Functional role of DNA spatial organization

Chaimae Mrabet

Department of Physics, Khalifa University

Abstract

This project discusses the functional role of genome organization in regulating gene expression and coordinating biological processes, with a focus on the spatial organization of chromatin in budding yeast. Two methods, microarray and RNA-seq, were used to analyze gene expression and spatial organization data, and GO enrichment analysis was used to investigate the relationship between chromosome structure and biological function. The study found that genes in spatial proximity cross-link together more often when they have similar expression profiles or are functionally related. The spatial organization of the yeast genome was found to be non-random and facilitates the coordinated expression of functionally related genes. RNA-seq gene expression data confirms the results, providing further evidence for the significant relationship between co-expression of genomic loci and proximity in nuclear space. The findings have implications for understanding the mechanisms of gene regulation and transcriptional control, with potential applications in disease diagnosis and therapeutic strategies.

Keywords: Bioinformatics, Saccharomyces cerevisiae, RNA-seq, Microarray, 4C analysis, Hi-C

1. Introduction

The recent advances in functional genomics studies and chromatin conformation-capturing techniques have provided many ways to study and analyse DNA structure and nuclear organization. These technologies allow a better understanding of the relationship between chromosome structure and biological function [1]. Furthermore, These techniques have provided new insights into the mechanisms of gene regulation and transcriptional control, and shed light on the importance of the spatial organization of the genome in these processes [2]. As these techniques continue to evolve and improve, we can expect to gain even deeper insights into the mysteries of the genome and its role in health and disease.

In recent years, researchers have been intensively studying the importance of genomic architecture and the arrangement of the genes within the genome. As well as the reasons behind the formation of this typical organization. It has been confirmed that

the 3D organization of the eukaryotic genomes and their chromosomal conformation plays a crucial role in the gene activity and its mechanism of working, which can be proved by looking at the chromosomal spatial distribution and the clustering of diverse genomic regions with similar expression patterns [3][4][5][6]. It became evident that the special arrangement of genes and chromosomes is non-random, and genomes tend to have specific organizations and regions during their cell cycle that is more efficiently transcribed [7]. The regulation of transcription in eukaryotes is a complex process that involves several levels of coordination. This process begins with the binding of regulatory proteins to specific DNA sequences, which can either enhance or repress transcription. These regulatory proteins can interact with each other to form protein complexes, which can further modify the activity of transcription factors [8][9].

Furthermore, the 3D organization of the genome plays a crucial role in regulating transcription. The way that DNA is packaged and organized in the nucleus can determine which genes are accessible to transcriptional machinery and which ones are not. Recent research, such as the study on the 3D organization of the yeast genome and its correlation with co-expression and functional relations between genes, has demonstrated the importance of this 3D organization in transcriptional regulation. Researchers found that genes with similar expression patterns tended to be located in close proximity to each other in the 3D space of the nucleus. This suggests that the 3D organization of the genome reflects functional relationships between genes and can be used to predict gene expression patterns. Additionally, it was found that genes that were physically clustered in the genome tended to be co-regulated, indicating that the 3D organization of the genome is directly involved in regulating transcription[4].

Studies using several experimental methods have provided unprecedented insights into the relationship between the co-expression of genes that are in spatial proximity which further showed a relation with genes' functional properties. For example. It was shown that regions like centromeres, telomeres, and tRNAs tend to be co-localized [8]. Also, co-localization was presented for highly expressed genes, genes that are functionally related (based on gene ontology, GO terms), and co-regulated genes. Moreover, genes with shared expression levels tend to be in special proximity to facilitate their co-transcription by shared transcription features and optimize chromatin remodeling [9]. The three-dimensional arrangement of DNA can influence the accessibility of genes to regulatory proteins and RNA polymerase, impacting their expression levels [10]. To gain insights into the functional role of DNA spatial organization, researchers often employ Gene Ontology (GO) terms and enrichment analysis. These methods allow for the identification of functional terms and biological processes that are associated with genes or gene sets, providing valuable insights into the underlying mechanisms of gene regulation and expression [8][11][12]. Enrichment analysis can be used to identify GO

terms that are associated with genes that are spatially co-localized within the nucleus, giving information about the functional roles of different nuclear compartments and their interactions.

Various high-throughput technologies have emerged for the analysis of the 3D organization of the genome, including 4C analysis, RNA-seq, and microarray. These technologies have provided valuable insights into the relationship between the 3D organization of the yeast genome and the co-expression and functional relations between genes.

The recent advances in chromatin conformation capture (3C) techniques and its derivative (4C and HI-C) allowed for more accurate investigation and description of genomic spatial organization. These techniques produced a huge amount of genomic data which can provide us with a better understanding of the cellular function and gene regulation in the 3D genome architecture. They enable us to study the chromosome organization in eukaryotic cells by looking at chromatin interactions in gene contact networks[13][14].

RNA-seq is another widely used technology for analyzing gene expression. It involves sequencing of RNA molecules in a sample, providing information about the expression levels of individual genes [15]. RNA-seq has been used to study the relationship between gene expression and the 3D organization of the yeast genome. In this project it will be shown that genes that are co-expressed tend to be located close to each other in the 3D space of the genome, suggesting that the spatial organization of the genome is important for gene co-expression.

Microarray analysis is also a powerful tool for analyzing gene expression. It involves the hybridization of labeled RNA or DNA to a microarray containing thousands of probes that are specific to individual genes[1]. Microarray analysis has been used to study the relationship between the 3D organization of the yeast genome and gene expression. and the results will show that genes that are located close to each other in the 3D space of the genome tend to have similar expression profiles.

