Paulsen et al., 2018). This tool is made to find statistically significant interaction in Hi-C datasets and is based on the non-central hypergeometric distribution. This tool calculates the probability of finding the number of interactions between loci observed in the dataset. It achieves this by finding the total number of interactions for the Hi-C matrix, the number of contacts between the two loci interacting and the linear genomic distance between these loci. This implementation of the non-central hypergeometric distribution is defined as:

$$P\left(n_{ij} \mid n, n_i, n_j, \omega_{ij}\right) = \frac{\binom{n_j}{n_{ij}}\binom{2n-n_i}{n_j-n_{ij}}{\omega_{ij}}^{n_{ij}}}{\sum_{{n'}_{ij}}\binom{n_j}{{n'}_{ij}}\binom{2n-n_i}{n_j-{n'}_{ij}}{\omega_{ij}}^{{n'}_{ij}}}$$

Here P denotes the probability of observing a given number of chromatin interactions between two loci n_{ij}. This probability depends on the total number of chromatin interaction in the chromosome n and the individual number of contacts to all regions for each locus involved in the interaction, n_i and n_j. The n' (n prime) term denotes the total number of interactions in the dataset. The last term, ω_{ij}, is the odds ratio for the loci involved in the interaction, and is given by this equation:

$$\omega_{ij} = \frac{\lambda_{ij}(2\lambda - \lambda_i - \lambda_j + \lambda_{ij})}{(\lambda_i - \lambda_{ij})(\lambda_j - \lambda_{ij})}$$

Where the odds ratio for the two loci involved in an interaction, ω_{ij}, is determined by the expected interaction between the two loci, λ_{ij}. The odds ratio is a measure of how much the actual interaction frequency deviates from the expected interaction frequency given the distance between two loci interacting. Lambda can be determined by the number of interactions between two loci by taking the genomic distance between these loci into account for intrachromosomal

data. For interchromosomal data a constant expected interaction frequency is determined since there is no distance dependent decay in interchromosomal data. The probability for the number of contacts for interchromosomal data is then transformed into a Fisher's exact test with a central hypergeometric distribution, since there is no distance dependent decay effect between loci on different chromosomes. The NCHG tool calculates p-values based on the odds ratio for every interaction in the dataset and returns each interaction with a corresponding p-value. The NCHG script is originally written to run sequentially, and the statistical testing is computationally expensive as it runs with time complexity of $O(n^2)$. The runtime increases cubically with each doubling of resolution in the matrices, meaning high resolution data is very computationally expensive to process. To increase performance, the pipeline enables splitting of the intrachromosomal input files on each chromosome. It then calls the NCHG script in parallel for all input files. There is no need for parallelism when processing interchromosomal data as the resolution (bin size) for these regions are 1 Mb or larger. After completion the single-chromosome output files are joined to generate a full-genome file once more. Each output fil is specific to the resolution that generated it, and intra- and interchromosomal datasets are kept separate. This is not only due to the practical limitations imposed by statistical tools, interchromosomal interactions occur much less frequently and are governed by different biological processes than the compartments found on single chromosomes.

One statistical test, as described above, is performed for each interaction present in the dataset. The null-hypothesis is thus the interaction frequency between two loci that is regarded as random, given the distance between them and the total number of interactions in the Hi-C matrix. Small p-values means the interaction frequency observed between the two loci is very unlikely to occur under the null hypothesis, indicating a significant interaction. Still, doing many multiple tests results in a portion of significant interactions being false positives. To account for this, false discovery rate detection was done using Benjamini-Hochberg's procedure (Benjamini & Hochberg, 1995). This results in the expected number of false positives being below the threshold, in this case 5%, meaning all p-values above 0.05 is excluded from the dataset and all p-values below 0.05 are deemed significant interactions. After this process is complete, the input data with raw counts is transformed to significant interactions and interactions with a significance below the threshold are removed.

3.2.4 Edgelists and Networks

After significant interactions are determined, the datasets are reformatted to edgelists. This is a simple data structure, where each line in the file represents one interaction. Interaction frequency between two loci is not considered after significance testing. This means that one locus can have several interactions to different loci, but the number of interactions between them is omitted. This means that the edge lists are used to generate unweighted, undirected networks. Networks are generated as an optional step in the pipeline, from the edgelist files, and automatically creates networks using the iGraph package (Csardi & Nepusz, 2005).

3.3 ICE Normalization

In the Hi-C to network pipeline the NCHG tool was used to find significant Hi-C interactions, and p-values were adjusted using Benjamini-Hochberg's procedure. When generating heatmaps, this was not the case, as the NCHG tool filters out too many interactions to generate complete heatmaps. Instead, raw data was normalized using the iterative correction and eigenvalue decomposition (ICE) method (Imakaev et al., 2012). This algorithm is implemented in Hi-C Pro but is also part of the python package Cooler (Abdennur & Mirny, 2020). The BEDPE files were converted to cooler files, and then the matrix was balanced to allow for clearer visualization of the Hi-C data. When defining the compartment status of each chromosome of each cell line, these balanced cooler files were used as input. The cooler and Scikit-learn python package was used to perform eigenvalue decomposition on the matrices (Pedregosa et al., 2011). The compartment status assigned to each genomic bin in the dataset, done on a per-chromosome basis. The direction of the sign of the first principal component of these matrices is what defines a compartment region. When this sign switches value, a compartment switch has occurred (Fortin & Hansen, 2015). These files were then converted back to BED files, allowing for integration with downstream analyses, such as network annotation.

3.4 Network metrics

Network calculations, centrality metrics and community detection was done using iGraph and the cdlib python packages (Rossetti et al., 2019). iGraph has an extensive library of built-in functions and algorithms that can be used to do calculate common network and centrality

metrics. There are also built in community detection algorithms, such as the fast greedy algorithm in iGraph and the fuzzy NMI algorithms, both present in. In addition to this, the cdlib library was used to compare networks as this package offers fuzzy algorithmic implementations of common comparison methods like NMI.

3.5 Visualization

All network visualization were generated using iGraph with matplotlib as the plotting backend (Hunter, 2007). Heatmaps were made using cooler, the remainder of plots not depicting networks nor heatmaps were made with matplotlib or the Seaborn python package (Waskom et al., 2014).

3.6 Code availability

All code written for and used in this project is available at:
https://github.com/Gabrielstav/mastercode

4. Results

4.1 Global network characterization

The Hi-C to network pipeline was used to process raw Hi-C data from five cell lines to obtain Hi-C interactions networks. The pipeline ran in parallel for all cell lines, processing data across a wide range of resolutions. MCF-7 and MCF-10A cell lines were processed at resolutions 50kb, 80kb, 120kb, 250kb, 500kb and 1Mb for intrachromosomal data. Another run was done for intrachromosomal data for the breast tissue cell lines, where the Hi-C matrices were normalized using iterative correction (ICE) matrices as an output from Hi-C Pro before finding significant interactions. The other three cell lines, IMR-90, HUVEC and HA-c (referred to as gsm2824367), were run at 120kb, 250kb and 1Mb for intrachromosomal interactions. Interchromosomal data was processed for all cell lines by the pipeline, at a 1Mb resolution. As the pipeline runs, the interactions counts are drastically reduced after multiple testing by NCHG and correcting for multiple testing by FDR, resulting in a decrease of interactions to around 2-5% of the original size of the network. The remaining bins are thus deemed statistically significant and are not associated with an interaction frequency. The output of the pipeline was then used to generate networks, yielding in total 38 networks representing Hi-C data. All networks analyzed in this work are intrachromosomal networks – only containing edges from nodes on the same chromosome. All networks analyzed in this work are intrachromosomal networks, where all nodes within a network have the same resolution.

These intrachromosomal networks range in size, increasing drastically in number of nodes for high resolutions. High resolution MCF-10A and MCF-7 networks not normalized by iterative correction before running the pipeline form smaller networks compared to the other cell lines at higher resolutions. The network size of the different datasets converges as resolution lowers – being the most similar at 1Mb resolution (see figure 11). As the resolution of Hi-C data halves (e.g., from 1Mb to 500kb) the number of bins doubles, the resulting increase in network size with higher resolution is thus to be expected.

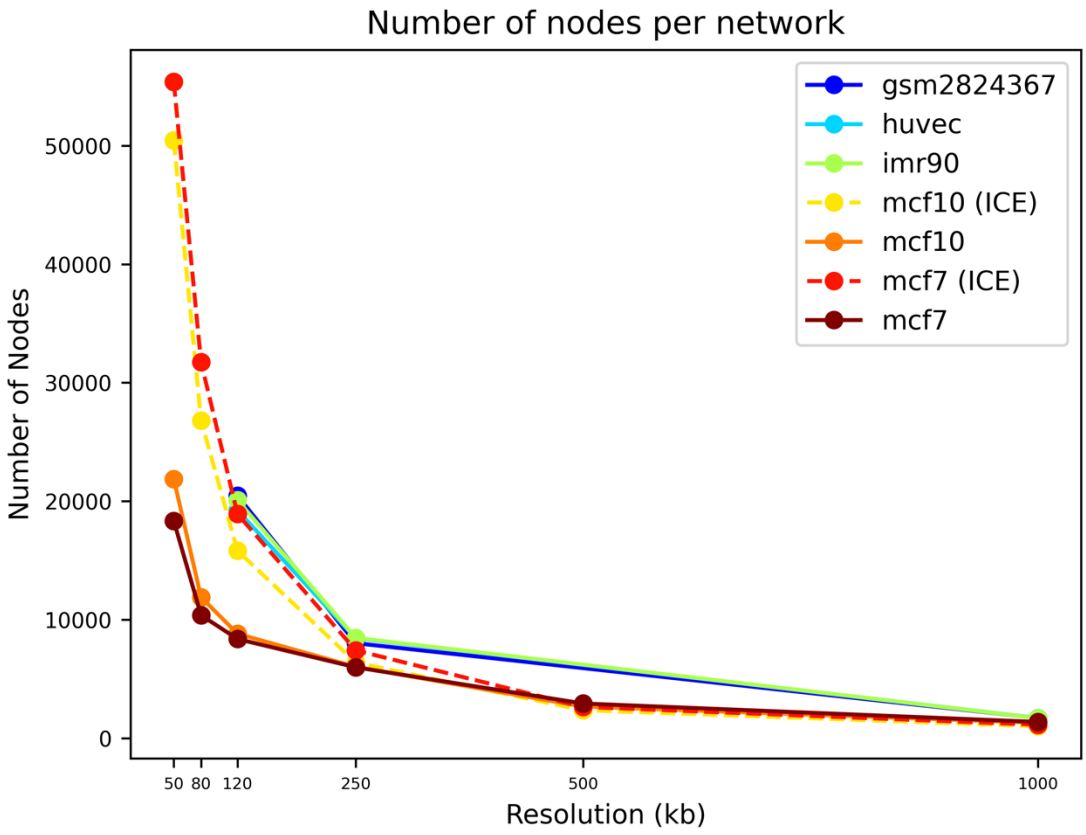


Figure 11: Relationship between resolution and network size (number of nodes). All solid lines represent raw data, while the dashed lines represent data corrected for bias by iterative correction matrix balancing before running the pipeline proper. The network size increases when resolution increases. The normalized MCF-10A and MCF-7 data is closer to the IMR-90, HUVEC and HA-c (gsm2824367) cell lines in respect to node size. MCF-10A and MCF-7 networks not generated from normalized data show a much lower network size than their counterpart. The spread in size increases with resolution, converging on similar values at 1Mb. Ignoring the raw MCF-10A and MCF-7 networks at high resolutions the general trend is the same for all cell lines and interaction counts increases slowly from 1Mb to 250kb resolution.

Similarly, the number of interactions (edges) per network across resolutions was determined. For higher resolutions, the MCF-7 and MCF-10A iteratively corrected (ICE) and raw data diverge. The ICE normalized data is more like the HUVEC, IMR-90 and GSM (HA-c) datasets for higher resolution data (120 kb). These patterns can be seen in Figure 12. At lower resolution (250kb – 1Mb) the interactions counts are similar for the raw and normalized MCF-10A and MCF-7 datasets, while the other cell lines have a higher interaction count at lower resolutions. At high resolutions, the iterative corrected (ICE) MCF-10A and MCF-7 cell lines diverge from their counterparts becoming more like the other cell lines (at 120 kb).

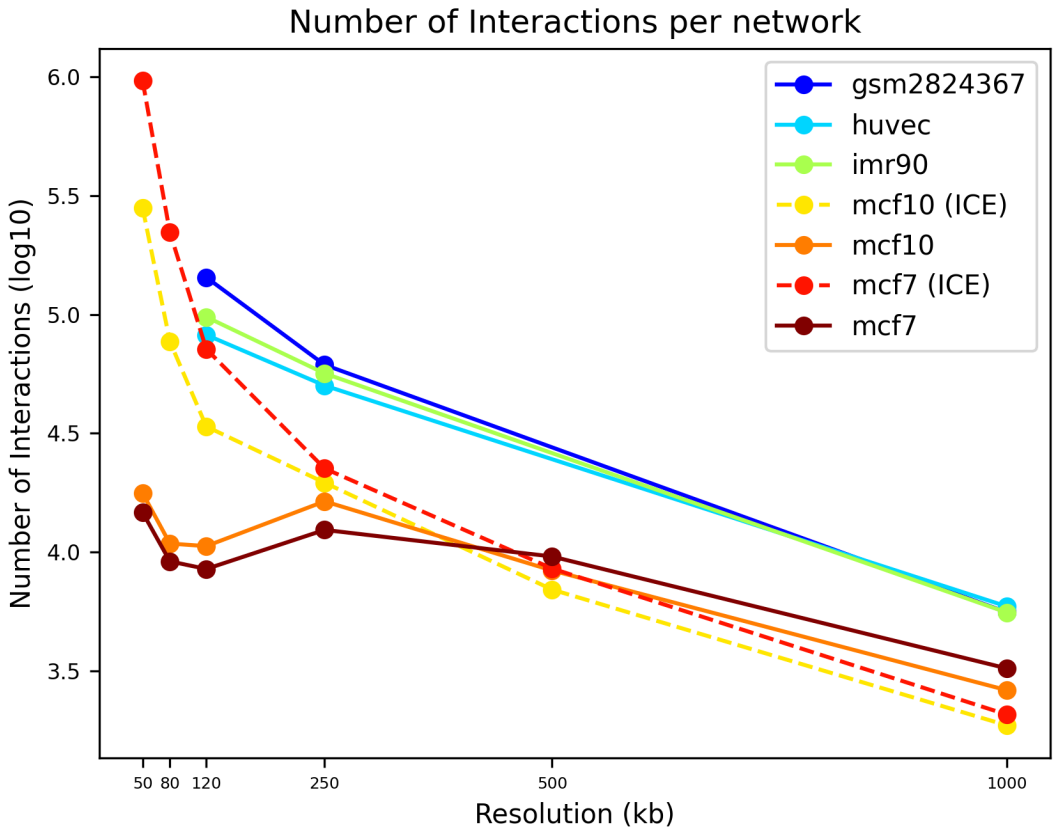


Figure 12: Relationship between number of interactions (edges) in log10 scale and the resolution of networks. ICE balanced matrices from MCF-10A and MCF-7 are in dashed lines, while the solid lines denote data not *normalized by iterative correction before running the pipeline. Again, the difference between ICE normalized data for MCF-10A and MCF-7 is large for high resolution data but smaller for low resolutions (250kb - 1 Mb). IMR-90, HUVEC and HA-c (GSM) interaction counts are very similar across all resolutions examined, and higher than the breast tissue cell lines (MCF-7 and MCF-10A).*

Next the number of connected components per network was determined, where the networks are the most similar low resolutions, as seen in the figure below. A higher number of connected components in the networks at the same resolution means that the networks have less overall connectivity. Comparison across resolutions needs to take the size of the network into account, as a larger network will have more nodes and might thus have more connected components. For 1Mb resolution data, the HA-c (gsm) and HUVEC data has the least number of connected components, while the MCF-10 and MCF-7 has the greatest number of connected components. There is a large discrepancy between the MCF-10A, and MCF-7 cell lines for the data normalized by matrix balancing (ICE) before running the pipeline, versus the raw data at high resolutions.

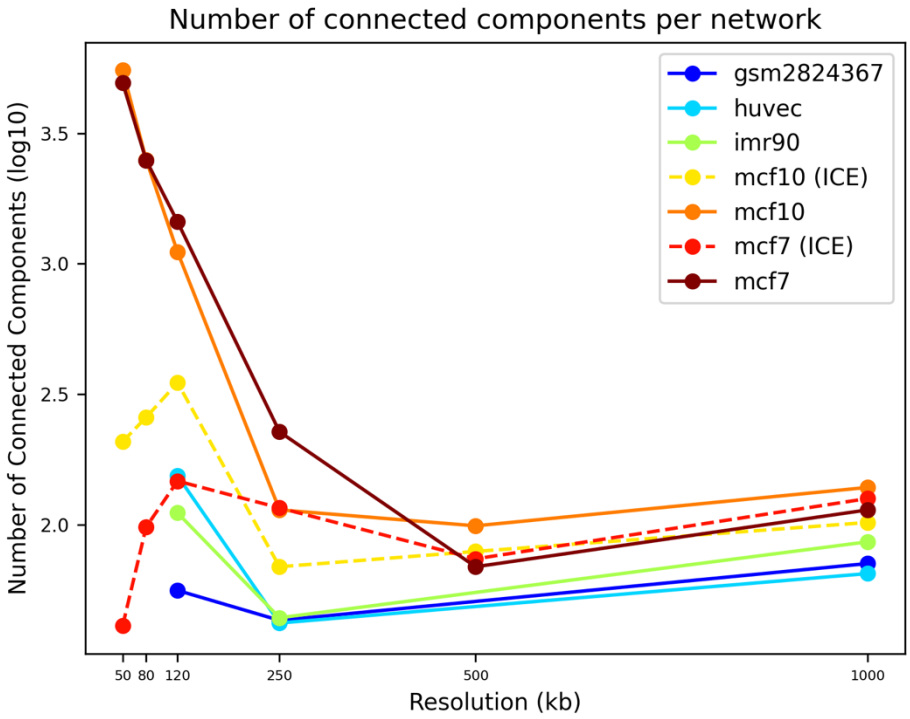


Figure 13: Number of connected components in networks across resolutions. Dashes lines represents the matrix balanced data for MCF-10A (yellow) and MCF-7 (red). Solid lines cell lines whose data was not normalized before running the pipeline. There is a clear divergence between the normalized data and the raw data for low resolutions. As resolution decreases so does the variance in the number of connected components between networks.

From these results the higher resolution data seems to be more prone to bias and is much more computationally expensive to analyze. All subsequent analyses were thus carried out on 1 Mb networks, containing intrachromosomal Hi-C interactions.

4.2 Centrality metrics

To determine if there were systemic differences in centrality between the 1Mb networks, centrality analysis was carried out on each cell line. The degree distribution seen in Figure 14 represents unique degree frequencies per network.

