

Identifying 3D expression domains by graph clustering

Master Thesis, 19-01-2016



Author: Jos Kuijpers

Student-ID: 1311530

Faculty: EWI

E-mail: j.s.f.kuijpers@student.tudelft.nl

Graduation Committee:

- Jeroen de Ridder (EWI)
- Lodewyk Wessels (NKI / EWI)
- Ahmed Mahfouz (EWI)
- Zaid al-Ars (EWI)

Abstract

In this work we explore 3D expression domains by using Hi-C contact information to construct an interaction network in combination with TRIP expression and position data. After clustering this network using a community finding algorithm, resulting clusters are analysed for enrichment of Gene Ontology terms. Our work demonstrates a link between the presence of expression measuring reporters and GO enrichments. Furthermore we find that the presence of an insignificant number of a single type of these reporters is generally not linked to these enrichments. In clusters containing a large number of these reporters, enrichment of GO terms is more prevalent. By performing analyses on presence of 3D expression domains we hope to gain more insight into permissive / non-permissive chromatin domains and how these domains interact over longer distances.

Introduction

The 3D conformation of chromosomes plays an important role in the regulation and expression of genes [29]. Studies have shown that genes are regulated not only by elements that are closeby on a linear scale, but also by elements which are located up to 2 million base pairs away [23]. Chromatin is a biological structure which consists of DNA and associated proteins [24]. It contains a complex of many repeating nucleosomes. The nucleosome is a DNA packaging unit, consisting of eight histone cores and a segment of DNA which is wound around it, comparable to a spool. At the base of the nucleosome there are linker histones. The structure of nucleosomes makes it possible to act as a general gene repressor by making the DNA inaccessible.

Histones are part of a larger structure (nucleosome). It acts as a scaffolding around which the double helix of DNA can coil. There are several types of histones known, divided into five major families: H1/H5, H2A, H2B, H3 and H4. H1 and H5 are linker histones which prevent mobility of nucleosomes [33], whilst the other ones are core histones. It is these core histones which assemble into the core of the nucleosome.

Together these two structures form a complex of many repeating nucleosomes which is the basis of eukaryotic chromatin. By winding the DNA around histones, the DNA becomes packaged into a much smaller volume, whilst at the same time preventing damage to occur. A second advantage to the winding of DNA around histones is that it becomes a more controlled environment regarding gene expression. If it is in a tight winding, there isn't room for transcription factors to bind (or at least becomes far more difficult).

It plays an important role in the regulation of transcription and other functions and is in either of two defined states: open or closed. The state corresponds with the ability of proteins being able to bind to a particular part of the chromatin. The state of chromatin is important in models regarding models for regulation and gene expression.

A study from *Filion et al* [6] on binding profiles of 100s of proteins across the genome concluded that chromatin can be divided into several principal types. This was done by analyzing the binding profiles of chromatin proteins using a 2-state HMM. For every protein the target and non-target loci are determined and identifies the most likely segmentation of 'bound' and 'unbound' loci. The types of chromatin were found using a PCA on the binding profiles and projecting the genomic sites onto the first 3 components.

The five types labeled in the work are identified as major types which may consist of subtypes. Each type is defined by unique combinations of proteins, with proteins not restricted to a single type of chromatin. One of the main findings is that roughly half the genome is covered by a repressive chromatin type (named black chromatin), which also lacks classic heterochromatin markers. Two types of identified chromatin appear to be similar (red and yellow), having most genes in an active state and overall expression levels are similar.

From this work it follows that the genome is divided into several regions of chromatin. There are several factors important for gene regulation: chromatin (proteins binding to DNA, packaging), regulatory proteins (Transcription Factors) and the 3D conformation of the genome.

The long-range interactions, based on the position on the linear genome have been studied extensively by means of chromosome conformation capture techniques, such as 3C, 4C and Hi-C. These methods measure interactions between chromatin and directly facilitate investigation of gene interactions.

Hi-C

Over the years several methods have been developed to study the 3D structure of chromatin based on the measurement of DNA-DNA interactions. These include Chromosome Conformation Capture (3C), 4C, which combines 3C with the use of inverse PCR and 5C (multiplexed ligation-mediated amplification). A more recent approach is the Hi-C method, which fundamentally differs from other methods as it does not require a set of target loci, providing a more unbiased genome wide analysis [2]. Hi-C is a method used to study chromatin interactions on a large scale via the use of contact information.

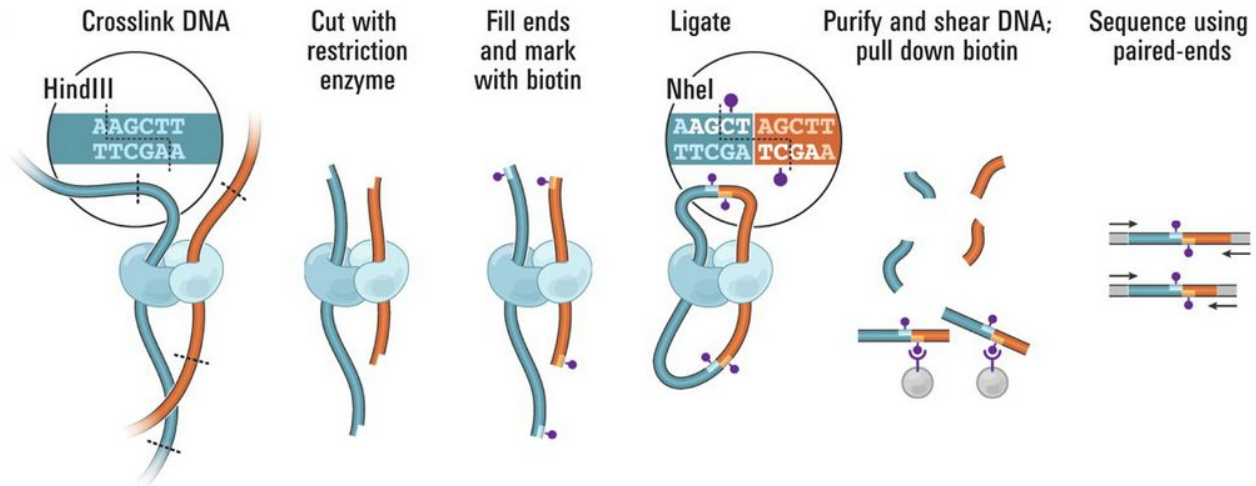


Figure 1: Overview of creation of Hi-C library [2]. 1) Cross-link two pieces of DNA together with an enzyme. 2) A restriction enzyme cuts the pieces. 3) DNA segment endings are filled and marked with biotin. 4) Segments are ligated. 5) Using a purify and shearing operation on the DNA, the fragments contain a ligation junction where contact was made 6) By using paired-end sequencing, the final interactions are obtained.

As depicted in Figure 1, a Hi-C library is created using a multi-step procedure which starts with the use of enzymes to bind (cross-link) two pieces of DNA together. Using a restriction enzyme the pieces are cut and the ends of the DNA segments are filled and marked with biotin. These segments are then ligated and using a purify and shearing operation on the DNA. DNA fragments are obtained that contain a ligation junction between two pieces of DNA that were in contact with each other in the nucleus.

The library is analyzed, by means of parallel sequencing, which in turn produces a catalog of interacting fragments. The next step is to construct a contact matrix by dividing the genome into regions (loci), where each entry m_{ij} stands for the number of products measured between locus i and j . Studies [31,32] have shown that there are three main biases affecting the final dataset: distance between restriction sites, the GC content of trimmed ligation junctions and sequence uniqueness.

A normalization step is required in order to account for linear distance bias, as loci can more often be linked if they are naturally close and thus less dependent of the folding of the genome.

A way this normalization can be done by dividing every entry m_{ij} by the average across the diagonal of matrix M . This comes with the downside where diagonals become shorter when the distance increases, which can be overcome by using adjacent diagonals to create a longer diagonal again. Besides a distance bias one can also normalize based on restriction site density. A downside to Hi-C resolution is increasing the resolution by n require n^2 additional reads.

The primary contribution of Hi-C in the biological field is that it shows that genomic regions which are linearly distant can share interactions by being in close proximity in space. By

understanding which linearly distant regions of the genome interact, we gain understanding over long-distance mechanics.

