Computational Mapping Of Nucleosomes In Human Genome:

Occupancy At The Elements Of A Regulatory Build

Master Thesis

Systems biology master program

Vilnius university

STUDENT NAME: Gabriele Stockunaite

STUDENT NUMBER:

SUPERVISOR: Erinija Pranckevičienė

SUPERVISOR DECISION: **……..**

FINAL GRADE **……**

DATE OF SUBMISSION: 22 May 2023

CONTENTS

LIST OF ABBREVIATIONS

Base pairs (bp)

CCCTC-binding factor (CTCF)

Gene expression omnibus (GEO)

Kyoto Encyclopaedia of Genes and Genomes (KEGGS)

Micrococcal nuclease sequencing (MNase-seq)

Principal component (PC)

Principal components analysis (PCA)

Transcription Factor (TF)

Transcription start site (TSS)

t-Distributed Stochastic Neighbour Embedding (t-SNE)

Uniform Manifold Approximation and Projection (UMAP)

INTRODUCTION

Nucleosomes are structures by which DNA is packed in the chromatin. They consist of 146 base pairs of DNA wrapped around a histone octamer. A formation of nucleosomes is controlled by many factors such as epigenetic marking, their remodeling by the protein complexes and a DNA sequence composition among other. It was shown that specific patterns of dinucleotide frequencies oscillating at the period of 10.2-10.4 base pairs exist in nucleosomal DNA sequences. These patterns are specific to organism, tissue and condition. The nucleosome occupancy is determined by both factors and nucleosome occupancy locations around regulatory elements and distances to it and the transcription start sites could play a role in gene regulation.

This work investigates patterns of nucleosome occupancy by measuring the base pair distances between nucleosomes and regulatory elements and TSS in the human genome identified by the Encode project and provided in the regulatory build of Ensemble. The nucleosome locations are taken from GEO database datasets produced by MNase-seq. The end goal of this work is to identify patterns in relation to the activity of the genes that are regulated.

AIM AND TASKS

The aim of this work is to identify patterns in nucleosome occupancy configurations in regulatory parts of human genome in relation to the activity of the genes that are regulated.

1. Collect coordinates of regulatory elements, TSS and nucleosome positions for human organism from selected existing data on academic databases.
2. Create a table of the regulatory elements, TSS and nucleosome coordinates in these genome regions and distances between them database.
3. Perform analysis on distance distributions between nucleosomes, closest TSS and closest regulatory element coordinates (or positions) to confirm hypothesis that patterns exist.
4. If classes are identified, perform gene overexpression analysis on the classified gene groups to identify any common themes.

LITERATURE REVIEW

Nucleosome structure and role in gene regulation

Nucleosomes are the basic unit of DNA packaging in eukaryotic cells. They consist of a section of DNA wound around a histone protein complex, which consists of two copies each of four different histone proteins: H2A, H2B, H3, and H4. The DNA wrapped around the histone complex forms a structure called a nucleosome(Alberts et al. 2002).

Nucleosomes provide measures of packaging and stabilize negative super coiling of DNA in vivo; provide epigenetic layer of information guiding interactions of trans-acting proteins with the genome through their histone modification (Iyer 2012)

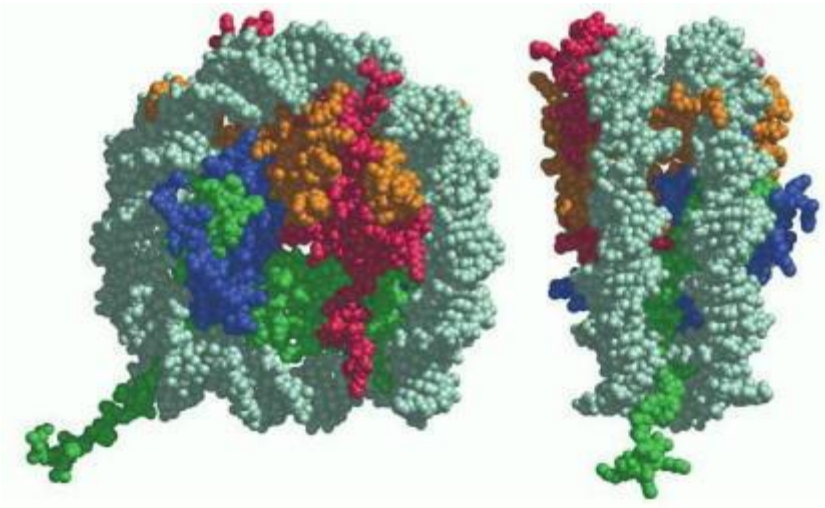

Figure 1. The DNA wrapped around the histone complex forms a structure called a nucleosome(Alberts et al. 2002)

The main function of nucleosomes is to compact the DNA in the nucleus of a cell. DNA is a long molecule that needs to be packaged in a compact form in order to fit inside the small space of the cell nucleus. Nucleosomes help to accomplish this by wrapping the DNA around the histone proteins, which allows the DNA to be more densely packed.

Nucleosomes also play an important role in the regulation of gene expression.(Bai and Morozov 2010). The arrangement of nucleosomes along the DNA molecule can influence whether a gene is able to be transcribed into RNA. For example, if a nucleosome is positioned over a promoter region of a gene, it can inhibit the binding of transcriptional machinery to the DNA and prevent the gene from being transcribed. Conversely, if a nucleosome is removed or repositioned, it can expose the promoter region and allow the gene to be transcribed.

Nucleosome positioning at human regulatory elements

Human regulatory elements are genomic regions that control gene expression and are involved in various cellular processes. They play critical roles in maintaining normal cellular functions and are implicated in various developmental and disease processes. The major types of human regulatory elements are:

- Promoters: Core promoters encompasses TSS within it and can be generally classified into 2 classes: sharp and broad promoters. The TSSs in sharp promoters are narrowly distributed, where transcription primarily initiates at closely positioned nucleotides, and TSSs in the broad promoters are distributed over a wide region.(Schoenfelder and Fraser 2019) Secondary promoters are DNA sequences located upstream of the transcription start site of a gene. They contain cis-regulatory elements, which regulate gene expression by recruiting the transcriptional machinery (Duttke et al. 2015).
- Enhancers: The primary function of enhancers is to enhance or increase the transcriptional activity of specific genes. They achieve this by binding to specific transcription factors and other regulatory proteins, which then interact with the transcriptional machinery at the promoter region to initiate gene transcription. Enhancers can influence gene expression in a context-dependent manner, allowing for precise temporal and spatial control of gene regulation(Field and Adelman 2020).
- Silencers: Silencers are DNA sequences that repress gene expression by recruiting transcriptional repressors and chromatin-modifying enzymes. They can be located upstream, downstream, or within the gene body and can act over long distances to silence gene expression(Pang and Snyder 2020).
- Insulators: Insulators are DNA sequences that separate chromatin domains and prevent the spread of regulatory signals between adjacent genes. They can also protect genes from the effects of neighbouring enhancers or repressors(Özdemir and Gambetta 2019).
- Chromatin domains: Chromatin domains are regions of the genome that have distinct chromatin structures and regulatory properties. They are characterized by specific combinations of histone modifications and transcription factor binding, which determine their regulatory potential and gene expression patterns(Sikorska and Sexton 2020).
- Transcription factors (TF) binding sites are specific DNA sequences within the regulatory regions of genes where transcription factors bind. Transcription factors are proteins that play a crucial role in regulating gene expression by interacting with specific DNA sequences and modulating the transcriptional activity of target genes.TF binding sites typically consist of short DNA sequences(Kribelbauer et al. 2019). CTCF is a highly conserved transcription factor that plays a critical role in chromatin organization and gene regulation, where it can act as an insulator as well as facilitate or repress interaction between enhancers and promoters(Hansen 2020).

Enhancers and promoters are enriched for histone modifications associated with active gene expression, and these regions have low nucleosome occupancy. Overall, nucleosome occupancy at some regulatory elements is strongly associated with nearby gene expression levels (Rhie et al. 2014), regions of the genome with low nucleosome occupancy tend to be associated with active gene expression and low levels of DNA methylation, while regions with high nucleosome occupancy tend to be associated with low gene expression and high levels of DNA methylation by (Lövkvist, Sneppen, and Haerter 2018).

For promoters, patterning of nucleosomes at individual promoters can vary substantially between cell-types, but that it is strikingly constant during stimulus-driven activation, and transcription initiation is linked to mainly transient remodelling events. (Oruba, Saccani, and van Essen 2020). However, during literature review no work regarding nucleosome – regulatory element – TSS distance configurations were found, which we will investigate in this work.

Types of nucleosome positioning factors

Nucleosomes are positioned along the DNA molecule in a way that is influenced by both the structure of the DNA itself (*Cis* factors) and the presence of certain proteins and regulatory elements (*Trans* factors)(Radman-Livaja and Rando 2010; Wright and Cui 2017).

Three major classes of trans factors have been identified in nucleosome positioning/occupancy that can bind to the specific parts of the DNA and affect a way nucleosomes are positioned: (1) transcription factors, (2) chromatin remodelers (Cairns 2005) and (3) RNA polymerase (Field et al., 2008; Mavrich et al., 2008b; Schones et al., 2008; Yuan et al., 2005) .For example, studies have found that ATP-dependent remodelers such as human SWI/SNF remodeling complex are capable of moving nucleosomes away from nucleosomes positioning sequences (NPSes)(Pham, He, and Schnitzler 2010).

Cis factor is the sequence-dependent differences in the physical properties of the DNA base-pair combinations that can influence the kinetics of nucleosome formation and their stability as well as the DNA association to the histone core(Gromiha 2000), even though the DNA and histones can form the nucleosomes solely based on physical features(Onufriev and Schiessel 2019). The special conformation properties of individual base-pairs, or the cooperative conformation of longer DNA tracts, predispose individual sequences to be in a particular 3D formation, or to be deformed in a particular manner upon interactions with proteins. (Cohanim and Haran 2009).

Due to the nature of these two classes of factors it is important to note that nucleosomes positions can differ in vitro vs. in vivo conditions. In vitro the nucleosome positions are solely determent by physical DNA and histone proteins kinetic factors whereas in vivo investigations cannot strictly

determine which mechanism caused the specific nucleosome positions. However studies show high similarity between in vivo and in vitro nucleosome maps, for example, both type of studies show similar nucleosome depletion at translational end sites (Kaplan et al. 2009)

Considering the DNA sequence coding of nucleosome formation interesting questions are raised in the literature on how come this mechanical information encoding is possible and not intervening with the amino acids codied by triplets, it is showed that multiplexing is only possible due to depletion of both coding mechanisms (Cohanim and Haran 2009; Eslami-Mossallam et al. 2016), which means that similar values can be encoded by different codes which is the case for synonymous amino acid codons and in the case of nucleosome base-pairs as will be presented in the next section.

Furthermore, it is thought that nucleosome positioning coding can directly influence sequence polymorphism and divergence (Langley, Karpen, and Langley 2014) as well as evolution through its interactions with DNA damage and repair mechanisms. These complex evolutionary dynamics between mutual influence of sequence evolution code and nucleosome positioning code evolution are reviewed in a study by (Langley, Karpen, and Langley 2014).

Overall, the positioning of nucleosomes along the DNA molecule is a complex process that is influenced by a variety of factors, including the structure of the DNA itself (Cis factors) and the presence of regulatory proteins and elements (Trans factors).

DNA sequence patterns that impact nucleosome positioning and rotation

The specific extend of the role of DNA sequence patterns in nucleosome positioning is a subject of ongoing research, however it is thought that certain nucleosomes in genomes are positioned by a preference of some DNA sequence patterns over the others – these preferable sequences are called Nucleosome positioning sequences (NPS) (Ioshikhes, Hosid, and Pugh 2011). It is showed that these nucleosome positioning preferences arise from sequence dependent mechanical properties of the DNA molecule (Eslami-Mossallam et al. 2016), because in order to form nucleosomes DNA needs to be bendable to go around sharp turns on a histone octamer in order to form nucleosome body and different nucleotides of the DNA sequence either promotes needed bending or obstructs it impacting the nucleosome position and rotation(Trifonov 2011; Cui et al. 2014).

Commonly investigated DNA sequence patters can be divided in two distinct dinucleotide groups, which in the literature are referred as WW/SS where W stands for Weak (A or T) and S stands for Strong (C or G) which was first notated this way in 1986 (Satchwell, Drew, and Travers 1986) and RR/YY, where R is purine and Y is pyrimidine. Sequences with high affinity to nucleosomes feature TT/AA/TA (WW) dinucleotides at negative roll positions (minor DNA groove facing inwards) and GC (SS) dinucleotides at positive roll positions (minor groove facing outwards) (Satchwell, Drew, and

Travers 1986; Segal et al. 2006). It has been shown that these patterns of dinucleotides at oscillating frequencies of the period of 10.2-10.4 base pairs exist in nucleosomal DNA sequences and have been observed in fruit flies (Mavrich et al. 2008), yeast (Albert et al. 2007), nematodes (Johnson et al. 2006) human and other mammalian cells (Ioshikhes, Hosid, and Pugh 2011; Wright and Cui 2019; Pranckeviciene et al. 2020). On the contrary, the regions between nucleosome called the linker regions has been found to include sequences that are stiff and obstructs DNA bending therefore negatively impacting nucleosome formation(Struhl and Segal 2013), please see the figure 1 .

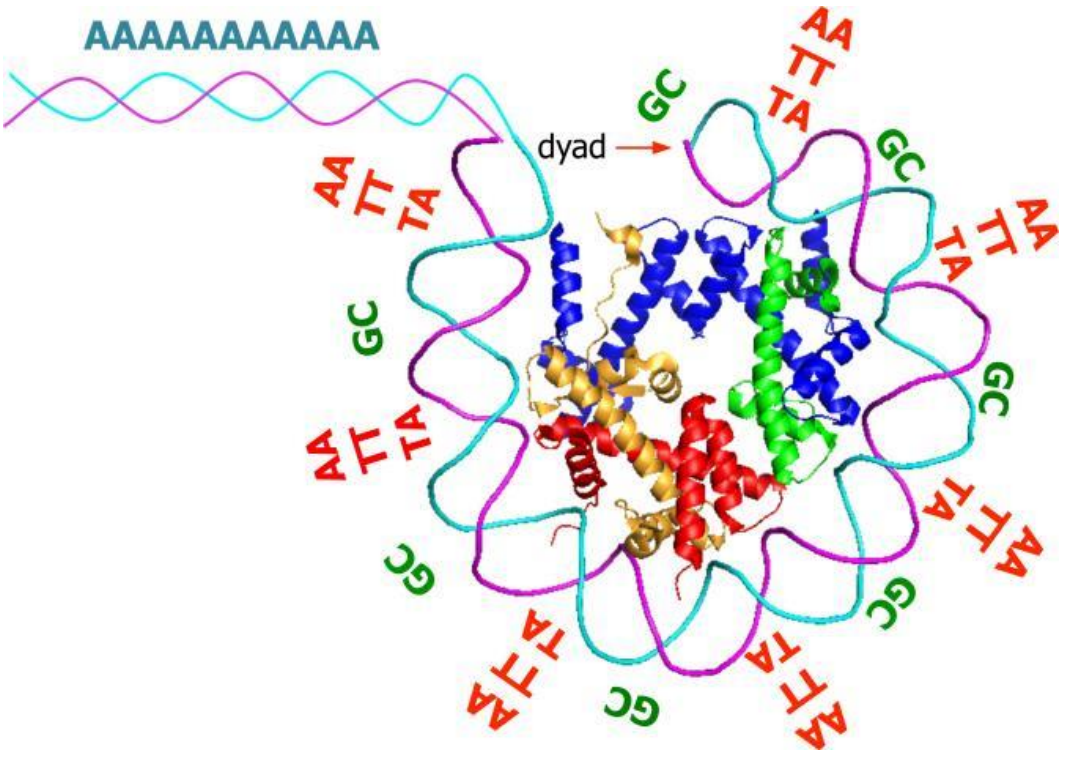

Figure 2 Illustration of nucleosome sequence preferences from (Struhl and Segal, 2013)

Although, significant differences do exist between the in vitro and in vivo nucleosome maps, the existing literature show that the genome explicitly encodes many aspects of the in vivo nucleosome organization through the nucleosomes' intrinsic DNA sequence preferences.(Kaplan et al. 2010). In higher eukaryotes genomic elements are closed by nucleosome but in unicellular organisms the genomic sites are open unless a nucleosome is repositioned there. Promoters of multicellular organisms are characterized by sequences favouring nucleosomes and in unicellular organisms by the disfavouring sequences(Tompitak, Vaillant, and Schiessel 2017).

As showed in (Pranckeviciene et al. 2020), the distributions of RR/YY peaks have similar characteristics in human CD4+ lymphocites and nucleus accumbens cells in mouse brains. and different positioning in yeast and human apoptotic cells. The RR2/YY2 pattern in yeast has RR

(Purine-Purine) peaks occurring in SHL minor zones and with YY (Pyrimidine-Pyrimidine) dinucleotides in all SHL zones; oppositely in human lymphocyte cells and mouse NAC both the RR and YY steps are found in major zones and are arranged in proximity opposite to each other. The very different RR2/YY2 pattern in human apoptotic cells is not persistent due to the stiffness and low bendability of RR and YY dinucleotides.

Genomic data databases available for academic use

The field of genomics and genetics research relies heavily on several key human DNA databases that are available for academic use. These databases offer a wealth of information and resources to researchers studying human genetics, genomics, and related fields.