The spatial arrangement of genes is a crucial factor to consider in bioinformatics studies, as it can have a significant impact on gene expression, regulation, and disease development. It can provide insights into the molecular mechanisms underlying various diseases. This is particularly true in diseases such as muscular dystrophy and Rett syndrome, where the conformation of the genome in the nucleus and rearrangements of the chromosome play a significant role. Mutations that affect the spatial organization of the genome can lead to abnormal gene expression and regulation, contributing to the development of these diseases. Additionally, the creation of the antibody repertoire during immunological development involves the rearrangement of the chromosome, and mutations that affect the spatial organization of the genome can disrupt this process,

leading to immune system dysfunction. Understanding the spatial organization of the genome is therefore critical to gaining important insights into the molecular mechanisms underlying various diseases and contributing to the development of effective treatments [8].

The project investigated the functional role of intra-phase genome organization by analyzing the spatial organization of chromatin in cell and gene-expression datasets using computational tools and software programs. The results showed that the measured expression levels of genes with contact links were highly correlated, indicating a significant relation between the co-expression of genomic loci and proximity in nuclear space. Additionally, the correlation was higher when genes had stronger links or higher count frequency. Further analysis demonstrated that genes in spatial proximity cross-link together more often when they have similar expression profiles or are functionally related. Results also demonstrated the enrichment of inter-chromosomal links connecting loci of genes with the same GO term. This suggests that the spatial organization of the yeast genome is non-random and facilitates the coordinated expression of functionally related genes. These findings provide further evidence for the functional role of genome organization in regulating gene expression and coordinating biological processes. The project's algorithm and code can be used to further investigate the relationship between genome organization and its role in gene expression and regulation and between the coordinated expression of functionally related genes in spatial proximity.

2. Methodology

This section will discuss our approach to analyse the experimental data of intra-chromosomal contacts in the budding yeast found by Duan et al. As well as identify the groups of genes that are co-expressed and may be involved in common biological processes.

2.1. ***Enrichment Analysis on Gene Sets using GO***

One of the primary applications of the Gene Ontology (GO) is to conduct enrichment analysis on sets of genes. This involves determining which GO terms are over-represented (or under-represented) based on annotations for a given gene set. For instance, if a group of genes is found to be up regulated under specific conditions, an enrichment analysis using GO can identify the relevant over-represented GO terms.

To gain insights into gene functions and their relationships, The enrichment analyses was directly performed from the home page of the Gene Ontology Consortium (GOC) website, which connects to the analysis tool from the PANTHER Classification System [16].The PANTHER Classification System is a comprehensive tool for biologists

to analyse genome-wide data from sequencing, proteomics, or gene expression experiments. It combines gene function, ontology, pathways, and statistical analysis tools, and is built using 82 complete genomes organized into gene families and subfamilies. The PANTHER system is maintained up-to-date with GO annotations, allowing for accurate and relevant analysis of gene sets [17].

The GO enrichment analysis tools were implemented as follows: Firstly, the tool was provided with the names of genes to be analyzed, which were obtained by forming gene pairs from the analysis of 4C data. To ensure the accuracy of the analysis, the gene pairs list was filtered using Linux commands to only include unique genes. The resulting list was then used for the GO enrichment analysis. Next, the GO aspect (molecular function, biological process, cellular component) was selected for the analysis, based on the species of the genes used in the project, which was Saccharomyces cerevisiae.The results table provided a detailed summary of significant shared GO terms or their parents that were used to describe the entered set of genes. It included information on the background and sample frequency, expected p-value, over/under-representation of each term, the Fold Enrichment of the genes observed in the uploaded list over the expected, and p-value [16].

Background frequency and sample frequency were defined as the number of genes annotated to a GO term in the entire background set and the number of genes annotated to that GO term in the input list, respectively. If there were more genes observed in the uploaded list than expected for a particular biological process, it indicates an over-representation (+) of genes. Conversely, if there were fewer genes observed than expected, it indicates an under-representation (-) of a term. In other words, If the Fold Enrichment(for a particular GO term is greater than 1, it suggests that the genes associated with that term are more common in the set of deferentially expressed genes than would be expected by chance [16].

P-value was defined as the probability or chance of seeing at least x number of genes out of the total n genes in the list annotated to a particular GO term, given the proportion of genes in the whole genome that are annotated to that GO Term. That is, the GO terms shared by the genes in the list were compared to the background distribution of annotation. The closer the p-value was to zero, the more significant the particular GO term associated with the group of genes was (the less likely the observed annotation of the particular GO term). A Small p-value indicates that the result is non-random and potentially interesting. P-Values were calculated by the Binomial statistic using the following equation:

$$P - value = \sum \binom{K}{k} p(c)^k (1 - p(c))^{K-k}$$

It is a statistical measure that helps determine the probability of obtaining a certain number of successes (i.e., genes annotated to a particular GO term) out of the total number of trials (i.e., genes in the user's list) based on the proportion of successes in the population (i.e., genes in the whole genome with GO annotations) [16].

The data was pre-processed using linux commands to include specific GO terms, and a Python code was used to interpret the results. The code incorporated both the ratio of background frequency to sample frequency and p-values to generate a heatmap. The code used for this analysis is accessible through the following link https://github.com/chaimae-mr/Go-term-analysis.git.