Certain mechanisms within the cell, such as transcription factories make good use of the way the DNA is folded [21]. The entire length of the DNA has to be packed in the small space of the nucleus, naturally forcing regions to come into contact via the specific way the folding of DNA occurs. Other important processes which depend heavily on the 3D folding of the genome are replication, recombination, repair, transcription and RNA processing machineries.[25]

A single transcription factor can therefore influence activity on several parts of the genome simultaneously through their close proximity.

These long-range interactions can be analyzed using a network topology. This can be achieved by using the Hi-C matrix as an adjacency matrix. Binned regions become the nodes of the network and the interactions work as edges for the network. In such a network, nodes represent a (set of) gene(s) or a more general region of the genome defined by a start base pair and an end base pair. Studies have been done that have transformed chromatin interaction data into networks [2,8,9,10,11] and have been used to study for example gene co-expression and disease-associated SNP enrichment as well as interchromosomal interactions. Such networks can be studied with the use of graph theory and clustering algorithms in order to extract smaller communities [5,8,9,10] which are useful for analysis as there remains a biological constraint that only a limited number of regions can truly interact with each other in the physical space.

Previous work [1] has shown the presence of expression / chromatin domains, with neighboring genomic regions having a similar expression pattern or chromatin state. In this work we investigate 3D expression environments, where multiple genomic regions show similar expression patterns based on the state of the chromatin. By investigating long range interactions between local expression domains we hope to establish the presence of 3D environments.

In order to do so we investigate long-range interactions via the Hi-C methodology to see how often genomic regions of similar expression patterns interact when they are linearly distant. We reason that regions of the genome that are in close proximity within the nucleus will show similar expression patterns for present genes as they reside in the same '3D environment', with the 3D environment corresponding with multiple sections of chromatin within close proximity within the nucleus. These regions will likely share more interactions with other regions within the same environment.

TRIP

A 2013 publication by *Akhtar et al* ([1]) introduced the methodology of TRIP (Thousands of Reporters Integrated in Parallel). Reporter genes are a powerful tool to investigate the influence of chromatin on gene expression. A reporter gene is a non-native gene which is inserted into a host genome and has its location mapped. By measuring the expression of the reporter gene

we get an idea of the expression level of the genomic location where the insertion took place. These expression levels help identify the state of the chromatin, as 'closed' chromatin will result in lower expression levels compared to 'open' chromatin.

Before the publication of the TRIP method this generally required low throughput analyses as reporters could only target specific genomic loci. The process of mapping the integration sites is extremely laborious and not suitable for genome-wide interrogation[22]. TRIP solves these problems by using a random barcodes of 16-20 bp to track each individual insertion. The barcode locations are mapped using sequencing and due to the size of the barcode, many potential markers can be produced for use during the same study. This leads to the ability of tracking multiple different insertion points and allows for genome-wide analysis.

This methodology uses insertional mutagenesis to insert reporters via the PiggyBac (PB, [26]) system into a host genome and is designed to study position effects genome-wide on a massive scale. The choice for the PB system was based on its high efficiency [3] and small size of terminal repeats [4]. These small terminal repeats limit the amount of 'helper DNA' which is inserted into the host genome, which helps prevent adverse effects such as gene silencing. The reporters are subsequently integrated into the host genome at random positions using the PB transposon plasmid library.

The TRIP dataset consists of locations defined at a certain base pair on the genome where reporters are integrated. For each reporter we have an expression value based on the normalized counts of transcription products measured in the cell pool. The normalization step is required as integrations are measured over a large pool of cells. This is done by determining the number of occurrences of a barcode in the cDNA using sequencing and normalizing this count against the corresponding barcode representation in the genomic DNA, where the location of the integration is determined using DNA sequencing. This yields an expression value for the reporter of a certain barcode at a specific genomic location. The normalized expression is calculated using the logarithm on the expression value. When there are less samples measured than expected (which results in a ratio smaller than 1), the logarithm returns a negative value for the expression. These negative expression signs are retained for further use as a possible indicator of repressive domains.

The primary contribution to the field of chromatin position effect studies gained from this methodology is the large scale monitoring of influences of local environment on expression. The research by *Akhtar et al* [1] shows that the expression of these reporters can be influenced up to ~20k base pair away.

Furthermore, this method shows that the integration site has an important influence on the expression of the integrated reporter gene, as otherwise each promoter would have a similar expression value. However *Akhtar et al* [1] show that this is not the case. The effect is caused by chromatin states; it is either in an "open" or "closed" state, which refers to the accessibility of said chromatin for proteins to bind on. The TRIP study has shown that the reporters integrate in groups of high (positive) or low (negative) expression, showing that chromatin domains exist where larger sections of chromatin will have similar expression patterns.

The establishing of pairwise interactions of a similar expression pattern forms the basis which we expand by translating all interactions to a network setup. We hypothesize that regions of the genome that are in close proximity will show similar expression patterns based on the presence in the same 3D environment. The expression value of a region is determined by the TRIP integrations in said regions (Fig 2B). By using Hi-C we start with pairwise interactions (Fig 2C) and to find more complex interacting regions we use a clustering approach on the densely connected network (Fig 2D and 2E). For the found clusters we perform Gene Ontology enrichment analysis and find that clusters with enrichment of TRIP show GO enrichments.

We investigate the effects of introducing a threshold on the Hi-C interaction scores on the fraction of Positive and Negative TRIP bin pairs to filter out possible noise from the data set. Next we investigate the folding of the chromosome and its resulting interactions by comparing the real network against randomly generated networks. We do this to verify that the interactions which happen in a real cell are not happening based on random chance, but because of the way the chromosome is folded. Finally we are interested to see if the biological functions are shared amongst these 3D environments. Biological functions which are shared between regions are an indication of chromosome organization, with functions separated over longer distances, but still being able to work together.