One of the most widely used databases is the NCBI GenBank, which is a public DNA sequence database managed by the National Center for Biotechnology Information (NCBI). It contains annotated DNA sequences and information about the organism from which the sequence was obtained. Another popular database is the ENSEMBL genome browser and annotation database, developed by the European Molecular Biology Laboratory (EMBL). This database provides access to annotated genomes of various species, including humans. The UCSC Genome Browser, developed by the University of California, Santa Cruz (UCSC), is another genome browser that provides access to annotated genomes of many species, including humans.
Furthermore, the NCBI Gene Expression Omnibus (GEO) is a public repository that contains data from various high-throughput gene expression profiling studies. This database offers researchers access to gene expression data from a variety of species, including humans, and provides advanced analytical tools and visualizations for studying gene expression patterns. The GEO database is a valuable resource for researchers studying gene expression, transcriptional regulation, and related areas of research, and has been used in many studies to identify genes and pathways associated with diseases and biological processes. In this review we are using this database to understand the research and data available on nucleosome positioning at regulatory elements in human genome.

Table 1 Summary of key human genomic data repositories

Database Name	Database Link	Database Description
NCBI Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/	A public repository that contains data from various high-throughput gene expression profiling studies. It provides access to gene expression data from a variety of species, including humans, and offers advanced analytical tools and visualizations for studying gene expression patterns.

NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/	A public DNA sequence database managed by the National Center for Biotechnology Information (NCBI). It contains annotated DNA sequences and information about the organism from which the sequence was obtained.
ENSEMBL	https://www.ensembl.org/index.html	A genome browser and annotation database developed by the European Molecular Biology Laboratory (EMBL). It provides access to annotated genomes of various species, including humans.
UCSC Genome Browser	https://genome.ucsc.edu/	A genome browser that provides access to annotated genomes of many species, including humans. It was developed by the University of California, Santa Cruz (UCSC).

METHODS

Data

Four types of data sets were used in this work:

- Nucleosome occupancy data set from series GSE139224 (hg19) (*GEO Accession viewer*) from genome binding/occupancy profiling obtained by high throughput sequencing from in a paper on MNase sensitivity of promoter chromatin in GM12878 cells during stimulation with heat-killed Salmonella typhimurium(Cole and Dennis 2020). Specific dataset used was "GSE139224_control_200U_NOHDI_140-200.mat", this dataset was chosen because our experiment required nucleosome positions obtained in control conditions and the data was from well researched GM12878 cell line.
- Human regulatory build elements dataset (hg38)
- Human genes Transcription Start Site coordinates dataset (hg38)
- Human genes annotation information from KEGG, Geneontology and Reactome databases accessed via Enrichr-KG online library(*Enrichr-KG*).

Methodology

In data preparation stage, first Nucleosome centre coordinates were identified from .mat files nucleosome centre locations probability in 2000 bp area closest to gene TSS with values representing likelihood of nucleosome centre for each of 10bp windows in the area. Then nucleosome centre coordinates were lifted from hg19 to hg38 genome coordinate system.
After the nucleosome coordinates were obtained, closest regulatory element to the nucleosome in +/- 5000 bp region was found and closest gene in +/- 2000 bp region were identified.

Three distances were calculated to be used as features in data classification: nucleosome to regulatory element, TSS to nucleosome and TSS to regulatory element. Data normalization was performed by removing outliers which are outside of 2 standard deviations region for each of the three distances.

For data classification unsupervised learning was selected to find similar data groups. In order to identify linear relationships Principal Component Analysis (PCA) method was performed, which is a dimensionality reduction technique that transforms a high-dimensional dataset into a lower-dimensional representation while retaining the most important information. It achieves this by creating new variables - principal components (PC), which are linear combinations of the original variables. The first principal component captures the largest amount of variance in the data, and each subsequent component explains as much of the remaining variance as possible. PCA results were inspected in scatter plot and underlying structures in the data where found.

Then, to explore possible nonlinear relationships and validate clustering data t-Distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection

(UMAP) analysis was done. T-SNE algorithm focuses on preservation of local similarities in the data by modelling the high-dimensional data points in such a way that similar points are represented by nearby points in a lower-dimensional space which we used to identify nonlinear patterns in the data. Similarly, UMAP was used to capture and visualize the underlying structure and relationships within complex l data. It takes a matrix of input values and reduces it to a lower-dimensional space while preserving the global and local structure of the data. UMAP achieves this by constructing a weighted graph that represents the pairwise similarities between data points, using fuzzy simplicial sets. It then optimizes the layout of this graph in a lower-dimensional space, such as two or three dimensions. Both t-SNE and UMAP are widely used in Single cell RNA sequencing data analysis to visualize and explore complex multimers data relationships, and identify cellular heterogeneity and gene expression patterns(Do and Canzar 2021; Wu et al. 2022; Zhou et al. 2023)
Data classification yielded 6 groups for which violin plots, which provide a visual representation of the distribution and variation of the data, were generated for each of the 3 features(distances) and depicted separately by regulatory element type.

Based on these plots, visual representations of nucleosome-regulatory element-TSS distance configuration profiles were drawn for each identified class.

Next, to investigate gene groups that fall into each of the identified profile classes gene enrichment analysis was performed for enrichment with GO (Gene ontology) biological process data, which is a widely-used resource that categorizes genes based on their involvement in specific biological processes, providing a standardized vocabulary to describe gene functions as well as KEGS (Kyoto Encyclopaedia of Genes and Genomes) database which maps genes to biological pathways and networks, offering insights into the interactions and functions of genes and finally Reactome data which provides information on pathways that gene products participate in.

Software used

The main work programming environment for **datasets preparation**, merging, cleaning and outlier removal was set up in Jupyter Notebooks with Anaconda Python installation. Key code libraries used were:

- Pandas – for easier large datasets manipulation in panda data frames
- Numpy – for statistical data analysis and efficient distances calculation on large data lists
- Liftover – for Nucleosome occupancy dataset coordinates lifting from hg19 to hg38 systems.

Next, for **data analysis** using unsupervised learning algorithms and **results visualisation** the Orange v3.35.0 data mining software was used. For identified classes gene groups enrichment analysis online tool Enrichr-KG (*Enrichr-KG*) was utilised.

RESULTS

Figure 3 PC1-PC2 plot by types of regulatory elements

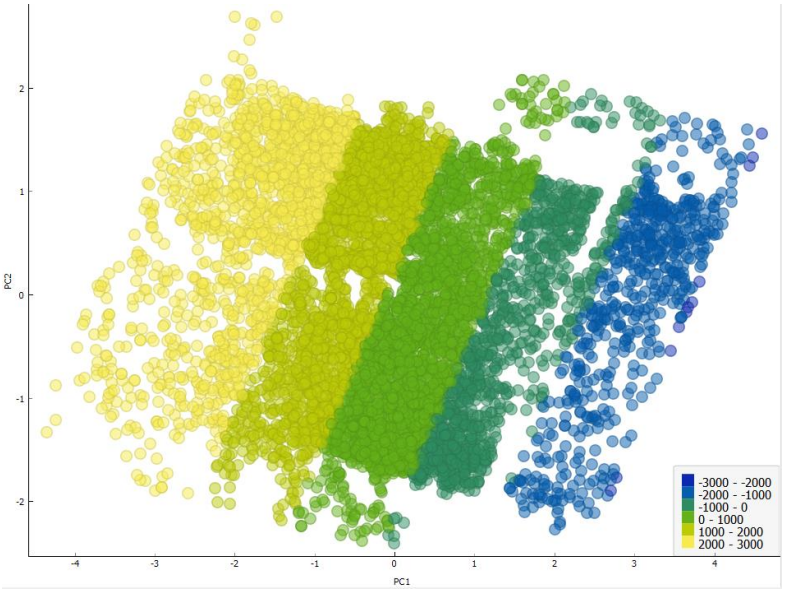

Figure 4 PC1 - PC2 plot by distance from nucleosome to regulatory element

Data exploration using unsupervised learning algorithms revealed structures present in the data based on three investigated features – distances between nucleosome, TSS and regulatory element. Distributions of each of individual distances were investigated.

For distances between regulatory elements and nucleosome, some tendencies were identified regarding two largest regulatory elements groups in our dataset – promoters and Transcription factor binding sites, please see Figure 6 and Figure 4.

In relation to nucleosome, promoters tend to be downstream of nucleosome such that distance N – Reg position is positive and tends to be between 1000 -3000 bps away from nucleosome.

On the other hand, largest concentration of TF binding site type regulatory elements is seen in +1000/-1000 region around the nucleosome**.**

However, our results show that various types of regulatory elements can have different types of distributions independently of their type.

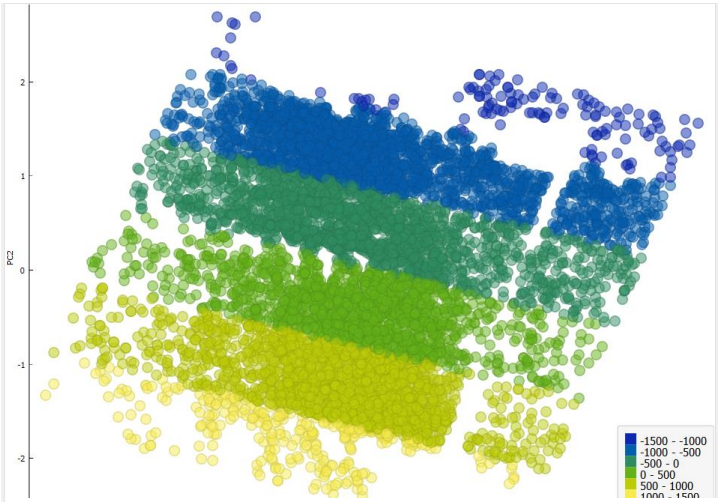

Figure 5 PC1 - PC2 plot by distance from TSS to nucleosome

Next, looking at the distances between TSS and nucleosome we can see that they are distributed evenly regardless of the type of regulatory element as can be seen from Figure 3 and Figure 5.

After initial investigation, a model of all three features (three distances) was created to obtain classes for different Nucleosome – regulatory element – TSS configurations from structures identified in t-SNE, UMAP and PCA analysis plots. Areas were selected separately and then combined list of linear and non-linear groups was created. These groups are presented in different colours in PCA and t-SNE and UMAP plots below in Figure 6 and Figure 7 respectively.

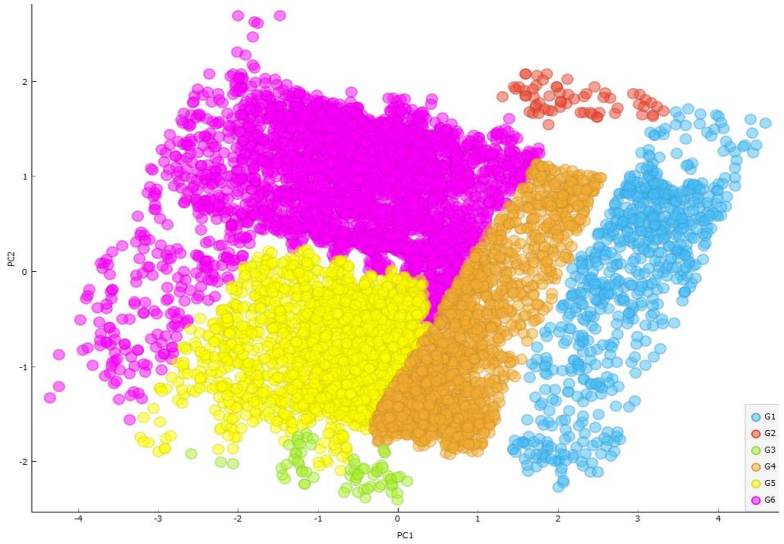

Figure 6 PC1-PC2 plot with coloured class groups

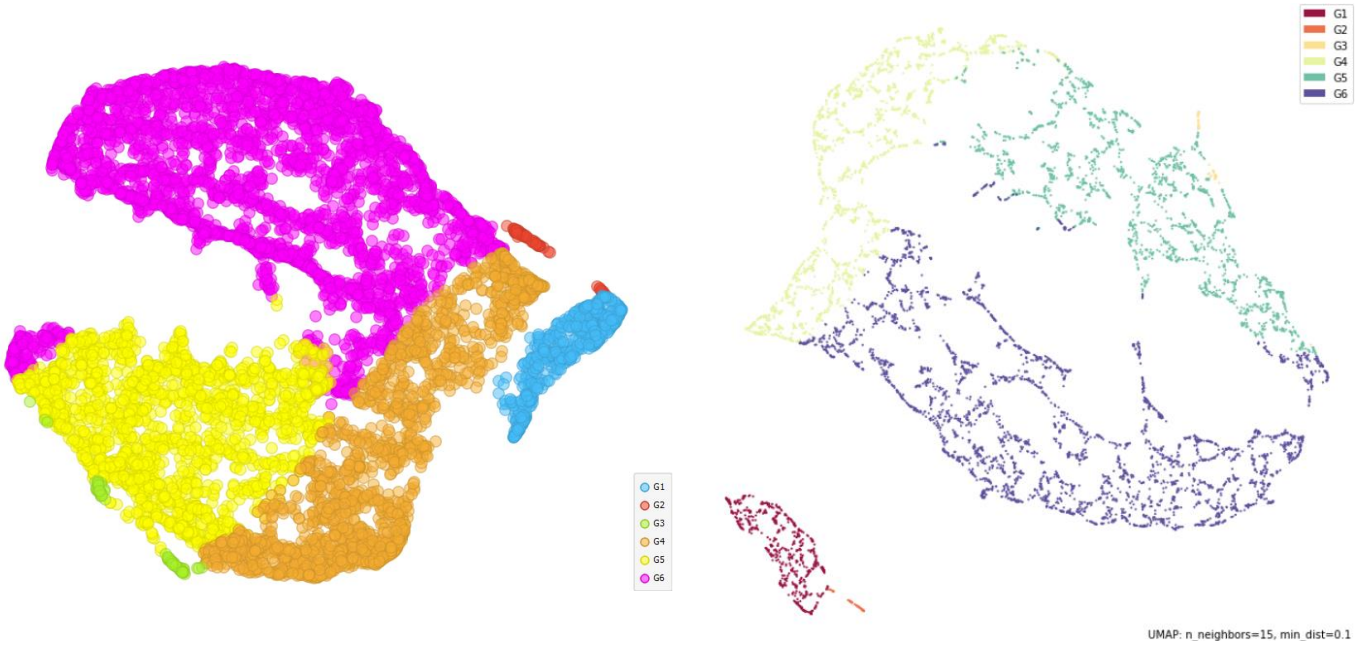

Figure 7 t-SNE and UMAP analysis plot with coloured class groups

t-SNE and UMAP analysis results was further used to verify our grouping and showed some group contamination in non-linear classification space. In order to obtain clear results, we deemed G3 points (depicted in green) as outlier group and removed them from further analysis as we did not have enough evidence of their grouping significance, G2 (red) points that were touching G1 cluster and G5(yellow) points that we in the G6(purple) area. The G6 group points in the G5 area were re-classified as G5. After the group validation we were left with 5 cleaned groups and their PCA analysis is shown in Figure 8:

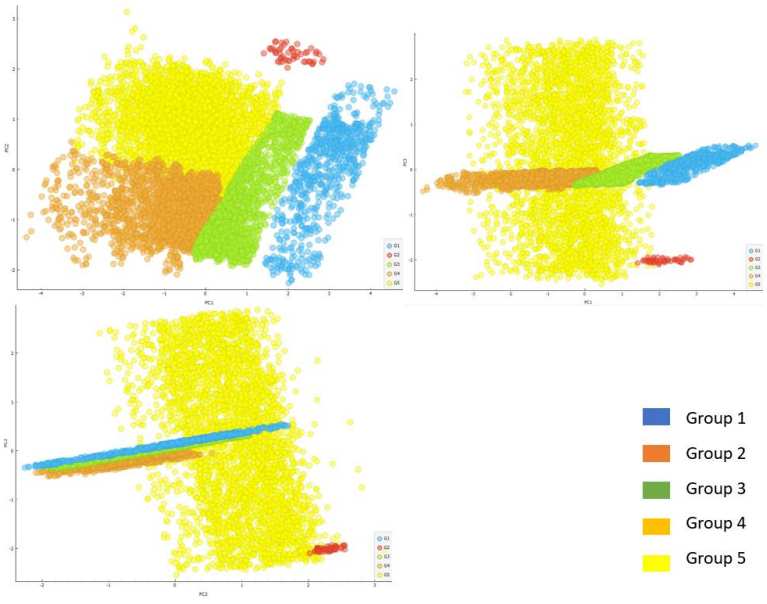

Figure 8 PCA analysis showing final classes: PC1-PC2, PC1-PC3, PC2-PC3

To characterise these groups violin plots were done for each of the clases depicting nucleosome distance to TSS and to regulatory elements as well as distance between TSS and regulatory element.

First group nucleosome positions analysis is shown in Figure 9 and shows that Transcription factor binding sites are characterised differently from and other regulatory elements. For TF binding sites subgroup, it‘s position from nucleosome is 1200 – 1600 bp upstream and TSS is in region around nucleosome from 1000 bp downstream to 600 bps upstream and disteance between TSS to TF binding site 1000-2300 bp , whereas other regulatory elements distance from nucleosome is 1100-1800 bp upstream and TSS is positioned closer to nucleosome - up to 900 bp downstream from the nucleosome, respectively TSS - Regulatory element distance is 1000-2500 bp.