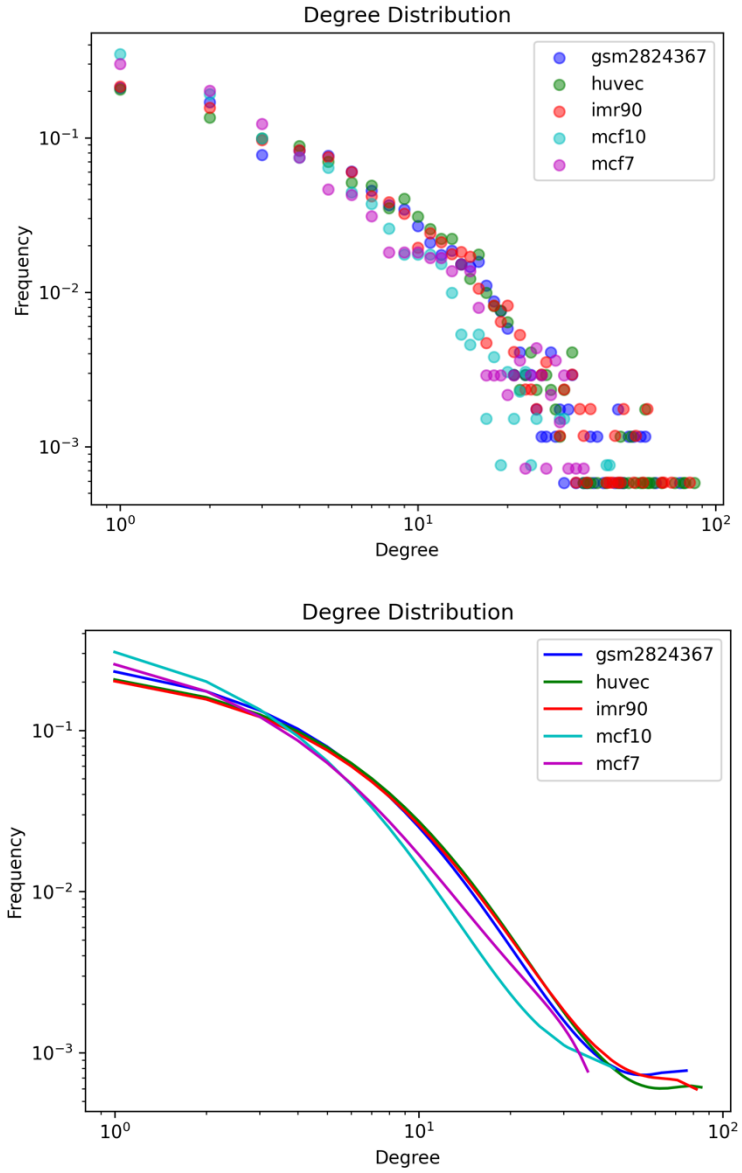


Figure 14: Degree distribution of all cell lines at 1Mb resolution, each data point in the scatter plot (top) is colored by cell line and represents each unique degree value for that cell line. There are no large-scale differences in degree distributions between cell lines (line plot, bottom). MCF-7 and MCF-10A do however have a larger frequency of low degree nodes compared to other cell lines. The highest degree nodes are all in HA-c (gsm), HUVEC and IMR-90 cell lines. The plot approaches a power-law curve, indicating a scale-free network.

There are no overall differences found between networks for in terms of degree. MCF-10A and MCF-7 has the highest frequency of low degree nodes and lowest frequency of high degree nodes. The degree ranges from 1 to 85 across the networks, with the highest degree frequency being 0.32 from MCF-7 and the lowest frequency being 6.3e-4. This frequency represents one node for the IMR-90, HUVEC and HA-c (gsm) networks, which is the reason for the collection of nodes at that frequency count. The MCF-10A and MCF-7 cell lines have a smaller number of nodes at their minimum frequency count, and a larger proportion of nodes of degree equaling one. This indicates that the MCF-7 and MCF-10A networks are less defined by hub nodes and have more spokes, that is nodes with one connection, compared to the other cell lines.
The distribution of the data remains similar between cell lines, and is power-law adjacent, which indicates that the networks are scale-free networks.

Next the closeness distribution of each network was determined and compared, as seen in Figure 15 below. The distribution of closeness scores is similar between networks on larger the whole-networks scale, as is the case with Degree.

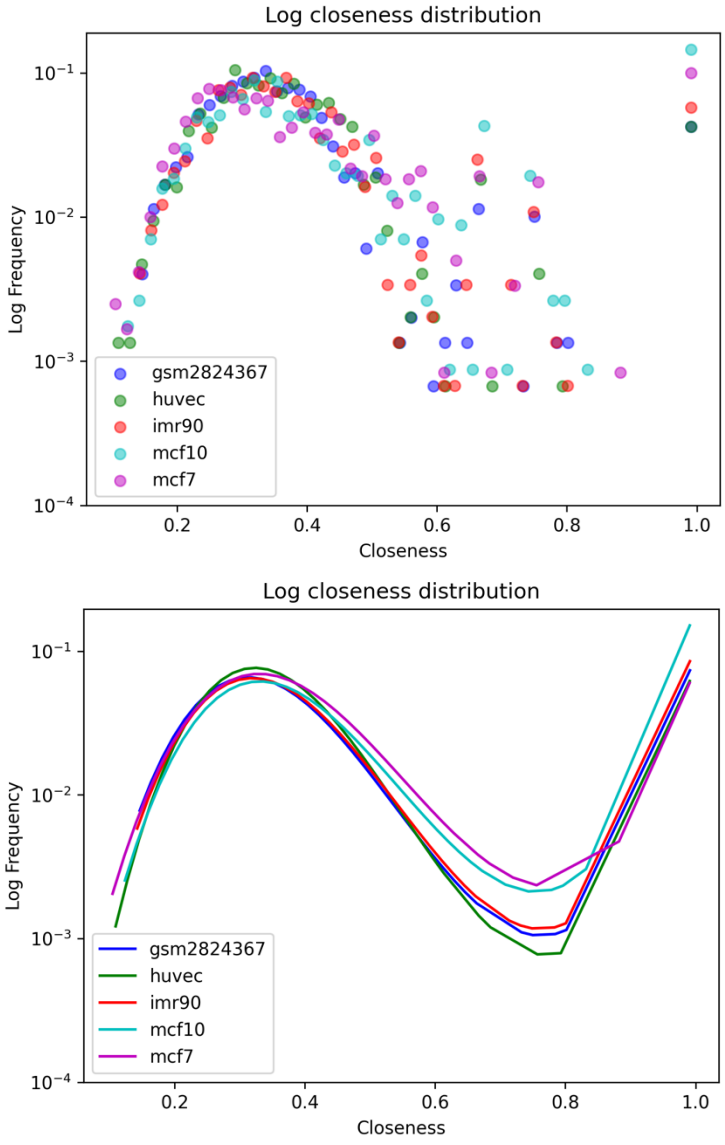


Figure 15*: Distribution of normalized closeness for all cell lines at 1Mb resolution. The distribution is similar (line plot, bottom) for all networks, with MCF-10A and MCF-7 cell lines having slightly more nodes are higher frequencies (magenta and cyan lines). Each data point in the scatterplot (top) is colored by cell line and represents a unique closeness value for that cell line. The four frequencies at closeness equaling 1.0 are the hub nodes found in chromosome 20, which is one connected component with five nodes in all networks but the HA-c (gsm) network. This results in the upwards trend in the slopes, not being reflective of the actual distribution of the closeness values across networks.*

Lastly the distribution of betweenness values was found for each network, yielding very similar distributions between networks from differing cell lines, as indicated by Figure 16 below.

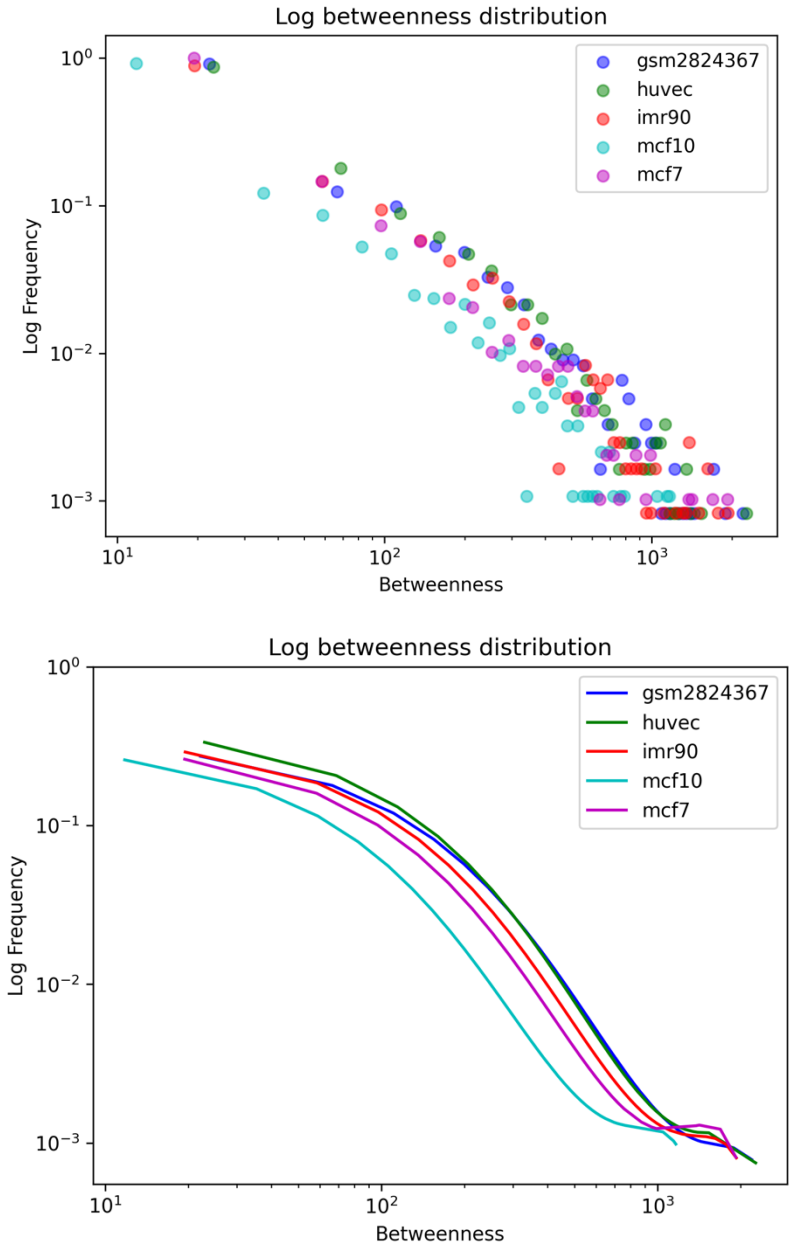
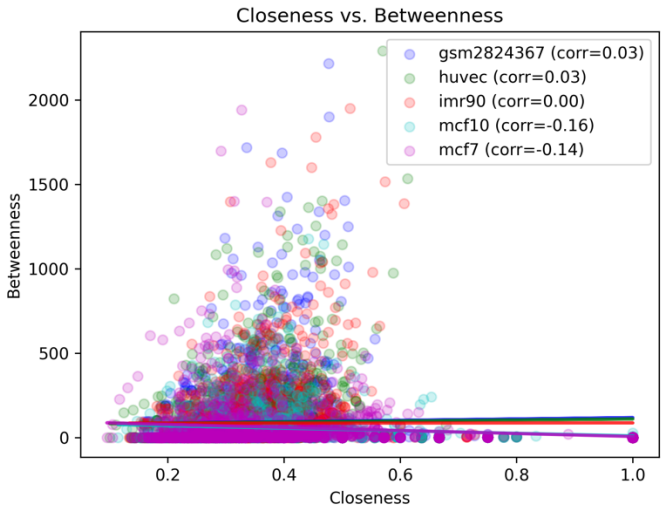
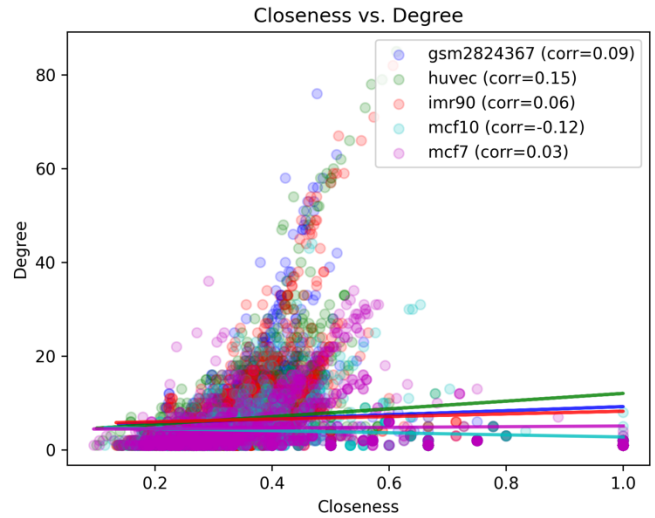


Figure 16: *Betweenness distribution for all cell lines at 1Mb resolution. The overall distribution is the same for all networks as seen with the line plot (bottom), with MCF-7 and MCF-10A having slightly smaller betweenness values than the other cell lines. Each data point in the scatter plot (top) is colored by cell line and represents a unique betweenness value found in that cell line network.*

As seen in the figures above, the overall centrality distributions for the networks do not differ by much. To determine if the centrality metrics calculated correlate at large scale – calculated on the whole networks – correlations plots were made, as seen in Figure 17 below. There was no

correlation between closeness and betweenness nor for closeness versus degree. Comparing degree and betweenness gave a different correlation, ranging from 0.43 in MCF-7 to 0.71 in MCF-10A. Nodes having a high degree are more likely to occupy a larger number of shortest paths between other nodes, explaining why two metrics correlates. This is not the case for closeness versus betweenness and degree, nodes in all networks with high degrees and betweenness values (hub and bottleneck nodes) do not have closeness values above 0.6.





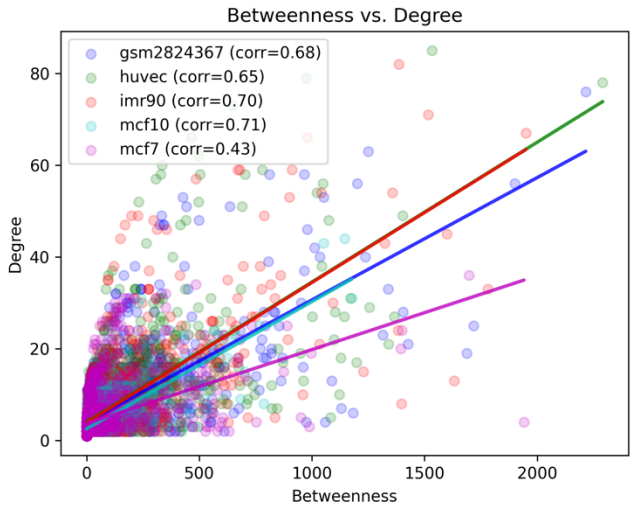


Figure 17: Correlation plots between centrality metrics for all networks at 1Mb resolution. No correlation between betweenness and closeness (top) nor closeness and degree (middle) were found. Degree and betweenness (bottom) do correlate, meaning that high degree nodes are more likely to also have high betweenness. High degree nodes have closeness values peaking around 0.4 – 0.6, while the nodes with the highest closeness have low degrees.

Lastly the Jaccard similarity between chromosomes for all networks was calculated. Here the set are the nodes and the similarity calculated is the proportion of nodes with the same genomic coordinates for each chromosome. Each node in the network spans a genomic region, and each chromosome is a collection of sub-graphs of the larger intra-chromosomal network from each cell line. The chromosomes were matched between cell lines, and all nodes were aligned in terms of genomic position and were compared between networks on a per-chromosome basis. The average proportion of nodes for each network overlapping – that is, being present in all networks – is thus the Jaccard score. This is a similarity score between networks as seen in Figure 18 below, identifying the proportion of nodes present across all networks.

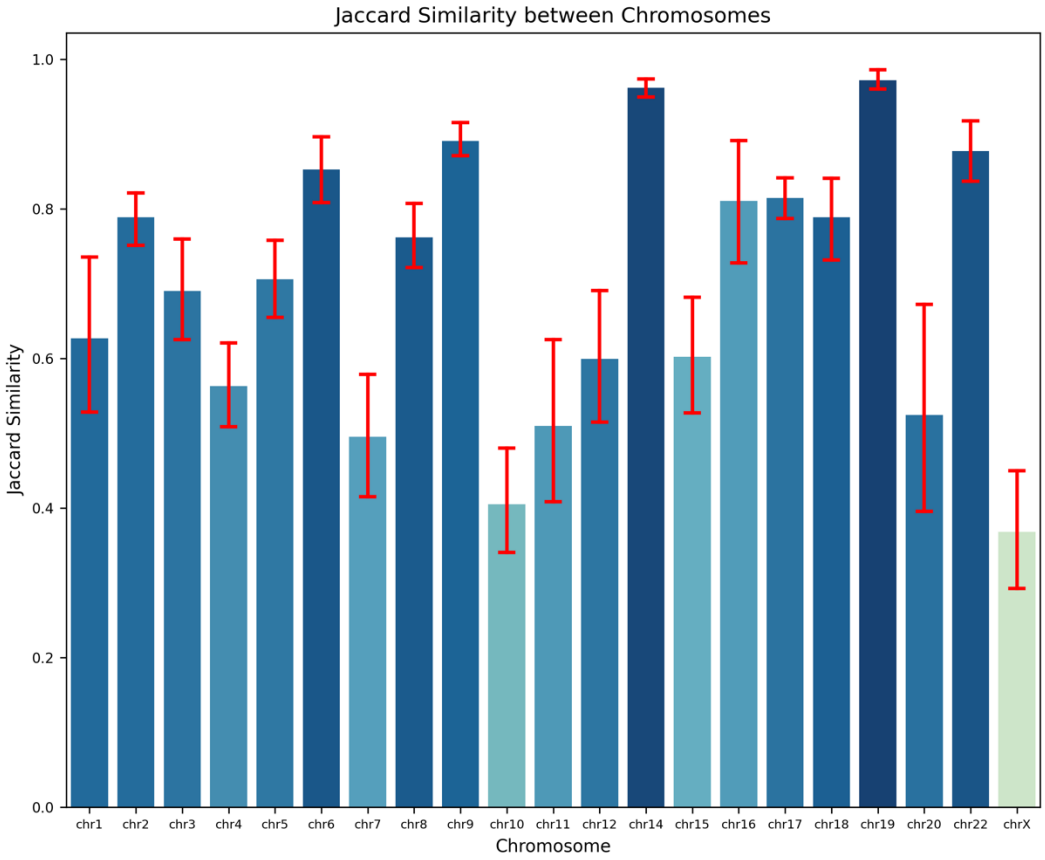


Figure 18: Jaccard similarities between all cell lines at 1Mb resolution, showing the Jaccard score ranging from 0.0 to 1.0, showing the average proportion of nodes at the same genomic location between all networks. Error bars show 95% confidence interval. This is an overall metric for comparing the node locations from different cell lines, enabled by the fact that each node is the exact same size and has a genomic location.

4.3 MCF-10A and MCF-7 comparison

After comparing global network metrics, MCF-10A and MCF-7 networks were selected for further analyses. The proportion of overlapping nodes was determined again and can be seen in Figure 19 on the next page. The overall similarity between MCF-10A and MCF-7 does not indicate large differences in terms of node overlap compared to the node overlap similarities found between all networks.