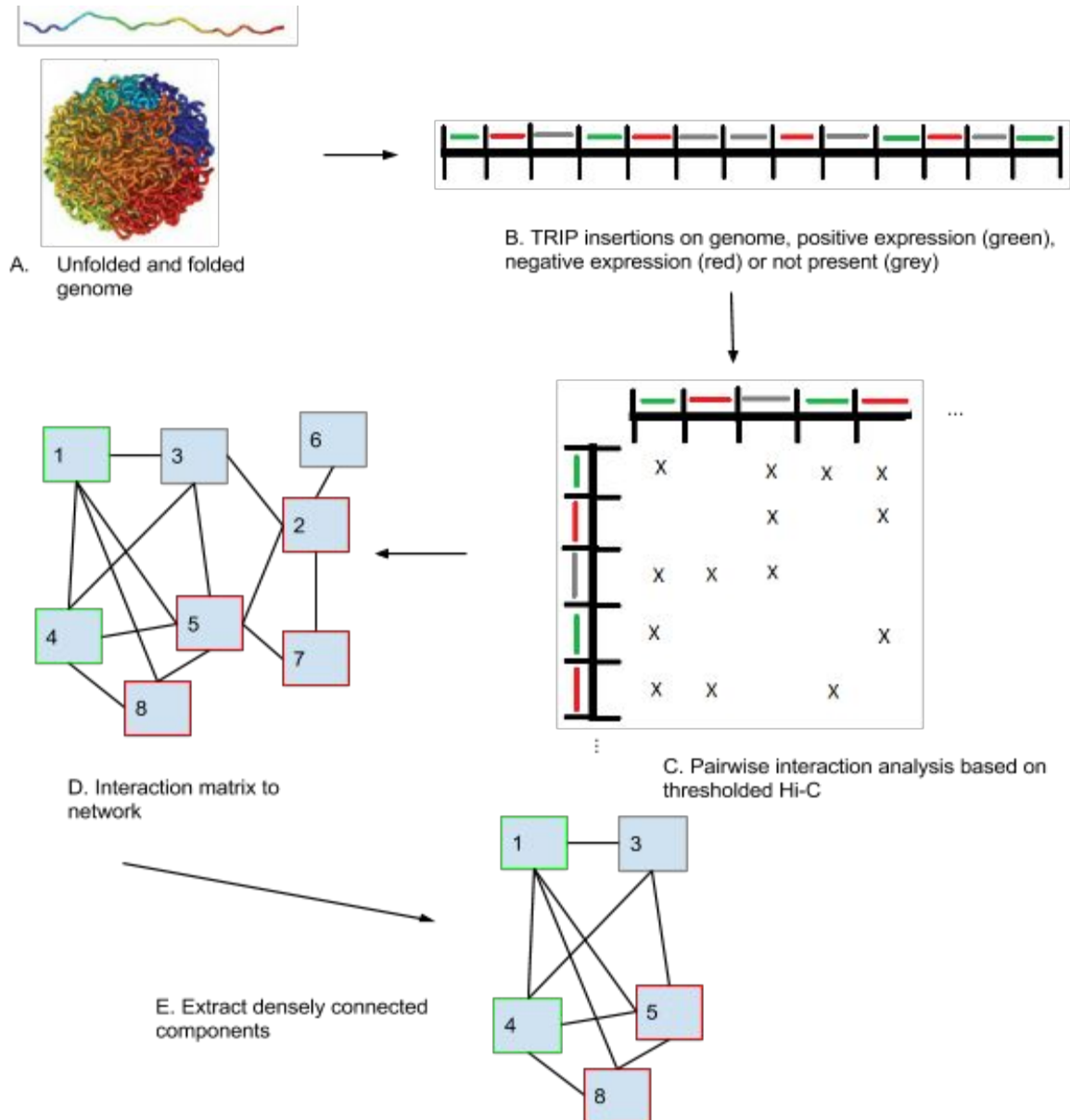


Figure 2: Starting with the densely folded genome in A, we investigate pairwise interactions and construct an interaction network. From this network we extract densely connected components and investigate the gene ontology term enrichments for each component.

Methodology

In order to gain understanding of 3D expression domains we look at the interactions between genomic loci. We are interested to see whether long range interacting domains show similar expression patterns and whether we can associate a functional role to such domains and interactions.

To do this we start with the Hi-C matrix, which is an indication of the interactions between pairs of regions of the genome through scoring. The Hi-C data is obtained using a 40k base pair window size, using the methodology of *Lieberman-Aiden* [2]. A Hi-C interaction i,j represents the number of interactions measured between the windows i and j . These interactions can be used to convert the symmetrical matrix into an undirected network by using the genomic bins as nodes and all interactions as the edges.

In order to map the TRIP reporter data onto the Hi-C network, the TRIP locations are mapped to the same 40k bp window. Due to the number of TRIP integrations in the dataset we end up with multiple data points in the same 40k bp regions. To provide each bin with an unambiguous expression sign (+ or -) we use a majority vote. In order to filter out low expression TRIP data points, we employ a small near-zero threshold on the TRIP reporter expression values ($[-0.5, 0.5]$). This threshold is used to remove very small (near zero) expression data points which could be the result of measuring errors or general noise. The resulting TRIP reporters can now be categorized in three distinct sets: Positive (+), Negative (-) and Neutral (0). In this case the neutral reporters count as there being no TRIP present at all.

Using a majority count we prevent a single large expression datapoint from deciding the sign of a bin. If a certain genomic region contains multiple reporters showing a similar chromatin state, based on the expression of the reporter, a single outlier can represent a measurement error. If multiple measurements for a region show a certain domain type, the outlier will likely not be a true measurement.

By using nonzero Hi-C interaction scores we find pairs of Hi-C bins which share a possible interaction within the nucleus. Based on the presence of TRIP reporters we define TRIP bin pairs as either Positive (++), with two Positive TRIP bins sharing an interaction, Negative(--), with two Negative TRIP bins sharing an interaction, and mixed(+-), with one Negative and one Positive TRIP bin in the interaction. The other combinations involving non-TRIP nodes (+0, -0 and 00) are not counted as TRIP bin pairs.

Hi-C Normalisation

As the contact probability is influenced by sequence proximity, the contact matrix is normalized by dividing each entry by the chromosome-wide average contact probability for loci at a similar distance (i.e. along a diagonal on the matrix) as done by *Lieberman-Aiden* [2]. A downside here is that as the distance from the main diagonal increases, the diagonals get shorter, which has an impact on the resulting normalization. As we are mostly interested in long-range interactions, the normalization step is mostly used to remove short-range bias.

Hi-C Thresholding

In Figure 3 we show what happens to the distribution of TRIP bin pair interactions when we introduce a threshold on the Hi-C interaction score. The black line depicts the percentage of Hi-C interactions which are at or above the threshold and would be an indicator for the expected number of TRIP bin pair interactions at or above the threshold. The other three lines (green, red and blue) depict the fraction of TRIP bin pair interactions at a certain Hi-C threshold. From the

plot we can see that the fraction of positive TRIP bin pairs above the threshold remains higher than the fraction of negative and mixed TRIP bin pairs. All bins that show interactions by HiC also show enrichment for all types of expression bin interactions (++,--, +-) up to the HiC thresholds indicated in table 1. The significance testing is performed by using the section of the matrix with interaction scores above the threshold as our sample size and compare the number of TRIP bin pair interactions of a certain type in this sample against the total number of TRIP bin pair interactions of the type in the entire matrix. The results of this test show that until the threshold values shown in table 1, the presence of TRIP bin pairs is significant with a p-value < 0.05.

TRIP bin pair combination	Hi-C threshold for insignificant presence
++	38
--	42
+-	51

Table 1: Hi-C thresholds for TRIP bin pair combinations which mark the point where there no longer is a significant number of TRIP bin pair combinations of a certain type left in the data.

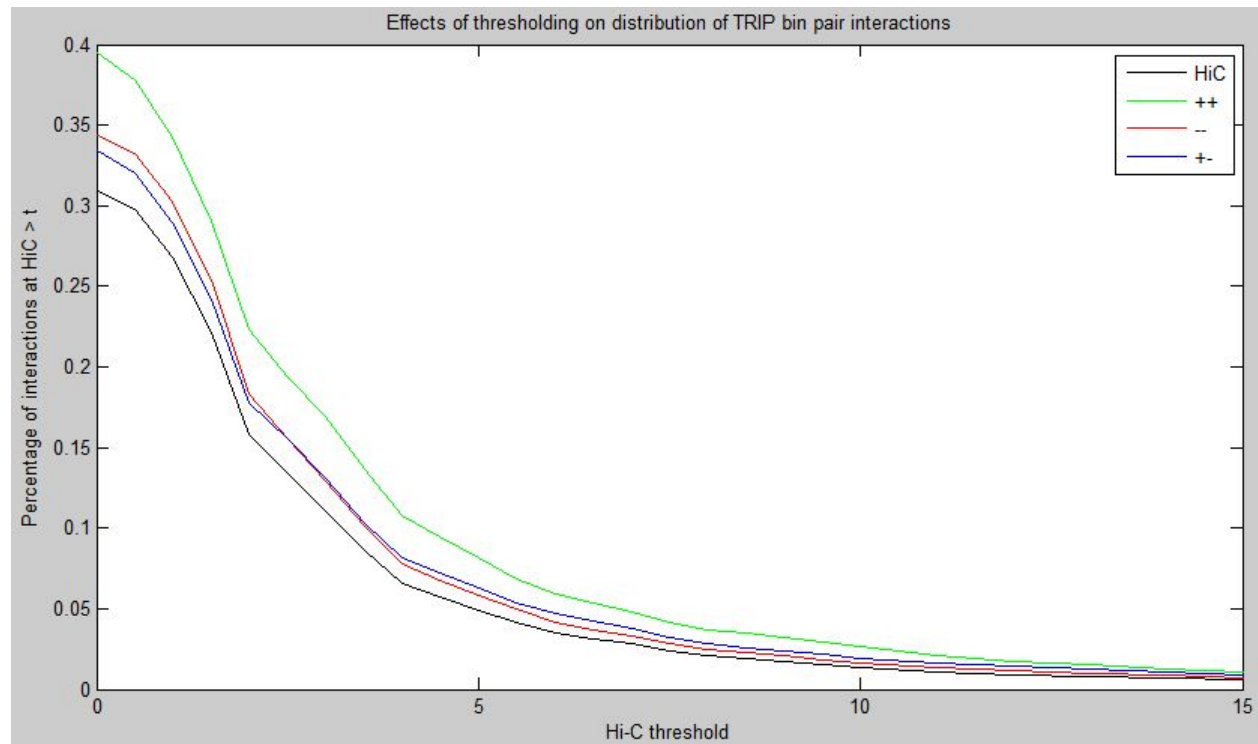


Figure 3: Effects of thresholding on distribution of TRIP bin pair interactions at Hi-C > t. Coloured lines show the fraction of TRIP bin pairs present at or above the thresholded Hi-C interaction score. It becomes clear that for all 3 combinations a higher fraction of interactions is present above the threshold than would be expected from random chance, which is depicted by the black line, the Hi-C interaction scores.