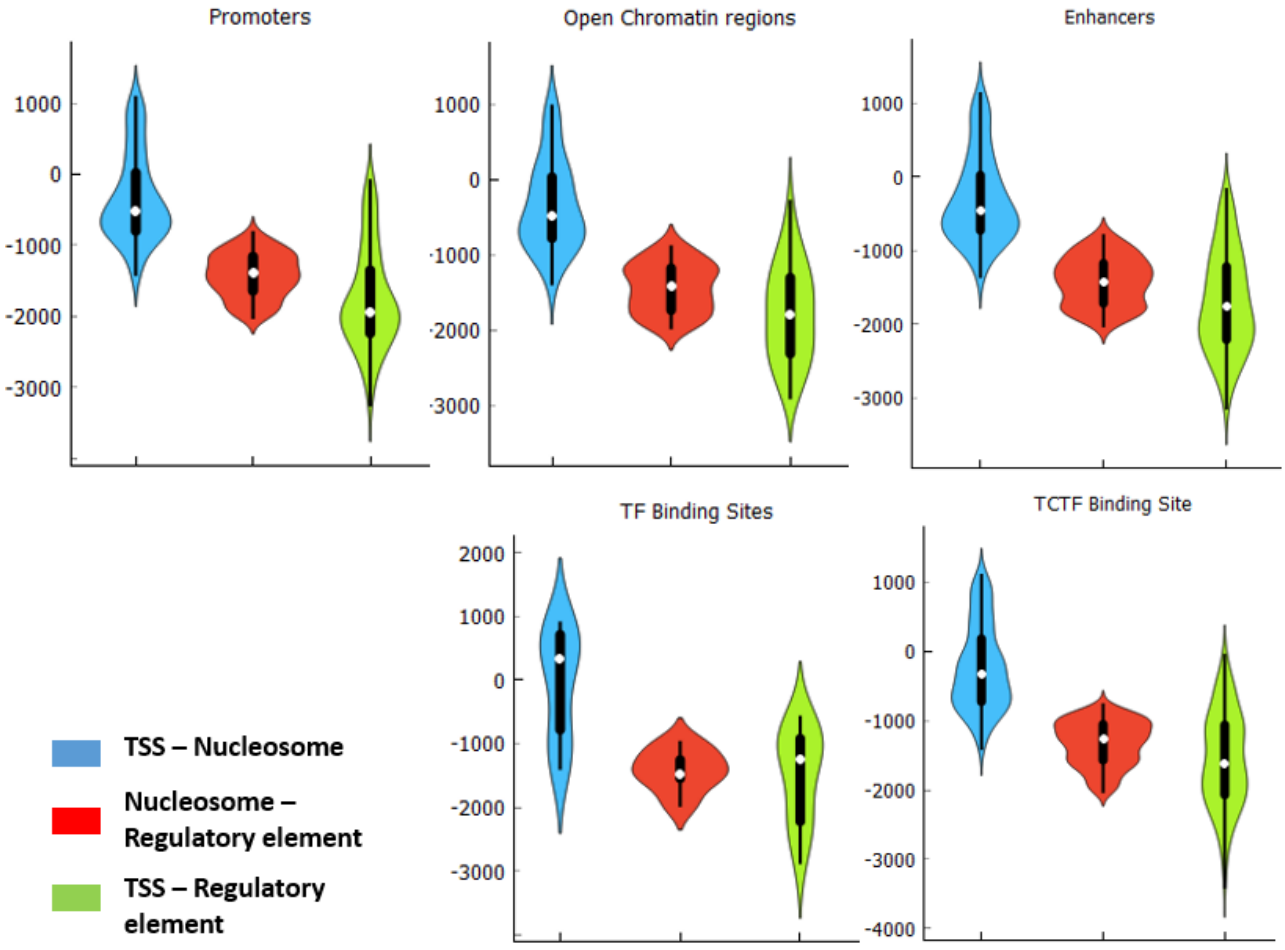

Figure 9 Group 1 distance distributions

Visual representation of the profile 1 nucleosome, regulatory element and TSS configuration is show in Figure 10 can be summarised as regulatory element far (>1000 bp) upstream of nucleosome and TSS close (<1000 bp) downstream of nucleosome, with exception of TF binding site sub-group, for which TSS can also be upstream.

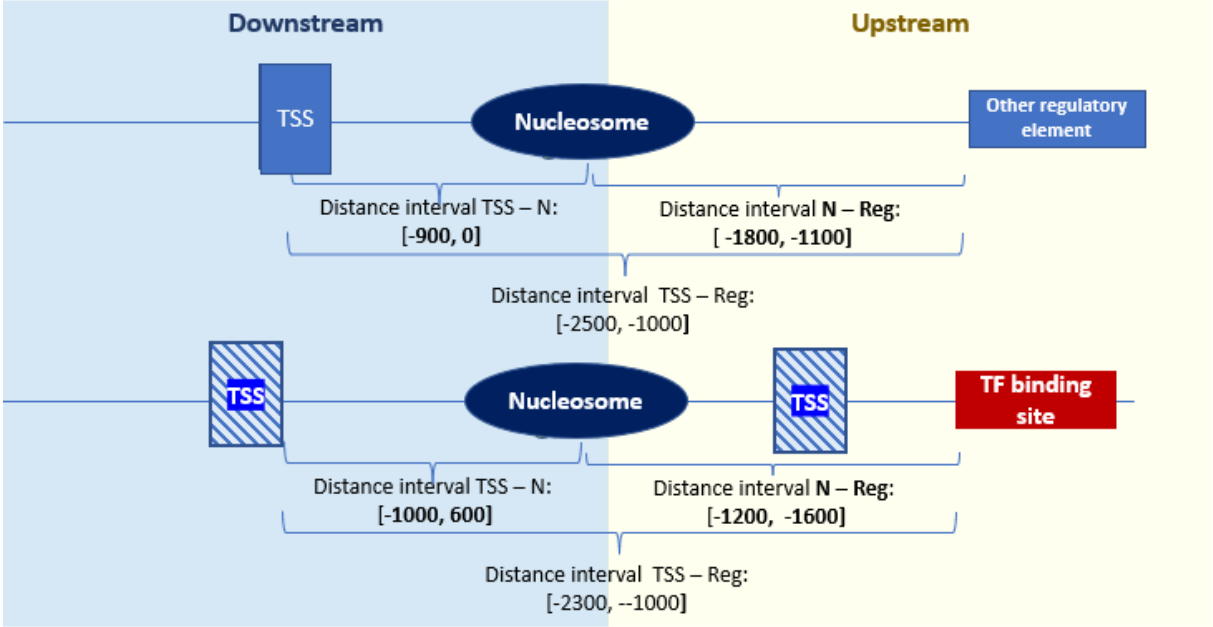

Figure 10 Nucleosome position profile 1

Second group nucleosome position analysis is shown in Figure 11. Regulatory element position is in +-400 bp region around the nucleosome and TSS is in 900-1600 bp downstream from the nucleosome and downstream from regulatory element with distance of 1100-1400 bp between them.

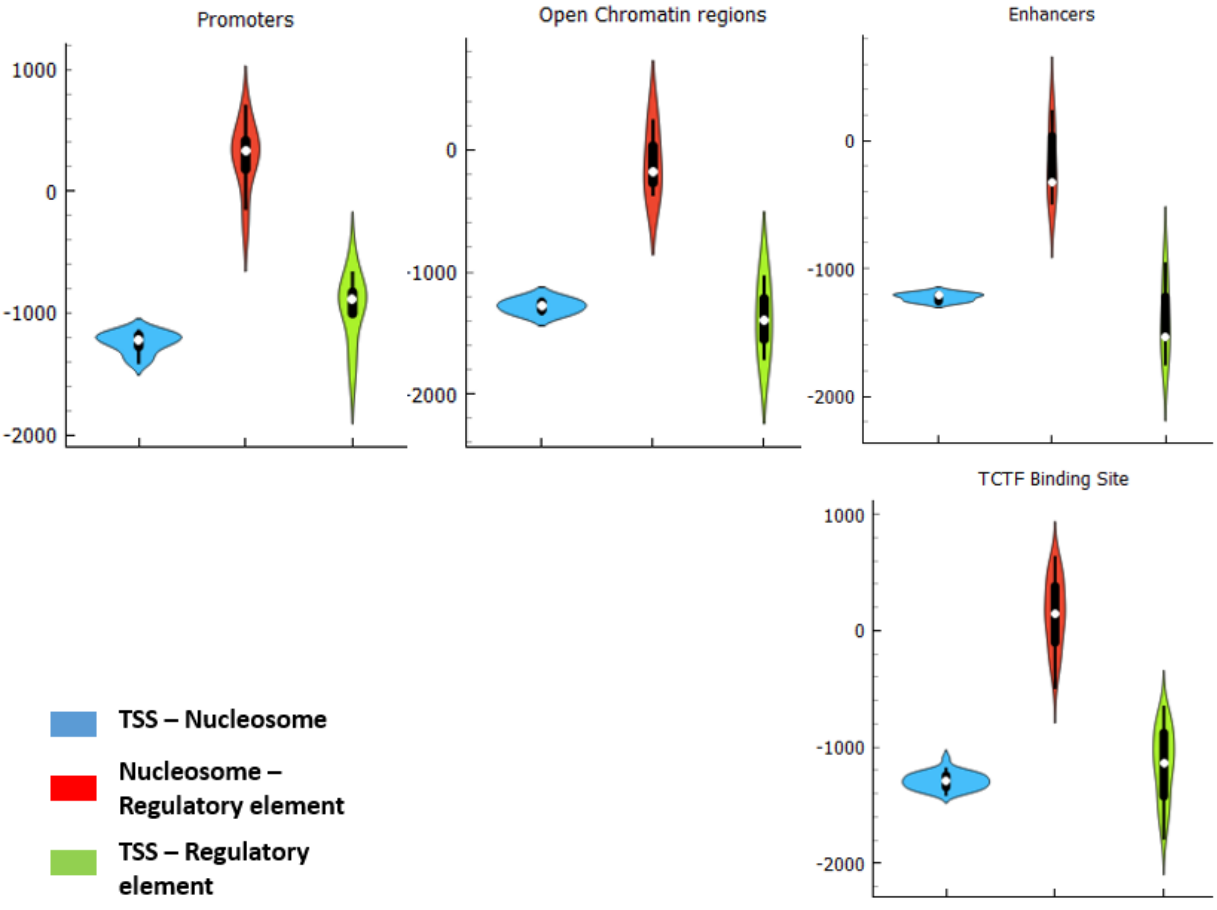

Figure 11 Group 2 distance distributions

Visual representation of the profile 2 nucleosome, regulatory element and TSS configuration is show in Figure 12 can be summarised as regulatory very close to (<400 bp) on the either sides of nucleosome and TSS is far (>900 bp) downstream of nucleosome.

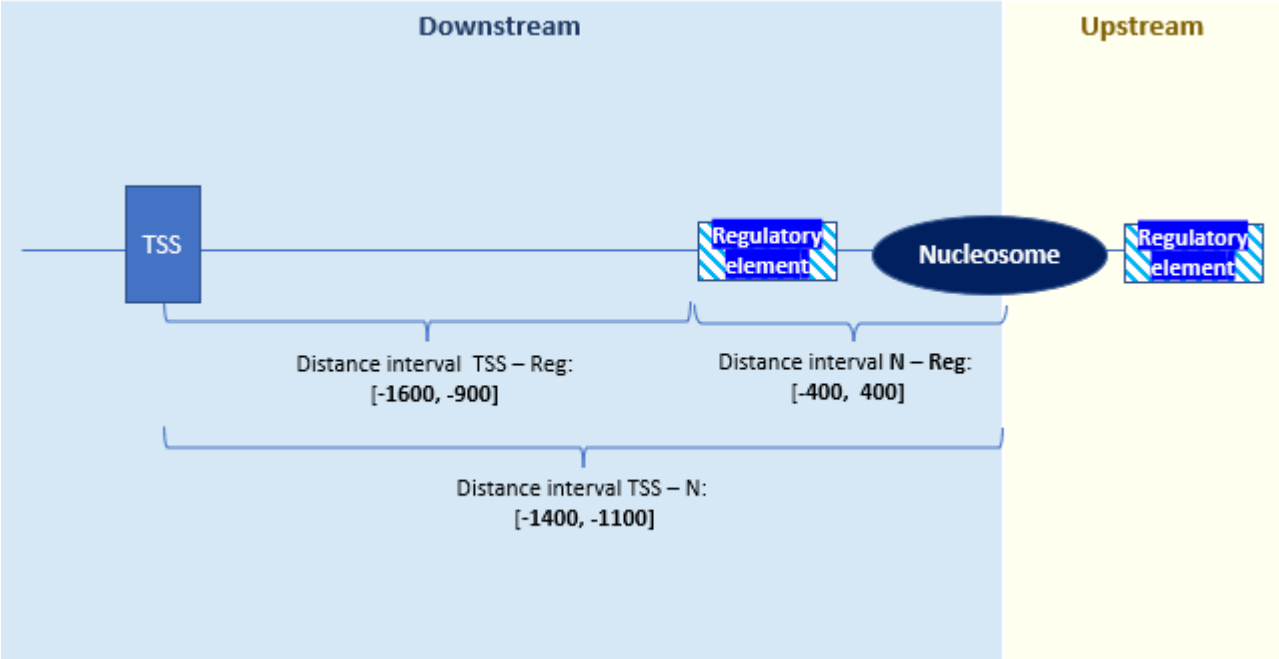

Figure 12 Nucleosome position profile 2

Third group nucleosome position analysis is shown in Figure 13 and is characterised differently for Promoter and CTCF binding site subgroup. Regulatory element position is from 200 downstream to 400 upstream region around the nucleosome for all types of regulatory elements. However TSS position differs between regulatory element types. For CTCF binding sites TSS is in wider region from 700 bp downstream to 600bp upstream of nucleosome, for promoter group TSS is always upstream by 400-900 bp but enhancers are only found downstream of nucleosome. For other regulatory elements TSS is from 100 downstream to 800 upstream of nucleosome. Distances between TSS and regulatory elements for promoters is from 400 to 1000 with promoter always downstream of the TSS, for CTCF binding sites the TSS can be on either side of the CTCF site by 600-700bp. For other regulatory elements, TSS site is closer to regulatory elements and can be 400 bp on either side of the regulatory element.

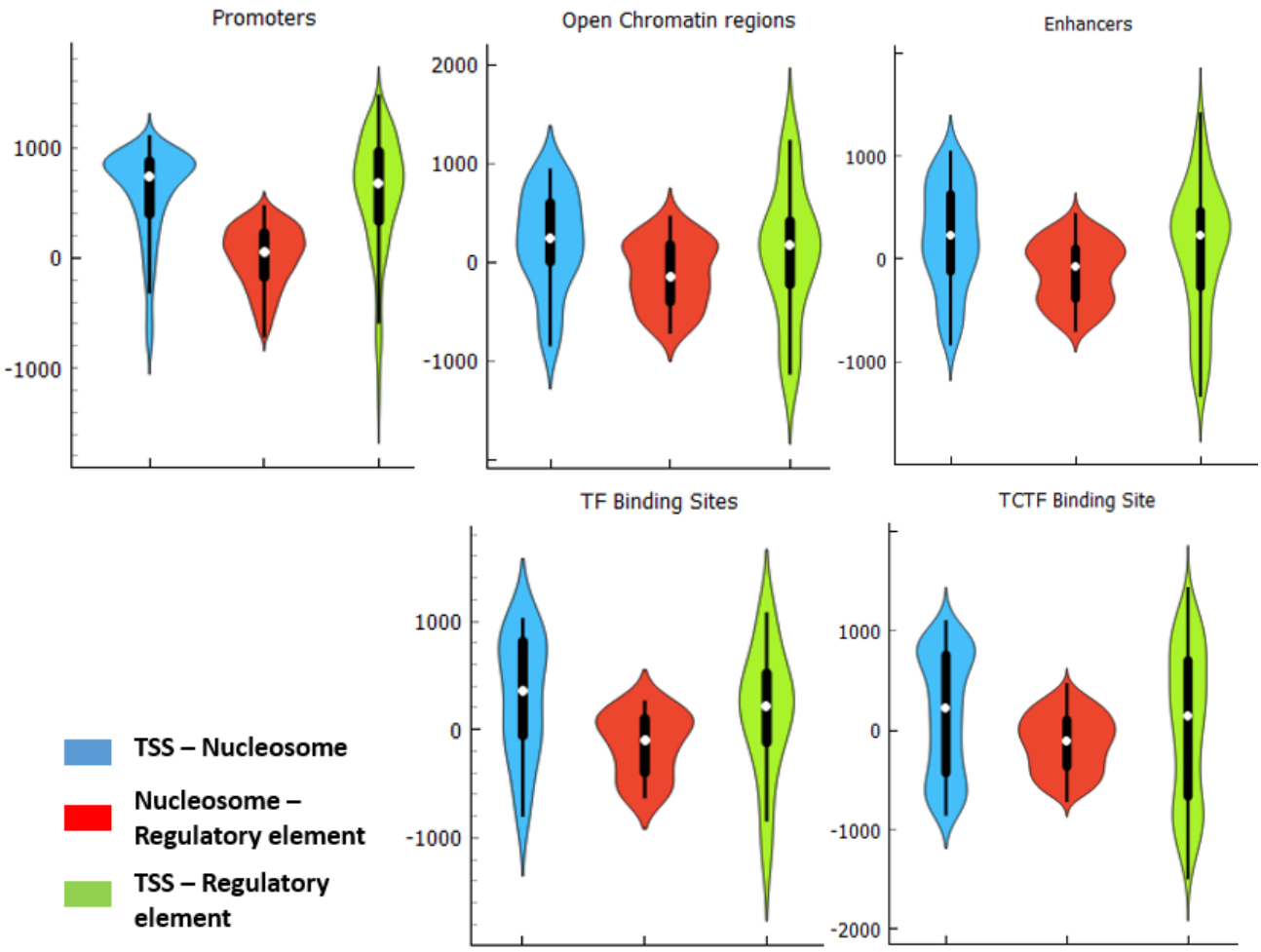

Figure 13 Group 3 distance distributions

Visual representation of the profile 3 nucleosome, regulatory element and TSS configuration is show in Figure 14 can be summarised as regulatory element very close to nucleosome (from 200 bp downstream to 400 bp upstream) and TSS position is close (<1000bp) upstream of nucleosome, except for CTCF binding sites where it can also be in a close downstream position, also in promoter group TSS is always upstream from promoter.