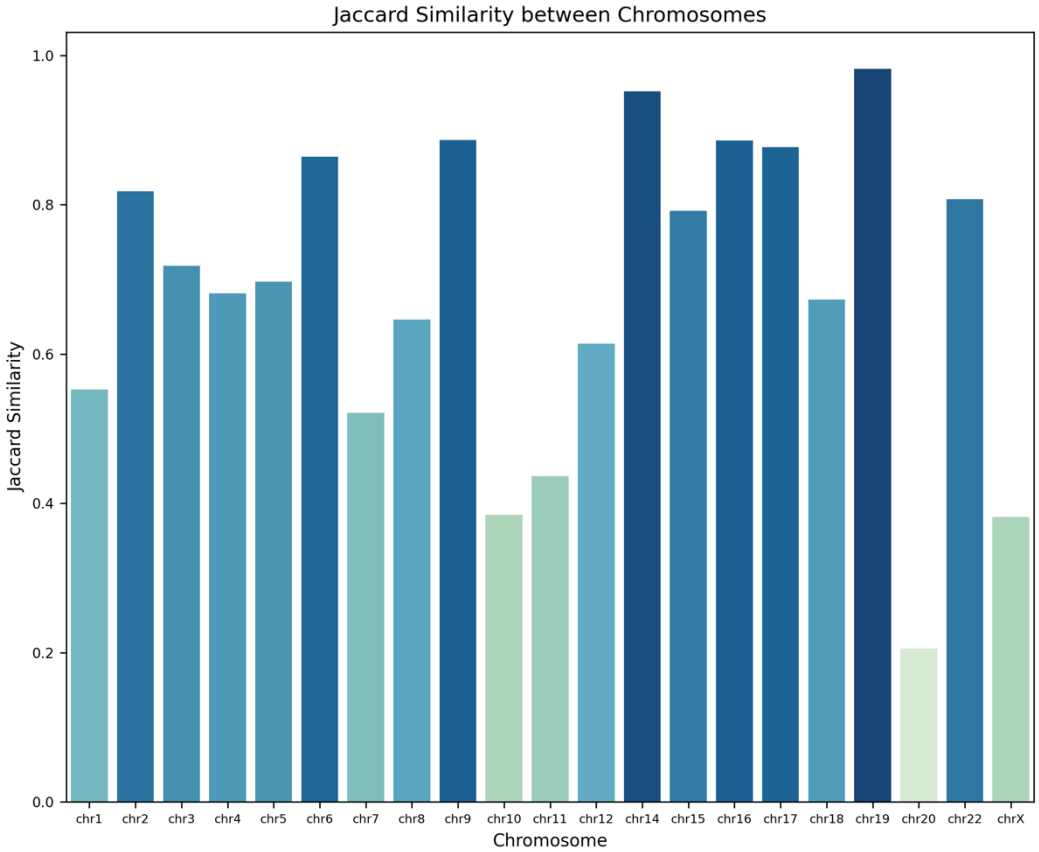


Figure 19: Jaccard similarity between nodes in the MCF-10A and MCF-7 networks on a per chromosome basis. Each node has the same size – 1Mb – and the networks are aligned per node to determine the overlap. The most dissimilar chromosome in terms of node overlap is the small chromosome 20, follow by chromosome X, 10 and 11.

Next the node positioning on the genome was plotted as ideograms, to show the distributions of nodes across the chromosomes, seen in Figures 20 and 21 on the next pages.

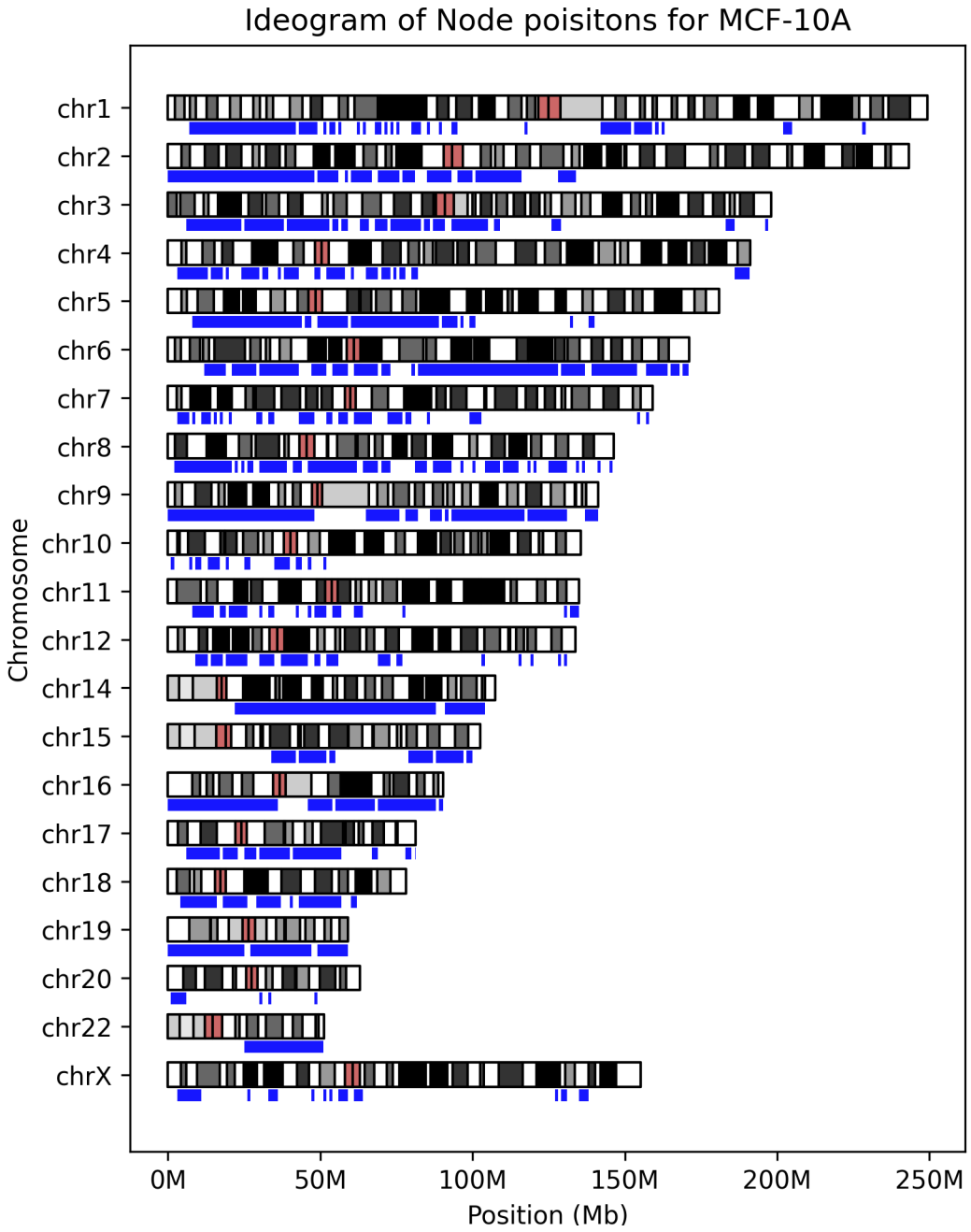


Figure 20: Ideogram of node position of the intrachromosomal 1Mb network of MCF-10A. The colored areas of the chromosome denote the cytogenic banding pattern. The white regions are Giemsa negative, light gray to black regions denotes Giemsa positive regions, where black is the highest staining density. Red regions are centromeres. Nodes are found in both gene poor (Giemsa negative) and gene dense (Giemsa positive) regions.

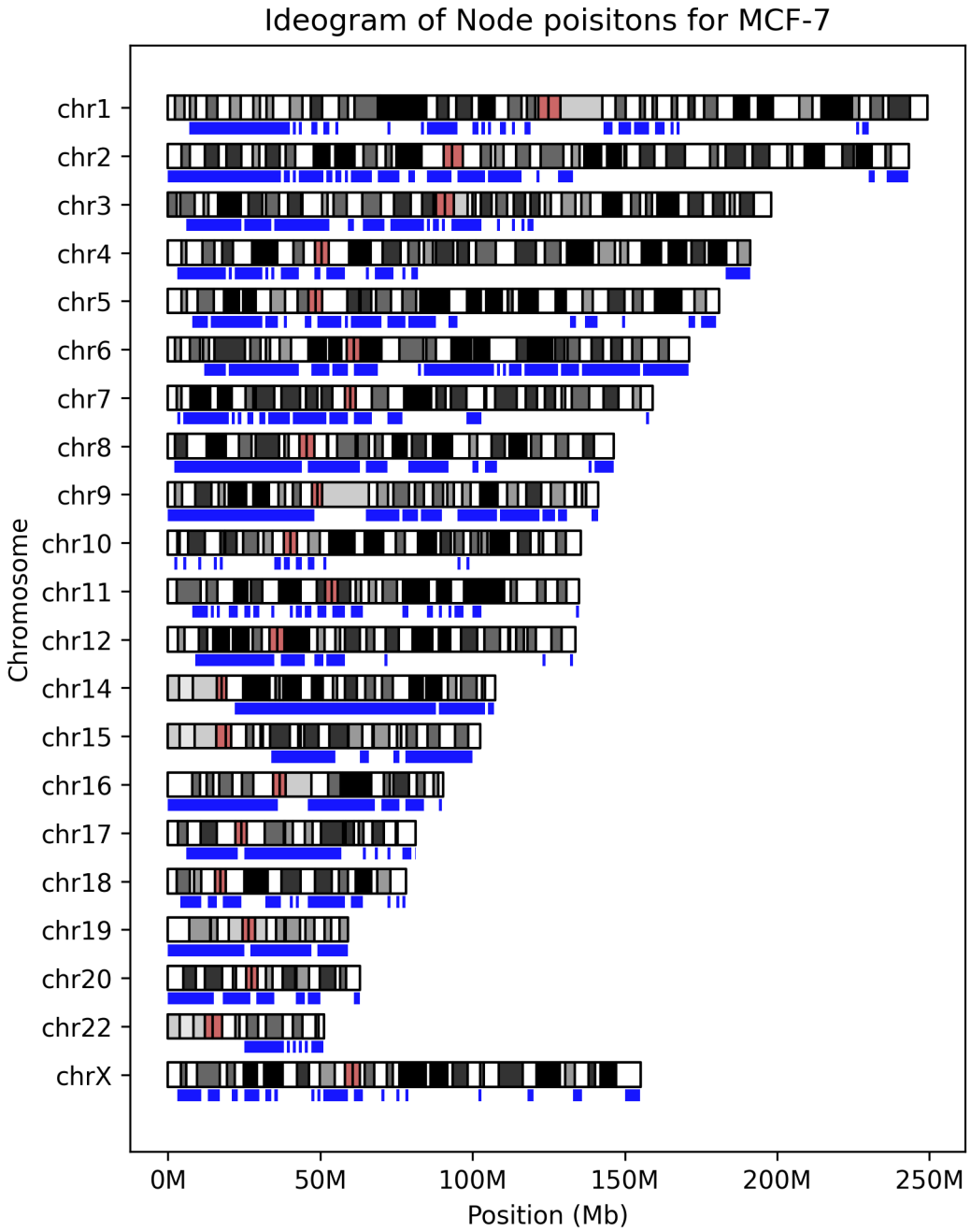


Figure 21: Ideogram of node position of the intrachromosomal 1Mb network of MCF-7. The colored areas of the chromosome denote the cytogenic banding pattern. The white regions are Giemsa negative, light gray to black regions denotes Giemsa positive regions, where black is the highest staining density. Red regions are centromeres. Nodes are found in both gene poor (Giemsa negative) and gene dense (Giemsa positive) regions.

For further centrality analysis on the MCF-10A and MCF-7 networks, it is important to isolate the largest connected components (LCC), such that the results are not skewed due to the disconnected nature of the intrachromosomal graphs. The size of the LCC for each network was calculated, as well as the LCC proportion of total network size, for each chromosome. The size of the LCC is quite similar between the MCF-7 and MCF-10A networks seen in Figure 22.

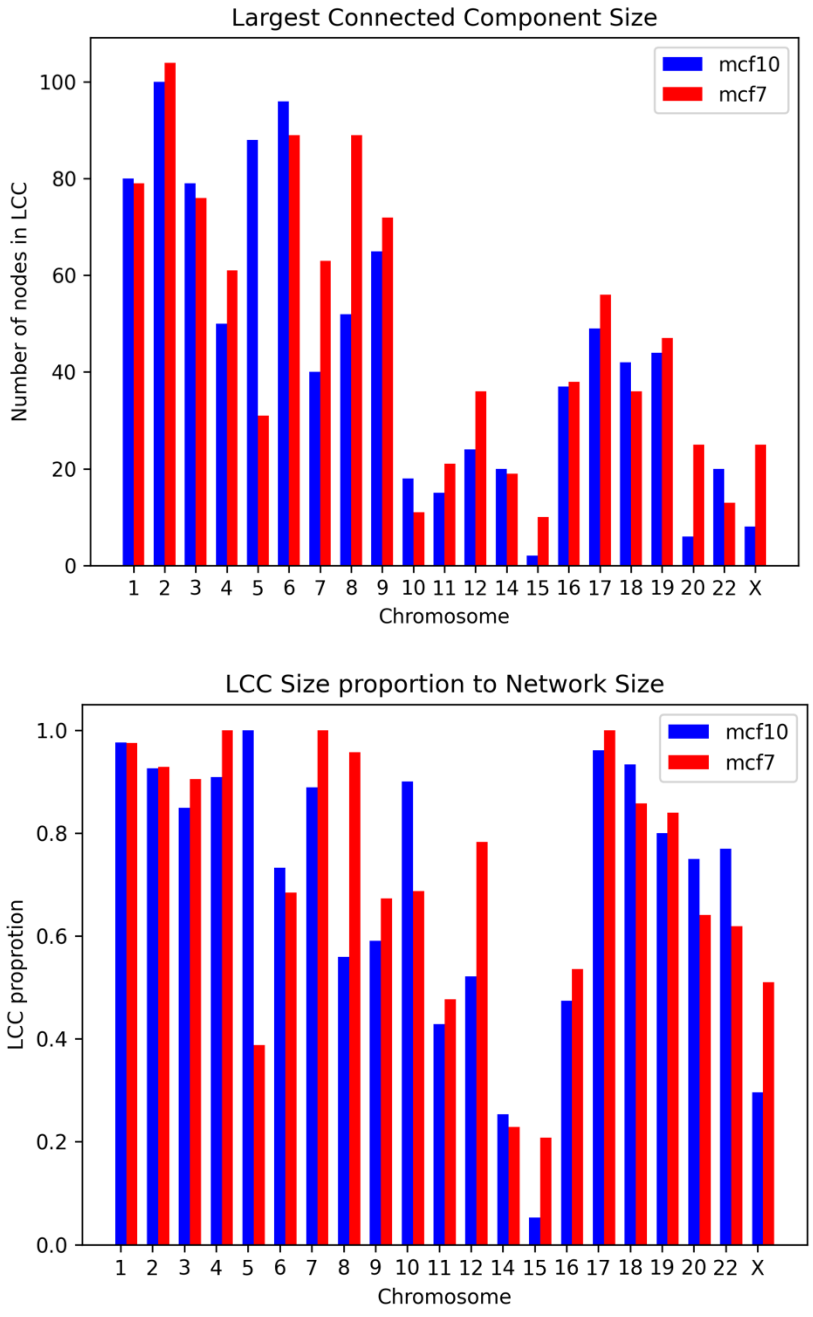


Figure 22: Size of largest connected component (top) and proportion of nodes in the largest connected component (bottom) for chromosomes found in the MCF-10A and MCF-7 networks.

4.3.1 Degree Centrality

After isolating LCCs for each graph in MCF-10 and MCF-7, the centrality metrics was compared. The top 100 ranked nodes in terms of degree were found in both cell lines, and plotted on a chromosome specific level, while also comparing node overlap as seen in Figure 23.

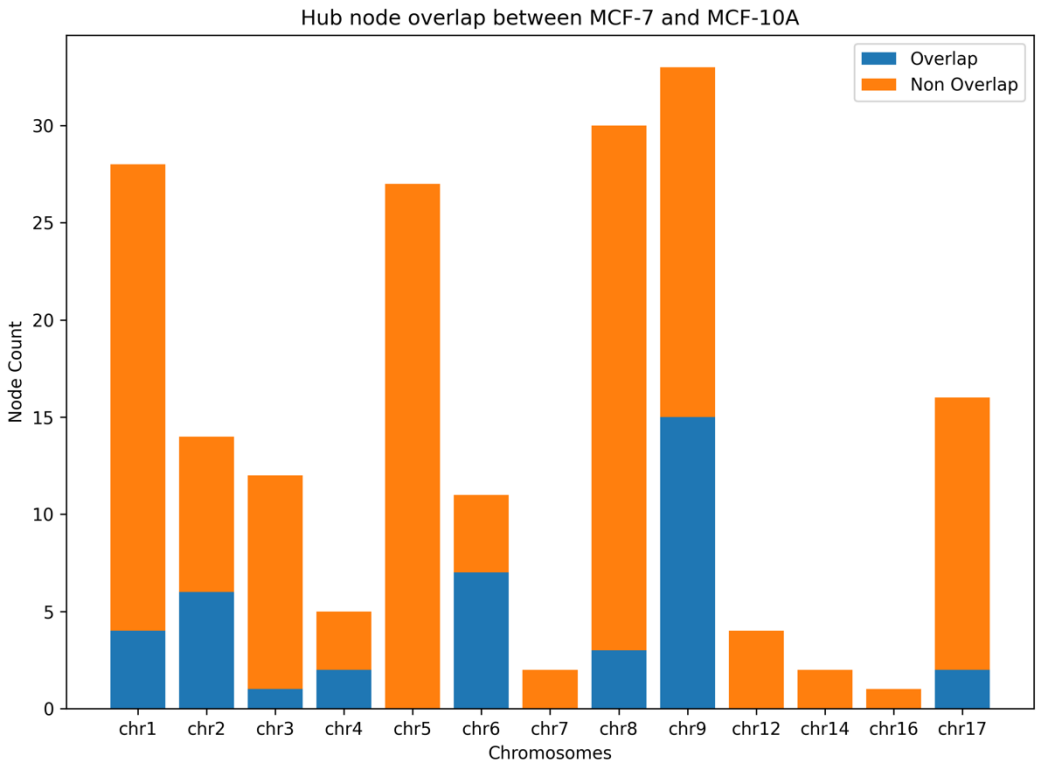


Figure 23: Overlap of top 100 hub nodes (nodes with highest degree) in MCF-7 and MCF-10A cell lines. Non-overlapping nodes are in the majority. Overlap is determined by a nodes genomic coordinate, resulting in chromosome 9, 6, 2 and 1 having the largest overlap in hub nodes. Chromosomes 5, 8, 1 and 9 have the highest amount of non-overlapping nodes.

Chromosome 6 and chromosome 9 were then compared between MCF-10A and MCF-7, where the nodes are colored by degree in their LCC networks. These networks represent the chromosomes with the highest amount of overlapping hub nodes. The same analysis was done for the least similar chromosomes in terms of hub node overlap – chromosome 1 and chromosome 5. These findings are represented as networks in the subsequent four figures (Figure 24 – 27).