Even for high thresholds, we still see that there remains a significant presence of TRIP bin pair interactions, a threshold of 36 has the result that ~0.3% of the Hi-C interactions are at or above the threshold.

Graph construction

Converting the entire Hi-C matrix into a network would mean that even the smallest of Hi-C values are counted as true interactions. Though importance of all smaller values cannot be ruled out with certainty, some minimum level of values should be imposed to eliminate measurement errors or noise in our dataset.

We have opted to pick a threshold which keeps a substantial number of edges in the network, whilst at the same time maintaining connectivity between all nodes. We hypothesize that a large portion of interactions are valid, and aim to avoid disconnecting nodes from a biological point of view as unconnected nodes appear to poorly model the possible number of interactions in a genomic region based on proximity to other genomic regions. To this end we apply a low threshold on the Hi-C interaction scores of 5, which was found by analysing the effects of the introduction of a Hi-C threshold on the data. If the threshold is set to a high level the resulting network will contain disconnected nodes. For some algorithms sparse networks would make analysis easier, however this comes with the risk of disconnection, which is not preferable.

As discussed before, we apply a threshold on the Hi-C interaction matrix to filter out noise. Figure 4 shows the number of interactions for each individual genomic bin that remain, with there being no disconnected nodes. It has to be noted that the plot is truncated as genomic bins 77 and 78 have a much higher interaction count, respectively 1748 and 1176. We immediately identify that multiple regions within the genome see far more interactions than others. These may represent hub-like environments. From literature we find that hubs can be associated with for example transcription factories [27,30], gene-silencing [28] or active chromatin [29,30]. These hubs show that a single region can have a large number of interacting neighbours and possibly form dense subnetworks within the graph. This leads us to the idea of using graph clustering to filter out these dense networks.

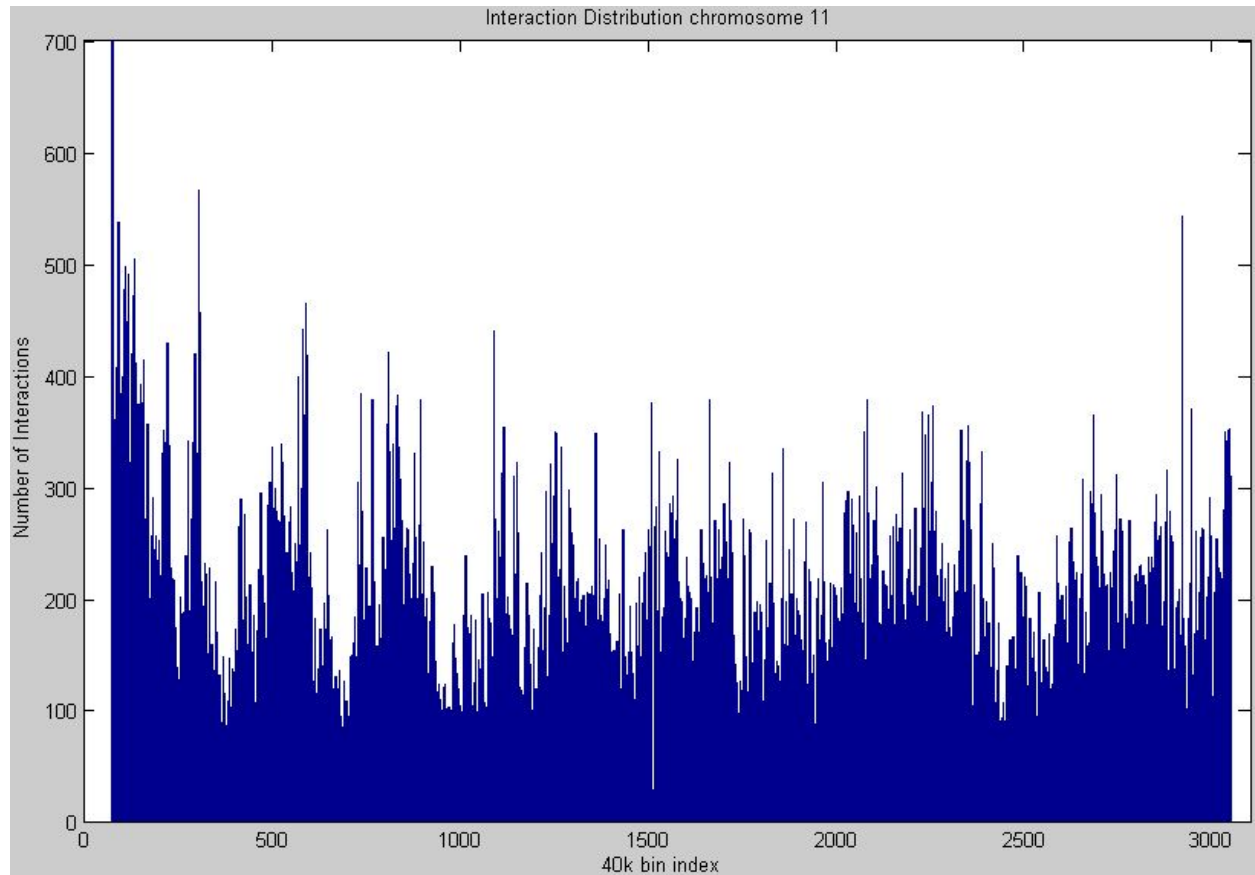


Figure 4: From the Hi-C interaction network we retrieve for every Hi-C bin the number of interactions. As can be seen we see a multitude of peaks, which are indicators of hub-like bins. The counts for Hi-C bins 77 and 78 are truncated and reach up to 1748 and 1176 respectively.

Clustering

Using the dense network, representing the folded genome within the cell, which we have created using Hi-C interaction data, we isolate the 3D environments for further analysis. The best way to isolate densely connected substructures within networks is the use of clustering techniques. In general clustering is used to segment networks, with sets of nodes showing certain properties such as a high connectivity compared to nodes which are not in the set. The reason for applying a clustering approach is that we have identified genomic hubs which can divide the broader network into smaller sub-networks.

There are a multitude of clustering techniques available in literature, each with their pros and cons. For this work we have evaluated several methods (using *Cytoscape* [19] to visualise the network and *Matlab* for the cluster creation and analysis), with varying results on our test chromosome (11). We briefly introduce each method below and explain their results.

Markov Clustering Algorithm (MCL)

The MCL algorithm was developed by *van Dongen* [7] and works by simulating flow in a network. It computes random walks and the process of expansion (longer paths) and inflation where walks within cluster environments will become more probable than walks outside the environment. Using iterations it will attempt to converge to the separation of clusters.

Using the Cytoscape implementation on chromosome 11 we found that it returns a single large cluster and a multitude of singular nodes. By varying the granularity setting we see some slightly different results (mostly small clusters of maximum 4 nodes), but we did not find the clusters of larger sizes. Large clusters are not probable due to the biological notion that a genomic locus cannot interact with a large number of other loci due to the way the DNA is packed in the nucleus. A possible explanation is the relatively low number of iterations used in combination with a non-sparse network, a necessary restriction to reduce the runtime. This may have left the algorithm unable to reach cluster sizes of our wish. Our cluster analysis required analysis of gene presence and therefore requires clusters of up to 25 nodes in size (which represents a total of 1 megabase worth of genomic regions which can interact). Clusters with 1-4 nodes cannot contain many possible interactions of genes and as research [9] has shown that genes do not necessarily have a direction connection, but can share common neighbours. These very small sets are likely not keeping these indirect connections intact and miss out on this concept.

Clique Detection

Cliques within a graph are sets of nodes which are fully connected. The main difficulty with clique detection is that it is a NP-complete problem. Combining this with the size of our network (2965 nodes and 229107 edges) and the only constraint of clusters being at or under 25 nodes, clique detection quickly becomes computationally expensive. Instead of looking for cliques we have tried an approach where for every node in the network the maximum neighbourhood component is calculated, which is similar to k-nearest neighbour algorithms. It takes the Hi-C interaction scores as a weight on the edge and locates clusters of nodes. A downside to this approach is that it is computationally expensive: chromosome 11 took 20 hours to be analyzed and resulted in only a few useable clusters, some of which were identical due to the fact that the algorithm started on every single node in the network. Due to the computationally expensive nature of this method we expect to see even longer runtimes for the chromosomes which are over twice the size of chromosome 11, making it an unfavorable method.