2.2. ***GO-slim terms Enrichment Analysis***

To determine the enrichment of GO-slim terms, the method involves counting the number of contacts (4C links) between all the genes belonging to each term and comparing it to the number expected for gene interactions that do not depend on functional category. The first step was to find a unique list of loci and their corresponding genes with a 500 offset. This list, along with the list of interacting loci from the 4C data analysis, was used in a C code to find all the interacting genes and save them to a file. This was done multiple times each time specifying a different frequency threshold (e.g., >=5).

The following analysis was then performed. First, all genes associated with a given GO term were found. Then, the number of links between genes in the original set was calculated. Following that, the expected number of links between the simulated set of genes was also calculated. The expected number of links was obtained through Monte Carlo simulations, where for each term, 100 groups of genes are randomly selected from the genome. The number of genes in each random group was equal to the number of genes annotated by the term of interest. Finally, the standard deviation of the simulated set of genes was computed. The 4C links are counted between all pairs of genes in each randomly selected group, and the average and distribution over the 100 simulations define the expected statistical properties of links for each GO category. All of the results were output to a file, including the GO term, the number of genes, the expected number of links, the simulated number of links, and the standard deviation. This allowed for easy access and interpretation of the results. The results were interpreted in heat maps based on the mean (the number of expected links divided by the number of simulated links) and z-score The source code along with examples for this analysis is available at the following link https://github.com/chaimae-mr/Go-term-analysis.git.

2.3. ***Yeast Microarray Gene Expression Data and Co-Expression Analysis***

The experimental data for genome-wide contacts in yeast were obtained from the work of Duan et al [10]. The experimental process of generating 4C contact data in-

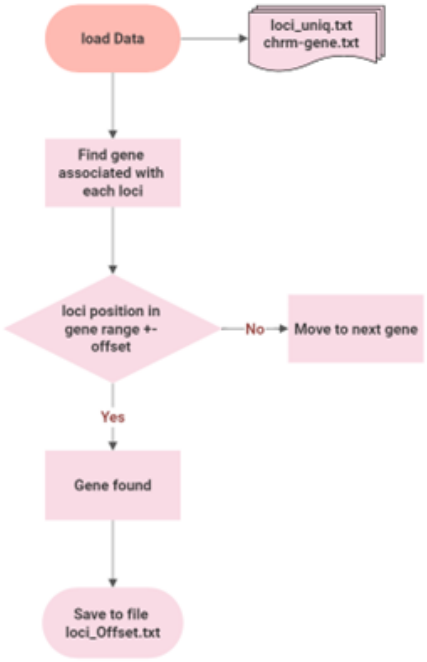
volves several steps. Firstly, the chromatin of interest is cross-linked with formaldehyde to capture spatially proximal genomic regions. The cross-linked chromatin is then digested using a restriction enzyme such as HindIII, which cuts the chromatin at specific recognition sites [8] [14]. Next, the digested chromatin is ligated together to form circular DNA fragments(fragments from both ends of the cross-linked interacting DNA pair), which contain the interacting genomic regions. The circularized DNA fragments are then amplified by PCR using locus-specific primers, which are designed to target the genomic region of interest. The resulting PCR products represent a library of DNA fragments that contain the interacting regions, which can be sequenced to obtain 4C contact data. The sequencing data is then processed and analyzed to identify the genomic regions that interact with the locus of interest, and to quantify the frequency of interactions between each pair of regions . The frequency is the number of sequenced fragments for each contact and is reported as the "count frequency" that is interpreted as a measure of spatial proximity between genomic loci [8] [14].

To ensure the accuracy and reliability of 4C contact data, quality control and normalization procedures are typically employed. For instance, PCR amplification efficiency and sequencing depth can introduce technical biases that need to be accounted for in the data analysis. Also , to eliminate pair reads that cannot be mapped to the genome or those that were ligated randomly in the experimental process Therefore, it is essential to include appropriate controls and normalization strategies to obtain high-quality 4C contact data [8].

Furthermore, The 4C data identifies spatial contacts in the yeast genome, including both intra- and inter-chromosomal links. The co-expression of genes on the same chromosome may be affected by cis-effects. And so, this analysis focuses on intrachromosomal links to eliminate this influence. The yeast gene expression data were obtained from the Gene Expression Omnibus (GEO) website, which is a public repository of gene expression data. Specifically, the data was collected using the Affymetrix yeast platform S98, which is a microarray technology that allows for the simultaneous measurement of the expression levels of thousands of genes. The dataset covers a wide range of experimental conditions, and a total of 1496 samples were used in the analysis [18].

To prepare the data for analysis, the samples were first normalized by converting the raw data to the linear scale and then dividing by the sample mean. This normalization step helps to correct for technical variation between samples, ensuring that the data is comparable across different experiments. To quantify the co-expression of two genes, the Pearson correlation coefficient was calculated between the corresponding probes across all 1496 samples. The Pearson correlation coefficient measures the linear relationship between two variables, in this case, the expression levels of two genes. A high correlation coefficient indicates that the expression levels of the two genes are closely related, while

(a) Workflow for calculating the average correlation between gene pairs



(b) Identifying Genes within a Specific Genomic Separation: Workflow for File Creation and Characterization by Offset Value

Figure 1: This presents the main working principle of the developed codes.

a low correlation coefficient indicates little or no relationship. This calculation was performed for all pairs of genes in the dataset, allowing for the calculation of the genome-wide correlation average, which provides a measure of the overall co-expression pattern in the yeast genome. Additionally, the average correlation for linked genes was calculated at different contact thresholds (offsets).