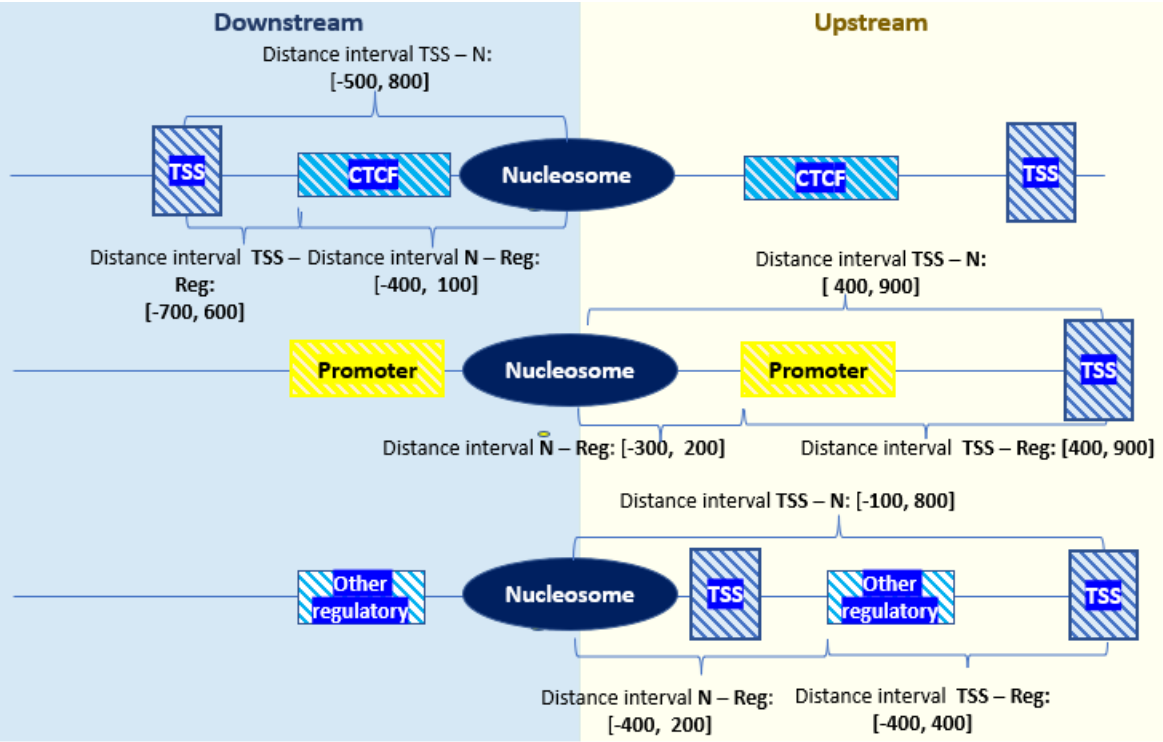

Figure 14 Nucleosome position profile 3

Fourth group nucleosome position analysis is shown in Figure 15 and is characterised differently for Open chromatin region subgroup. Regulatory element position is downstream in this group 500-1500 bp from nucleosome for Open chromatin regions and 500-2400 bp for other regulatory elements. TSS is 200-800 bp downstream from nucleosome for Open Chromatin region subgroup and 200-800 bp upstream from nucleosome for other regulatory types subgroups. Distance between regulatory element and TSS for open chromatin regions is up to 1000 bps and between 1000-3000 bp for other regulatory elements subgroup, with TSS always upstream of the regulatory element.

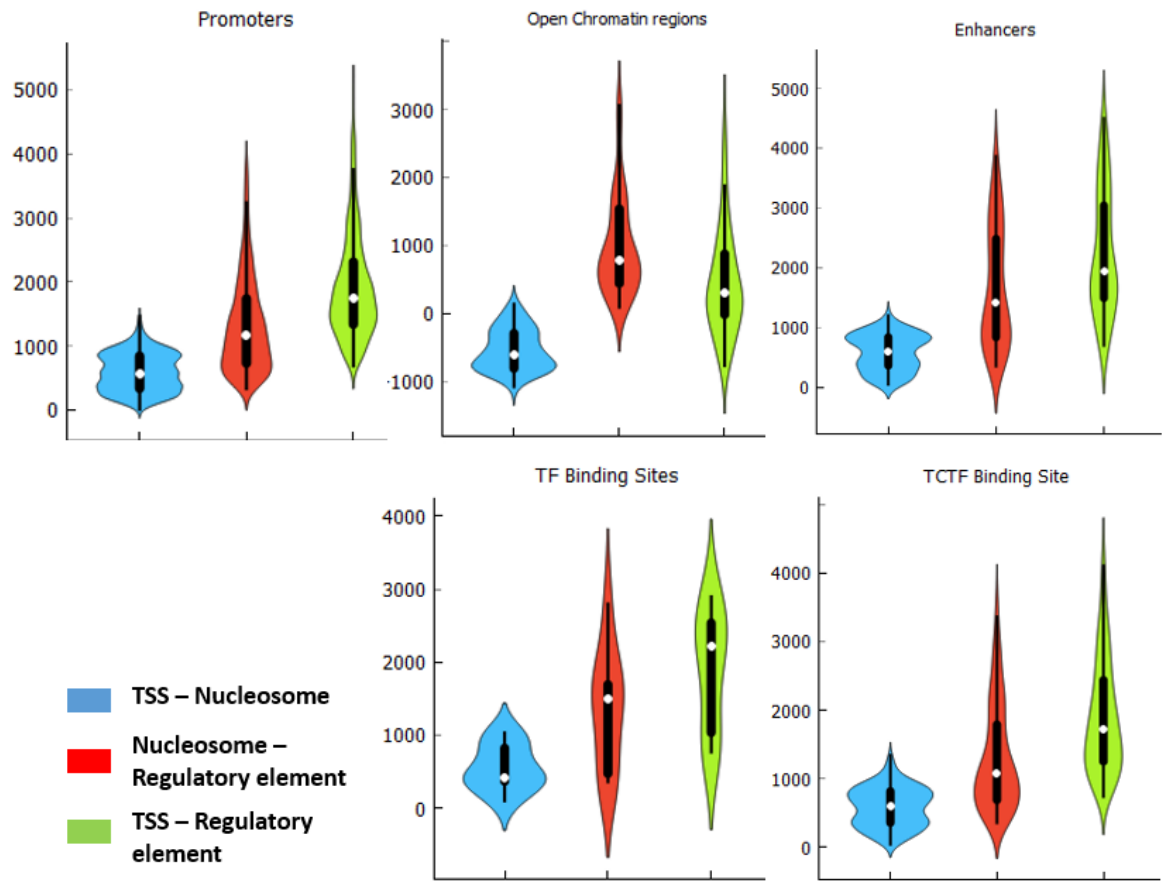

Figure 15 Group 4 distance distributions

Visual representation of the profile 4 for nucleosome, regulatory element and TSS configuration is show in Figure 16 and can be summarised as regulatory element further downstream of nucleosome (500-2400 bp) and TSS position is close (<1000bp) upstream of nucleosome, except open chromatin regions which are close downstream of the nucleosome. .

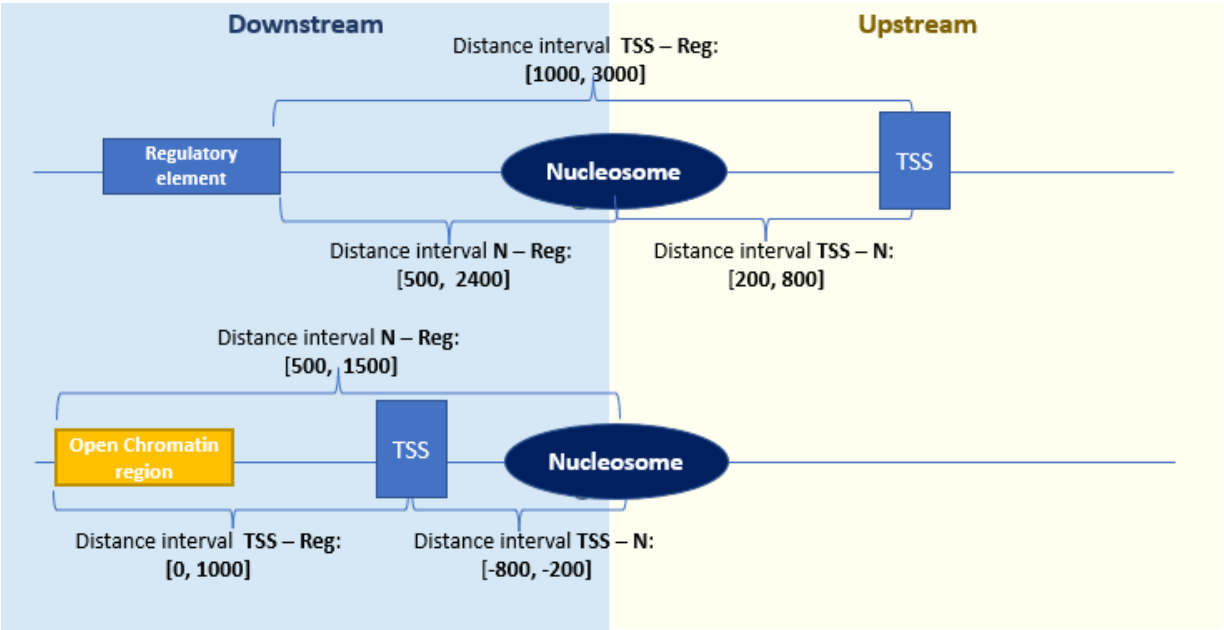

Figure 16 Nucleosome position profile 4

Fifth group nucleosome position analysis is shown in Figure 17. Regulatory element position is downstream in this group 400-2200 bp from nucleosome. TSS is 100-800 bp downstream from nucleosome and distance between regulatory element and TSS can be up to 100 bp, when TSS is upstream of regulatory element and up to 1800 bp when TSS is downstream from the regulatory element.

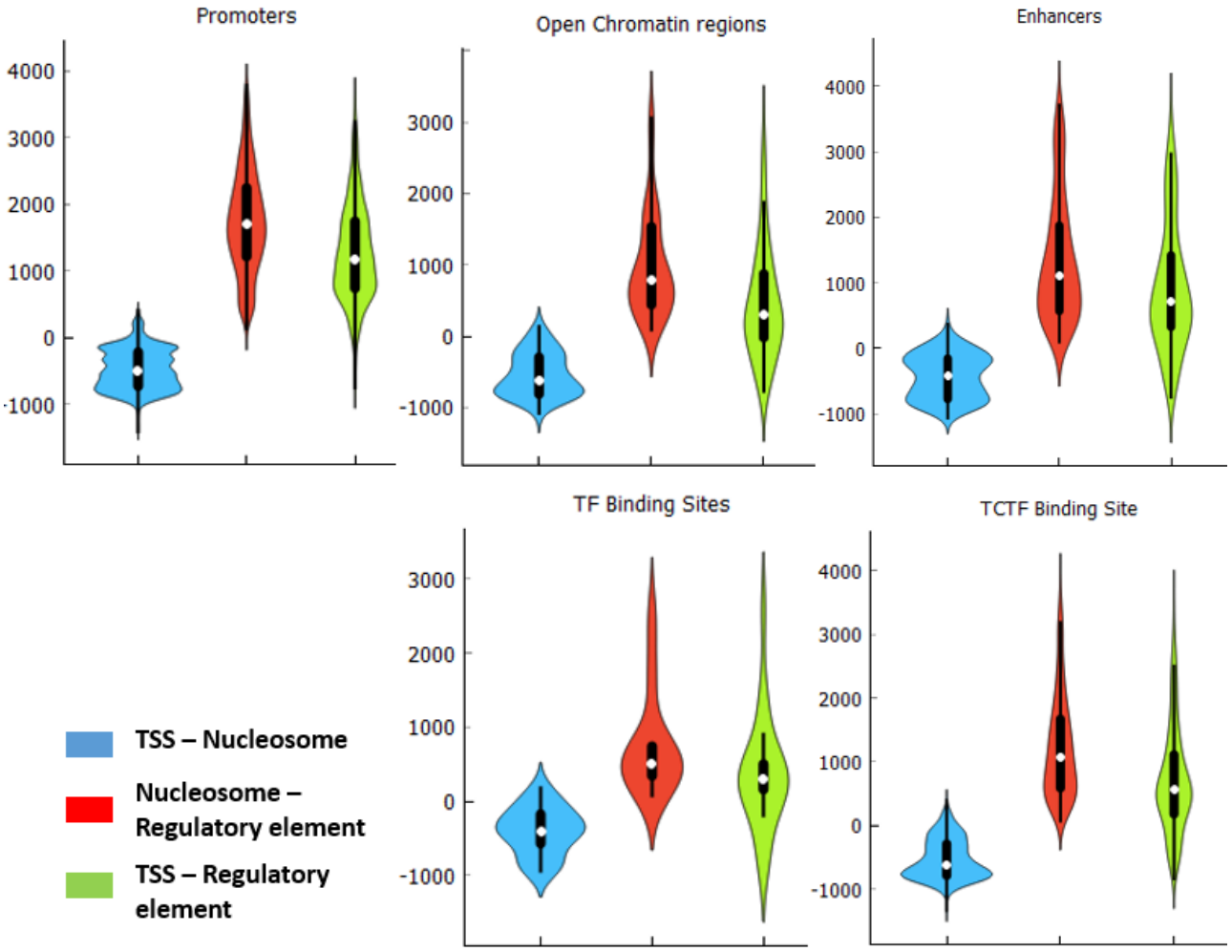

Figure 17 Group 5 distance distributions

Visual representation of the profile 5 for nucleosome, regulatory element and TSS configuration is show in Figure 18 and can be summarised as regulatory element further downstream of nucleosome (400-2200 bp) and TSS position is close (<1000bp) downstream of nucleosome. With TSS possibly being on either side of the regulatory element.

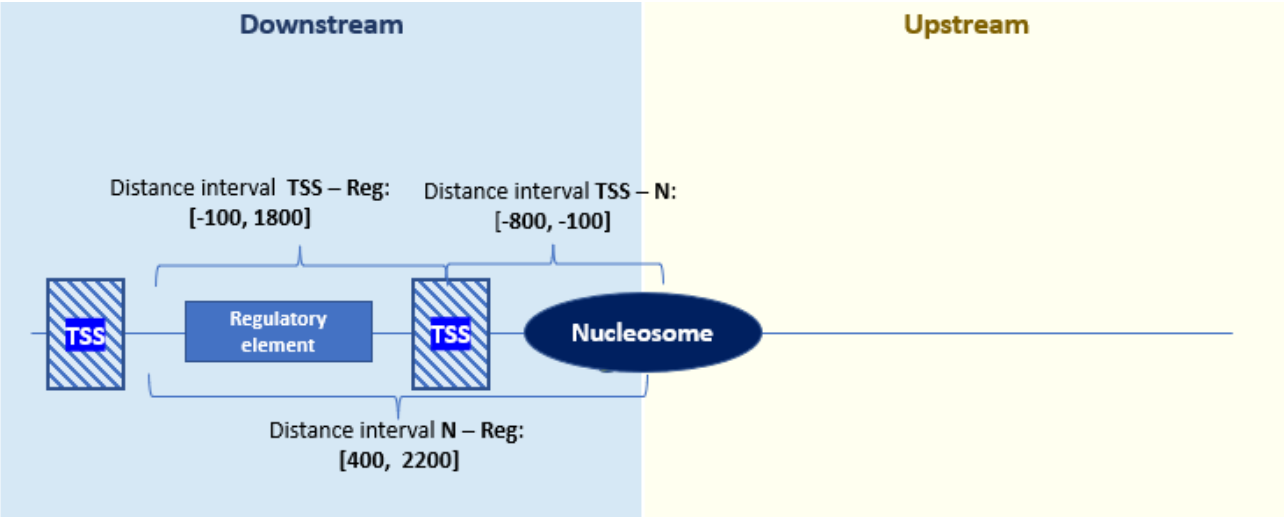

Figure 18 Nucleosome position profile 5

Gene overexpression analysis with data from GO biological processes, KEGGS and Reactome libraries was done for each set of genes in nucleosome profile classes and showed that there are some related gene groups present in each of the classes and frequently these groups are connected to each other thematically. Enrichment analysis results gene network for Profile 1 gene group is shown in Figure 19Figure 19 and complete list of genes for all profiles is provided in annex 2.