Community Finding

The community finding algorithm proposed by *Girvan and Newman* [10] tries to find subgraphs within the network which connect densely internally, but sparsely with other clusters. It works by calculating betweenness of all edges, removes the edge with the highest betweenness and subsequently recalculates the betweenness for all affected edges. Due to the computationally high burden of re-calculating betweenness after every edge removal, a more greedy variation of

the algorithm was developed by *Newman* [11], calculating a Q score which decides how much a given division in communities alters from randomness.

This means that the algorithm stops sooner based on a Q value threshold and needs less iterations and calculations. For chromosome 11 we find 6 communities with a Q value of 0.16 which lies below the regular threshold. Based on the size restrictions an altered algorithm was made which iteratively re-clusters all cluster results from the Girvan-Newman algorithm. These iterations continue until result clusters are reached with a maximum size of 25 nodes. For every cluster we calculate if the presence of TRIP nodes is significant. In comparison to other algorithms this gives us more control over the size of the communities we are looking for, as well as providing a decent runtime. Chromosome 11 was dissected in roughly 76 minutes, whereas for example the MCL algorithm was not completed within 4 hours, whilst not providing reasonable results when it did finish. The alternative to the clique approach was done in roughly 20 hours.

Also compared to Clique detection the algorithm is more suited, as communities do not require graphs to be fully interconnected, which would be very limited from a biological viewpoint. Groups of genes do not necessarily directly influence all other genes in the pathway. Within the nucleus the interacting genomic regions will unlikely form a perfect sphere where each region can freely interact with all other regions.

Cluster Significance

The reason why we want to classify our clusters as being part of a domain is that we can look for the possibility of genomic functionalities being present in these domains. The clustering algorithm produces a set of result clusters and how informative these clusters are is based on the distribution of types of nodes within each cluster. Informative clusters are clusters having a presence of TRIP bins which makes them likely candidates for further analysis. The first type is a cluster without any TRIP bins at all, from which we cannot deduce what kind of expression domain it lies in, as this is determined by the presence of TRIP bins. The second type contain some TRIP bins, but not enough to allow us to conclude that the cluster represents a certain domain, such clusters can also contain a mixture of some TRIP bins of both types. The third type contains a significant presence of TRIP bins, which is calculated using hypergeometric testing for the individual TRIP signs (+ or -). The test uses the full set of nodes as the population and the number of TRIP nodes of a certain sign as the number of success states within population. For the draw we take the nodes from a cluster and again use the number of TRIP nodes as the number of success states within the draw.

Based on the testing we state that a cluster which contains a significant number of positive or negative TRIP bins is part of a particular expression domain, based on the sign of the TRIP bins. If a cluster is both significant for positive *and* negative TRIP bins, this is an indication of a boundary between two expression domains.

GO Term Enrichment Analysis

All genes per chromosome were downloaded from the Biomart section of the *Ensembl* website [16] and provide us with the start- and end-base pair for each gene. The entire length of the gene is mapped onto the 40k bp bins we use for Hi-C and TRIP and this returns a list of bins a gene is present in. As shown in figure 5 this can lead to a situation where a gene is present in more than one bin. In order to conserve as much information as possible, the decision was made to assign a gene to every bin where it has a presence in, no matter how small the part of the gene is actually present within a bin.

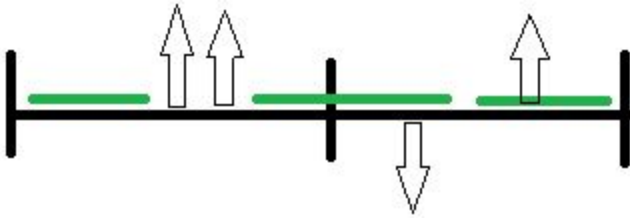


Figure 5: Illustration of Genes in combination with TRIP data points, by assigning a gene to all bins it has a presence in we conserve as much information as possible, due to the limited number of TRIP integrations in the data.

For every cluster we retrieve the gene list, which is imported into the *DAVID* web tool [15] and it returns a list of GO terms which are found to be enriched when comparing the gene list against the background genome (mouse embryonic stem cells, mm9 reference genome). For every GO term we employ a 25% FDR to decide if there is an actual enrichment or not.

Results

Hypothesis 1: Regions of the genome that are in close proximity will show similar expression patterns as they reside in the same '3D environment'

We reasoned that due the way the genome is packed inside a cell [2, 3, 17, 18], areas of the genome which might be linearly distant may come into close proximity. Following this it has been suggested that this happens with a specific purpose, for example to construct transcription factories in nuclei [13, 14]. In this case we expect to see that these regions show similar expression patterns, as they reside in the same '3D environment'. The '3D environment' is determined to be sections of the genome that share interactions based on proximity, as suggested by Lieberman-Aiden [3]. From a biological viewpoint we expect this hypothesis to be true, as it collaborates the idea of these mechanisms, as all interacting regions would show an increase in gene expression levels.

In order to test this hypothesis we perform a statistical analysis of the Hi-C interaction data contained in matrix m . Hi-C interactions can be classified as non-existing, where $m_{ij} = 0$, or

possible, with $m_{ij} > 0$. A higher score shows a higher certainty of the interaction happening in a cell. If regions tend show similar expression patterns, we expect to find statistically more regions of the same expression pattern at Hi-C interaction scores higher than zero. We found that ~40% of all ++ pairs and ~34% of all -- pairs are present in Hi-C regions with a score higher than zero. At the same time we find 33.44% of all mismatching pairs present in Hi-C regions with a score higher than zero. Based on the distribution of the Hi-C data (~31% of all data points have a score higher than zero) we see a higher presence of ++ pairs compared to -- and +- pairs. This shows that Positive TRIP bins are more likely to interact with other Positive TRIP bins and have a larger presence at higher Hi-C values on average compared to Negative TRIP bin pairs.

Hypothesis 2: Introducing a threshold on Hi-C interaction scores, to filter out noise and less probable interactions, will result in larger fractions of Positive (++) and Negative (--) TRIP bin pairs.

Due to noise inherent to all measurements, some interactions shown in the dataset may not necessarily occur in living cells. We wondered at which point we no longer find a significant presence of TRIP bin pairs of the same sign. By iteratively introducing a higher threshold on the Hi-C interaction score we find that up to a relatively high score, we keep seeing a significant presence of TRIP bin pair interactions. At the same time the number of the data points with an interaction score higher than the threshold decreases drastically. With no threshold we start with ~31% of the interactions having a score higher than zero. At a threshold of 5 this is reduced to only ~5%, at a threshold of 12 we have < 1% of the interactions left. By raising the threshold we reach a point where we no longer have a significant presence of TRIP bin pairs (38 for ++, 42 for -- and 51 for +-), at which point we have respectively 3, 5 and 15 interactions left in total. For the rest of this project we chose a threshold of 5, which reduces the amount of possible interactions by 95% whilst still keeping a densely connected network. As we saw back in figure 3, the introduction of a threshold does not result in larger fractions of TRIP bin pairs, but at all points there are larger fractions present at $\text{Hi-C} > \text{threshold}$ than expected from the distribution of Hi-C.

Finding that we have a significant presence of matching expression pairs leads us to the question if this is a random effect, or if the folding of the genome drives these similarly expressed pairs.

Hypothesis 3: The natural folding of the genome results in a network with more interactions between regions of similar expression patterns compared to non-natural (random) networks.

We saw enrichment for TRIP bin pair interactions in our network and wondered if similar patterns would show up if we simulate a different folding of the genome. This is tested using a randomization setup of our interaction matrix. In order to retain the general topology of the network we opted for a degree-preserving randomization approach, wherein every node retains its degree, but edges are randomized. In order to prevent alterations to the degree during the iteration, thresholds on Hi-C scores were introduced beforehand. In biological terms this means

that we shuffle genomic regions around, but making sure that every region retains the number of neighbours it had before the shuffle. The resulting network will retain a similar structure, but with the TRIP nodes shuffled around we expect to see less interactions between similar expression patterns.