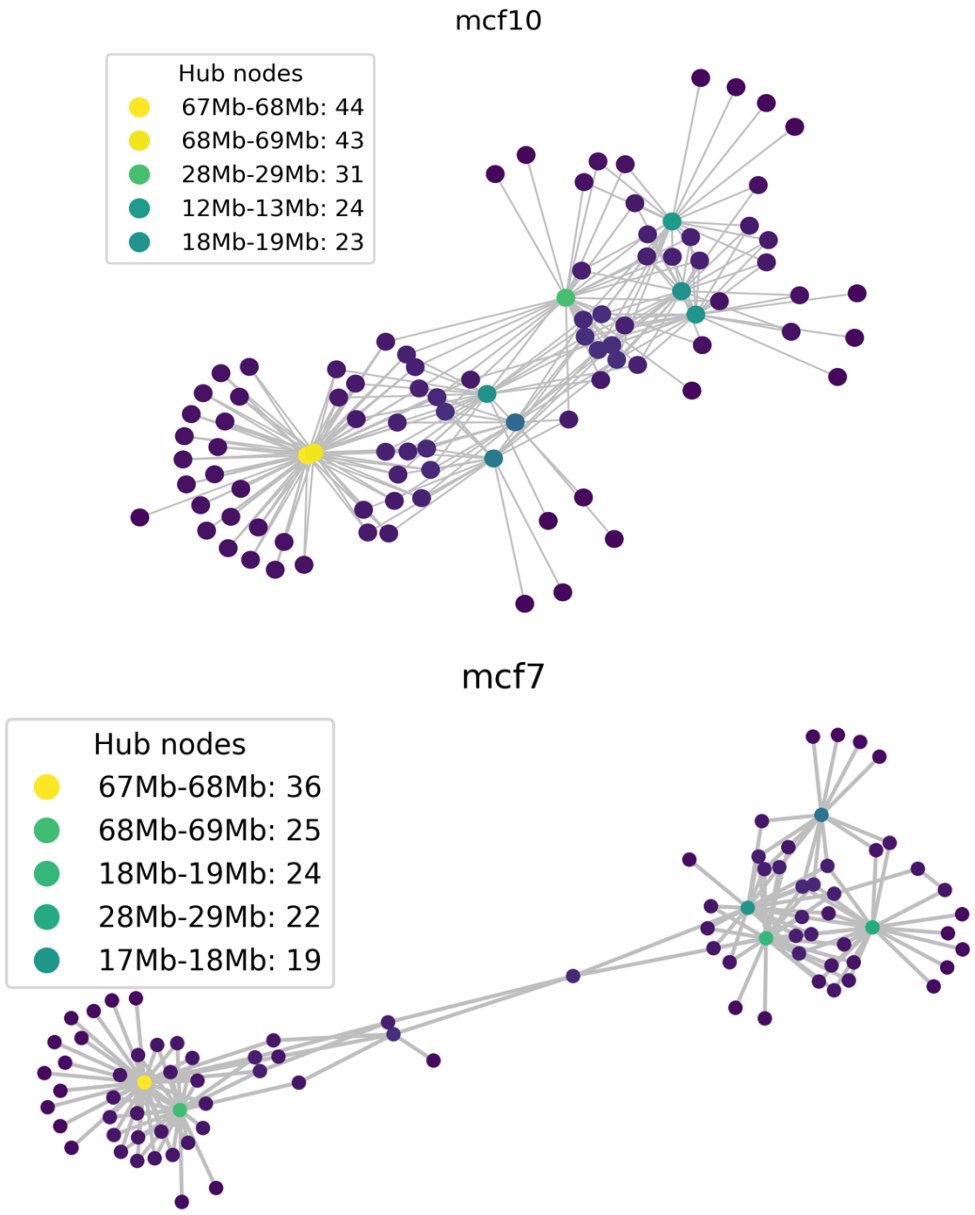


Figure 24: Largest connected component plot of chromosome 6 from MCF-10A and MCF-7. The top 5 nodes hub nodes for each cell line are highlighted in the legend along with their genomic positions. Both networks have many low degree nodes connecting to a few hub nodes, these hub nodes are largely conserved. The MCF-7 network has a rearrangement in its structure – there are fewer interactions between domains, showing a clear division of chromatin structure into two domains separated by one bottleneck node. MCF-10A on the other hand is more interconnected with an additional region of hub nodes.

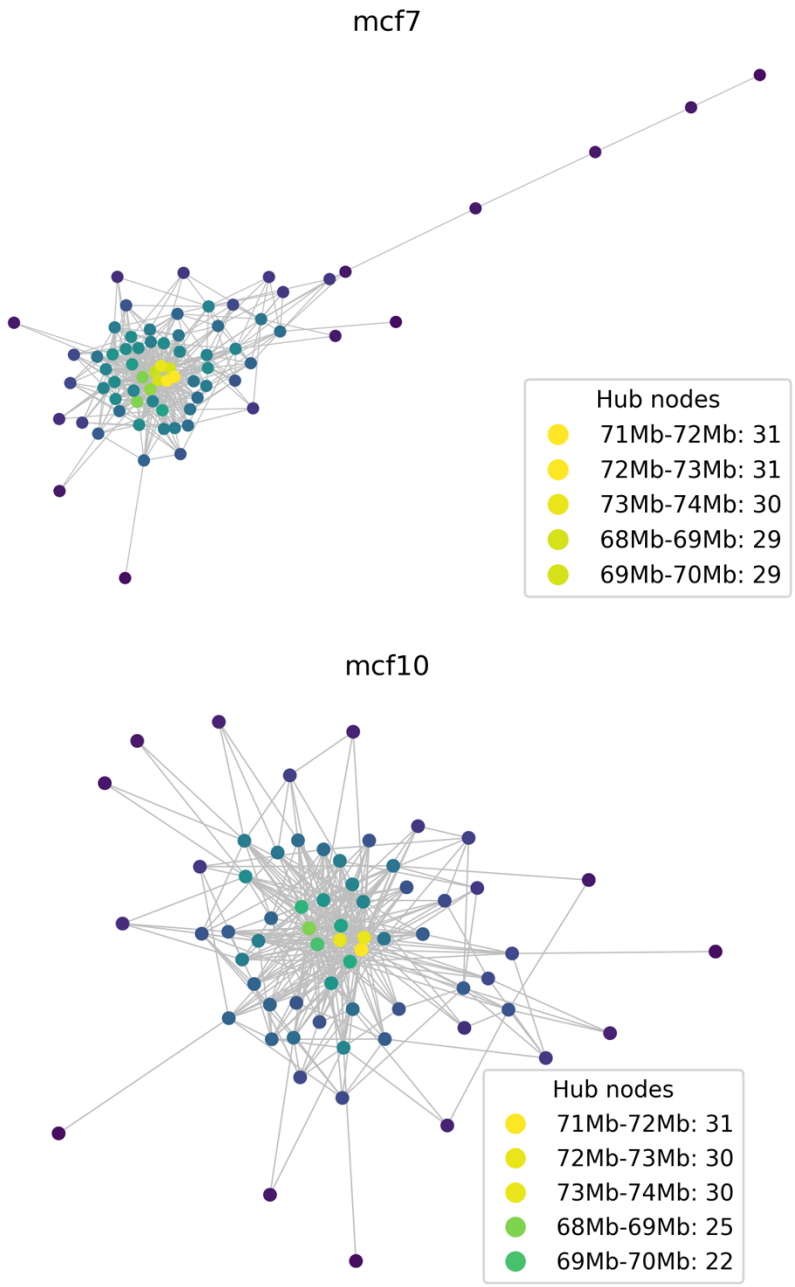


Figure 25: MCF-10A and MCF-7 LCC networks with top hub nodes in chromosome 9. These networks have very similar structures – and all the top hub nodes are conserved between cell lines. The hub nodes are in a single central position in both networks – showing a clear core-periphery structure for both networks.

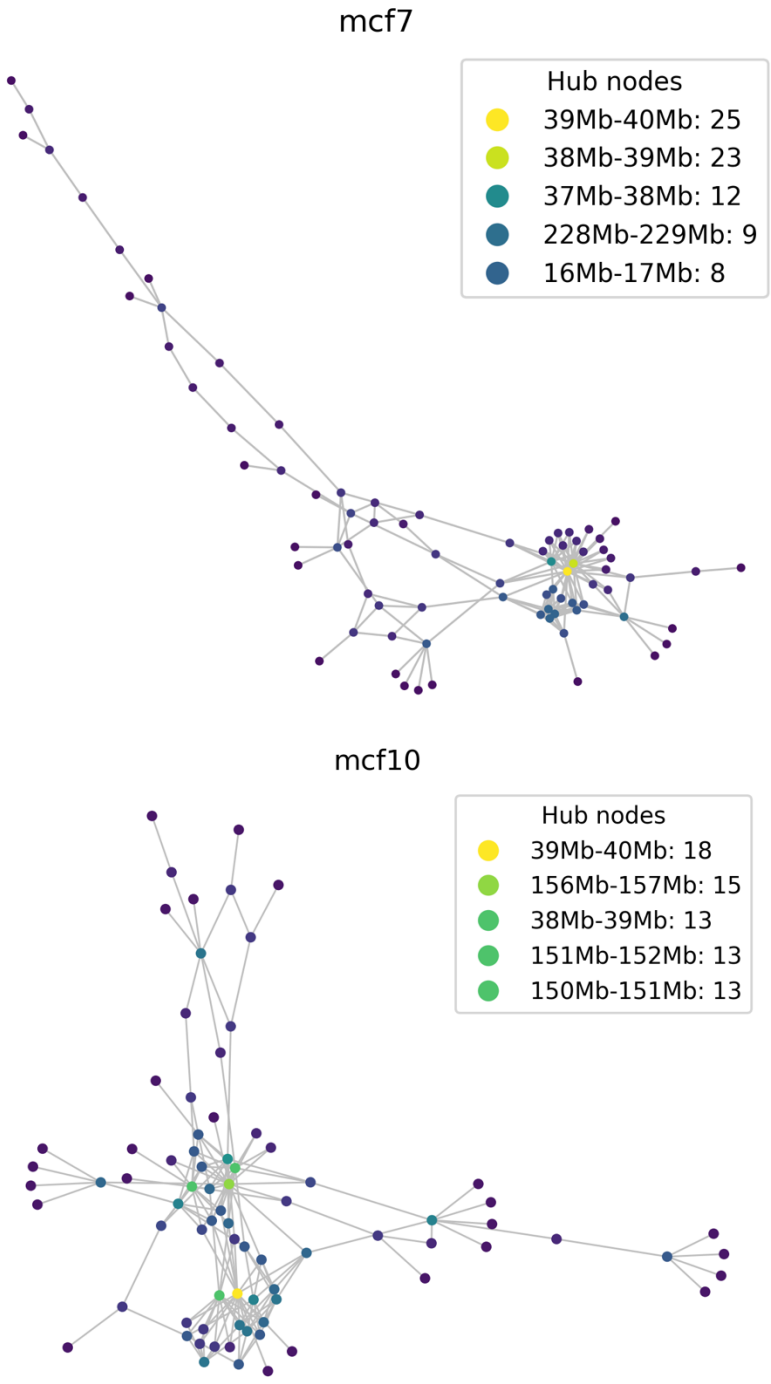


Figure 26: MCF-10A and MCF-7 LCC networks with top hub nodes in chromosome 1. There is a shift in networks structure and hub nodes between MCF-10A to MCF-7: There are only two high degree nodes in the MCF-7 LCC, which are also present in MCF-10A. The 150 Mb nodes are not hub nodes, or not present at all, in the MCF-7 network. The MCF-7 network has one hub region, whereas MCF-10A has two distinct regions of high degree nodes.

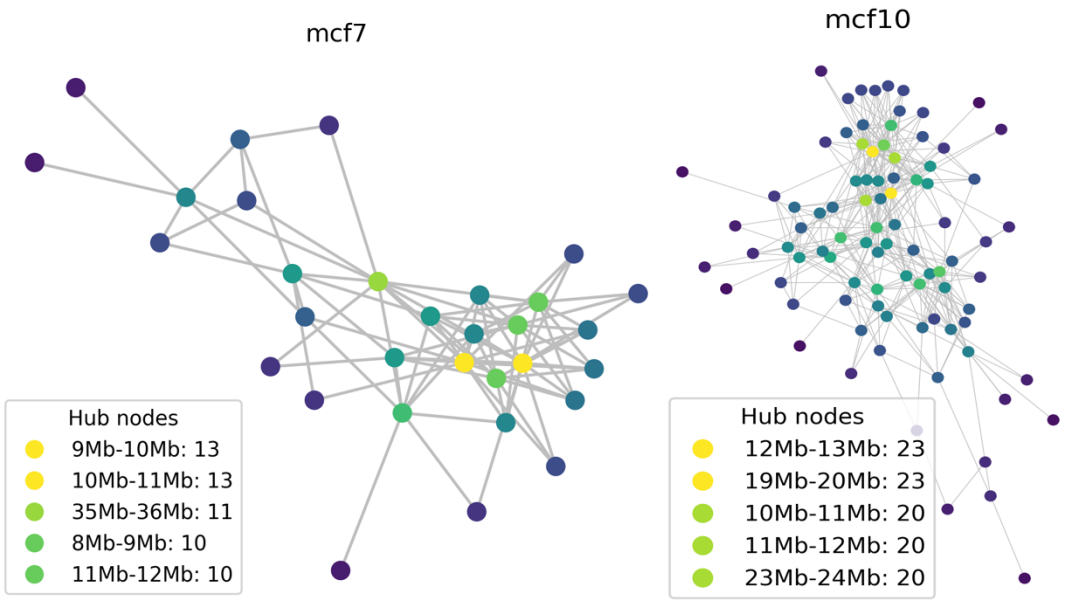


Figure 27: MCF-10A and MCF-7 LCC networks with top hub nodes in chromosome 5. The MCF-10A network is much larger than the MCF7 networks, where the MCF7 network has loss of interactions between hub regions – resulting the smaller LCC for chromosome 5 in MCF-7. Within this LCC several of hub nodes change but are in the chromosomal region as the hub nodes of the corresponding region in MCF-10A.

Overall, there are large differences in network structure and degree centrality between MCF-10A and MCF-7 for several chromosomes. Treating MCF-10A as a baseline or reference network the MCF-7 networks show less connectivity between hub regions and differences in top hub nodes. This feature chromosome specific, as some chromosome networks retain a very similar topology – as seen in chromosome 5 with its core-periphery structure.

4.3.2 Betweenness centrality

The betweenness centrality was calculated for each LCC on each chromosome for both MCF-10A and MCF-7 networks. This was done to determine the overlap of bottleneck nodes between networks. The top 100 betweenness centrality nodes were found for each network and the overlap proportion was calculated, as seen in Figure 28. Chromosome 6 had the higher proportion of overlap in bottleneck nodes, while chromosome 2 had the highest absolute amount followed closely by chromosome 1. These closeness networks are compared between cell lines on the proceeding pages (Figures 28 – 30).

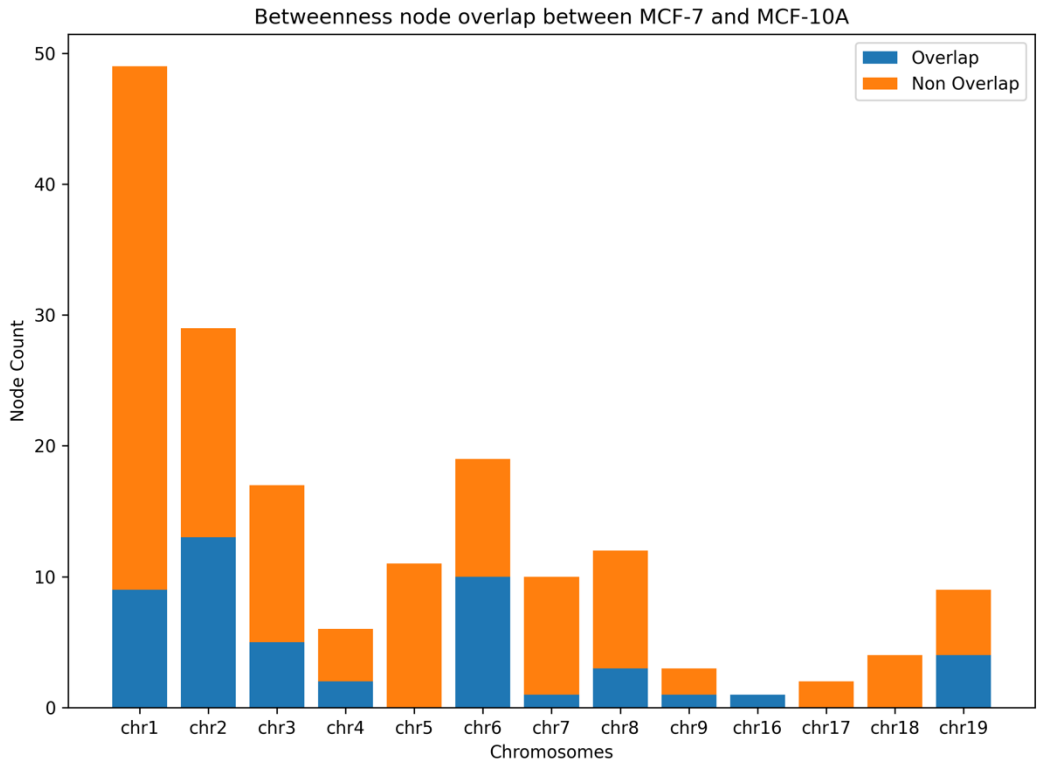


Figure 28: Node overlap for betweenness in LCC top 100 bottleneck nodes for MCF-10 and MCF-7. This pattern of bottleneck node overlap is similar to the hub node overlap for chromosome 6 and chromosome 5. Most nodes with a high betweenness score are found in networks with a larger connected component – chromosome 1, 2 and 6.

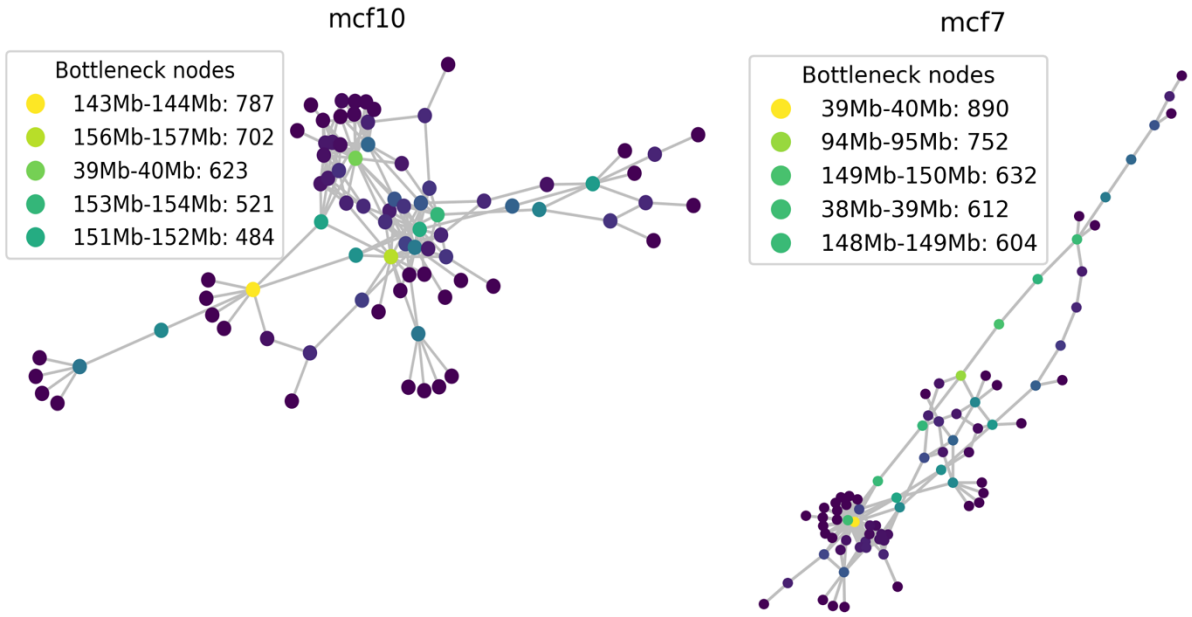


Figure 29: Betweenness centrality network plot for LCC in chromosome 1 between MCF-10A and MCF-7. The LCC networks have overall similar structures, but loss of interactions occurs in MCF-7, with less bifurcating paths between nodes in general. The top bottleneck nodes are different between the networks.