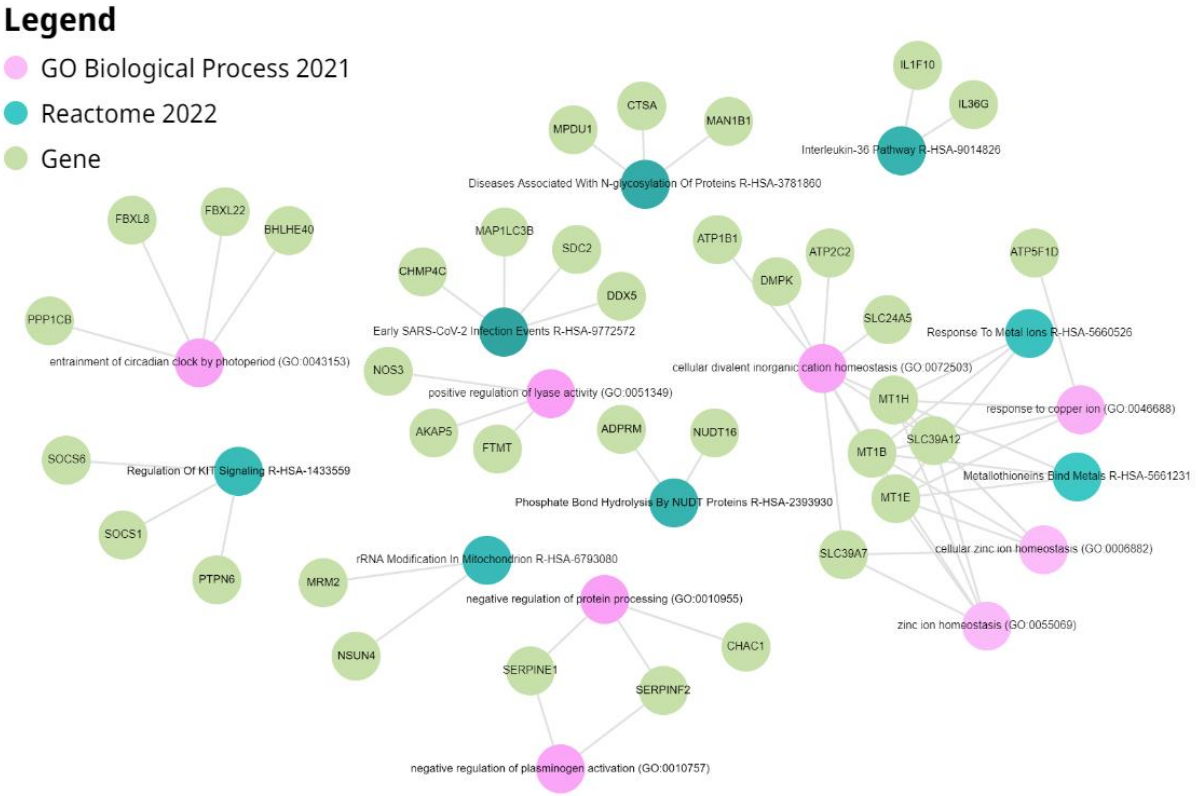

Figure 19 Profile 1 genes enrichment analysis results

The identified gene subgroups for profile 1 are ordered by statistical significance in Figure 20 and it is showed that a common theme in the most significant associated gene groups is **metal ions, namely - zinc and copper ion related processes and pathways** from GO Human biological processes and Reactome libraries, however no significant associations were found within KEGGs gene products data base.

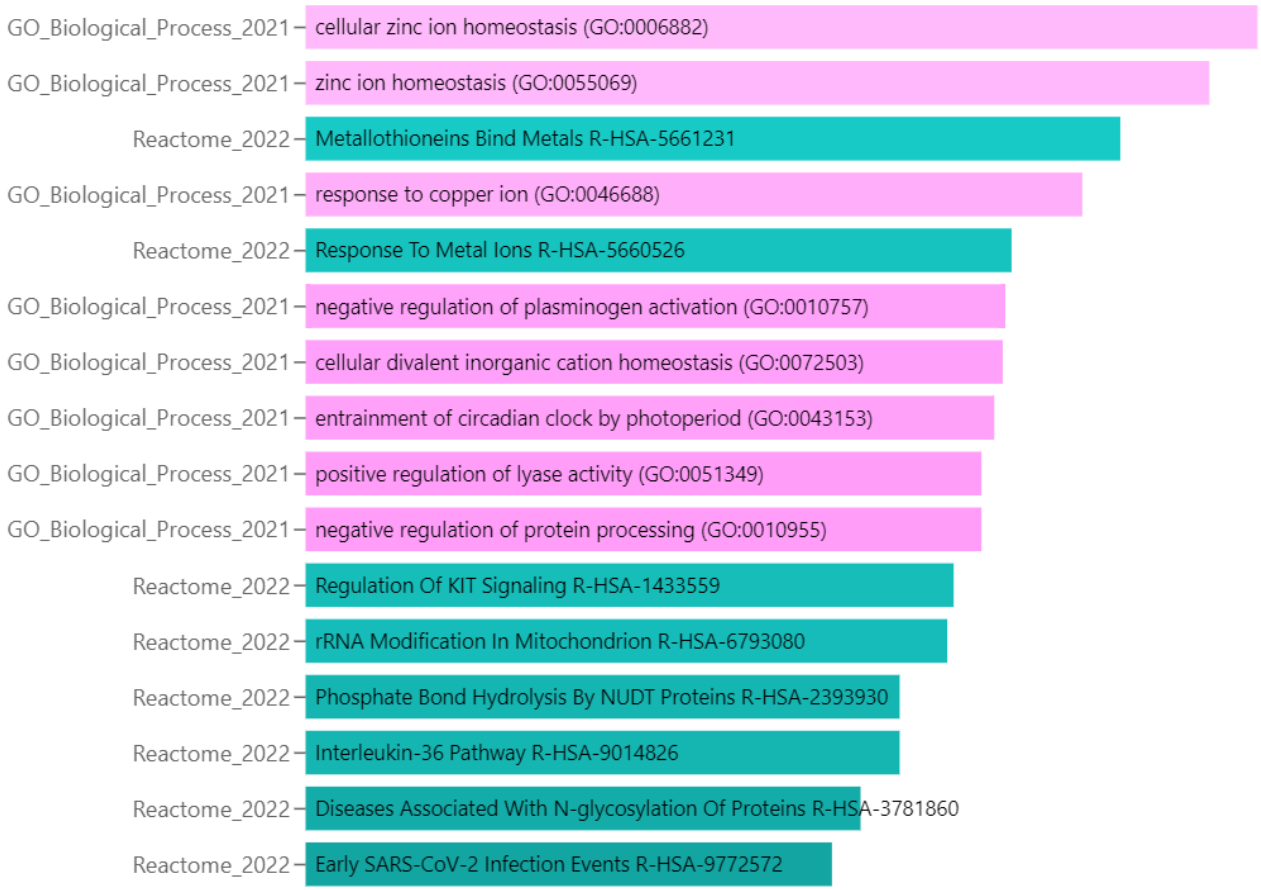

Figure 20 Profile 1 gene associations by significance

Enrichment analysis results gene network for nucleosome positioning Profile 2 gene group is shown below in Figure 21, this group contained the least number of genes and the analysis identified least associations.

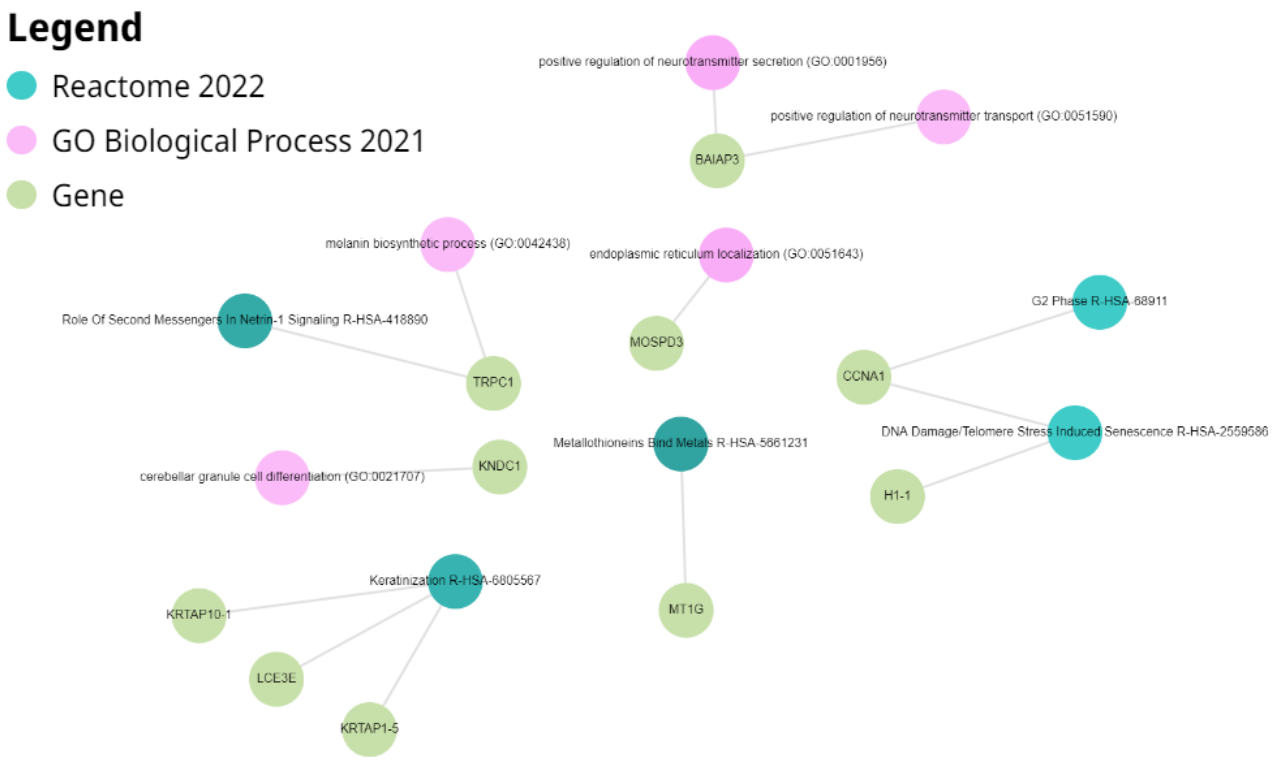

Figure 21 Profile 2 genes enrichment analysis results

The identified gene subgroups for profile 2 are ordered by statistical significance in Figure 22 and it is showed that the most significant associated biological processes are DNA Damage/Telomere Stress Induced Senescence, melanin biosynthetic process, positive regulation of neurotransmitter transport, cerebellar granule cell differentiation, G2 Phase from GO Human biological processes and Reactome libraries, however no significant associations were found within KEGGs gene products data base. All identified processes are directly or indirectly related **to cell cycle**.

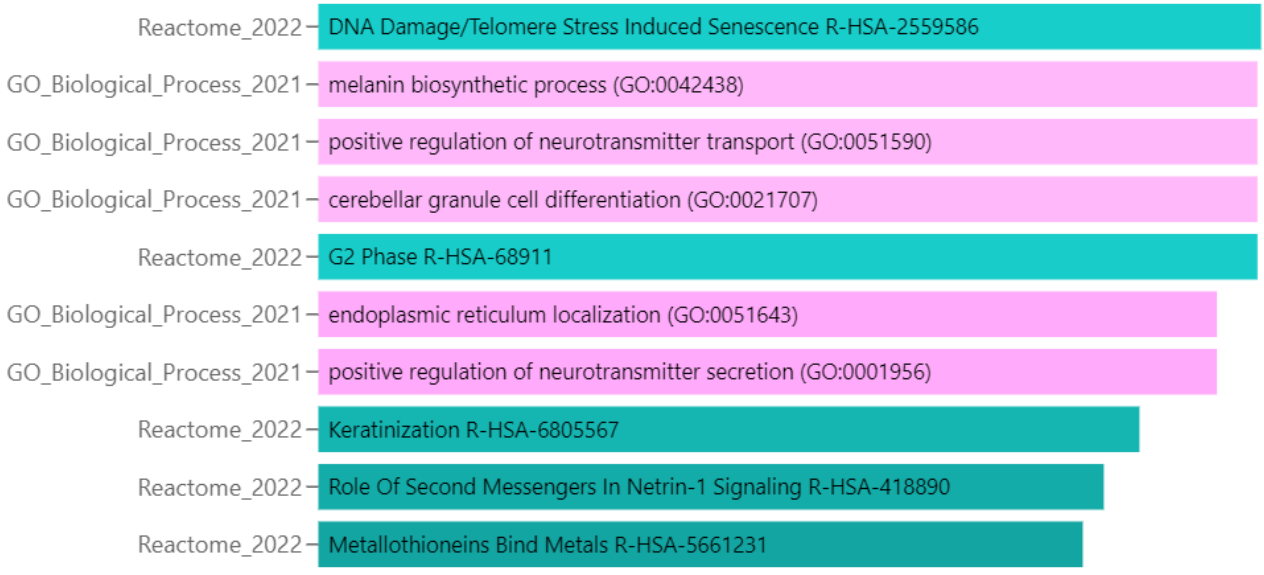

Figure 22 Profile 2 gene associations by significance

Enrichment analysis results gene network for nucleosome positioning Profile 3 gene group is shown below in Figure 23.

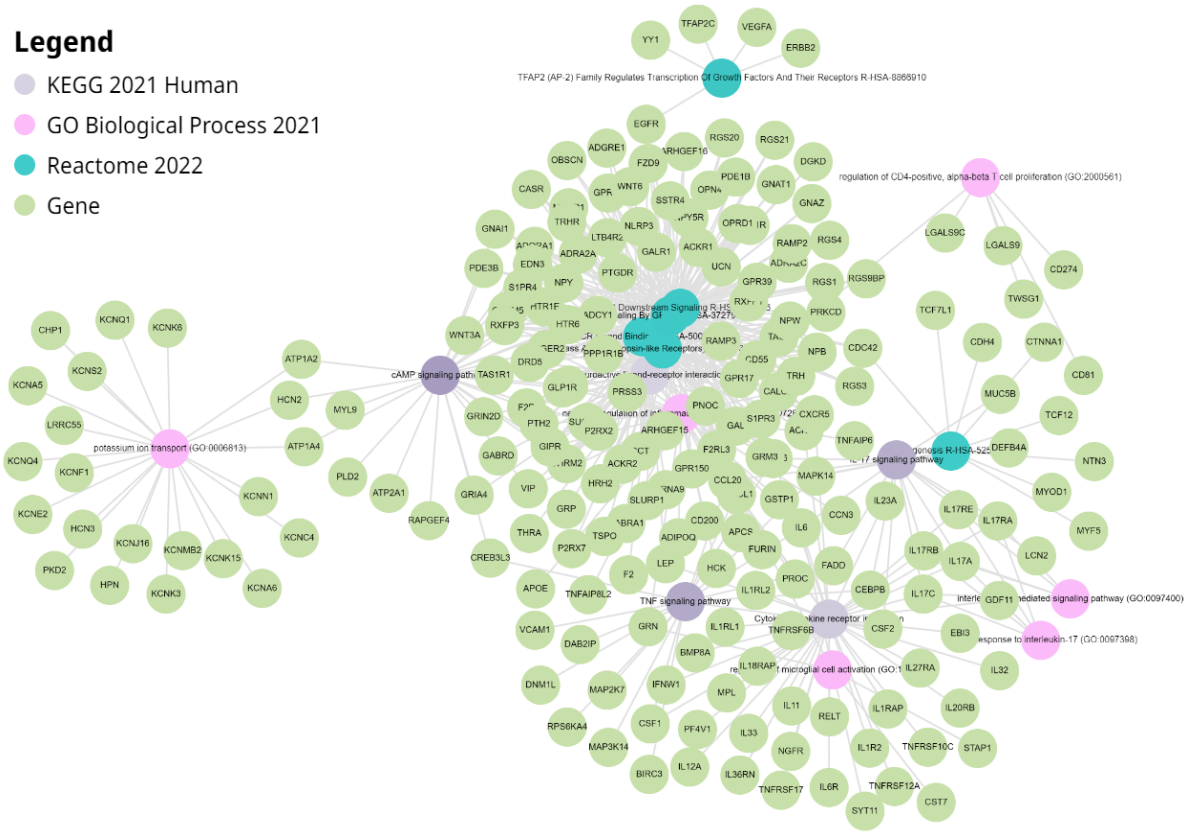

Figure 23 Profile 3 genes enrichment analysis results

Gene analysis results for Profile 3 are ordered by statistical significance in Figure 24 and it is showed that the most significant associated biological processes are Neuroactive ligand-receptor interaction, regulation of CD4-positive, alpha-beta T cell proliferation, GPCR Ligand Binding, cellular response to interleukin-17, interleukin-17-mediated signaling pathway, negative regulation of inflammatory response, Myogenesis, Class A/1 (Rhodopsin-like Receptors) from which common theme of **immune system functions** can be seen.