The randomization algorithm uses a probability of 50% to redistribute an edge and iterates 4 times the number of edges over the matrix. During every iteration, two sets of interacting pairs, a-b and c-d, are selected at random and based on the probability the pairs are switched (a-c, b-d). The full algorithm was run 100 times to produce 100 random networks which still have a similar topology based on the original network. As the positive TRIP bin pairs were shown to be significantly enriched for all Hi-C values above zero as well as having a larger presence than the negative and mixed TRIP bin pairs, we tested our hypothesis by investigating the enrichment of Positive TRIP bin pairs for our randomized networks.

In figure 6 we show the result of this analysis in the form of a boxplot. As we can see, in general there is a spread in values, with a high number of networks not being significant for Positive TRIP bin pair presence. The significance values are again calculated using the same hypergeometric setup for the regular analysis. For every threshold level we calculate the significant presence of Positive TRIP bin pairs. At the same time we can see that for the second half of the threshold spectrum (starting at 3, the 7th plot) on average the distribution shows a mean around the significance threshold, meaning roughly 50% of the networks contain a significant presence of Positive TRIP bin pairs. With larger thresholds this fraction shrinks again. We conclude that the folding of the genome plays a very important part in the presence of interactions between genomic regions with a similar expression pattern, based on the large spread in values. Especially for low thresholds on the Hi-C interaction scores we find a large spread in the presence of a significant number of Positive TRIP bin interactions which indicates that the natural structure brings these regions together more often than seen from random generated networks.

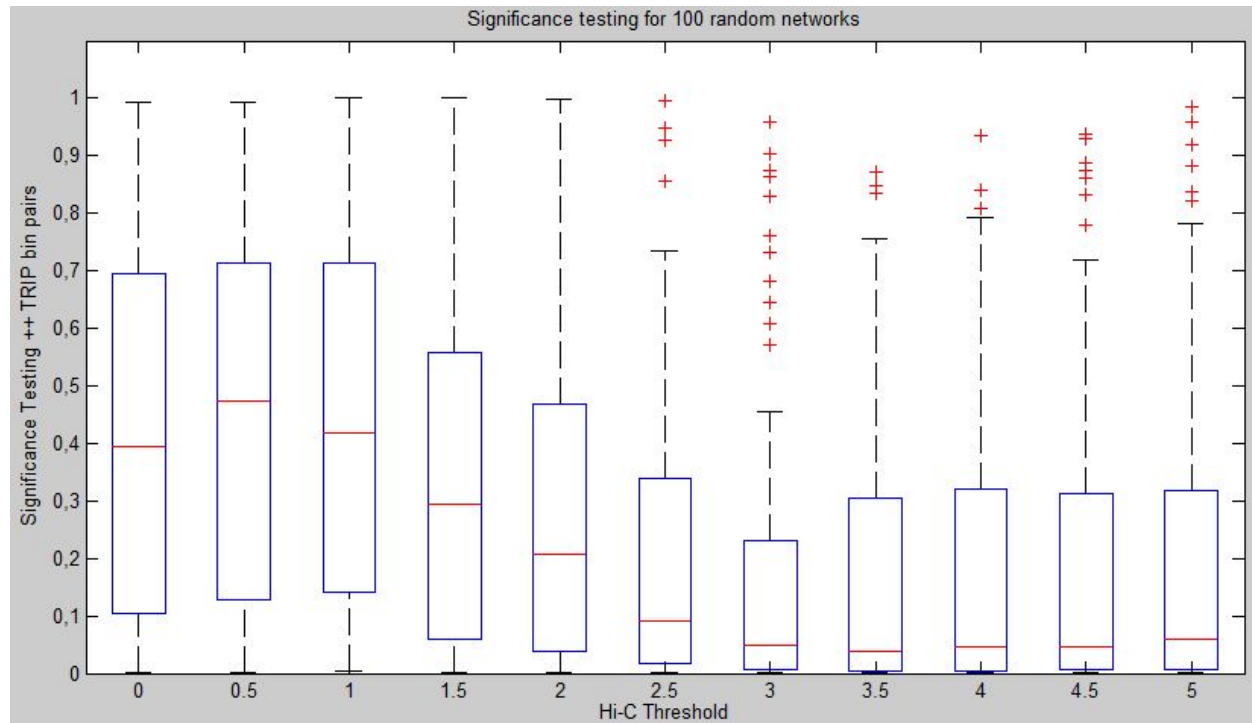


Figure 6: 100 random networks tested for significant presence of Positive TRIP bin pairs (++). For every network we calculate if the presence of Positive TRIP bin pairs is statistically significant for Hi-C thresholds of 0 - 5, with incremental steps of 0.5. Especially for the lower thresholds we see that there are hardly any significant presences of ++ TRIP bin pairs. After a threshold of 3 (entry 6) do we see the median closing in on the 5% significance threshold. However even then we see a large spread and conclude the natural folding is a more favorable folding for these interactions.

As we have shown that on a local pair-wise scale we see matching expression domains we wondered if these local expression patterns are part of larger domains, being 3D environments. The nodes in our network represent the domains and the connections between the nodes are the Hi-C interactions. We are interested to go to higher order interactions (higher than 2D). In the graph this is represented by a set of densely connected nodes and to find these we use graph clustering.

In Figure 7 we show the distribution of clusters for all chromosomes. As we can see all chromosomes except 20 (chromosome X) show a majority of clusters with no TRIP presence whatsoever. Larger chromosomes obviously result in a larger number of clusters. The number of significant clusters appears less linked, as for example chromosome 1 has half the number of significant clusters compared to chromosome 2 but is a larger chromosome. Furthermore chromosome 12 shows a fairly large number of significant TRIP clusters for its size.

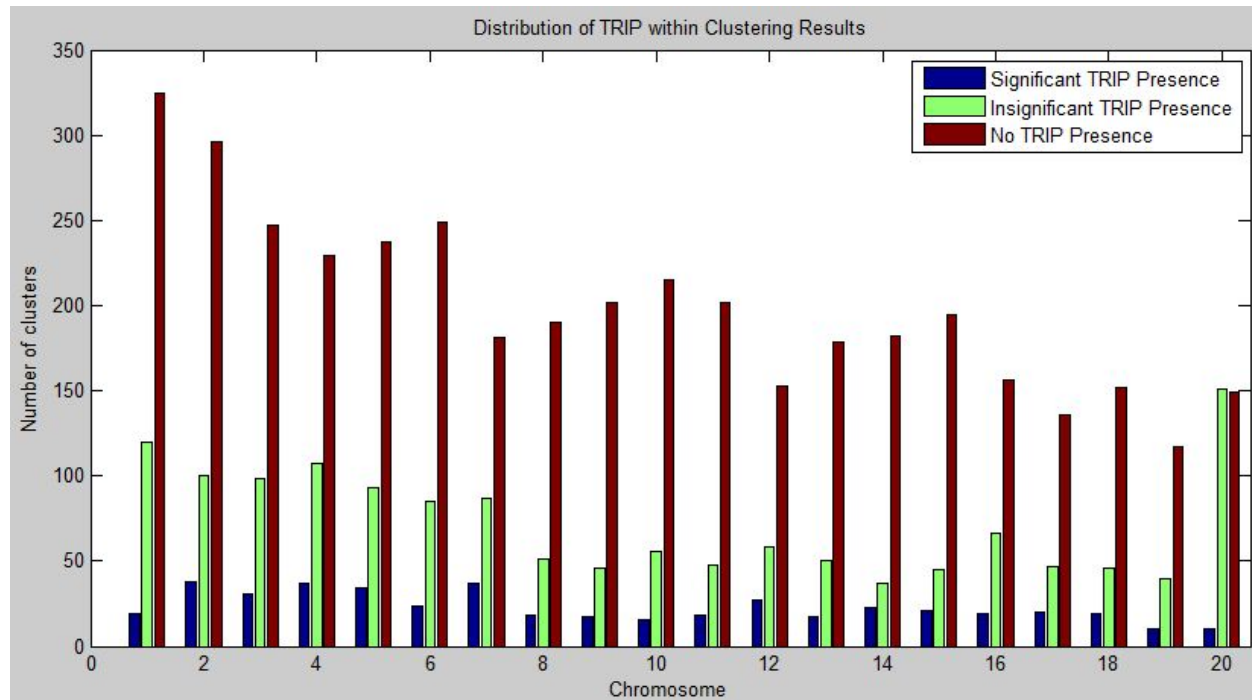


Figure 7: Clustering results for all 20 chromosomes. As we can see the number of significant clusters is fairly independent compared to chromosome size. The number of clusters with an insignificant number of TRIP nodes or no TRIP nodes at all is more in line with the size of the chromosome. Chromosome 20 has a curious result where the clustering results in more clusters with some TRIP presence compared to all other chromosomes.