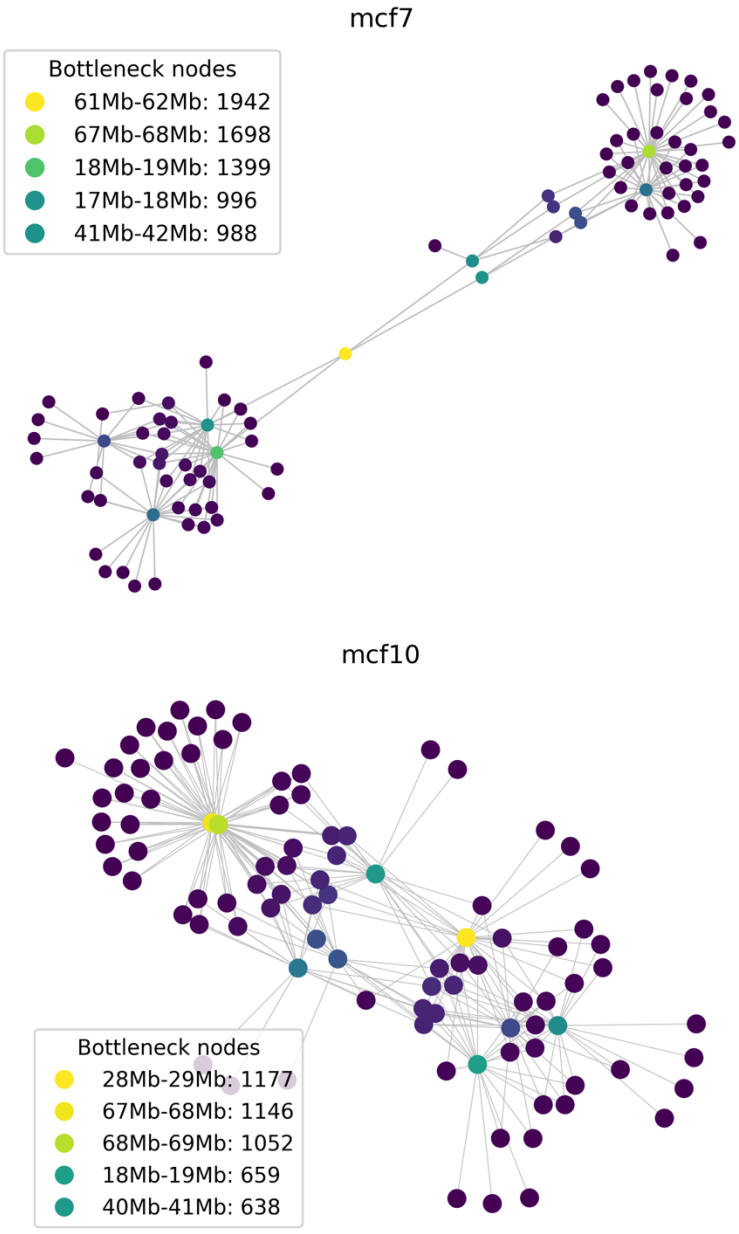


Figure 30: LCC Betweenness network plot for chromosome 6 between MCF-10A and MCF-7. The bottleneck nodes are a result of the different topology seen in MCF-7, where one node connected two domains. This pattern is not seen in MCF-10A, where several paths through multiple bottleneck nodes exists between the different domains of the network.

4.3.3 Closeness Centrality

A per chromosome analysis of closeness centrality was done similarly to betweenness and degree centralities. The top 100 highest closeness nodes for each network and their overlap are seen in Figure 31. High closeness nodes have a different overall overlap pattern than the one seen in the betweenness and degree centralities, where chromosomes 1 and 5 had the most hub and bottleneck node overlap. In terms of closeness, the largest overlaps are at chromosomes 1, 9 and 15. Chromosome 15 in both MCF-10A and MCF-7 has a small LCC. Chromosome 1 and 8 was compared as they contain larger overlaps and non-overlapping regions while having a larger LCC.

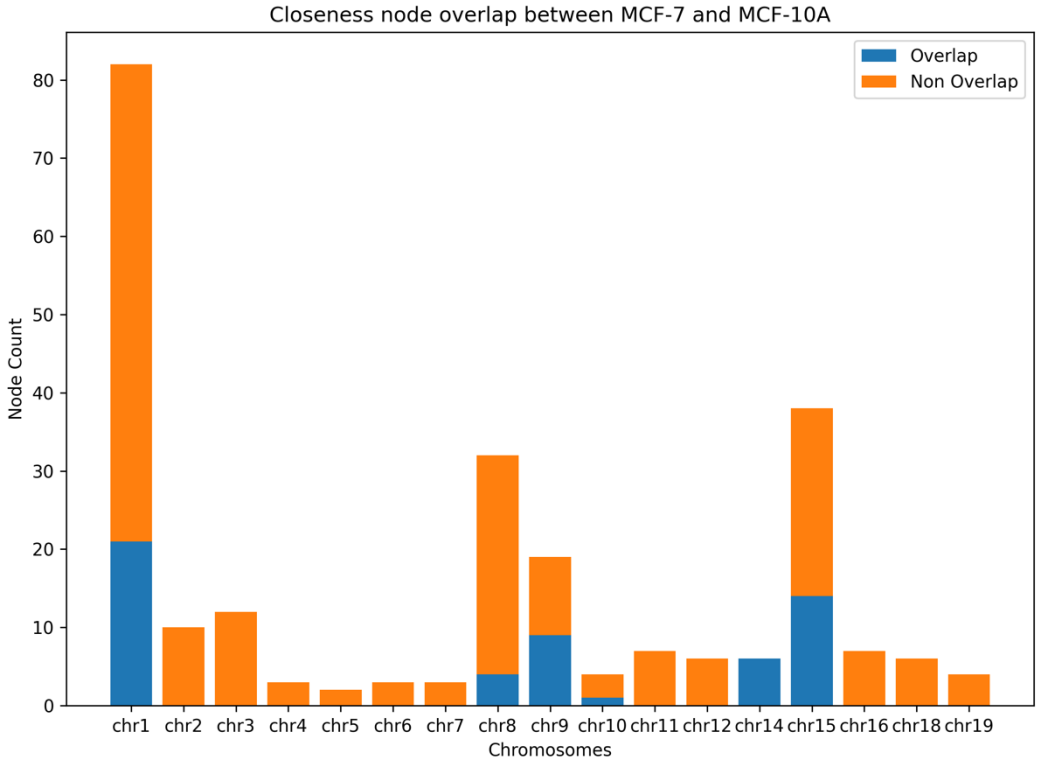


Figure 31: Closeness overlap in top 100 ranked nodes for closeness between MCF-10A and MCF-7. There are four distinct chromosomes, chromosomes 1, 8, 9, and 15, which all contain a higher proportion of the overlapping closeness nodes, indicating that these chromosomes are more similar in terms of closeness.

Closeness ranked networks were generated and visualized and are summarized in figures below in Figure 32 and Figure 33. Chromosome 1 and chromosome 8 was compared between the MCF-10A and MCF-7 cell lines, and the closeness nodes with the highest values are in the legends.

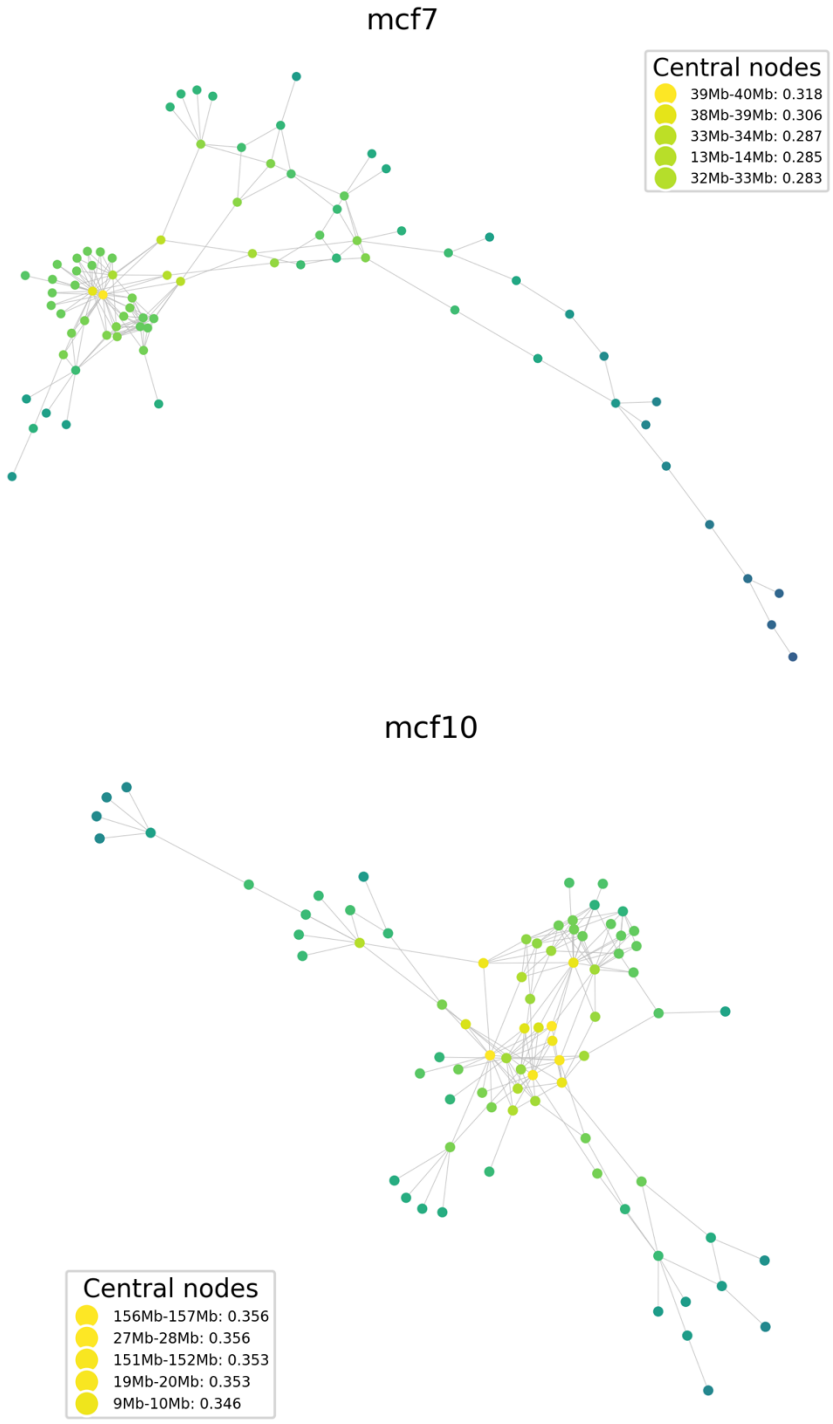


Figure 32: Closeness centrality LCC networks for chromosome 1 in MCF-10A and MCF-7 cell lines. There is no overlap in the top closeness (central) nodes. MCF-7 has lost the central region of interactions between close nodes – reflecting a change in interaction patterns between the networks.

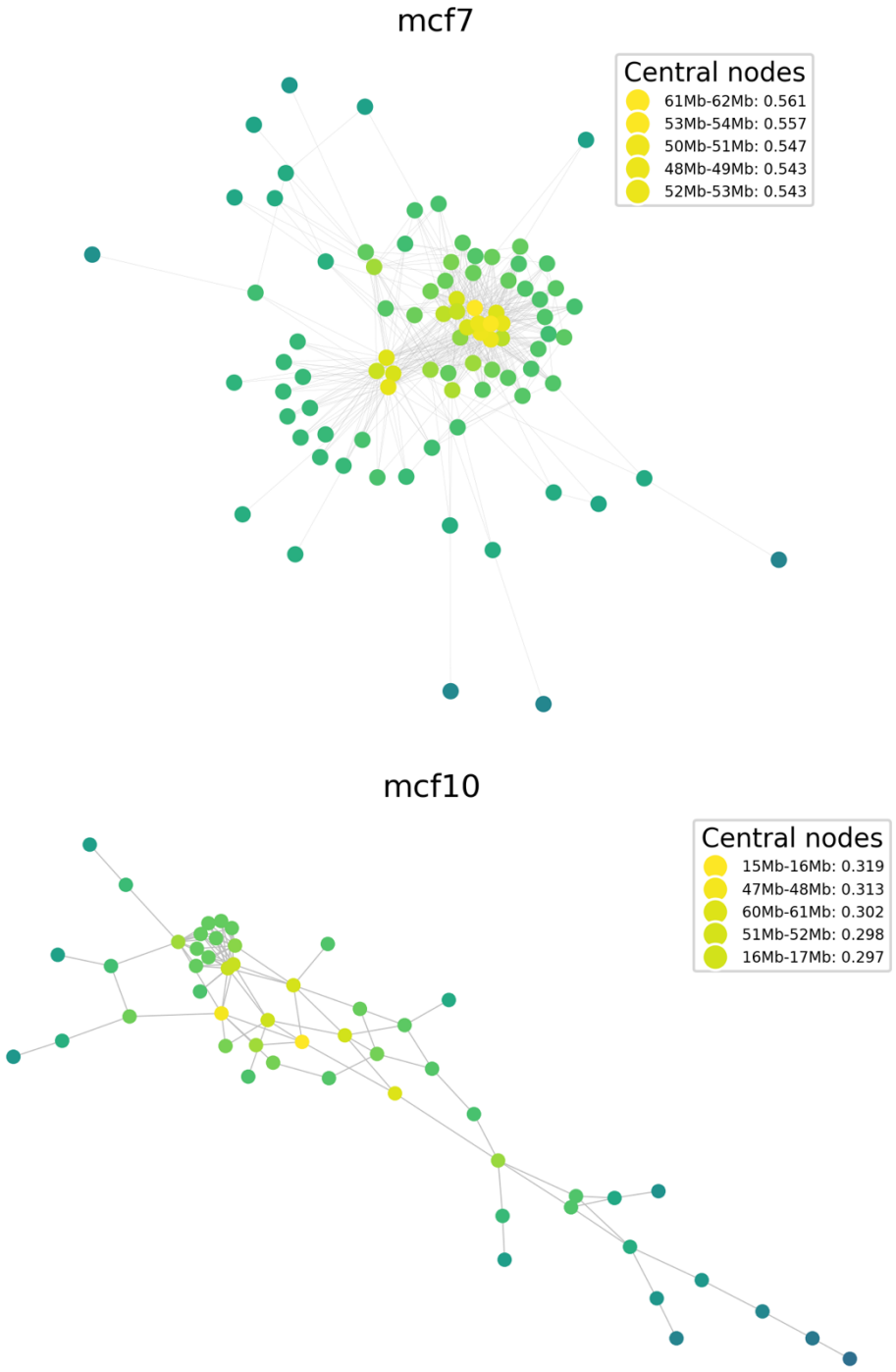


Figure 33: Closeness network for chromosome 8 in MCF-10A and MCF-7 networks. There is only one direct overlap in the top hub nodes – but the overall region of highest closeness is the same. The network structures are quite different, as the MCF-7 LCC is organized into a core-periphery network with two central regions of high closeness. MCF-10A on the other hand has a less distinct core-periphery organization with fewer spokes (low degree nodes).

4.3.4 Community detection in MCF-10A & MCF-7

After determining centralities of MCF-10A and MCF-7 cell lines, and examining networks for nodes with high centralities, the topology of the networks was explored by community detection. At the full genome level, the intrachromosomal MCF-10A and MCF-7 networks have similar overall topologies as determined by fast greedy community detection. As seen in Figure 34.

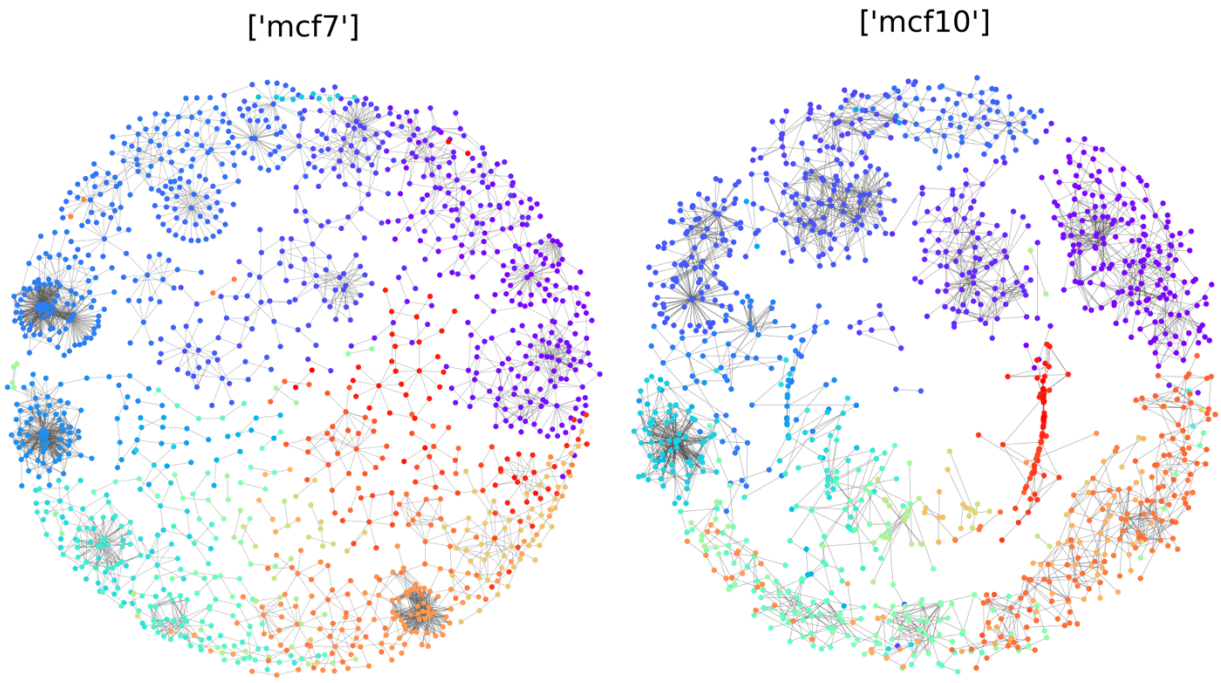


Figure 34: Full genome community detection using fast greedy for MCF-10A and MCF-7. The networks have a highly similar modularity: MCF-10A modularity = 0.909 and MCF-7 modularity = 0.914. On full genome level there are no systematic differences.

To determine communities and their differences on a chromosomal level, each network was split into separate chromosomes, and the communities was determined for each chromosomal network with the fast greedy algorithm. The number of communities are highly similar for each chromosome, except for chromosome 8 which has a much higher community count in MCF-10A compared to MCF-7.