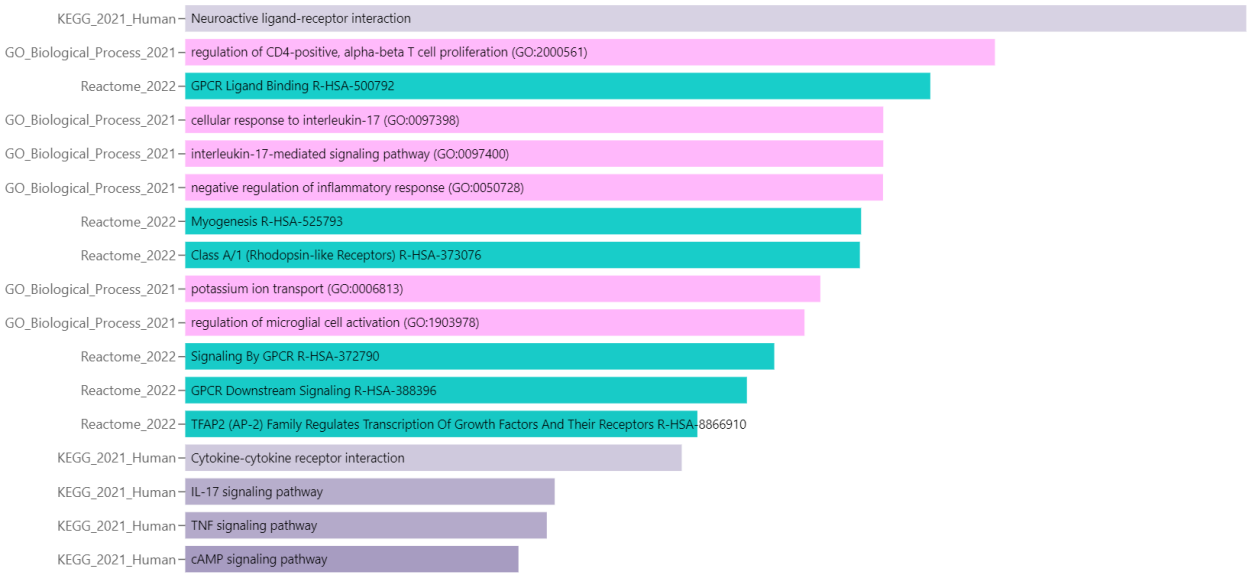

Figure 24 Profile 3 gene associations by significance

Enrichment analysis results gene network for nucleosome positioning Profile 4 gene group is shown below in Figure 25.

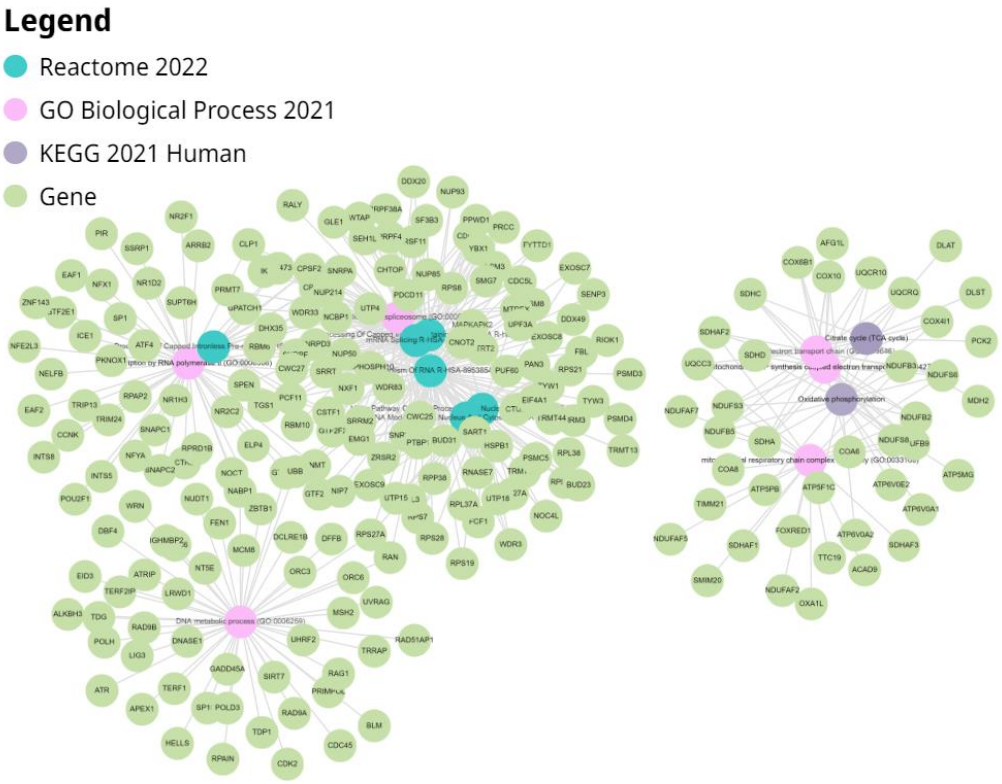

Figure 25 Profile 4 genes enrichment analysis results

The identified gene subgroups for Profile 4 are ordered by statistical significance in Figure 26 and it is showed that the most significant associated biological processes are from Reactome and GO biological process libraries and include: Metabolism Of RNA, DNA metabolic process, mitochondrial ATP synthesis coupled electron transport, Processing Of Capped Intron-Containing Pre-mRNA, aerobic electron transport chain, mitochondrial respiratory chain complex assembly, snRNA transcription by RNA polymerase II, RNA Polymerase II Transcribes snRNA Genes, mRNA Splicing, Citric Acid (TCA) Cycle And Respiratory Electron Transport, Respiratory Electron Transport, ATP Synthesis By Chemiosmotic Coupling, Heat Production By Uncoupling Proteins, Oxidative phosphorylation, Citrate cycle (TCA cycle). Three main themes can be seen: **DNA metabolic process**, **RNA processing** and **cellular respiration process**.

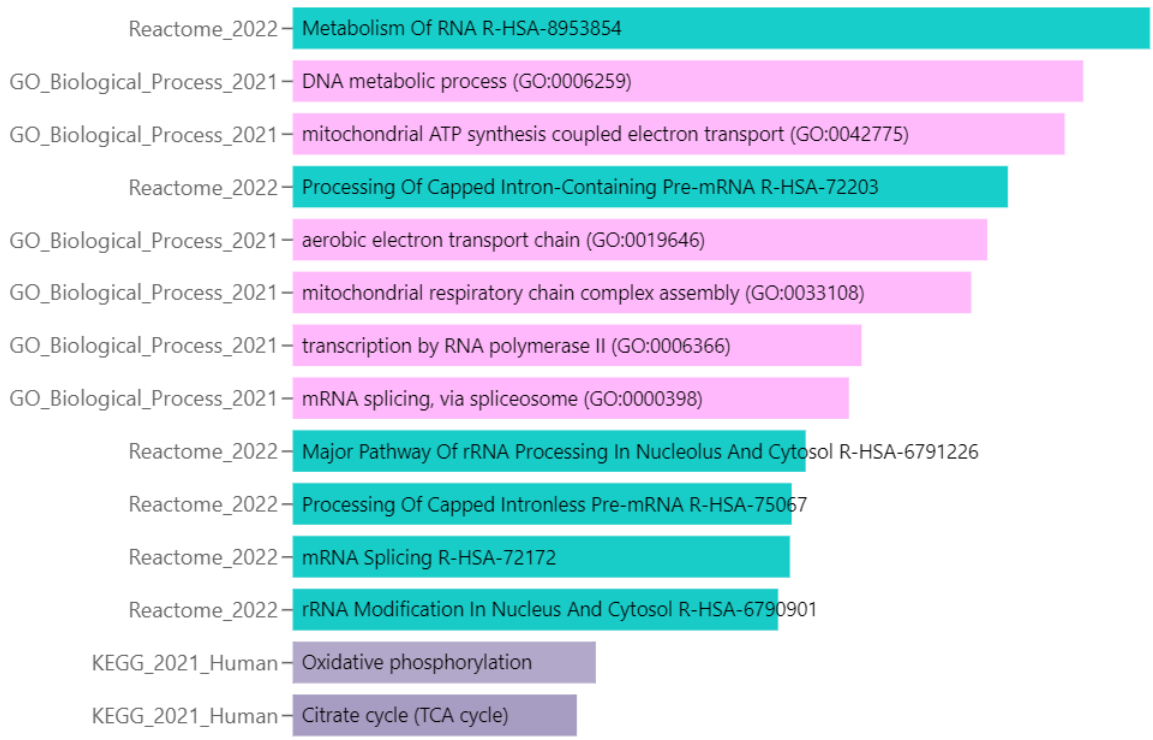

Figure 26 Profile 4 gene associations by significance

Enrichment analysis results gene network for nucleosome positioning Profile 5 gene group is shown below in Figure 27.

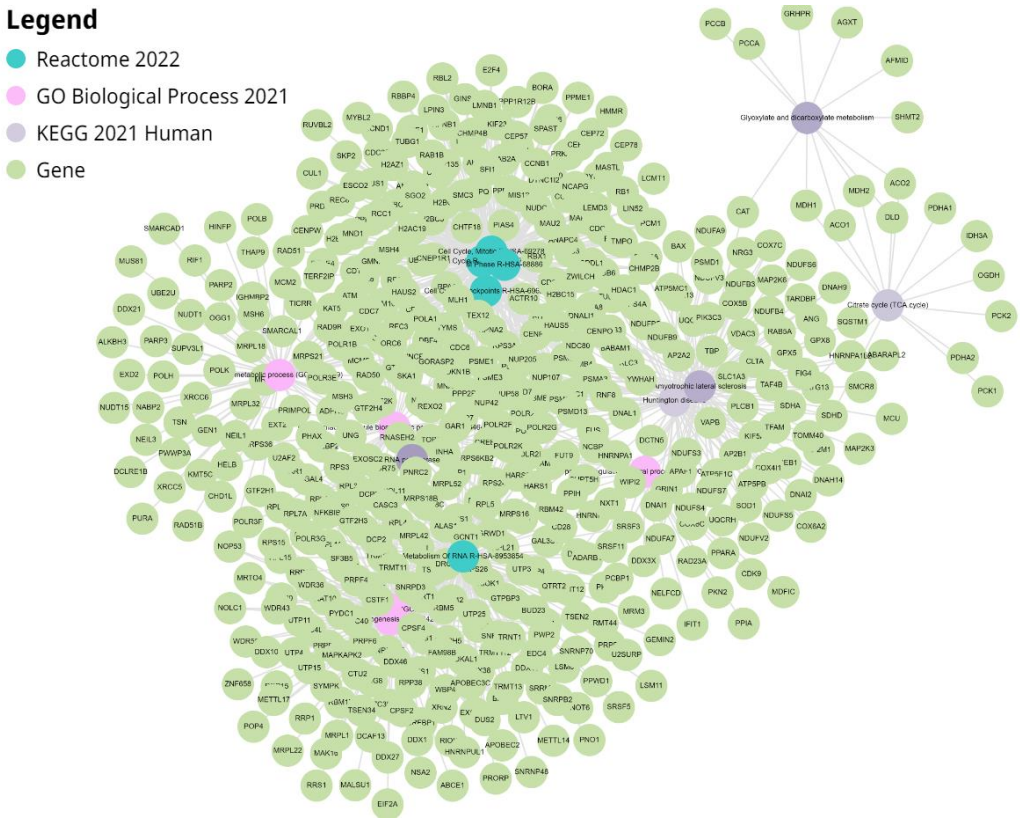

Figure 27 Profile 5 gene enrichment analysis results

The identified gene subgroups for Profile 5 are ordered by statistical significance in Figure 28 and it is showed that the most significant associated biological processes are from Reactome and GO biological process libraries and include: Cell Cycle, Cell Cycle-Mitotic, rRNA processing, Metabolism Of RNA, M Phase, positive regulation of viral process, Cell Cycle Checkpoints. Two main themes can be seen: **Cell Cycle regulation** and **RNA processing**.

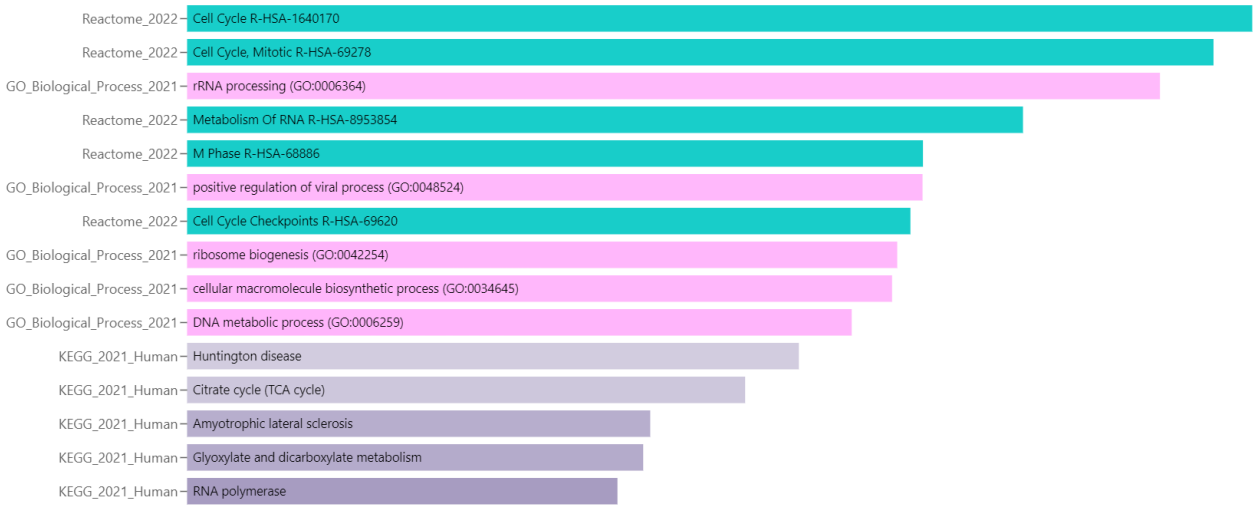

Figure 28 Profile 5 gene associations by significance

DISCUSSION

As nucleosome occupation patterns in relation to closest TSS and closest regulatory element were identified a further horizon investigation continuing this work could be investigation in the common SNP's in the intergenic locations caused by nucleotide deletion, insertion or substitution that result in non-silent mutation that occur in the nucleosome occupied regions with a hypothesis that these mutations would not have negative/disease inducing effect on the individual as they are covered by a nucleosome. For this work the dbSNP database could be used as a data source as it is a public database of single nucleotide polymorphisms (SNPs) and other variations in human DNA.

However, one limitation of this work is that we have explored and identified possible nucleosome position configurations based on data from one experiment, to further investigate stability of these nucleosome configurations datasets from other tissues and conditions could be analysed.

CONCLUSIONS

For this work, the nucleosome dataset was selected from GSE database based on well research cell line GM12878, control conditions of experiment and accessible nucleosome data format. Closest regulatory element and closest transcription start site was computationally found for each nucleosome and enriched dataset was prepared with calculated distances between these elements. The analysis showed that there exist different structures in Reg-N-TSS distances data, and its distribution is not random. Most likely distances have been identified for promoters, which in relation to nucleosome, tend to be downstream such that distance N – Reg is positive and tends to be between 1000 to 3000 bps away from nucleosome. Second tendency is that largest concentration of Transcription factor binding site type regulatory elements is seen in +1000/-1000 region around the nucleosome**.** However, contrary to what was expected the results show that various types of regulatory elements can have different types of distance distributions independently from their type. Next, 5 profiles of Reg-N-TSS distance distributions were identified:

- Profile 1 is characterised as regulatory element far (>1000 bp) upstream of nucleosome and TSS close (<1000 bp) downstream of nucleosome, with exception of TF binding site sub-group, for which TSS can also be upstream.
- Profile 2 is characterised as regulatory element very close to (<400 bp) on the either side of nucleosome and TSS is far (>900 bp) downstream of nucleosome.
- Profile 3 is characterised as regulatory element very close to nucleosome (from 200 bp downstream to 400 bp upstream) and TSS position is close (<1000bp) upstream of nucleosome, except for CTCF binding sites where it can also be in a close downstream position, also in promoter group TSS is always upstream from promoter.
- Profile 4 is characterised as regulatory element further downstream of nucleosome (500-2400 bp) and TSS position is close (<1000bp) upstream of nucleosome, except open chromatin regions which are close downstream of the nucleosome.
- Profile 5 is characterised as regulatory element further downstream of nucleosome (400-2200 bp) and TSS position is close (<1000bp) downstream of nucleosome. With TSS possibly being on either side of the regulatory element

Gene overexpression analysis revealed that groups of genes with these distribution profiles participate in common biological processes:

- In Profile 1 gene group that a common theme in the most significant associated gene groups is metal ions, namely - zinc and copper ion related processes and pathways.
- In Profile 2 gene group two main themes can be seen: identified processes are directly or indirectly related to cell cycle and ageing.
- Profile 3 gene group analysis results showed that common theme of immune system functions can be seen.

- In Profile 4 gene group three main themes appeared: DNA metabolic process, RNA processing and cellular respiration process.
- In Profile 5 gene group two main themes can be seen: Cell Cycle regulation and RNA processing.

RECOMMENDATION

If your project leads to some recommendations, please include them separately under Recommendations section. It should be a short structured one or two paragraphs.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Erinija Pranckeviciene for academic and moral support in writing this work, contagious excitement towards the nucleosome occupancy questions and ability to provide clear and interesting answers to complex topics.

I would also like to thank my girlfriend and my family for support in this stressful but very interesting period.