In figure 8 we show a partial result of our clustering of chromosome 11, where individual clusters have been iteratively broken down until the threshold of 25 nodes is reached. Every node represents a cluster in hierarchical order, where all nodes linking to a node are subclusters of that particular node.

The coloring in the figure shows enrichment for TRIP nodes within each cluster and is a functionality used in Cytoscape to apply colour to a node based on the value. From the figure we learn that a large number of clusters do not show a significant presence of TRIP bins as well as that multiple clusters can be further broken down into smaller subclusters and still have a significant presence of TRIP bins. This is an indication that a single significant cluster contains multiple separate regions which are densely connected and each of these regions contains a significant number of TRIP bins.

The significance testing was done using hypergeometric testing. As input for the test we use the total number of nodes in the entire network, the total number of TRIP bins of a certain sign, the size of a cluster and the number of TRIP bins of a certain sign within the cluster. The test then shows if the number of TRIP bins in the cluster is statistically significant or not. The significance calculation has been done for both Positive and Negative TRIP bin presence and the colour represents the result. Green nodes represent clusters with an enriched presence of Positive TRIP bin pairs, Red nodes the enriched presence of Negative TRIP bin pairs and the yellow

node is a cluster enriched in both Positive and Negative TRIP bins. For example in chromosome 11 we find 268 clusters sized at or below 25 nodes. These individual clusters are, according to the algorithm, more densely connected and are thus clusters of interest and represent the 3D environments within the cell.

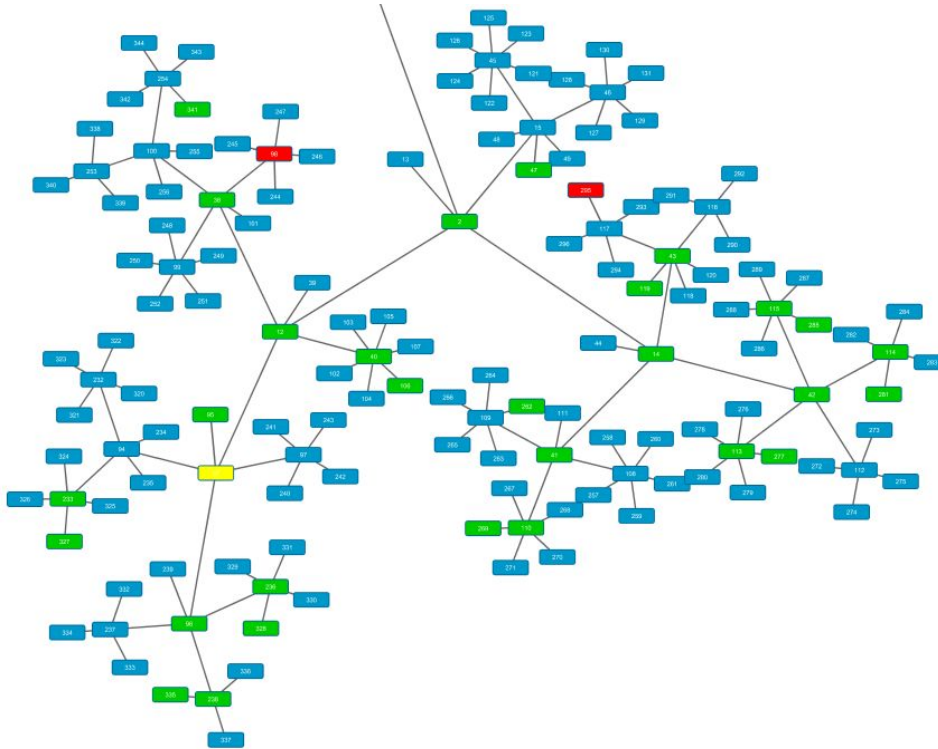


Figure 8: partial clustering result of Chromosome 11. The tree structure represents reclustering done by the algorithm and every node represents a clustering result. Green nodes are enriched for Positive TRIP bins, red nodes for Negative TRIP bins, the blue nodes are not enriched for either and the yellow node is enriched for both.

Hypothesis 4: Regions of the genome, which are close in proximity in the 3D environment, share biological functionalities based on present genes on these regions

By investigating whether the identified environments share biological functionalities, we aim to demonstrate that genes are organized on the genome in such a way that even though they are far apart on a linear level, their closeness in 3D environments allows interactions. This follows for example the suggestion of transcription factories [13, 14]. We perform GO enrichment analysis using the DAVID [15] online database where we compare the gene list of a cluster against the background genome, mouse embryonic stem cells, mm9. The expectation is to find a statistically significant enrichment, which show that these regions do not come into contact with each other based on chance, but rather share biological functions.

Based on our GO term enrichment analysis we found that from our 268 clusters, we have 154 clusters which have an enrichment in at least 1 GO term. Based on these findings we conclude

that our clustering has gathered nodes of similar functionalities together and that it is a representation of parts of the natural cell. At the same time we see that 114 clusters are not enriched for any GO term at all. A partial explanation can be that there are left-over clusters, i.e. a cluster is broken up into several enriched clusters and some nodes that did not fit in any other cluster and are still grouped together as well as clusters which are gene-deprived. By investigating if our identified GO term enrichments cover both TRIP and non-TRIP nodes we can make better predictions for these neutral nodes as to what domain they belong to. Furthermore we found that cluster size and the significant presence of GO term are positively correlated at 0.37. The size of the geneset for a cluster has a large correlation with GO term enrichment (0.41) and cluster size (0.71). It is interesting to see that larger gene sets will in general show more GO term enrichments, but to a lesser extent than would be expected, otherwise the correlation would be higher.

Amongst our findings we see that not all identified environments seem to be enriched with genes with known common biological tasks. Both large and small node sets, as well as large and small gene sets can result into enrichment of GO terms. An example of GO term enrichment is shown below in Figure 9 and Figure 10, for a cluster of 9 nodes with an enrichment of Negative TRIP nodes, as seen in the red nodes. We observe a neutral node (307) with a high number of interactions with the red nodes. By analyzing the GO term enrichments in the set we can find if the nodes that do not have a TRIP integrations (neutral nodes) can indeed be marked as being in the same non-permissive domain.