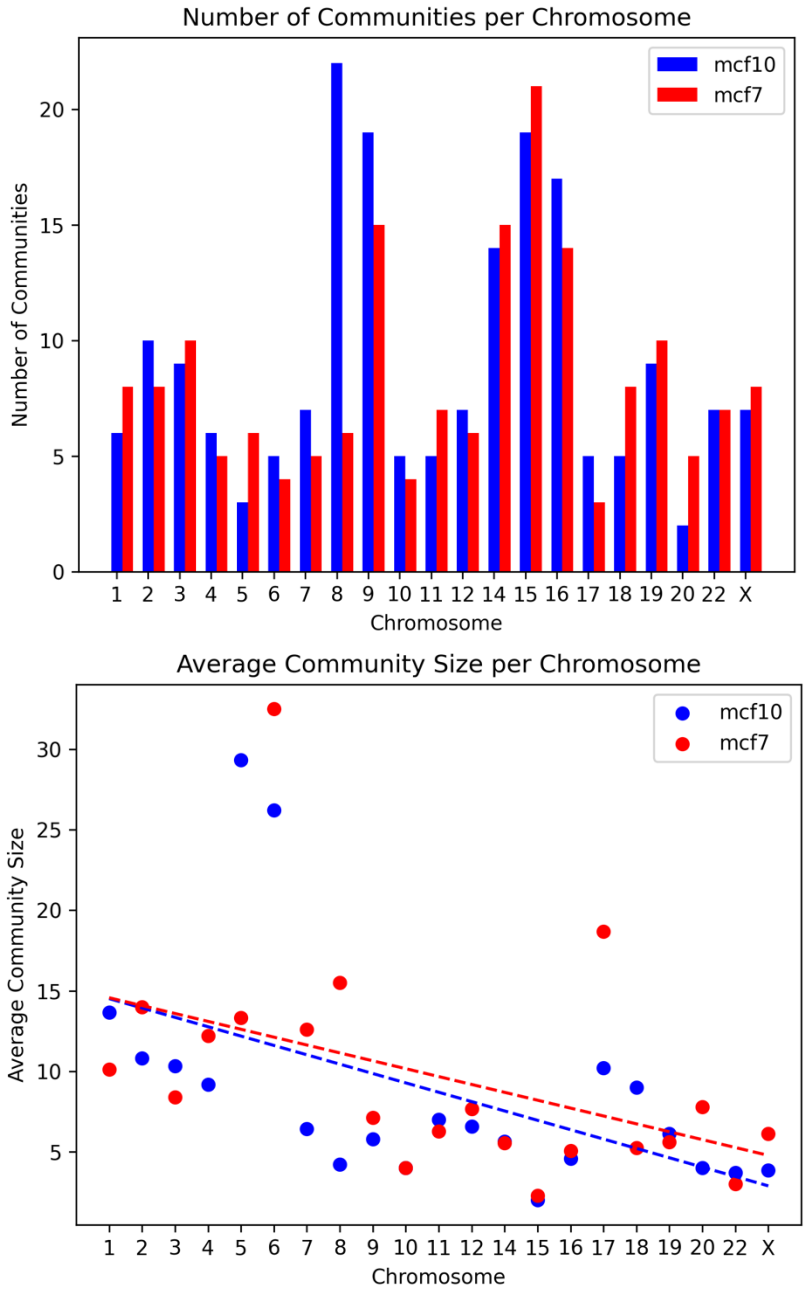


Figure 35: Number of communities found by the fast greedy algorithm for MCF-10A and MCF-7 intrachromosomal networks (top), and the average community size for each chromosome in MCF-10A and MCF-7 networks (bottom). The number of communities between networks are similar for most chromosomes – the largest deviation is found in chromosome 8. The average community size has large deviation between networks for chromosomes 5, 8, 17 while the average community size is more similar between other chromosomes. There is a slight decrease in community size as network size decreases due to chromosome length.

To compare communities with networks of varying community size and different node locations, the NMI fuzzy metric was used to compare communities across networks on a chromosome

specific level, seen in Figure 36. The NMI score ranges from 0.0 to 1.0 and increases for small chromosomes, except for chromosome 17 and 20. The larger chromosomes (chr1-15) have a more dissimilar community structure compared to the small chromosomes, with some exceptions.

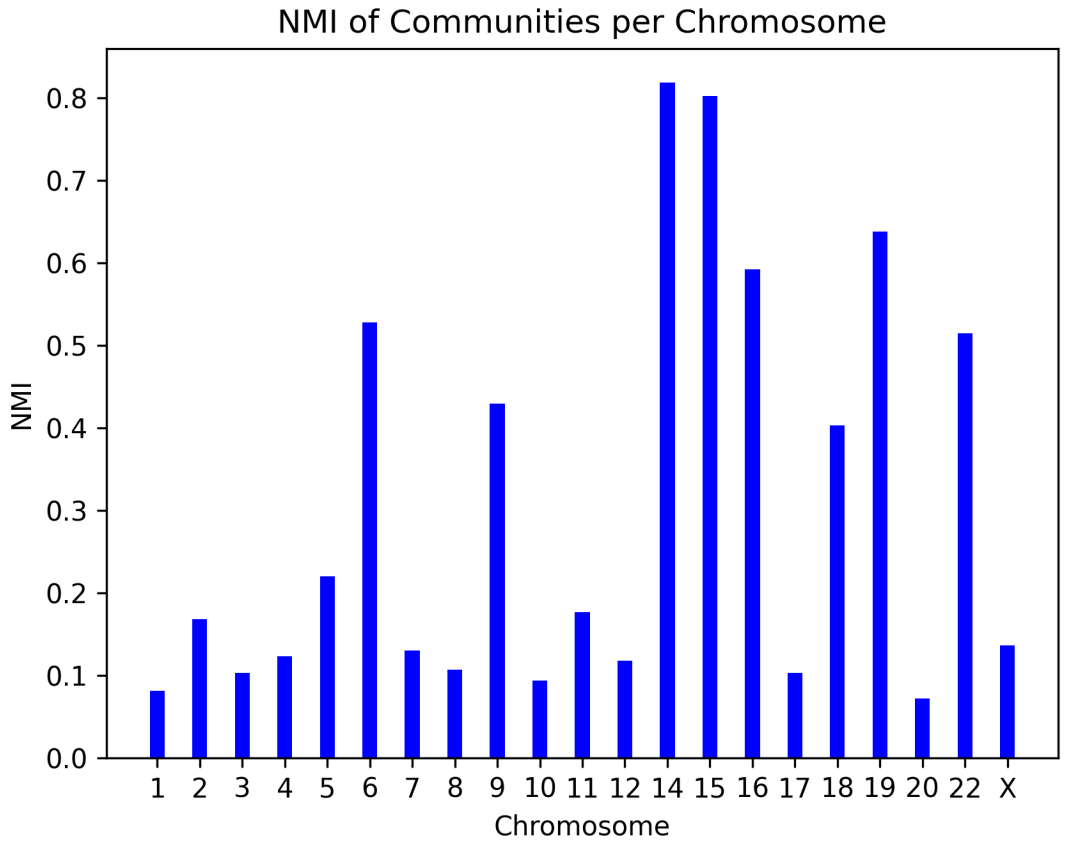


Figure 36: NMI of communities per chromosome for MCF-10A and MCF-7 networks. Communities are first detected by fast greedy modularization, then each community is compared per chromosome. MCF-10A is treated as the reference the MCF-7 communities are compared to. A higher score means the community structure is likely more similar between the networks, as the NMI in this case is a metric of how much information is gained about the community structure of MCF-7 from knowing the community structure in MCF-10A.

Chromosome 1 is the largest chromosome and is found to be the most dissimilar between MCF10-A and MCF-7. These communities for chromosome 1 between cell lines are seen in Figure 37.

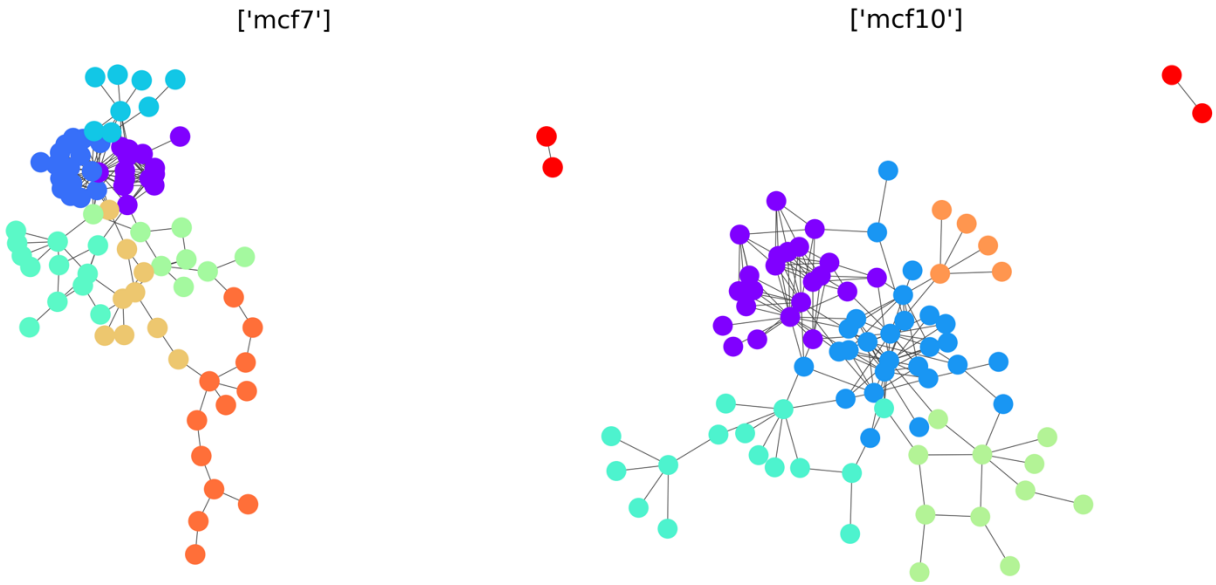


Figure 37: Chromosome 1 with fast greedy community structure, 8 communities in MCF-7 and 6 in MCF-10A. There are two connected components in each cell line, but their structure differs as well as the community assignments. These community structures were determined to be the most different by the NMI metric.

Due to the nested structure of communities and the nature of the intrachromosomal networks, some communities within connected components might not be detected. Calculating NMI once more but for each chromosome with LCC only yielded highly similar result for networks with fewer connected components. For networks with many connected components, the metric changes drastically when isolating the LCC before calculating NMI.

Figure 38 shows an example of the LCC isolation, in chromosome 14 for MCF-10A and MCF-7, which was determined to be the most similar chromosomes in terms of community structure by the NMI metric.

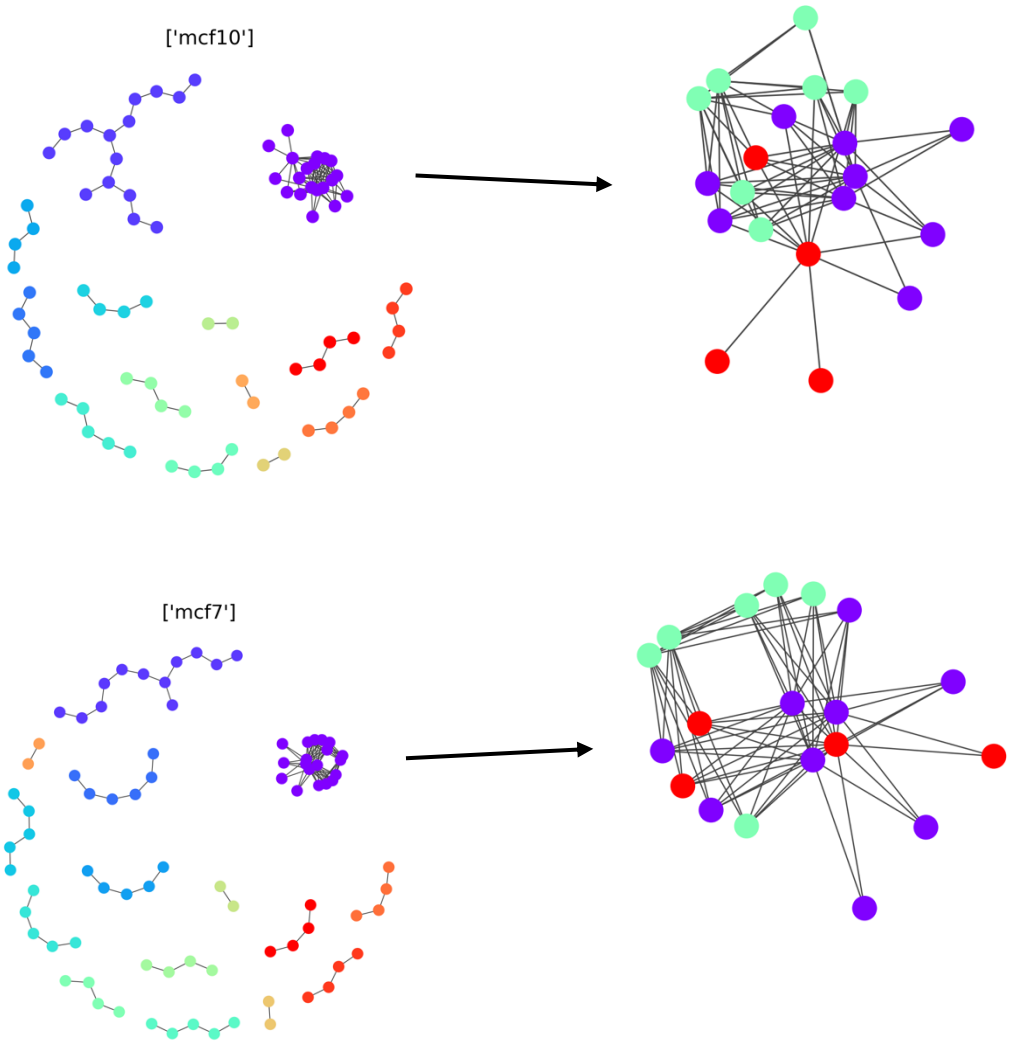


Figure 38: MCF-10A and MCF-7 chromosome 14 community detection by fast greedy. This chromosome was the most similar regarding NMI of community structure between MCF-10A (left) and MCF7 (right). Each component is assigned to a separate community (top left, bottom left). When isolating the largest connected component from the whole network (purple, top left and bottom left) and determining communities for this largest connected component (top right and bottom right), several smaller communities are detected.

Isolation of the LCC from chromosome 14 in MCF-10A and MCF-7 to detect was done and compared with NMI once more. This changed the NMI score from 0.81 to 0.14, indicating that the relationship between community similarity is highly dependent on the number of LCC present in the networks being compared. The same was done for chromosome 1, the most dissimilar chromosome the according to the NMI metric, the score was unchanged (0.9).

Chromosome 1 has two connected components, further indicating that connected component numbers influence the NMI score significantly. The analysis was done again, but only for the LCCs of MCF10-A and MCF-7.

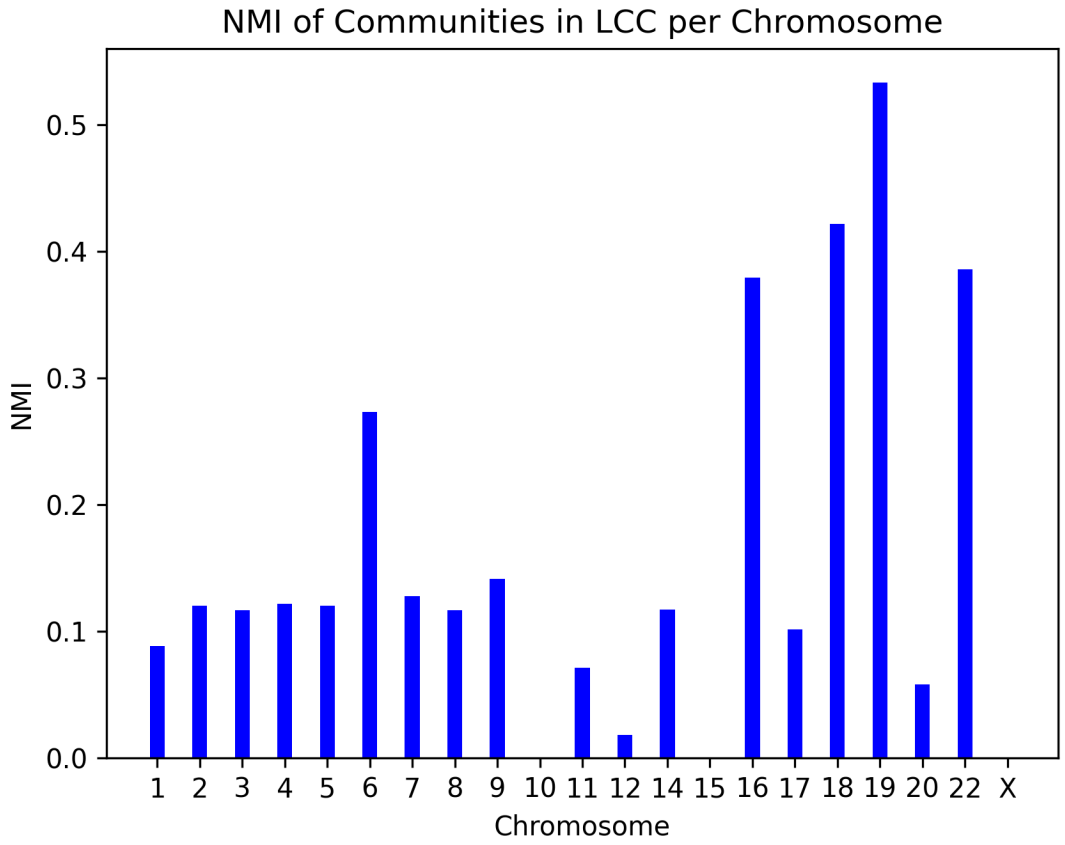


Figure 39: NMI of communities on the LCC in each chromosome for MCF-10A and MCF-7 networks. Communities are detected by fast greedy modularization, then each LCC on each chromosome is compared. MCF-10A is treated as the reference the MCF-7 communities are compared to. A higher score means the community structure in LCCs is more similar between the datasets. NMI in this case measures how much information is gained about the community structure of the LCC in each chromosome on MCF-7 from knowing the community structure of the corresponding LCC per chromosome in MCF-10A.

When selecting the LCC on each chromosome the NMI scores change more for chromosomes with different numbers of LCCs and for chromosomes with many smaller LCCs. From this analysis chromosomes 16, 18, 19 and 22 have the most similar community structures when compared by NMI. For the full chromosome comparison this was not the case, in this analysis chromosomes 14 and 15 were the most similar.

4.3.5 Chromosomal compartments in MCF-10A and MCF-7

The final part of the network analysis and comparison between MCF-10A and MCF-7 was compartment calling by ICE (iterative correction and eigenvalue decomposition). Compartments were called at 1Mb resolution using the balanced matrix files for each chromosome (for each cell line). Ideograms were generated, where the resulting compartments are overlaid on the chromosomes and nodes are positioned below the chromosome.

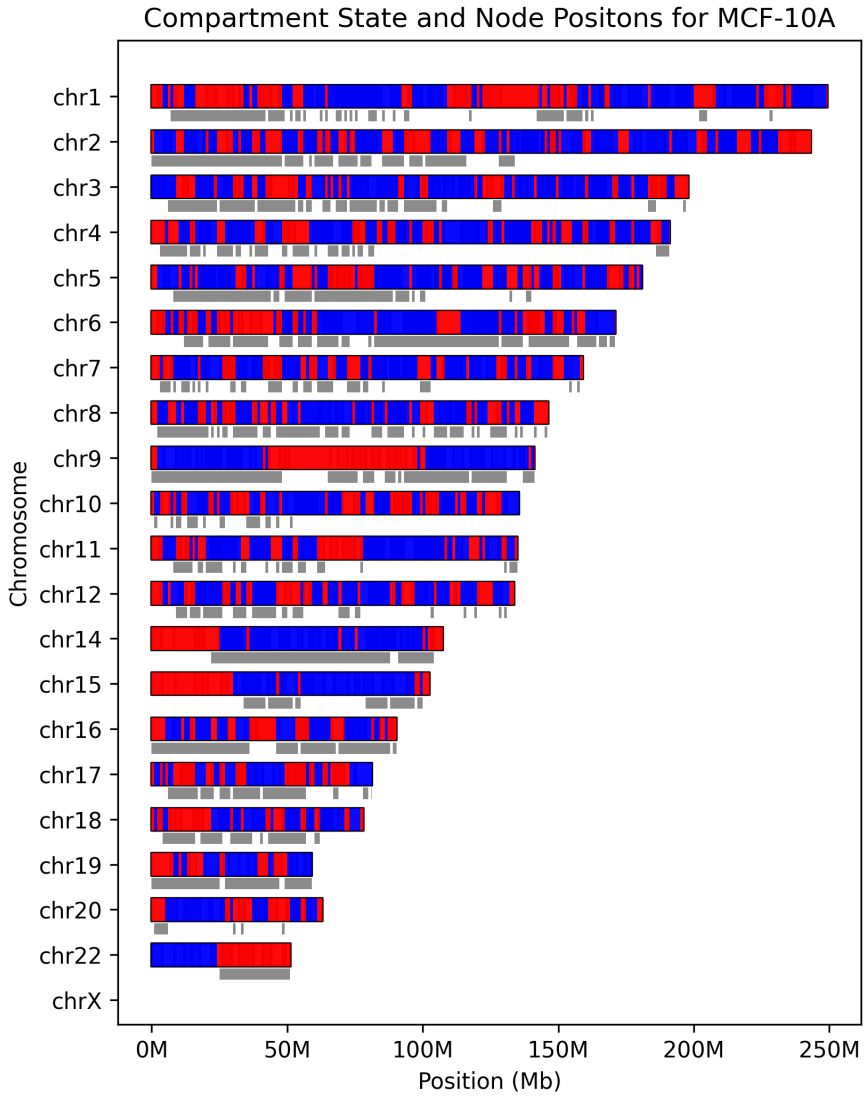


Figure 40: Compartment status and node positioning (gray blocks) for MCF-7. Compartment A is colored red and compartment B is colored blue. Compartment A correlates with higher gene density and euchromatin, while compartment B correlates with heterochromatin and lower gene density.