REFERENCES

Albert, Istvan, Travis N. Mavrich, Lynn P. Tomsho, Ji Qi, Sara J. Zanton, Stephan C. Schuster, and B. Franklin Pugh. 2007. “Translational and Rotational Settings of H2A.Z Nucleosomes across the Saccharomyces Cerevisiae Genome.” *Nature* 446 (7135): 572–76. https://doi.org/10.1038/nature05632.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. “Chromosomal DNA and Its Packaging in the Chromatin Fiber.” *Molecular Biology of the Cell. 4th Edition*. https://www.ncbi.nlm.nih.gov/books/NBK26834/.

Bai, Lu, and Alexandre V. Morozov. 2010. “Gene Regulation by Nucleosome Positioning.” *Trends in Genetics* 26 (11): 476–83. https://doi.org/10.1016/j.tig.2010.08.003.

Cairns, Bradley R. 2005. “Chromatin Remodeling Complexes: Strength in Diversity, Precision through Specialization.” *Current Opinion in Genetics & Development*, Chromosomes and expression mechanisms, 15 (2): 185–90. https://doi.org/10.1016/j.gde.2005.01.003.

Cohanim, Amir B., and Tali E. Haran. 2009. “The Coexistence of the Nucleosome Positioning Code with the Genetic Code on Eukaryotic Genomes.” *Nucleic Acids Research* 37 (19): 6466–76. https://doi.org/10.1093/nar/gkp689.

Cole, Lauren, and Jonathan Dennis. 2020. “MNase Profiling of Promoter Chromatin in *Salmonella Typhimurium*-Stimulated GM12878 Cells Reveals Dynamic and Response-Specific Nucleosome Architecture.” *G3 (Bethesda, Md.)* 10 (7): 2171–78. https://doi.org/10.1534/g3.120.401266.

Cui, Feng, Linlin Chen, Peter R LoVerso, and Victor B Zhurkin. 2014. “Prediction of Nucleosome Rotational Positioning in Yeast and Human Genomes Based on Sequence-Dependent DNA Anisotropy.” *BMC Bioinformatics* 15 (1): 313. https://doi.org/10.1186/1471-2105-15-313.

Do, Van Hoan, and Stefan Canzar. 2021. “A Generalization of T-SNE and UMAP to Single-Cell Multimodal Omics.” *Genome Biology* 22 (1): 130. https://doi.org/10.1186/s13059-021-02356-5.

Duttke, Sascha H. C., Scott A. Lacadie, Mahmoud M. Ibrahim, Christopher K. Glass, David L. Corcoran, Christopher Benner, Sven Heinz, James T. Kadonaga, and Uwe Ohler. 2015. “Human Promoters Are Intrinsically Directional.” *Molecular Cell* 57 (4): 674–84. https://doi.org/10.1016/j.molcel.2014.12.029.

“Enrichr-KG.” n.d. Accessed May 13, 2023. https://maayanlab.cloud/enrichr-kg.

Eslami-Mossallam, Behrouz, Raoul D. Schram, Marco Tompitak, John van Noort, and Helmut Schiessel. 2016. “Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach.” *PLOS ONE* 11 (6): e0156905. https://doi.org/10.1371/journal.pone.0156905.

Field, Andrew, and Karen Adelman. 2020. “Evaluating Enhancer Function and Transcription.” *Annual Review of Biochemistry* 89 (1): 213–34. https://doi.org/10.1146/annurev-biochem-011420-095916.

“GEO Accession Viewer.” n.d. Accessed May 13, 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139224.

Gromiha, M. Michael. 2000. “Structure Based Sequence Dependent Stiffness Scale for Trinucleotides: A Direct Method.” *Journal of Biological Physics* 26 (1): 43–50. https://doi.org/10.1023/A:1005250718139.

Hansen, Anders S. 2020. “CTCF as a Boundary Factor for Cohesin-Mediated Loop Extrusion: Evidence for a Multi-Step Mechanism.” *Nucleus* 11 (1): 132–48. https://doi.org/10.1080/19491034.2020.1782024.

Ioshikhes, Ilya, Sergey Hosid, and B. Franklin Pugh. 2011. “Variety of Genomic DNA Patterns for Nucleosome Positioning.” *Genome Research* 21 (11): 1863–71. https://doi.org/10.1101/gr.116228.110.

Iyer, Vishwanath R. 2012. “Nucleosome Positioning: Bringing Order to the Eukaryotic Genome.” *Trends in Cell Biology* 22 (5): 250–56. https://doi.org/10.1016/j.tcb.2012.02.004.

Johnson, Steven M., Frederick J. Tan, Heather L. McCullough, Daniel P. Riordan, and Andrew Z. Fire. 2006. “Flexibility and Constraint in the Nucleosome Core Landscape of Caenorhabditis Elegans Chromatin.” *Genome Research* 16 (12): 1505–16. https://doi.org/10.1101/gr.5560806.

Kaplan, Noam, Irene Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Desiree Tillo, Yair Field, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. 2010. “Nucleosome Sequence Preferences Influence in Vivo Nucleosome Organization.” *Nature Structural & Molecular Biology* 17 (8): 918–22. https://doi.org/10.1038/nsmb0810-918.

Kaplan, Noam, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, et al. 2009. “The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome.” *Nature* 458 (7236): 362–66. https://doi.org/10.1038/nature07667.

Kribelbauer, Judith F., Chaitanya Rastogi, Harmen J. Bussemaker, and Richard S. Mann. 2019. “Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes.” *Annual Review of Cell and Developmental Biology* 35 (1): 357–79. https://doi.org/10.1146/annurev-cellbio-100617-062719.

Langley, Sasha A., Gary H. Karpen, and Charles H. Langley. 2014. “Nucleosomes Shape DNA Polymorphism and Divergence.” *PLoS Genetics* 10 (7): e1004457. https://doi.org/10.1371/journal.pgen.1004457.

Lövkvist, Cecilia, Kim Sneppen, and Jan O. Haerter. 2018. “Exploring the Link between Nucleosome Occupancy and DNA Methylation.” *Frontiers in Genetics* 8 (January): 232. https://doi.org/10.3389/fgene.2017.00232.

Mavrich, Travis N., Cizhong Jiang, Ilya P. Ioshikhes, Xiaoyong Li, Bryan J. Venters, Sara J. Zanton, Lynn P. Tomsho, et al. 2008. “Nucleosome Organization in the Drosophila Genome.” *Nature* 453 (7193): 358–62. https://doi.org/10.1038/nature06929.

Onufriev, Alexey V, and Helmut Schiessel. 2019. “The Nucleosome: From Structure to Function through Physics.” *Current Opinion in Structural Biology* 56 (June): 119–30. https://doi.org/10.1016/j.sbi.2018.11.003.

Oruba, Agata, Simona Saccani, and Dominic van Essen. 2020. “Role of Cell-Type Specific Nucleosome Positioning in Inducible Activation of Mammalian Promoters.” *Nature Communications* 11 (1): 1075. https://doi.org/10.1038/s41467-020-14950-5.

Özdemir, Isa, and Maria Cristina Gambetta. 2019. “The Role of Insulation in Patterning Gene Expression.” *Genes* 10 (10): 767. https://doi.org/10.3390/genes10100767.

Pang, Baoxu, and Michael P. Snyder. 2020. “Systematic Identification of Silencers in Human Cells.” *Nature Genetics* 52 (3): 254–63. https://doi.org/10.1038/s41588-020-0578-5.

Pham, Chuong D., Xi He, and Gavin R. Schnitzler. 2010. “Divergent Human Remodeling Complexes Remove Nucleosomes from Strong Positioning Sequences.” *Nucleic Acids Research* 38 (2): 400–413. https://doi.org/10.1093/nar/gkp1030.

Pranckeviciene, Erinija, Sergey Hosid, Nathan Liang, and Ilya Ioshikhes. 2020. “Nucleosome Positioning Sequence Patterns as Packing or Regulatory.” *PLOS Computational Biology* 16 (1): e1007365. https://doi.org/10.1371/journal.pcbi.1007365.

Radman-Livaja, Marta, and Oliver J. Rando. 2010. “Nucleosome Positioning: How Is It Established, and Why Does It Matter?” *Developmental Biology* 339 (2): 258–66. https://doi.org/10.1016/j.ydbio.2009.06.012.

Rhie, Suhn Kyong, Dennis J. Hazelett, Simon G. Coetzee, Chunli Yan, Houtan Noushmehr, and Gerhard A. Coetzee. 2014. “Nucleosome Positioning and Histone Modifications Define Relationships between Regulatory Elements and Nearby Gene Expression in Breast Epithelial Cells.” *BMC Genomics* 15 (1): 331. https://doi.org/10.1186/1471-2164-15-331.

Satchwell, S. C., H. R. Drew, and A. A. Travers. 1986. “Sequence Periodicities in Chicken Nucleosome Core DNA.” *Journal of Molecular Biology* 191 (4): 659–75. https://doi.org/10.1016/0022-2836(86)90452-3.

Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer–Promoter Contacts in Gene Expression Control.” *Nature Reviews Genetics* 20 (8): 437–55. https://doi.org/10.1038/s41576-019-0128-0.

Segal, Eran, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. 2006. “A Genomic Code for Nucleosome Positioning.” *Nature* 442 (7104): 772–78. https://doi.org/10.1038/nature04979.

Sikorska, Natalia, and Tom Sexton. 2020. “Defining Functionally Relevant Spatial Chromatin Domains: It Is a TAD Complicated.” *Journal of Molecular Biology*, Perspectives on Chromosome Folding, 432 (3): 653–64. https://doi.org/10.1016/j.jmb.2019.12.006.

Struhl, Kevin, and Eran Segal. 2013. “Determinants of Nucleosome Positioning.” *Nature Structural & Molecular Biology* 20 (3): 267–73. https://doi.org/10.1038/nsmb.2506.

Tompitak, Marco, Cédric Vaillant, and Helmut Schiessel. 2017. “Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions.” *Biophysical Journal* 112 (3): 505–11. https://doi.org/10.1016/j.bpj.2016.12.041.

Travers, A., and H. Drew. 1997. “DNA Recognition and Nucleosome Organization.” *Biopolymers* 44 (4): 423–33. https://doi.org/10.1002/(SICI)1097-0282(1997)44:4<423::AID-BIP6>3.0.CO;2-M.

Trifonov, Edward N. 2011. “Thirty Years of Multiple Sequence Codes.” *Genomics, Proteomics & Bioinformatics* 9 (1–2): 1–6. https://doi.org/10.1016/S1672-0229(11)60001-6.

Varga-Weisz, Patrick D., and Peter B. Becker. 1995. “Transcription Factor-Mediated Chromatin Remodelling: Mechanisms and Models.” *FEBS Letters* 369 (1): 118–21. https://doi.org/10.1016/0014-5793(95)00549-O.

Wright, Gregory M., and Feng Cui. 2017. “Nucleosome Rotational Positioning Is Influenced by Cis- and Trans-Acting Factors.” bioRxiv. https://doi.org/10.1101/207167.

Wright, Gregory M, and Feng Cui. 2019. “The Nucleosome Position-Encoding WW/SS Sequence Pattern Is Depleted in Mammalian Genes Relative to Other Eukaryotes.” *Nucleic Acids Research* 47 (15): 7942–54. https://doi.org/10.1093/nar/gkz544.

Wu, Cheng, Jiaqi Yang, Wei Xiao, Zehang Jiang, Shuxia Chen, Dianlei Guo, Ping Zhang, Chunqiao Liu, Huasheng Yang, and Zhi Xie. 2022. “Single-Cell Characterization of Malignant Phenotypes and Microenvironment Alteration in Retinoblastoma.” *Cell Death & Disease* 13 (5): 1–12. https://doi.org/10.1038/s41419-022-04904-8.

Zhou, Tai Lai, Heng Xin Chen, Yin Zhao Wang, Si Jie Wen, Ping Hong Dao, Yu Hang Wang, and Min Feng Chen. 2023. “Single-Cell RNA Sequencing Reveals the Immune Microenvironment and Signaling Networks in Cystitis Glandularis.” *Frontiers in Immunology* 14 (February): 1083598. https://doi.org/10.3389/fimmu.2023.1083598.

SUMMARY

Computational Mapping Of Nucleosomes In Human Genome: Occupancy At The Elements Of A Regulatory Build

Gabriele Stockunaite
Master Thesis
Systems biology master program
Vilnius university

This work investigates patterns of nucleosome occupancy by measuring the base pair distances between nucleosomes and regulatory elements and TSS in the human genome. The goal of this work is to identify nucleosome occupancy patterns in regulatory regions in relation to the activity of the genes that are regulated.

It was done by collecting coordinates of regulatory elements, TSS and nucleosome positions for human organism from selected existing data on academic databases. Then, creating a table of distances between them. Then, performing analysis of distance distributions between nucleosomes, closest TSS and closest regulatory element coordinates (or positions) to confirm hypothesis that patterns exist by using unsupervised learning algorithms – PCA, UMAP and t-SNE. Finally, perform gene overexpression analysis on the classified gene groups to identify any common biological process themes.

The conclusions were made that most likely distance for promoters, which in relation to nucleosome, tend to be downstream such that distance N – Reg is positive and between 1000 to 3000 bps. Second tendency is that largest concentration of Transcription factor binding site regulatory elements is seen in +1000/-1000 region around the nucleosome. However, contrary to what was expected the results show that various types of regulatory elements can have different types of distance distributions independently from their type.

Finally, 5 profiles of Reg-N-TSS distance distributions were found and for each gene group an overexpression analysis was done finding dominant topics for identified profiles such as participation in metal ions metabolism processes or immune system related biological processes and gene product pathways.

SUMMARY IN LITHUANIAN

Nukleosomų išsidėstymo identifikavimas skaičiuojamuoju būdu reguliacinių genomo elementų žmogaus genome atžvilgiu

Gabriele Stockunaite
Magistro darbas
Sistemų Biologjos magistro strudijos
Vilniaus universitetas

Šiame darbe tiriami nukleosomų išsidėstymo modeliai, pagal atstumus (bazių poromis) tarp nukleosomų ir reguliacinių elementų bei transkripcijos pradžios vietų žmogaus genome. Šio darbo tikslas yra nustatyti nukleosomų išsidėstymo modelius/dėsningumus reguliaciniuose regionuose ir jų sąąsajas su gretimų genų savybėmis.

Tai atlikta surenkant žmogaus organizmo reguliavimo elementų koordinates, TSS ir nukleosomų pozicijas iš pasirinktų duomenų akademinėse duomenų bazėse. Tuomet duomenys buvo paruošti ir komputaciniu būdu parengta šių elementų atstumų lentelė. Atlikta atstumų tarp nukleosomų, artimiausių TSS ir artimiausių reguliacinių elementų koordinačių (arba pozicijų) analizė naudojant nepriklausomo mašininio mokymosi algoritmus – PCA, UMAP ir t-SNE, siekiant patvirtinti hipotezę, kad egzistuoja nukleosomų išsidėstymo dėsningumai. Galiausiai atlikta nustatytų genų grupių

ekspresijos analizė, įvertinti ar egzistuoja bendri biologiniai procesai, kuriuose dalyvauja šių grupių genai.
Buvo padarytos išvados, kad labiausiai tikėtinas promotorių išsidėstymas, yra prieš nukleosomą, todėl atstumas N – Reg yra teigiamas ir siekia nuo 1000 iki 3000 bazių porų. Antroji tendencija yra ta, kad didžiausia transkripcijos tipo reguliacinių elementų koncentracija yra +1000/-1000 srityje aplink nukleosomą. Tačiau, priešingai nei tikėtasi, rezultatai rodo, kad įvairių tipų reguliaciniai elementai gali turėti skirtingus atstumų išsidėstymus aplink nukleosomą, nepriklausomai nuo jų tipo. Galiausiai, buvo rasti 5 Reg-N-TSS atstumų išsidėstymo profiliai / klasės ir kiekvienai genų grupei buvo atlikta genų ekspresijos analizė ir identifikuotos dominuojančios biologinių procesų temos kiekvienam profiliui / klasei, tokios kaip dalyvavimas metalo jonų apykaitos procesuose arba imuninės sistemos veikloje.