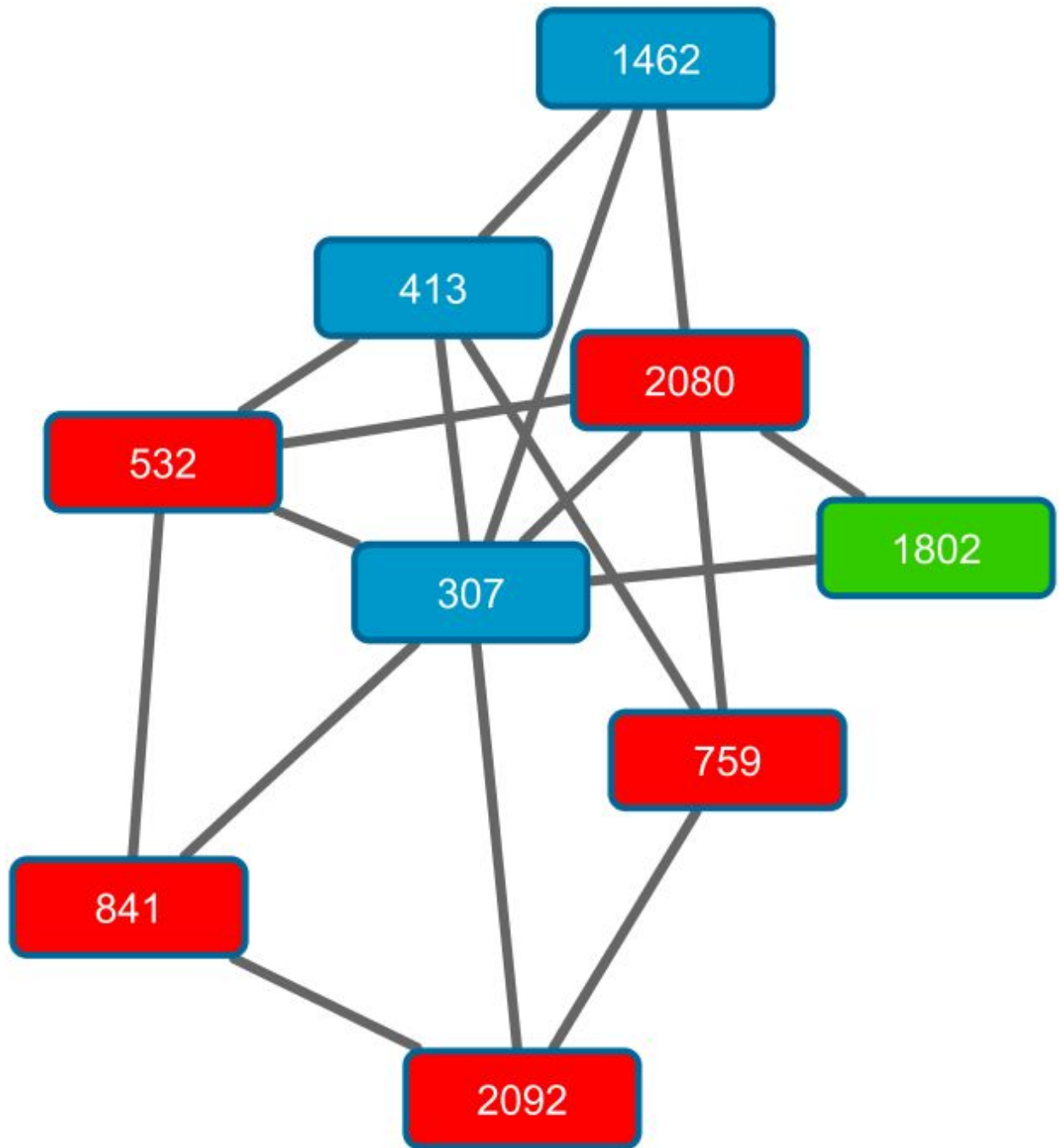


Figure 9: Result cluster 18 containing a set of 5 negative TRIP nodes, a single positive TRIP node and 3 neutral nodes. Red nodes are bins with an enrichment of Negative TRIP points, blue nodes have no TRIP enrichment. The numeric indicators are the bin ID's based on the linear genome. The cluster is enriched for negative TRIP nodes.

the result that if there is a mixed presence of TRIP nodes there are more often than not GO term enrichments, which could be attributed to either side of the TRIP sign. The presence of only one of the groups does not influence the presence of enriched GO terms, unless the TRIP bins have an enriched presence.

	GO Term Enrichment	No GO term Enrichment
Enriched for TRIP	17	0
Not enriched, but only has Positive TRIP presence	7	9
Not enriched, but only has Negative TRIP presence	66	67
Not enriched, has mixed TRIP presence	42	12
No TRIP presence at all	22	26

Table 2: GO Analysis shows that clusters with enrichment of TRIP also have enrichment for GO terms.

Discussion

In this work we have investigated 3D environments and a link between expression patterns and gene ontology enrichment. Using an iterative community finding algorithm on Hi-C interaction data we found a large selection of small communities which depict interacting regions on the genome which can be linearly distant and come in close proximity due to how the chromosomes are folded. Our goal was to find a division of the folded genome which indicates interactions between linear domains of permissive and nonpermissive chromatin domains in the form of 3D environments.

Using the found communities we performed gene ontology enrichment analysis and found that over half of our clusters have an enrichment of a biological function. We then wondered if there is a link between the presence of an enrichment and expression patterns and found that there is a possible link between positive expression patterns and biological functionalities. At the same time we found that there is also a very small correlation between negative expression patterns and these biological functionalities. A strong negative correlation (-0.54) was found between the significant presence of Positive TRIP bins and the number of genes in a bin in the enriched set and a weaker negative correlation (-0.36) for the unenriched set. This shows that the presence of TRIP bins is linked to the presence of gene sets which are enriched within the environment.

The size of our environments share a direct correlation with the enrichment of GO terms and the number of genes, with larger environments having an enrichment more often as well as larger gene sets having an enrichment. Nonsignificant 'pure' clusters (only one of the TRIP types present) do not show a nonrandom distribution of GO enrichments, at the same time 'mixed' clusters (where both positive and negative TRIP bins are present) do show an increase in GO enrichments.

Our clustering results likely include 'residue' clusters, which are less informative subclusters obtained by breaking down a larger cluster. We reasoned that larger clusters can contain multiple clusters of interest and naturally this will also result in sets of nodes which are not interesting, but do end up in our resultlist. We found a possible link between the enrichment of TRIP nodes within a cluster and the presence of enriched Gene Ontology terms. Further research can be done to see if there is a link between a specific type of TRIP node and the types of enrichment more prevalent for these types.

Acknowledgements

This work was carried out as part of the BioInformatics Master track at the Delft University of Technology under the supervision of Jeroen de Ridder (TUD) and Lodewyk Wessels (NKI). During the course of the project they have provided valuable feedback and ideas on how to tackle this project. The primary TRIP data source was provided by the NKI and Waseem Akhtar and Johann de Jong helped by providing insight in the data.

References

- [1] Akhtar, W., et al, 2013, Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel
- [2] Lieberman-Aiden, E., et al, 2009, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.
- [3] Cadiñanos and Bradley, 2007, Generation of an inducible and optimized piggyBac transposon system.
- [4] Meir et al., 2011, Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy
- [5] Boulos, R.E., et al, 2013, Revealing Long-Range Interconnected Hubs in Human Chromatin Interaction Data Using Graph Theory.
- [6] Fillion, J., et al, 2010, Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells.
- [7] Stijn van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.
- [8] Sandhu, et al, 2012, Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks

- [9] Babaei, et al, 2015, Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex
- [10] Girvan, M. and Newman, M.E.J., 2002, Community Structure in social and biological networks
- [11] Newman, M.E.J., 2003, Fast algorithm for detecting community structure in networks
- [12] Balcan, M., Gupta, P., 2010, Robust Hierarchical Clustering
- [13] Rieder, D., et al, 2012, Transcription Factories. *Frontiers in Genetics*
- [14] Papantonis, A, 2013, Transcription Factories: genome organization and gene regulation.
- [15] DAVID webtool kit, <http://david.abcc.ncifcrf.gov/>
- [16] Ensembl BioMart, <http://www.ensembl.org/biomart/>
- [17] Sexton, T., et al, 2012, Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome
- [18] Peric-Hupkes et al., 2010, Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.
- [19] Cytoscape, <http://www.cytoscape.org/>
- [20] de Jong, J., et al, 2014, Chromatin Landscapes of Retroviral and Transposon Integration Profiles
- [21] Rao, S., et al, Cell 2014, A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping
- [22] Gierman, et al., 2007, Domain-wide regulation of gene expression in the human genome
- [23] van Heyningen, V., Bickmore, W, 2013, Regulation from a distance: long-range control of gene expression in development and disease
- [24] Annunziato, A. (2008) DNA Packaging: Nucleosomes and Chromatin. *Nature Education* 1(1):26
- [25] Gilbert, D. and Fraser, P, 2015, Three Dimensional Organization of the Nucleus: adding DNA sequences to the big picture
- [26] Piggybac Transposon System, <https://www.systembio.com/piggybac>
- [27] Razin, S., et al, 2011, Transcription factories in the context of the nuclear and genome organization
- [28] Shevelyov, Y. and Nurminsky, D., 2011, The nuclear Lamina as a Gene-silencing Hub
- [29] Razin, S. et al, 2013, Communication of genome regulatory elements in a folded chromosome
- [30] Zhou, G-L., et al, 2006, Active Chromatin Hub of the Mouse α -Globin Locus Forms in a Transcription Factory of Clustered Housekeeping Genes.
- [31] Cournac, A., et al, 2012, Normalization of a chromosomal contact map.
- [32] Yaffe, E., Tanay, A., 2011, Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.
- [33] Pennings, s. et al, 1994, Linker histones H1 and H5 prevent the mobility of positioned nucleosomes.