The code for this analysis can be found in the given linkhttps://github.com/chaimae-mr/SeniorProject-II.git. The following describes the main working principle to find the average correlation for linked genes and interpret the results: ( )

- The list of locus pairs that are in contact, along with their position and frequency count information was obtained. It was created using the 4C analysis method based on the HINDIII library of experimental data.
- The 4C list was noticed to have 240,628 loci pairs in which there are only 4010 unique loci values. Using Linux commands, these values were extracted for better and more efficient data processing.
- A list of genes was obtained from the GEO Website of the gene's position in the chromosome and coordinate range. the inter-chromosomal contacts were processed by mapping them to the corresponding genes.

- All the genes that fall within a specific genomic separation (offset) from the contact position were identified and saved into files. These files are characterized by the offset value. A simple workflow of this step is shown in Figure 1b.

- The co-expression (correlations between expression levels) of the corresponding genes of the locus pair was then determined, and the average correlation at that specific offset for the list of locus pairs was calculated. A simple workflow of this step is shown in Figure 5.

- The previous step was performed multiple times for different offset values ranging from 100 to 10,000 bp to understand the dependence of average correlation on the size of the offset.

- The results were then interpreted using Python packages like matplotlib . Average correlation vs. offset and Average correlation vs. count frequency (specifically at 500 offset)

2.4. ***RNA-seq gene expression data and co-expression analysis***

RNA-seq is a high-throughput sequencing technique used to investigate and quantify gene expression at the transcriptome level. RNA-seq works by converting RNA molecules into complementary DNA (cDNA) fragments, which are then sequenced using high-throughput sequencing platforms such as Illumina [19]. The resulting sequence data is then mapped to a reference genome or transcriptome to determine the abundance of transcripts in the sample. we performed RNA-Seq data analysis using a pipeline that involves several steps. The used pipeline can be accessed through this linkthttps://github.com/chaimae-mr/SeniorProject-II.git.

In order to obtain the RNA-seq expression data for S. cerevisiae, the Gene Expression Omnibus (GEO) database was used [18]. The data was downloaded from three GEO platforms, each containing different experimentation conditions and various samples for each series. A total of approximately 1500 SRA accession IDs were randomly selected and downloaded. The RNA-seq data was processed using a collection of tools known as the SRA-Toolkit. This software suite, designed by the National Center for Biotechnology Information (NCBI), provides a means of accessing and utilizing sequence data stored in the Sequence Read Archive (SRA). The SRA-Toolkit was used to download and retrieve SRA data in the compressed SRA format. The main two commands used are prefetch and fast-dump. To download the sequence files in compressed SRA format, we used the "prefetch" operation with the SRA accession IDs obtained from GEO, while fastq-dump retrieved the SRA fastQ files, which were used for mapping and transcript quantification. For gene quantification, the RNA-seq by Expectation

Maximization (RSEM) package tool was used, which is an open-source software tool for gene quantification using single-end or paired-end RNA-seq data. RSEM uses the powerful and efficient alignment software Bowtie to map the fastQ files to the yeast reference genome [19].

Two main functions of RSEM were used in our analysis: rsem-prepare-reference and rsem-calculate-expression. First, the yeast reference genome was prepared by obtaining the FASTA-formatted file from Ensembl genome browser release 82. Then, RSEM was used to calculate the gene expression levels from the mapped reads, with the output given as gene read counts. The read counts were normalized using two different methods, Transcripts Per Kilobase Million (TPM) and Fragments Per Kilobase Million (FPKM) [19]. To create a matrix table of genes (rows) and sample counts (columns), all the read counts according to their gene correspondence were merged. Using the Output file that consists of 6041 genes each has 1500 read counts and was normalized by (TPM) method, the correlation coefficient was calculated between each pair of genes. The following link https://github.com/chaimae-mr/SeniorProject-II.git includes the code that was used to process gene expression data and calculate the pairwise correlation coefficients (co-expression) between each pair of genes using the Pearson correlation coefficient method.

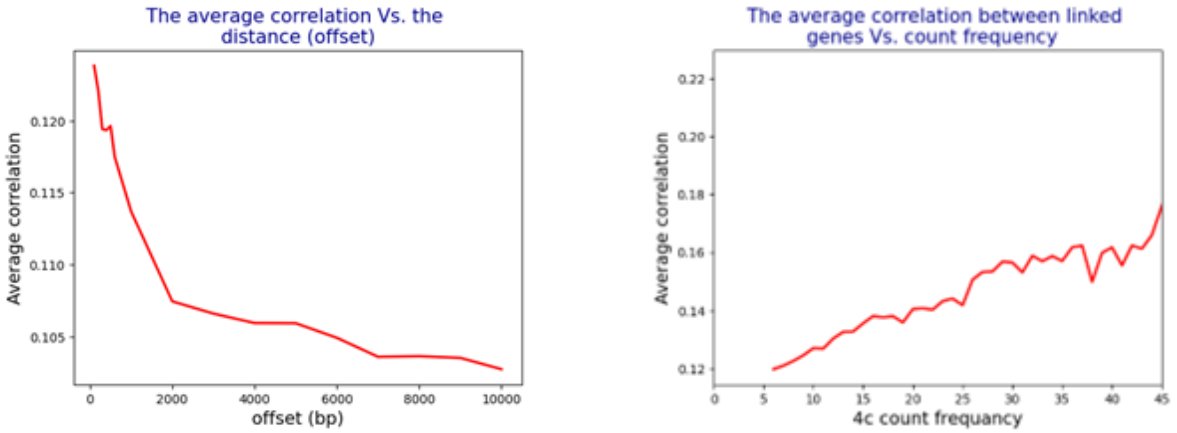
The output file consisted of 18,243,820 pairs of genes with their standardized correlation. This was then used to perform the co-expression analysis using 4C contact data, the same way it was done using microarray gene expression data.