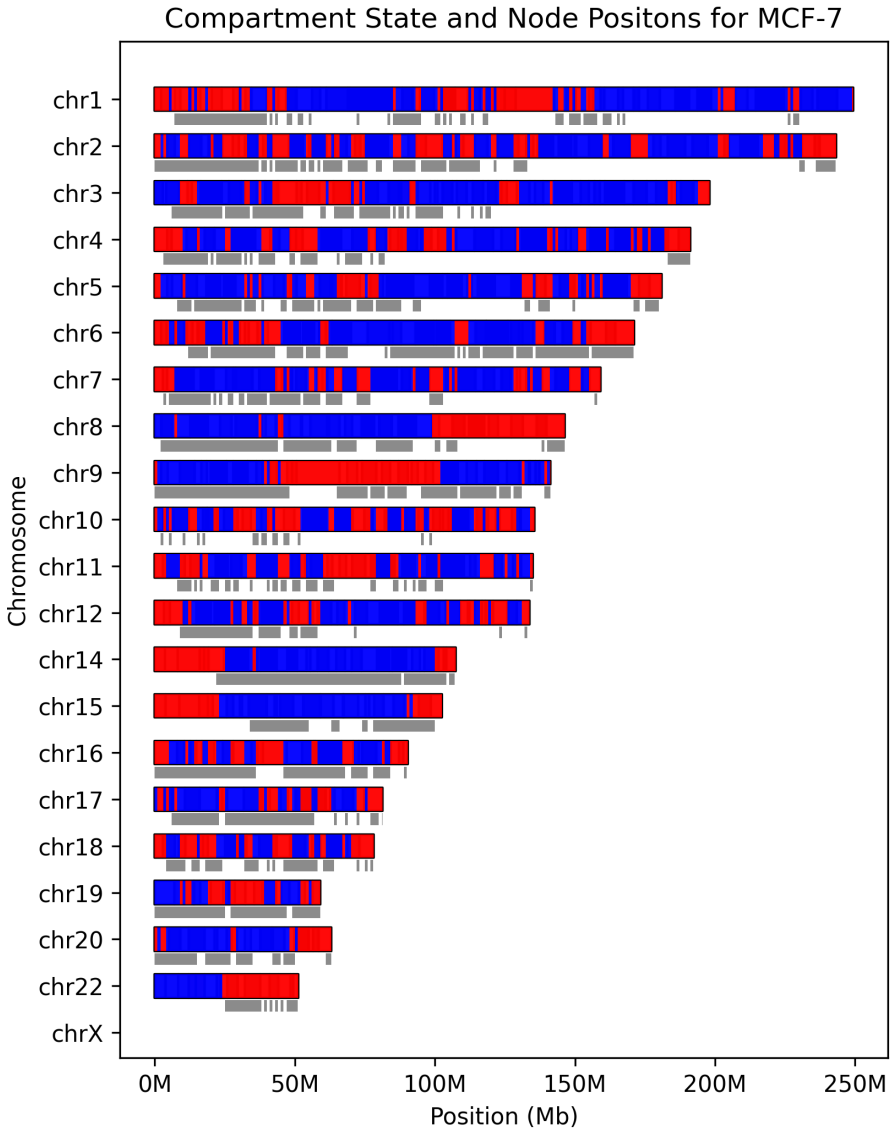


Figure 41: Compartment status and node positioning (gray blocks) for MCF-7. Compartment A is colored red and compartment B is colored blue. Compartment A correlates with higher gene density and euchromatin, while compartment B correlates with heterochromatin and lower gene density.

The MCF-10A and MCF-7 networks were thus annotated with this compartment data, as each eigenvalue sign switch correlates with compartment switching, and these compartments might change between cell lines.

The proportion of nodes in the full genome intrachromosomal networks switching compartments between MCF-10A to MCF-7 was determined. Here, MCF-10A is treated as a reference network and MCF-7 as the changed network. Nodes overlapping between the two networks are calculated per chromosome, and for each overlapping node the compartment state is compared (Figure 42).

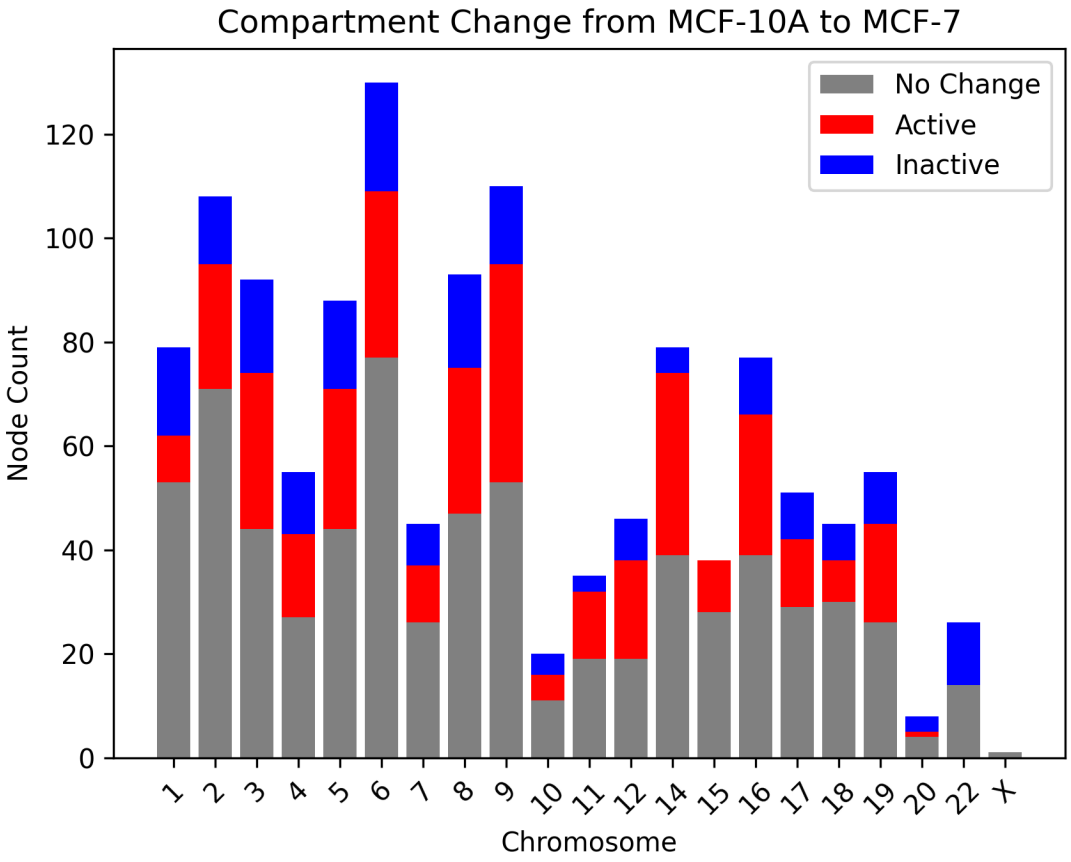


Figure 42: Compartment switch from MCF-10A to MCF-7 for overlapping nodes. MCF-10A was determined as the reference network that MCF-7 was compared to. The Y axis is the total number of overlapping nodes for each chromosome between MCF-10A and MCF-7. Most nodes do not switch compartments (gray). The active nodes in the bar plot (red) are nodes that switch from compartment B (MCF-10A) to compartment A (MCF-7). The inactive nodes (blue) are nodes that switch from compartment A (MCF-10) to compartment B in MCF-7. There is in general a higher proportion of active compartment switching from MCF-10A to MCF-7.

An extensive part of chromosome switch compartment states, and in general the MCF-7 cell line has more nodes in open compartments (A). Chromosome 1, 20 and 22 and deviations from this trend. Below are networks from MCF-10A and MCF-7 annotated with compartment status (Figure 43).

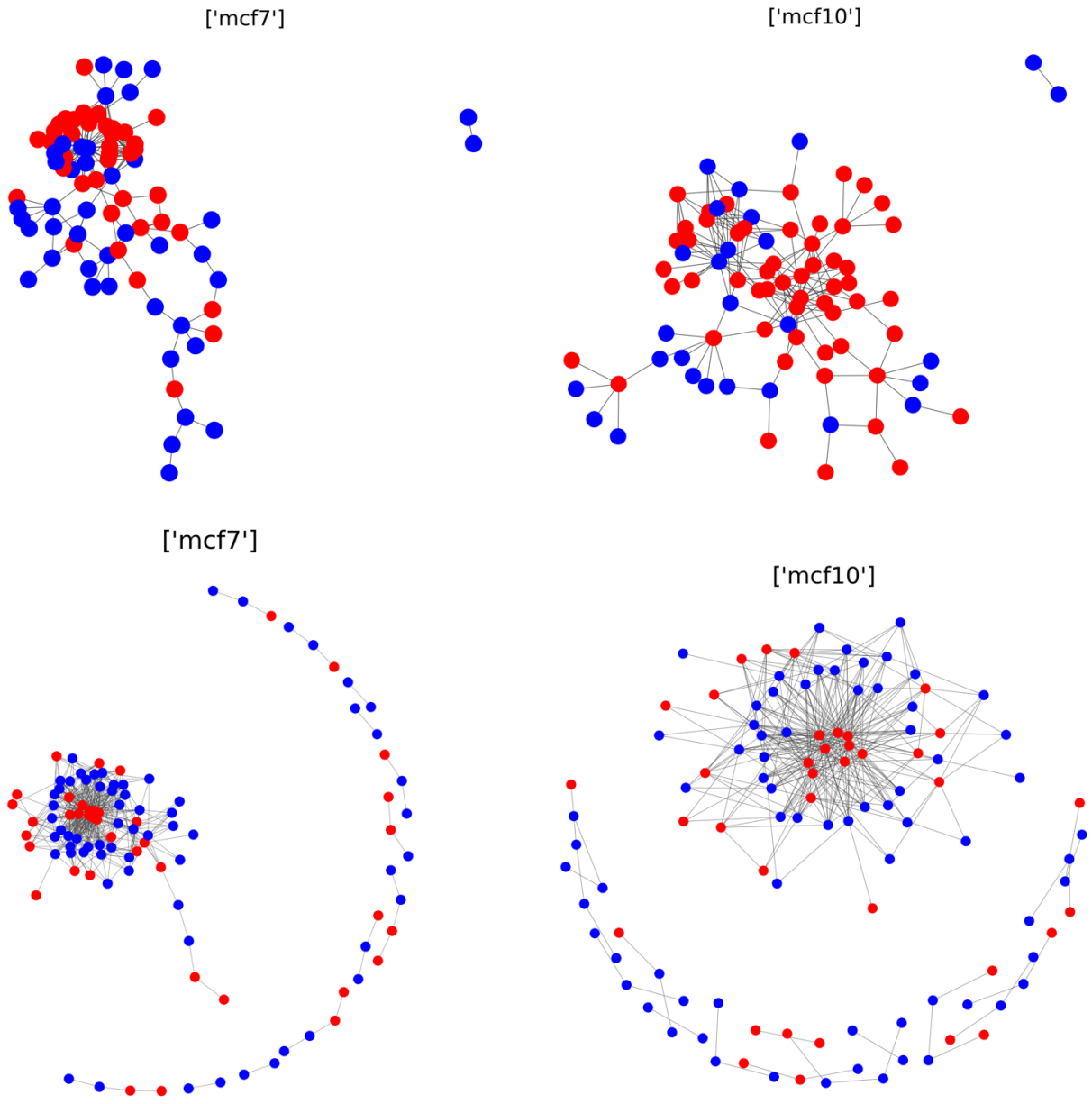


Figure 43: MCF-7 (left) and MCF-10A networks annotated with compartment status. Chromosome 1 (top) has more nodes in compartment B in MCF-7 compared to MCF-10A. Chromosome 9 (bottom) has the largest number of nodes switching compartments from MCF-10A to MCF-7.

5. Discussion

This project has involved two main parts, the first of which involved development of pipeline that processes raw Hi-C data to networks and the exploration of these networks. The pipeline works across a range of resolutions and runs automatically to generate networks from this data. The Hi-C to network pipeline allows for analyses of deeper complexities than what has been carried out in this work. Networks of any resolutions can be generated, annotated, filtered upon, and combined with a few function calls. This allows for annotation not just of compartments – but any omics data type such as promoter capture Hi-C, ChIP-Seq of histone proteins and gene names and location. In this regard the pipeline is a powerful framework that can be used to expand upon the analyses done in this work.

The second part of the project involved selecting specific Hi-C networks generated by the pipeline and exploring their properties using network approaches. MCF-10A and MCF-7 was compared in terms of centrality metrics, community structure and compartment status. It is however challenging to draw specific conclusions from these analyses – as they show statistical trends in the networks not easily linked to the underlying genome state from low-resolution Hi-C data alone. It was observed that there are no meaningful differences in centrality metrics on a genome-wide level for intrachromosomal networks, but this difference is greatest when the networks (MCF-10A and MCF-7) are compared on a chromosome-to-chromosome basis. The resolution of the data analyzed in this study is low, yielding very small networks from the small chromosomes (chromosomes 16 – 22). This means that for many chromosomes the sample size is low and thus might be less informative in terms of chromatin structure comparison across networks. In addition to the size limitation, the nodes of 1Mb often contain many genes which are much smaller than the nodes themselves. Since the nodes are created without calling TADs from the same HI-C datasets, they can overlap many genomic features and cut through TADs. Genes within these TADs are again even smaller, and the much larger bins network metric (i.e., degree) might not correlate with the underlying genes expression status. This problem could be mitigated with lower resolution data, TAD calling and annotation of the data with chromatin marks or promoter-capture HI-C.

Degree centralities were determined by analyzing the LCCs for each chromosome.
Degree centrality showed similar patterns across chromosomes between MCF-10A and MCF-7, as seen in Figure 24 and Figure 25. While other chromosomes have differing hub patterns across the cell lines - for instance in chromosome 1 in Figure 26 the MCF-7 network does not have hub nodes at ≈ 150 Mb as the MCF-10A cell line does. Chromosomes 9, 6,1 and 2 are the most similar in terms of hub node overlap. Chromosomes 6, 2 and 1 are again the most similar in terms of overlapping betweenness nodes - which can be expected as betweenness correlated with degree but not with closeness.

Networks show rearrangement of structure - in Figure 24 for instance - there is loss of hub nodes in MCF-7, creating two separate domains separated by one node (chr6: 61 - 62 Mb). This node was later determined to be part of the centromere of chromosome 6 for hg19 - meaning that the pipeline processing failed to remove all centromeric regions. This was found to be the case for target nodes as the pipeline step fails to remove nodes receiving interactions, and thus the nodes in centromeric regions are artifacts of errors in the processing pipeline.
This might be the case for other high closeness nodes as well - undermining the network structure comparisons between cell lines. The wrongful inclusion of centromeres means the MCF-7 network in Figure 30 is in fact two separate LCCs, as this is the only node connecting the two network modules. This means that the resulting network structures would be even more different if that node was deleted in both networks. This specific processing error leads to a large effect on the analysis of the network as the node included leads to two LCCs when there should only be one in the MCF-7 LCC of chromosome 6. Despite this error the structures analyzed across networks shows that MCF-7 has distinct structures varying from MCF-10A - there are different hub nodes, bottleneck nodes and nodes with high closeness between networks. Some regions are more conserved than others, which might be chromatin structures important for the normal functioning of the cell. The regions that are the most different in MCF-7 might play regulatory roles and be more involved in specialized cellular functions and differentiation. This is speculation without verification of other omics datasets - but demonstrates an important point: There are specific structural differences between the normal breast cell line MCF-10A and the cancer tissue breast cell line (MCF-7).

Closeness centrality did not correlate with betweenness nor degree on a genome level for any cell line. Chromosomes 15, 9 and 1 are the most similar chromosomes in terms of hub nodes. The most similar chromosome in terms of closeness was chromosome 1 - which has no overlap in top hub nodes between the cell lines, reflecting the structural change of the chromosome 1 LCC in MCF-7. Chromosome 8 has a large difference in the network topology - showing a much higher closeness value for top nodes.

Community detection was done next - to analyze the topological similarities and differences between the networks. The number of communities did not differ much between each chromosome between MCF-10A and MCF-7, except for chromosome 8. This chromosome has a very different topology in the two cell lines - which might indicate large-scale chromatin structural changes associated with the transformation of naive breast cancer cells to malignant cells. The average community size per chromosome was then found, where chromosomes 5, 8 and 17 had the largest differences between cell lines.
Due to the nature of the networks, classical metrics like rand index cannot be used without deleting nodes because of the often-large differences in network sizes and node locations. Deleting nodes thus risking removing nodes important for the network structure and the removal of many nodes. A better approach could be to re-run the processing pipeline and include singleton nodes in the network, to ensure that the network sizes are the same.
Using cdlibs fuzzy implementation of the NMI metric, the comparison of node similarities could still be carried out despite these features of the networks. This was first done on each chromosome, without selecting the LCC. most chromosomes were quite dissimilar and had low NMI values, while smaller chromosomes were more similar. This might reflect that smaller chromosomes have fewer nodes and generally are less connected, having more components. This analysis was redone on the LCC of each network, giving similar results for chromosomes with larger connected components and a decrease in similarities for chromosomes with many small components. The chromosomal NMI approach (Figure 36) gave the most similar community structures of chromosomes 6, 14, 15, 16 and 6. The LCC NMI approach (Figure 39) resulted in chromosomes 19, 18, 16, 6 being the most similar, while chromosomes 14 and 15 were not similar at all when comparing the LCC. The overall most similar chromosomes were thus chromosome 19, chromosome 16 and chromosome 6.

Lastly the chromosomal compartments for MCF-10A and MCF-7 were determined by ICE (iterative correction and eigenvalue decomposition). The matrices were balanced, and then decomposed, the eigenvalues switch sign (+ / -) along the chromosome and this correlates with compartment switching. The corresponding compartment plots are pictured in Figures 40 and 41. There is a big difference in the granularity of compartments called between chromosomes. In chromosome 14 for instance, the first chromosome arm is determined as compartment A while in fact this region is associated with heterochromatin formation and gene silencing. Similar patterns are seen in chromosome 15, where the short arm is determined as compartment A. This indicates from error in compartment calling process or the matrix balancing analyses. Nonetheless, there were compartment differences found between cell lines, as seen in Figure 42. Chromosome 9 had the largest switch to active state (A), while chromosome 1 had the largest switch to inactive state (B).