3. Results

3.1. ***Inter-chromosomal Contacts and Co-expression of Genes***

To investigate the relationship between co-expressions of interacting genes and inter-chromosomal contact distance, we conducted an analysis of the correlation coefficient between interacting loci and the whole-genome average. We measured the average Pearson correlation coefficient of expression level between genes using 1496 Affymetrix Yeast S98 microarray samples covering a wide range of experimental conditions obtained from the GEO database. Using 4c experimental data from the HINDIII library, we calculated the average correlation of gene pairs located within a specific offset between inter-chromosomal contacts. Figure 2a shows the correlation coefficient plotted against different offset sizes to examine the effect of the offset size on the correlation coefficient.

Results shows that the correlation between linked genes was highest for small offsets (<500 bp) and decreased as the offset increased. Additionally, we found that the correlation coefficient between linked genes (0.1196) was significantly higher than the genome-wide average (0.0843), indicating that linked genes have a higher correlation than genes located farther apart on the genome.



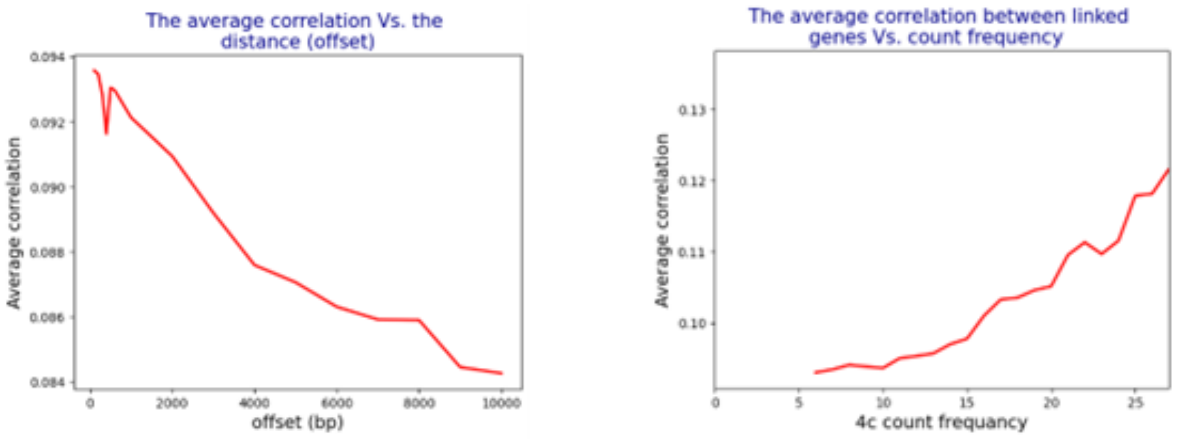
(a) (b)

Figure 2: (a) The average correlation of genes as a function of the distance (offset) from an inter-chromosomal contact using HINDIII experimental data from 1496 Affymetrix Yeast S98 microarray samples .(b) The average correlation between linked genes as a function experimental count frequency threshold of the corresponding links.

Furthermore, we also examined the relationship between gene co-expression and proximity in the nuclear space by analyzing the average co-expression for genes connected by links as a function of the threshold count frequency in the 4C experiment Figure 2b. The results indicates that there is a significant association between gene co-expression and proximity in the nuclear space, which is demonstrated by the increase in correlation with the frequency of the experimental fragment count for the contact. In other words, the correlation between linked genes increases monotonically with the frequency of the experimental fragment count for the contact. This implies that the likelihood of gene co-expression is higher for genes that are in close proximity to each other in the nuclear space, as indicated by the higher correlation between linked genes.

To confirm our results, the same analysis was performed on different data-sets using RNA-seq gene expression data instead of yeast gene expression data from Microarray to calculate the average correlation between linked genes. As shown in 3a , the correlation between linked genes was highest for small offsets (< 500 bp) and decreased as the offset increased. Furthermore, we discovered that the correlation coefficient between linked genes (0.0930) was significantly higher than the genome-wide average (0.0843) which is comparable to the genome-wide average (0.0841) from Microarray. Figure 3b also illustrates that the correlation between linked genes increases monotonically with the frequency of the experimental fragment count for the contact, conforming that the average correlation between linked genes will be higher when genes have higher count frequency and stronger links.

These results successfully validate the findings reported in, which analyzed the rela-



(a) (b)

Figure 3: (a) The average correlation of genes as a function of the distance (offset) from an inter-chromosomal contact using RNA-Seq gene expression data. (b) The average correlation between linked genes as a function experimental count frequency threshold of the corresponding links.

tionship between the experimental data of intrachromosomal contacts in budding yeast using Microarray Gene expression data by Duan et al. In addition, performing the same analysis on a different gene expression dataset obtained using RNA-seq provides further evidence that the measured expression levels of genes with contact links are highly correlated, indicating a significant relationship between co-expression of genomic loci and proximity in nuclear space.

3.2. ***Inter-Chromosomal Contacts and Go-slim Terms***

To further demonstrate that genes in spatial proximity cross-link together more often when they have similar expression profiles or are functionally related, we investigated whether the observed chromosomal contacts were related to the biological functions of the affected loci. To achieve this goal, we analyzed the distribution of inter-chromosomal contacts within groups of genes that shared similar Gene Ontology (GO) annotations.

The Gene Ontology (GO) classification system provides a standardized vocabulary for describing the functions, processes, and components of genes and their products across different species. GO terms are categorized into three main domains: Biological Process, Molecular Function, and Cellular Component. The Biological Process domain describes the biological objective to which the gene or gene product contributes, The Molecular Function domain is defined as the biochemical activity of a gene product, including specific binding to ligands or structures, and The Cellular Component domain refers to the place in the cell where a gene product is active and reflects our understanding of eukaryotic cell structure.

In our analysis, Each gene is assigned one or more GO terms that describe its biological function or the biological process it participates in. By comparing the number

of contacts within groups of genes with similar GO annotations to randomly selected groups with the same number of genes, we were able to assess whether the observed inter-chromosomal links were enriched or depleted among different GO terms. results showed that most of the terms in each of the three main domains of molecular function, biological process, and cellular component were significantly enriched with inter-chromosomal contacts.

The enrichment of inter-chromosomal links among different Gene Ontology (GO) terms at different threshold count frequencies is shown in Figure 4. The data used for this analysis was obtained from both HINDIII and EcoRI libraries. The figure shows the ratio of the observed number of linked genes for each GO term to the number predicted from Monte Carlo simulation. The color represents the ratio, where blue and purple indicate high ratios (enriched terms), and orange and red indicate low ratios (depleted terms). The saturation of the color indicates the significance of the ratio (z-score), as shown in the legend. The GO terms are classified into the three main domains - biological process, molecular function, and cellular component - and ordered based on the number of genes in each domain. The results show that most of the terms in each of the three domains are significantly enriched with inter-chromosomal contacts, and the enrichment ratio of different terms is more pronounced at higher threshold frequencies defining the strengths of contacts. This analysis provides valuable insights into the biological functions associated with the chromosomal contacts and helps to elucidate the complex regulatory mechanisms underlying gene expression.

In Enrichment Analysis on Gene Sets using GO, a heat map is yet to generated to visualize the distribution of data. However, from he data obtained for functional enrichment analysis using The PANTHER Classification System indicates that the uploaded list contained a larger number of genes related to a particular biological process than would be expected by chance. The ratio of observed to expected genes was found to be high, and the corresponding p-values were small, suggesting that the enrichment of these genes in the biological process is statistically significant for the terms that were analysised in Figure 4.

Both Enrichment analyses provided evidence that the interchromosomal DNA interactions are not random at the scale of individual genes. This implies that the regulation of gene expression and the organization of chromosomal architecture are intricately linked. The observed non-random interactions suggest that certain biological functions or processes require the coordination of gene expression from multiple chromosomal regions. This finding has significant implications for our understanding of gene regulation and the organization of the genome, and it provides a foundation for future studies aimed at elucidating the underlying mechanisms governing these interactions.

Our results partially validate the findings reported in , as we observed a similar