Most nodes are not associated with regions that switch compartments. Of the nodes that are, most switch from inactive in MCF-10A to active in MCF-7. Networks annotated with compartment status were visualized (Figure 43) for chromosome 1 and chromosome 9. MCF-7 chromosome 9 has more hub nodes in the center of the network, all belonging to compartment A, while in MCF-10A the structure is similar but with fewer nodes. It is difficult to say if these approaches lead to conclusive results without validation from external compartment status datasets - or annotation with genomic data that correlates strongly with compartment state such as histone modification data. Additionally, compartments A and B are simplifications of the actual chromatin interaction patterns and were called at low resolution (1Mb). The results might have different were a lower resolution dataset used. Hi-C data might suffer from biases and the datasets were not normalized for the differences between them, i.e., the Hi-C to network pipeline does not compare the raw datasets in the significance steps to find significantly different interactions. This might be an issue, as the different node sets and structures might be a feature of the data and not the cancer state. Replicates and further comparison between datasets would be needed to further analyze this. To increase the robustness of the pipeline, such methods could be introducing, as well as TAD calling directly on the data while processing it such that nodes are informed by TAD positioning directly. This might lead to nodes representing the chromatin

structure more accurately but would introduce additional complexity and normalization for equal node sizes across networks.

The lack of validation and annotation with external data is a strong limiting factor. Nodes have a location but are not analyzed for genomic feature on that location, this is also a limiting factor for this analysis. Many chromatin network analyses are done with annotation of several addition datasets. Hub nodes have been found to not only occupy more central positions, which was seen in the MCF-10A and MCF-7 networks but tend to be more important genomic feature such as super enhancers (Sandhu et al., 2012). In ChIA-PET and promoter capture Hi-C chromatin networks, around 40% of nodes are in the LCC, as seen in (Sandhu et al., 2012) and (Lace et al., 2020). These ChIA-Pet CINs in MCF-7 – and the K562 cancer cell line – were found to form scale-free networks (Sandhu et al., 2012). The degree distribution of MCF-10A and MCF-7, analyzed in this body of work, approached a power-law curve, indicating that the networks are indeed scale free. In another annotated CIN, it was found that the MCF-7 cell line had a different network structure for oncogenes (specifically EMP2) compared to other cell lines (Thibodeau et al., 2017).

6. Conclusion & Future perspectives

In this project a novel pipeline was developed to process raw Hi-C data to chromatin networks. The Hi-C to network pipeline handles datasets ranging in resolution from 50kb to 1Mb, processing both inter- and intrachromosomal data to significant chromatin interactions. The networks generated from this pipeline were analyzed at whole-genome scales. No significant difference between centrality metrics were found between the scale-free like Hi-C networks at this scale. The healthy breast cell line MCF-10A was then compared to the breast cancer cell line MCF-7. Centrality metrics were compared at a chromosome level between these cell lines, finding that hub and betweenness nodes vary to large extent between cell lines. High closeness centrality was found to be less conserved between cell lines than degree and betweenness. Communities were then detected by the fast greedy algorithm to compare topologies between cell lines. A normalized mutual information algorithm was used to compare the community structure on chromosome and largest connected component level. Communities between MCF-10A and MCF-7 differ at both levels, with the largest difference in community structure being at the largest chromosomes. Lastly, the compartment status of the Hi-C data that generated the networks was determined. Most nodes that switch compartments from MCF-10A to MCF-7 are active switches.

The limitations to this project regarding the network analysis are that networks were not annotated with external datasets to validate the findings. Furthermore, no replicates were used to rule out the possibility that differences observed are not biases inherent in the data but actual representative differences. Furthermore, cell lines were compared at a low resolution which led to small network sizes compared to higher resolution data. The Hi-C to network processing pipeline does not consider differences between the datasets being compared, it processes each cell line separately. Last is the limitation of all Hi-C data - networks represent population average chromatin structures, and do not capture the heterogeneity within cell populations.

Further analyses might include correlations between compartment states and centrality values. Networks with higher resolution could be used to call TADs and be annotated with gene accession numbers. Nodes with a high centrality or differing centrality score across cell lines

could then be used to associate the network metrics directly to genomic features found in breast cancer. Furthermore, communities in cell lines could be mapped back to the linear genome and compared with structural variations involved in breast cancer. Additionally, more sophisticated, and specialized approaches used to determine differential community structures could be implemented – such as the Alpaca algorithm. Weighted networks could be created automatically during the processing pipeline, since raw Hi-C data has interaction counts. This would open many new network analysis possibilities.

Modeling data as networks represents not only the data itself but its relation to the system it is part of. Network analysis approaches in Hi-C networks thus allow for a detailed representation of the complex biological system they represent. Annotation of networks allows for a multi-omics approach to integrate several data types into one network. Representing networks across multiple scales can be achieved in one network – nodes can represent any relationship between the data. With the increased power of *in silico* machine learning technologies enabled by higher quality data, paired with network theory approaches, new discoveries about the 3D structure of chromatin and its relation to the epigenetic landscape are bound to be made.

References

Abdalla, M. O. A., Yamamoto, T., Maehara, K., Nogami, J., Ohkawa, Y., Miura, H., Poonperm, R., Hiratani, I., Nakayama, H., Nakao, M., & Saitoh, N. (2019). The Eleanor ncRNAs activate the topological domain of the ESR1 locus to balance against apoptosis. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11378-4

Abdennur, N., & Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, *36*(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), Article 6794. https://doi.org/10.1038/35019019

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-45839-z

Ay, F., & Noble, W. (2015). Analysis methods for studying the 3D architecture of the genome. *Genome Biology*, *16*, 183. https://doi.org/10.1186/s13059-015-0745-7

Babaei, S., Mahfouz, A., Hulsman, M., Lelieveldt, B. P. F., Ridder, J. de, & Reinders, M. (2015). Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex. *PLOS Computational Biology*, *11*(5), e1004221. https://doi.org/10.1371/journal.pcbi.1004221

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., Browne, G., van Wijnen, A. J., Lian, J. B., Stein, J. L., Dekker, J., Imbalzano, A. N., & Stein, G. S. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biology*, *16*(1), 214. https://doi.org/10.1186/s13059-015-0768-0

Baudin, B., Bruneel, A., Bosselut, N., & Vaubourdolle, M. (2007). A protocol for isolation and culture of human umbilical vein endothelial cells. *Nature Protocols*, *2*(3), 481–485. https://doi.org/10.1038/nprot.2007.54

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature*, *447*(7143), Article 7143. https://doi.org/10.1038/nature05915

Briand, N., & Collas, P. (2020). Lamina-associated domains: Peripheral matters and internal affairs. *Genome Biology*, *21*(1), 85. https://doi.org/10.1186/s13059-020-02003-5

Buitrago, D., Labrador, M., Arcon, J. P., Lema, R., Flores, O., Esteve-Codina, A., Blanc, J., Villegas, N., Bellido, D., Gut, M., Dans, P. D., Heath, S. C., Gut, I. G., Brun Heath, I., & Orozco, M. (2021). Impact of DNA methylation on 3D genome structure. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-23142-8

Chovanec, P., Collier, A. J., Krueger, C., Várnai, C., Semprich, C. I., Schoenfelder, S., Corcoran, A. E., & Rugg-Gunn, P. J. (2021). Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-22201-4

Clapier, C. R., Iwasa, J., Cairns, B. R., & Peterson, C. L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, *18*(7), Article 7. https://doi.org/10.1038/nrm.2017.26

Cremer, T., & Cremer, M. (2010). Chromosome Territories. *Cold Spring Harbor Perspectives in Biology*, *2*(3), a003889. https://doi.org/10.1101/cshperspect.a003889

Cremer, T., Cremer, M., Dietzel, S., Müller, S., Solovei, I., & Fakan, S. (2006). Chromosome territories – a functional nuclear landscape. *Current Opinion in Cell Biology*, *18*(3), 307–316. https://doi.org/10.1016/j.ceb.2006.04.007

Csardi, G., & Nepusz, T. (2005). The Igraph Software Package for Complex Network Research. *InterJournal*, *Complex Systems*, 1695.

Dale, R. K., Pedersen, B. S., & Quinlan, A. R. (2011). Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, *27*(24), 3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Dali, R., & Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*, *45*(6), 2994–3005. https://doi.org/10.1093/nar/gkx145

Das, K., Samanta, S., & Pal, M. (2018). Study on centrality measures in social networks: A survey. *Social Network Analysis and Mining*, *8*(1), 13. https://doi.org/10.1007/s13278-018-0493-2

Davidson, I. F., & Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology*, *22*(7), Article 7. https://doi.org/10.1038/s41580-021-00349-7

Dekker, J. (2017). *ENCSR011GNI* [Data set]. https://doi.org/10.17989/ENCSR011GNI

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

Deng, S., Feng, Y., & Pauklin, S. (2022). 3D chromatin architecture and transcription regulation in cancer. *Journal of Hematology & Oncology*, *15*(1), 49. https://doi.org/10.1186/s13045-022-01271-x

Deng, W., Rupon, J. W., Krivega, I., Breda, L., Motta, I., Jahn, K. S., Reik, A., Gregory, P. D., Rivella, S., Dean, A., & Blobel, G. A. (2014). Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell*, *158*(4), 849–860. https://doi.org/10.1016/j.cell.2014.05.050

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), Article 7398. https://doi.org/10.1038/nature11082

Ehler, E., Babiychuk, E., & Draeger, A. (1996). Human foetal lung (IMR-90) cells: Myofibroblasts with smooth muscle-like contractile properties. *Cell Motility and the Cytoskeleton*, *34*(4), 288–298. https://doi.org/10.1002/(SICI)1097-0169(1996)34:4<288::AID-CM4>3.0.CO;2-4

Fortin, J.-P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, *16*(1), 180. https://doi.org/10.1186/s13059-015-0741-y

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics: TIG*, *29*(3), 150–159. https://doi.org/10.1016/j.tig.2012.11.004

Fussner, E., Ching, R. W., & Bazett-Jones, D. P. (2011). Living without 30nm chromatin fibers. *Trends in Biochemical Sciences*, *36*(1), 1–6. https://doi.org/10.1016/j.tibs.2010.09.002

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., & Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science*, *360*(6384), 102–105. https://doi.org/10.1126/science.aar7831

Han, J., Zhang, Z., & Wang, K. (2018). 3C and 3C-based techniques: The powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, *11*(1), 21. https://doi.org/10.1186/s13039-018-0368-2

Hansen, A. S., Cattoglio, C., Darzacq, X., & Tjian, R. (2018). Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*, *9*(1), 20–32. https://doi.org/10.1080/19491034.2017.1389365

Huang, H., Chen, S. T., Titus, K. R., Emerson, D. J., Bassett, D. S., & Phillips-Cremins, J. E. (2019). A subset of topologically associating domains fold into mesoscale core-periphery networks. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-45457-9

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(03), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ideker, T., & Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, *8*(1), 565. https://doi.org/10.1038/msb.2011.99

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, *9*(10), Article 10. https://doi.org/10.1038/nmeth.2148

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, *108*, 87–111. https://doi.org/10.1016/j.jnca.2018.02.011

Jerkovic´, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, *22*(8), Article 8. https://doi.org/10.1038/s41580-021-00362-w

Kalluchi, A., Harris, H. L., Reznicek, T. E., & Rowley, M. J. (2023). Considerations and caveats for analyzing chromatin compartments. *Frontiers in Molecular Biosciences*, *10*. https://www.frontiersin.org/articles/10.3389/fmolb.2023.1168562

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00034

Lace, L., Melkus, G., Rucevskis, P., Celms, E., Čerāns, K., Kikusts, P., Opmanis, M., Rituma, D., & Viksna, J. (2020). Characteristic Topological Features of Promoter Capture Hi-C Interaction Networks. In A. Roque, A. Tomczyk, E. De Maria, F. Putze, R. Moucek, A. Fred, & H. Gamboa (Eds.), *Biomedical Engineering Systems and Technologies* (pp. 192–215). Springer International Publishing. https://doi.org/10.1007/978-3-030-46970-2_10

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods (San Diego, Calif.)*, *72*, 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031

Lammerding, J. (2011). Mechanics of the Nucleus. *Comprehensive Physiology*, *1*(2), 783–807. https://doi.org/10.1002/cphy.c100038

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), 056117. https://doi.org/10.1103/PhysRevE.80.056117

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, *11*(3), 033015. https://doi.org/10.1088/1367-2630/11/3/033015

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), Article 4. https://doi.org/10.1038/nmeth.1923

Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289. https://doi.org/10.1126/science.1181369

Lomberk, G., Wallrath, L., & Urrutia, R. (2006). The Heterochromatin Protein 1 family. *Genome Biology*, *7*(7), 228. https://doi.org/10.1186/gb-2006-7-7-228

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Madrid-Mencía, M., Raineri, E., Cao, T. B. N., & Pancaldi, V. (2020). Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks. *Nucleic Acids Research*, *48*(8), 4066–4080. https://doi.org/10.1093/nar/gkaa159

Maeshima, K., Hihara, S., & Eltsov, M. (2010). Chromatin structure: Does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, *22*(3), 291–297. https://doi.org/10.1016/j.ceb.2010.03.001

Malod-Dognin, N., Pancaldi, V., Valencia, A., & Pržulj, N. (2020). Chromatin network markers of leukemia. *Bioinformatics*, *36*(Suppl 1), i455–i463. https://doi.org/10.1093/bioinformatics/btaa445

McArthur, E., & Capra, J. A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *The American Journal of Human Genetics*, *108*(2), 269–283. https://doi.org/10.1016/j.ajhg.2021.01.001

McDaid, A., Greene, D., & Hurley, N. (2011). Normalized Mutual Information to evaluate overlapping community finding algorithms. *CoRR*.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nichols, M. H., & Corces, V. G. (2021). Principles of 3D compartmentalization of the human genome. *Cell Reports*, *35*(13), 109330. https://doi.org/10.1016/j.celrep.2021.109330

Nishino, Y., Eltsov, M., Joti, Y., Ito, K., Takata, H., Takahashi, Y., Hihara, S., Frangakis, A. S., Imamoto, N., Ishikawa, T., & Maeshima, K. (2012). Human mitotic chromosomes consist

predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO Journal*, *31*(7), 1644–1653. https://doi.org/10.1038/emboj.2012.35

Norton, H. K., Emerson, D. J., Huang, H., Kim, J., Titus, K. R., Gu, S., Bassett, D. S., & Phillips-Cremins, J. E. (2018). Detecting hierarchical genome folding with network modularity. *Nature Methods*, *15*(2), Article 2. https://doi.org/10.1038/nmeth.4560

Nugoli, M., Chuchana, P., Vendrell, J., Orsetti, B., Ursule, L., Nguyen, C., Birnbaum, D., Douzery, E. J., Cohen, P., & Theillet, C. (2003). Genetic variability in MCF-7 sublines: Evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer*, *3*, 13. https://doi.org/10.1186/1471-2407-3-13

Ou, H. D., Phan, S., Deerinck, T. J., Thor, A., Ellisman, M. H., & O'Shea, C. C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, *357*(6349), eaag0025. https://doi.org/10.1126/science.aag0025

Padi, M., & Quackenbush, J. (2018). Detecting phenotype-driven transitions in regulatory network structure. *Npj Systems Biology and Applications*, *4*(1), Article 1. https://doi.org/10.1038/s41540-018-0052-5

Pancaldi, V. (2023). Network models of chromatin structure. *Current Opinion in Genetics & Development*, *80*, 102051. https://doi.org/10.1016/j.gde.2023.102051

Paulsen, J., Liyakat Ali, T. M., & Collas, P. (2018). Computational 3D genome modeling using Chrom3D. *Nature Protocols*, *13*(5), 1137–1152. https://doi.org/10.1038/nprot.2018.009

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, *4*(1), 10. https://doi.org/10.1186/1756-0381-4-10

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*(null), 2825–2830.

Qu, Y., Han, B., Yu, Y., Yao, W., Bose, S., Karlan, B. Y., Giuliano, A. E., & Cui, X. (2015). Evaluation of MCF10A as a Reliable Model for Normal Human Mammary Epithelial Cells. *PLoS ONE*, *10*(7), e0131285. https://doi.org/10.1371/journal.pone.0131285

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014a). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014b). A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Rossetti, G., Milli, L., & Cazabet, R. (2019). CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, *4*(1), Article 1. https://doi.org/10.1007/s41109-019-0165-9

Rossetti, G., Pappalardo, L., & Rinzivillo, S. (2016, March 24). *A novel approach to evaluate community detection algorithms on ground truth*.

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, *19*(12), Article 12. https://doi.org/10.1038/s41576-018-0060-8

Sandhu, K. S., Li, G., Poh, H. M., Quek, Y. L. K., Sia, Y. Y., Peh, S. Q., Mulawadi, F. H., Lim, J., Sikic, M., Menghi, F., Thalamuthu, A., Sung, W. K., Ruan, X., Fullwood, M. J., Liu, E., Csermely, P., & Ruan, Y. (2012). Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. *Cell Reports*, *2*(5), 1207–1219. https://doi.org/10.1016/j.celrep.2012.09.022

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*(1), 259. https://doi.org/10.1186/s13059-015-0831-x

Shadeo, A., & Lam, W. L. (2006). Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Research*, *8*(1), R9. https://doi.org/10.1186/bcr1370

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), Article 6765. https://doi.org/10.1038/47412

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-53708-y

Thibodeau, A., Márquez, E. J., Shin, D.-G., Vera-Licona, P., & Ucar, D. (2017). Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-14389-7

Vantangoli, M. M., Madnick, S. J., Huse, S. M., Weston, P., & Boekelheide, K. (2015). MCF-7 Human Breast Cancer Cells Form Differentiated Microtissues in Scaffold-Free Hydrogels. *PLoS ONE*, *10*(8), e0135426. https://doi.org/10.1371/journal.pone.0135426

Waskom, M., Botvinnik, O., Hobson, P., Cole, J. B., Halchenko, Y., Hoyer, S., Miles, A., Augspurger, T., Yarkoni, T., Megies, T., Coelho, L. P., Wehner, D., cynddl, Ziegler, E.,

diego0020, Zaytsev, Y. V., Hoppe, T., Seabold, S., Cloud, P., … Allan, D. (2014). *seaborn: V0.5.0 (November 2014)*. Zenodo. https://doi.org/10.5281/zenodo.12710

Yan, K.-K., Lou, S., & Gerstein, M. (2017). MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLOS Computational Biology*, *13*(7), e1005647. https://doi.org/10.1371/journal.pcbi.1005647

Zakirov, A. N., Sosnovskaya, S., Ryumina, E. D., Kharybina, E., Strelkova, O. S., Zhironkina, O. A., Golyshev, S. A., Moiseenko, A., & Kireev, I. I. (2022). Fiber-Like Organization as a Basic Principle for Euchromatin Higher-Order Structure. *Frontiers in Cell and Developmental Biology*, *9*. https://www.frontiersin.org/articles/10.3389/fcell.2021.784440

Zhang, Y., Sun, Z., Jia, J., Du, T., Zhang, N., Tang, Y., Fang, Y., & Fang, D. (2021). Overview of Histone Modification. In D. Fang & J. Han (Eds.), *Histone Mutations and Cancer* (pp. 1–16). Springer. https://doi.org/10.1007/978-981-15-8104-5_1

Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature Genetics*, *54*(5), Article 5. https://doi.org/10.1038/s41588-022-01065-4