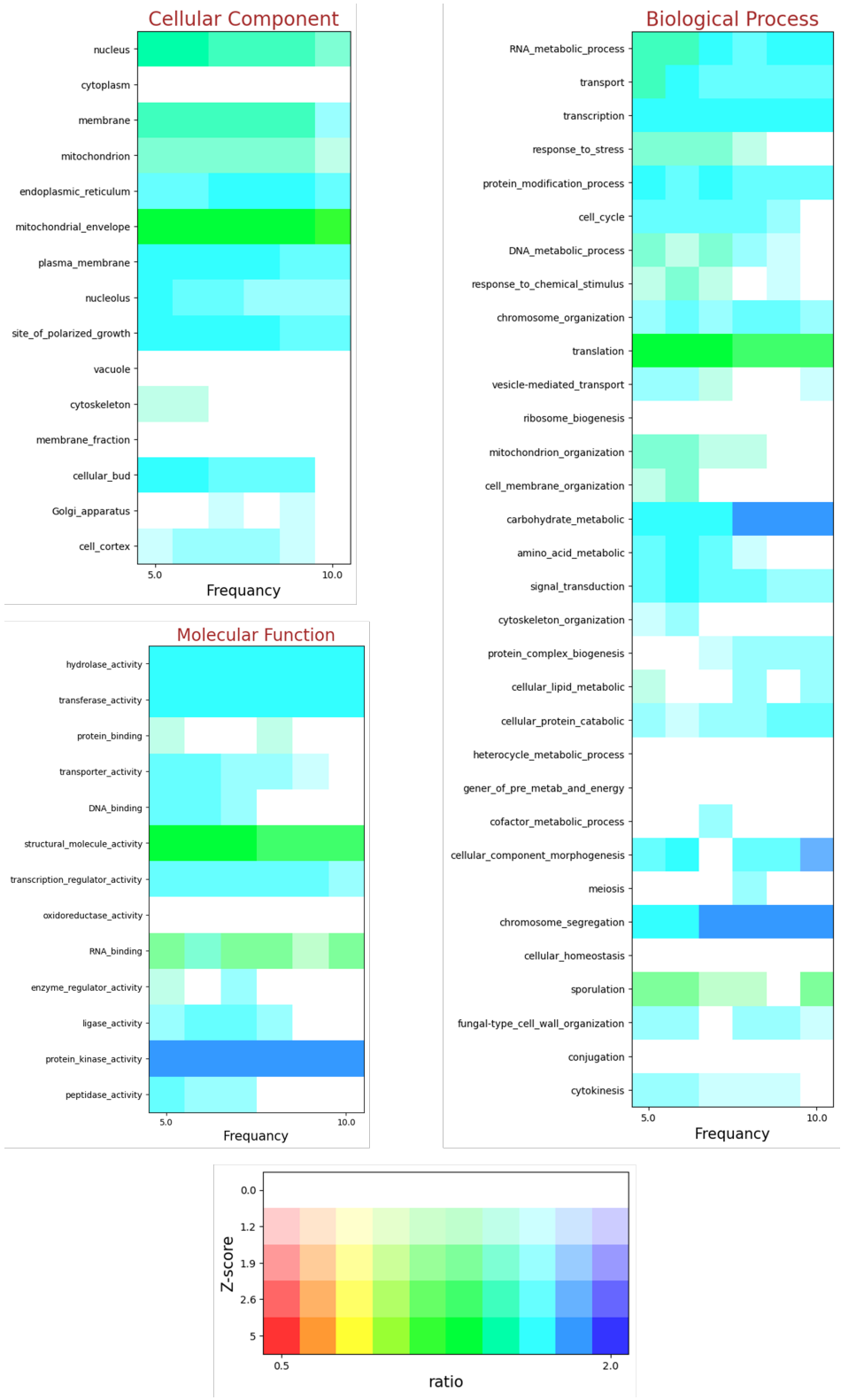


Figure 4: The Heatmaps shows the distribution of inter-chromosomal contacts between groups of genes based on Gene Ontology (GO)-slim terms. Colors represent the ratio of observed linked genes to predicted numbers from Monte Carlo simulation, with blue/purple indicating enriched terms and orange/red indicating depleted terms. The saturation indicates the significance.The GO terms are categorized into three main domains and are sorted by the number of genes.

trend in Go enrichment analysis. However, some differences were noted, which may be attributed to errors in the code used for the analysis or the gene list obtained from the GEO Website, including the gene's position in the chromosome and coordinate range

4. Future work

To further discusses the functional role of genome organization in regulating gene expression. We can extend the work to compare the studied budding yeast genome to other related genomes using Hi-C method.

The Hi-C contact data is a type of genomic data that describes the three-dimensional structure of chromatin in the nucleus. This information is obtained experimentally through a protocol that involves several steps. The first step is cross-linking, where cells are treated with formaldehyde to create covalent bonds between proteins and DNA. This step ensures that the chromatin structure is preserved and that the interactions between different regions of the genome are captured.After cross-linking, the chromatin is fragmented into smaller pieces using either restriction enzymes. This step breaks the chromatin into smaller pieces that can be sequenced more easily. The fragmented chromatin is then treated with ligase, which joins the cross-linked DNA fragments together. This step creates chimeric DNA molecules that represent the interactions between different regions of the genome [14][20][13]. The cross-links are then reversed, and the chromatin is purified to remove any contaminants. The purified chromatin is then processed to create a sequencing library. The library is sequenced on a high-throughput sequencing platform, which generates millions of short reads that represent the interactions between different regions of the genome[20]. The sequenced reads are then mapped to a reference genome, and Hi-C contact maps are generated using bioinformatics tools. The Hi-c contact data/matrix was obtained as shown in Figure 3. It summarizes the main working principle in analysing the raw experimental data. The processing pipeline can be accessed thought the following link:https://github.com/chaimae-mr/SeniorProject-II.git.

According to [21] Hi-C contact data can be processed as follow :

- Raw sequencing reads along with their quality scores are obtained from the Gene Expression Omnibus (GEO) database in FASTQ format, using SRA IDs and SRA-Toolkit.

- The raw sequences are aligned to the reference genome sequence and the results are often stored in the Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) format. To store Hi-C data at the alignment level, two important aspects are often considered, chimeric alignments (increase the number of informative

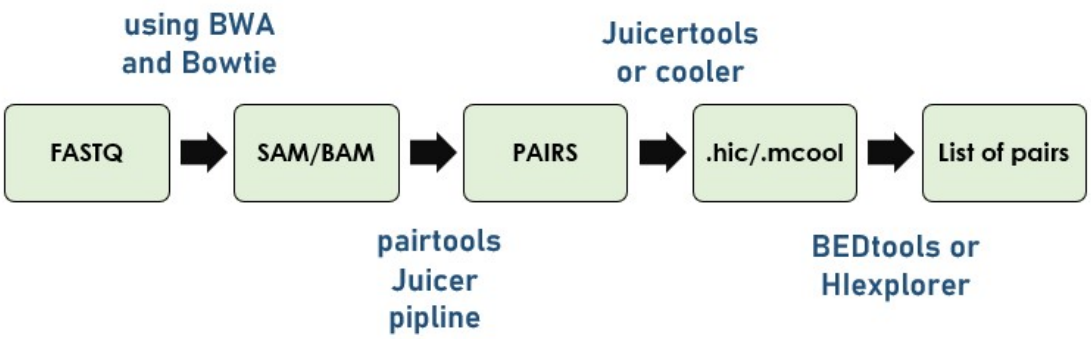


Figure 5: Workflow to Obtain a Hi-C contact matrix.The raw sequences are first aligned to the reference genome sequence and the results are often stored in the Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) format, The alignments in the SAM/BAM files are processed further into a contact list file that records all the pairs of aligned positions representing valid interacting loci. The entries in the contact list file are aggregated to genomic bins to create a contact matrix file. The latest formats like .hic and .mcool are normalized contact matrices

reads going into the contact matrix) and independent alignment of the two ends (map the two ends separately as if they were unrelated single-end data, and merge them later in some way).

- A contact list, or a "pairs" file, stores filtered pairwise interacting loci at the read alignment level. A pair file is typically created from a SAM/BAM file using specialized software (pairtools) that can extract paired-end reads from and generate a separate file that contains information about the pairs.

- A contact matrix is created by counting the read pairs in a contact list file in .hic format into genomic bins (at a given resolution). The genomic bins or resolutions of a matrix can be kilobases to megabases long, or they may correspond to restriction fragments. The rows and columns of the matrix represents genomic loci.The two dimensions of the matrix correspond to two interacting loci, each axis representing a genome. Each element of the matrix (also called a pixel) reflects the frequency of interactions between two genomic regions. The dense matrix form is used for high-resolution matrices for analyses focusing on a small region. A sparse matrix form stores only the elements with a nonzero value, in the format of ¡bin index i¿¡bin index j¿¡value¿ for each element

- Convert the Hi-C contact matrix into a format that includes the positions of the loci. This can be done using BEDtools or HIcexplorer, which can generate .bed files containing the genomic coordinates of each locus in the matrix.

- Use of awk and midpoint functions (linux command) to get the exact interacting

position .

5. Conclusion

In conclusion, our study demonstrates the functional role of genome organization in regulating gene expression and coordinating biological processes. We have shown that genes in spatial proximity cross-link together more often when they have similar expression profiles or are functionally related. Furthermore, we have found that the spatial organization of the yeast genome is non-random and facilitates the coordinated expression of functionally related genes. Our analysis also reveals that the correlation between linked genes is highest for small offsets and decreases as the offset increases, indicating that linked genes have a higher correlation than genes located farther apart on the genome. Additionally, we have confirmed our results using RNA-seq gene expression data, providing further evidence for the significant relationship between co-expression of genomic loci and proximity in nuclear space. These findings have implications for our understanding of the mechanisms of gene regulation and transcriptional control and may help to elucidate the role of genome organization in health and disease. Overall, our study provides a framework for further investigation into the relationship between genome organization and gene expression and regulation.As more high-resolution genome-wide data becomes available, our understanding of the spatial aspect of genome regulation will undoubtedly continue to deepen.

References

[1] EMBL-EBI, "References — functional genomics ii." [Online]. Available: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/references/

[2] P. Fraser and W. Bickmore, "Nuclear organization of the genome and the potential for gene regulation," *Nature*, vol. 447, pp. 413–417, 05 2007.

[3] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen, "Genome architectures revealed by tethered chromosome conformation capture and population-based modeling," *Nature Biotechnology*, vol. 30, pp. 90–98, 12 2011.

[4] D. Homouz and A. S. Kudlicki, "The 3d organization of the yeast genome correlates with co-expression and reflects functional relations between genes," *PLoS ONE*, vol. 8, p. e54699, 01 2013.

[5] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data," *Nature Reviews Genetics*, vol. 14, pp. 390–403, 05 2013.

[6] N. Y. Rodriguez-Granados, J. S. Ramirez-Prado, A. Veluchamy, D. Latrasse, C. Raynaud, M. Crespi, F. Ariel, and M. Benhamed, "Put your 3d glasses on: plant chromatin is on show," *Journal of Experimental Botany*, vol. 67, pp. 3205–3221, 04 2016.

[7] A. Papantonis and P. R. Cook, "Genome architecture and the role of transcription," *Current Opinion in Cell Biology*, vol. 22, pp. 271–276, 06 2010.

[8] Y. Shavit, I. Merelli, L. Milanesi, and P. Lio', "How computer science can help in understanding the 3d genome architecture," *Briefings in Bioinformatics*, vol. 17, pp. 733–744, 10 2015.

[9] M. M. Babu, S. C. Janga, I. de Santiago, and A. Pombo, "Eukaryotic gene regulation in three dimensions and its impact on genome evolution," *Current Opinion in Genetics Development*, 12 2008.

[10] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, pp. 363–367, 05 2010.

[11] T. G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic Acids Research*, vol. 47, pp. D330–D338, 11 2018.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 05 2000.

[13] J. Dekker, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, 02 2002.

[14] S. Sati and G. Cavalli, "Chromosome conformation capture technologies and their impact in understanding genome function," *Chromosoma*, vol. 126, pp. 33–44, 04 2016.

[15] S. Lamarre, P. Frasse, M. Zouine, D. Labourdette, E. Sainderichin, G. Hu, V. Le Berre-Anton, M. Bouzayen, and E. Maza, "Optimization of an rna-seq dif-

ferential gene expression analysis depending on biological replicate number and library size," *Frontiers in Plant Science*, vol. 9, 02 2018.

[16] "Go enrichment analysis," Gene Ontology Resource. [Online]. Available: http://geneontology.org/docs/go-enrichment-analysis/

[17] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "Large-scale gene function analysis with the panther classification system," *Nature Protocols*, vol. 8, pp. 1551–1566, 07 2013. [Online]. Available: https://www.nature.com/articles/nprot.2013.092

[18] R. Edgar, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, pp. 207–210, 01 2002.

[19] A. Al-Aamri, A. S. Kudlicki, M. Maalouf, K. Taha, and D. Homouz, "Inferring gene regulatory networks from rna-seq data using kernel classification," *Biology*, vol. 12, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2079-7737/12/4/518

[20] v. Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander, "Hi-c: A method to study the three-dimensional architecture of genomes." *Journal of Visualized Experiments*, 11 2021.

[21] S. Bicciato and F. Ferrari, *Hi-C Data Analysis*. Humana, 09 2022.