APPENDICES

Appendix 1: Reviewed Nucleosome datasets in GSE database

GEO ID	Cell type	Description	Citations
GSE35586	the GM12878 and K562 ENCODE cell lines	To isolate mononucleosome core DNA fragments from the GM12878 and K562 ENCODE cell lines we followed the micrococcal nuclease (MNase) digestion and isolation protocol	
GSE120857	Mnase-seq of nucleosome positioning in ctrl/N-myc_kd neuroblastoma cell line BE2C	We knock down MYCN in human neuroblastoma cell line by RNAi. MYCN knock-down caused global epigenome change, including nucleosome gain on key DNA-repair genes and promoter H3K9ac down regulation.	Hu X, Zheng W, Zhu Q, Gu L et al. Increase in DNA Damage by MYCN Knockdown Through Regulating Nucleosome Organization and Chromatin State in Neuroblastoma. Front Genet 2019;10:684. PMID: 31396265
GSE135551	MNase-seq analyses for Jurkat cells infected with HIV (NL4-3 WT or IND116A)	Genome binding/occupancy profiling by high throughput sequencingWe used the technology of MNase-seq and capture-based target enrichment to analyze nucleosome positioning on the HIV DNA.	Machida S, Depierre D, Chen HC, Thenin-Houssier S et al. Exploring histone loading on HIV DNA reveals a dynamic nucleosome positioning between unintegrated and integrated viral genome. Proc Natl Acad Sci U S A 2020 Mar 24;117(12):6822-6830. PMID: 32161134
GSE76344	Histone modification and nucleosome mapping in human liver cancer cells	The nucleosome is a fundamental unit of chromatin architecture, but it has been under-evaluated during the analysis of histone modification using ChIP-Seq. Thus, we developed a novel approach for defining histone modification at single nucleosome resolution by synergistic analyses of histone modification data from ChIP-Seq and high-resolution nucleosome mapping data from native MNase-Seq using a PCR-free strategy. With this approach, we have generated histone modification data at the single nucleosome resolution in human embryonic stem cells, normal hepatocytes, and liver cancer cells of both genders, and performed quantitative analysis of epigenetic regulation in stem cell differentiation into hepatocytes and neoplastic transformation of liver cancer.	Zheng D, Trynda J, Sun Z, Li Z. NUCLIZE for quantifying epigenome: generating histone modification data at single-nucleosome resolution using genuine nucleosome positions. BMC Genomics 2019 Jul 2;20(1):541. PMID: 31266464

GEO ID	Cell type	Description	Citations
GSE214212	RNA-seq, ChIP-seq and Mnase-Seq analysis of CHD6-silenced prostate cancer cells C4-2	Purpose: Study of the role of CHD6 linking nucleosome ejection in castration-resistant prostate cancer(CRPC) Method: The expression of CHD6 or E2F1 was silenced by 2 shRNAs (3 replicates) targeted at CHD6 or E2F1 in C4-2 cells, scramble RNA as control. mRNA profiles and genome-wide chromatin-state maps were generated by deep sequencing. CHD6 and IgG ChIP was conducted in C4-2 cells. MNase before and after CHD6-silenced was conducted in C4-2 cells. Results: E2F1 and E2F1 downstream genes were significantly down regulated by CHD6 knockdown. The binding of CHD6 on chromatin is required for nucleosome eviction in transcriptional activation of oncogenic pathways.	Zhao D, Zhang M, Huang S, Liu Q et al. CHD6 promotes broad nucleosome eviction for transcriptional activation in prostate cancer cells. Nucleic Acids Res 2022 Nov 28;50(21):12186-12201. PMID: 36408932
GSE176336	The Enrichment and Regulation of High-methylated CpG sites in CGI Border in ARPE-19 [MNase-seq]	table maintenance of cytosine methylation is essential for mammalian biological processes such as gene transcription. To investigate the relationship of stably high methylation sites with gene expression, we cultured the ARPE-19 for lasting 10 generations and performed reduced representation bisulfite sequencing (RRBS) to obtain the highly methylated sites. Then we analyzed the characters of highly methylated sites and detected the transcription level of the cells.	
GSE148935	LNCaP cells was used as cell model, Mnase-seq and Mnase-ChIP-seq were applied for mono-nucleosome profiling and ChIP-exo of GATA2 compariing in vehicle and DHT conditions.	Methods: LNCaP cells between passage number 30-35 were used for assay. cell nucleus was extracted, digested by MNase to Mono-nucleosome and ChIP/ChIP-exo was performed, the ChIP products were further used to generate library with illumina ChIP-seq kit. Hi-seq 3000 was used for sequencing and the data was analyzed by MACS2 for peaks.	Li T, Liu Q, Chen Z, Fang K et al. Dynamic nucleosome landscape elicits a noncanonical GATA2 pioneer model. Nat Commun 2022 Jun 7;13(1):3145. PMID: 35672415
GSE134297	Multiple roles of H2A.Z in regulating promoter chromatin architecture in human cells (Mnase-seq & ChIP-seq)	We employed an MNase-Transcription Start Site Sequence Capture method to map and determine the accessibility of all nucleosomes, including those that contain H2A.Z, at high coverage for all human Pol II promoters.	Cole L, Kurscheid S, Nekrasov M, Domaschenz R et al. Multiple roles of H2A.Z in regulating promoter chromatin architecture in human cells. Nat Commun 2021 May 5;12(1):2524. PMID: 33953180

GEO ID	Cell type	Description	Citations
GSE139224	Nucleosome mapping via MNase-seq during heat-killed Salmonella typhimurium (HKST) stimulation in GM12878 cells.	Genome binding/occupancy profiling by high throughput sequencing Summary We employed an MNase-Transcription Start Site Sequence Capture method to map and determine the accessibility of all nucleosomes during immune stimulus, at high coverage for all human Pol II promoters.	Cole L, Dennis J. MNase Profiling of Promoter Chromatin in Salmonella typhimurium-Stimulated GM12878 Cells Reveals Dynamic and Response-Specific Nucleosome Architecture. G3 (Bethesda) 2020 Jul 7;10(7):2171-2178. PMID: 32404364

Appendix 2: Gene overexpression analysis results

Profile 1	
GO Biological Process	**Genes**
Cellular divalent inorganic cation homeostasis	SLC39A7, MT1H, SLC39A12, SLC24A5, ATP2C2, DMPK, MT1E, MT1B, ATP1B1
Zinc ion homeostasis	SLC39A12, SLC39A7, MT1E, MT1H, MT1B
Cellular zinc ion homeostasis	SLC39A12, MT1B, MT1E, MT1H, SLC39A7
Response to copper ion (GO:0046688	MT1H, ATP5F1D, MT1E, and MT1B
Negative regulation of plasminogen activation	SERPINF2, SERPINE1
Positive regulation of lyase activity	FTMT, AKAP5, and NOS3
Negative regulation of protein processing	SERPINF2, CHAC1, and SERPINE1
Entrainment of circadian clock by photoperiod	BHLHE40, FBXL22, PPP1CB, and FBXL8
Reactome pathway	**Genes**
Regulation Of KIT Signalling	PTPN6, SOCS1, and SOCS6
Response To Metal Ions	MT1B, MT1E, and MT1H

Early SARS-CoV-2 Infection Event	DDX5, CHMP4C, SDC2, and MAP1LC3B
Phosphate Bond Hydrolysis By NUDT Proteins.	NUDT16, and ADPRM
Diseases Associated With N-glycosylation Of Proteins	MAN1B1, CTSA, and MPDU1
rRNA Modification In Mitochondrion	NSUN4, and MRM2
Interleukin-36 Pathway	IL36G, and IL1F10
Metallothioneins Bind Metals	MT1E, MT1B, and MT1H

Profile 2	
GO Biological Process	**Genes**
Melanin biosynthetic process (GO:0042438)	TRPC1
Positive regulation of neurotransmitter transport (GO:0051590)	BAIAP3
Cerebellar granule cell differentiation (GO:0021707)	KNDC1
Endoplasmic reticulum localization (GO:0051643)	MOSPD3
Positive regulation of neurotransmitter secretion (GO:0001956)	BAIAP3
Reactome pathway	**Genes**
G2 Phase R-HSA-68911	CCNA1
Role Of Second Messengers In Netrin-1 Signaling R-HSA-418890	TRPC1
Keratinization R-HSA-6805567	KRTAP1-5, LCE3E, and KRTAP10-1
Metallothioneins Bind Metals R-HSA-5661231	MT1G

DNA Damage/Telomere Stress Induced Senescence R-HSA-2559586	CCNA1, and H1-1

Profile 3	
GO Biological Process	**Genes**
Regulation of microglial cell activation (GO:1903978)	IL6, GRN, STAP1, CST7, SYT11, and CX3CL1
Regulation of CD4-positive, alpha-beta T cell proliferation (GO:2000561)	CD81, TWSG1, CD274, LGALS9, CD55, and LGALS9C
Cellular response to interleukin-17 (GO:0097398)	IL17RA, IL17RB, IL17RE, IL17A, and IL17C
Potassium ion transport (GO:0006813)	KCNA6, KCNK15, KCNMB2, KCNK3, ATP1A2, KCNC4, PKD2, HPN, KCNJ16, HCN3, KCNN1, KCNQ4, KCNF1, LRRC55, KCNE2, ATP1A4, KCNS2, KCNK6, KCNA5, KCNQ1, HCN2, and CHP1
GPCR Downstream Signaling R-HSA-388396	HTR6, CALCB, IL6, RAMP2, GPR150, SCT, APCS, GPR39, PROC, PRKCD, GRN, ADCY1, PTGDR, PTH2, FURIN, NLRP3, GSTP1, DRD5, PTGER2, CCN3, TNFAIP8L2, HCK, HRH2, TNFAIP6, GLP1R, CD200, MAPK14, VIP, RXFP1, ADIPOQ, APOE, GIPR, RAMP3, and TSHR
Interleukin-17-mediated signaling pathway (GO:0097400)	IL17A, IL17RA, IL17RE, IL17RB, and IL17C
Reactome pathway	**Genes**
TFAP2 (AP-2) Family Regulates Transcription Of	ERBB2, EGFR, VEGFA, TFAP2C, and YY1

Growth Factors And Their Receptors R-HSA-8866910	
Myogenesis R-HSA-525793	MYF5, NTN3, MAPK14, CDC42, MYOD1, TCF12, CTNNA1, CDH4, and TCF7L1
Class A/1 (Rhodopsin-like Receptors) R-HSA-373076	PNOC, EDN3, ACKR2, CXCR5, ACKR1, TAC1, HTR1E, ADRA2A, MCHR1, F2RL3, CHRM2, RXFP3, DRD5, NPW, SUCNR1, KISS1R, CHRM5, GALR1, TSHR, PTGDR, F2R, TRH, ADRA2C, PTGER2, SSTR4, ADORA1, S1PR3, LTB4R2, S1PR4, NPY5R, TRHR, RXFP1, HTR6, HRH2, OPN4, GPR65, GPR39, GPR17, ACKR4, NPY, GAL, OPRD1, CXCL6, CX3CL1, CCL20, NPB, and NLRP3
GPCR Downstream Signaling R-HSA-388396	ADCY1, LTB4R2, RGS4, TRH, OPN4, RGS1, S1PR4, NPB, RGS21, TSHR, PRKCD, DRD5, PNOC, ADRA2C, SCT, RXFP1, CXCL6, CCL20, RGS3, CHRM2, GALR1, MCHR1, NPY, PDE1B, DGKD, PTH2, CXCR5, PTGDR, S1PR3, SUCNR1, F2R, HTR6, ARHGEF15, CDC42, GPR150, CASR, GRM3, CHRM5, ADORA1, GNAI1, GPR17, EDN3, OPRD1, RGS20, GAL, SSTR4, GPR65, NPY5R, TAS1R1, GNAZ, NLRP3, GLP1R, RAMP2, GNAT1, TRHR, PDE3B, RGS9BP, GPR39, NPW, HTR1E, ACKR4, GIPR, ADRA2A, ARHGEF16, CX3CL1, EGFR, RXFP3, OBSCN, HRH2, F2RL3, TAC1, RAMP3, KISS1R, PTGER2, PPP1R1B, and CALCB
GPCR Ligand Binding R-HSA-500792	F2RL3, SCT, TSHR, CX3CL1, OPN4, KISS1R, DRD5, ACKR2, HRH2, GAL, TRH, PTH2, PTGDR, WNT6, OPRD1, GPR65, ADORA1, GLP1R, GPR17, CCL20, GPR39, RXFP3, F2R, WNT3A, ACKR4, CHRM5, TAC1, ACKR1, PTGER2, GIPR, UCN, NPY, CALCB, CD55, SUCNR1, NPB, NPY5R, TRHR, HTR6, CHRM2, GALR1, RAMP3, CXCR5, ADRA2C, PNOC, ADGRE1, EDN3, ADRA2A, NLRP3, HTR1E, MCHR1, CXCL6, CASR, S1PR3, LTB4R2, NPW, FZD9, RAMP2, TAS1R1, RXFP1, S1PR4, SSTR4, and GRM3
Signaling By GPCR R-HSA-372790.	RGS4, ACKR4, CD55, WNT6, PTH2, TRH, CDC42, TRHR, GRM3, GPR65, OPRD1, NPY, PNOC, SCT, EGFR, RAMP3, RGS3, HTR1E, CASR, PRKCD, ACKR1, CXCR5, RGS20, RGS1, TAC1, GNAT1, S1PR3, KISS1R, DRD5, RXFP1, F2RL3, SSTR4, ADGRE1, OBSCN, ARHGEF15, WNT3A, GIPR, NPB, CHRM2, RXFP3, ADRA2C, CHRM5, F2R, HTR6, TSHR, NLRP3, PDE3B, CCL20, ADCY1, CX3CL1, DGKD, NPW, UCN, GNAI1, RAMP2, ACKR2, GPR39, EDN3, GALR1, PTGDR, PTGER2, GPR150, GAL, ADORA1, GPR17, OPN4, RGS21,

	ARHGEF16, PDE1B, TAS1R1, SUCNR1, LTB4R2, GNAZ, ADRA2A, FZD9, CALCB, HRH2, MCHR1, S1PR4, PPP1R1B, RGS9BP, CXCL6, NPY5R, and GLP1R
KEGG Pathway	**Genes**
Cytokine-cytokine receptor interaction	IL20RB, IL32, CSF1, TNFRSF12A, EBI3, IL6R, IL27RA, TNFRSF10C, ACKR4, LEP, IL1RAP, IL11, IL1R2, CX3CL1, RELT, CSF2, GDF11, TNFRSF17, IL12A, CCL20, IL17C, IL23A, CXCR5, NGFR, IL6, IL17RE, IL33, IL17A, TNFRSF6B, IL18RAP, IL36RN, CXCL6, IFNW1, IL17RB, IL1RL2, IL1RL1, PF4V1, MPL, IL17RA, and BMP8A
TNF signaling pathway	MAPK14, BIRC3, CX3CL1, CEBPB, MAP3K14, FADD, IL6, MAP2K7, RPS6KA4, CCL20, CSF2, DAB2IP, DNM1L, CSF1, CXCL6, CREB3L3, and VCAM1
IL-17 signaling pathway	IL17RB, LCN2, IL6, IL17C, CCL20, CXCL6, DEFB4A, IL17A, FADD, MUC5B, IL17RA, CEBPB, MAPK14, IL17RE, and CSF2
cAMP signaling pathway	RAPGEF4, PTGER2, CHRM2, ADORA1, ATP1A4, DRD5, SUCNR1, ADCY1, NPY, HTR1E, HCN2, TSHR, VIP, ATP2A1, PLD2, HTR6, EDN3, PPP1R1B, PDE3B, GNAI1, GLP1R, GIPR, CREB3L3, GRIA4, MYL9, ATP1A2, F2R, and GRIN2D
Neuroactive ligand-receptor interaction	DRD5, GRIN2D, SCT, PTGDR, S1PR4, CHRNA9, GLP1R, CALCB, LEP, ADRA2A, F2RL3, THRA, F2, GABRA1, NPB, HTR1E, UCN, TAC1, GIPR, TSPO, PTGER2, CHRM5, P2RX7, SLURP1, VIP, CHRM2, GRIA4, TSHR, F2R, NPY, KISS1R, EDN3, GABRD, ADORA1, GAL, TRH, P2RX2, NPY5R, GRM3, HTR6, LTB4R2, GALR1, NPW, GRP, PTH2, RXFP1, S1PR3, MCHR1, OPRD1, ADRA2C, SSTR4, TRHR, HRH2, RXFP3, and PRSS3

Profile 4	
GO Biological Process	**Genes**
Aerobic Electron Transport Chain (GO:0019646)	NDUFB3, NDUFS3, SDHC, UQCR10, SDHA, NDUFB2, NDUFS6, SDHD, NDUFB9, NDUFS8, COX4I1, AFG1L, COX10, UQCRQ, UQCC3, NDUFB5, SDHAF2, COX6B1, SDHD, AFG1L, NDUFS3, COA6
Mitochondrial ATP Synthesis Coupled	UQCC3, COX4I1, SDHC, NDUFB2, NDUFS8, UQCRQ, NDUFB9, SDHA, NDUFS6, UQCR10, NDUFB3, COX10, SDHAF2, NDUFB5, COX6B1, SDHD, AFG1L, NDUFS3, COA6

Electron Transport (GO:0042775)	
Transcription by RNA Polymerase II (GO:0006366)	NOCT, CTR9, WDR33, SRRT, ELP4, NR2C2, CPSF2, RPRD1B, SNAPC2, INTS8, NR1H3, NFYA, GTF2H3, PCF11, SNRPD3, RPAP2, GTF2F2, SNRPF, ATF4, SNAPC1, PKNOX1, NR1D2, EAF2, CPSF3, ARRB2, INTS5, NR2F1, ZNF473, SUPT6H, CCNK, POU2F1, NELFB, TRIM24, NFE2L3, SSRP1, ZBTB1, TRIP13, CSTF1, RNMT, NABP1, SP1, ICE1, CLP1, NFX1, GTF2E1, GTF2H5, NCBP1, PIR, ZNF143, EAF1
mRNA Splicing, via Spliceosome (GO:0000398)	CWC25, WDR33, CDC5L, CDC40, SART1, CSTF1, YBX1, SNRPF, SNRNP25, GTF2F2, PCF11, RBM10, LSM8, CPSF3, PRPF38A, RBM6, SRRT, CPSF2, ZRSR2, WDR83, DDX20, GPATCH1, SNRPD3, PTBP1, MTREX, PRCC, SPEN, TGS1, SRRM2, SRSF11, IK, CWC27, PRPF4, PPWD1, CLP1, DHX35, PRMT7, PUF60, BUD31, NCBP1, SNRPA, SF3B3, LSM3, RALY
Mitochondrial Respiratory Chain Complex Assembly (GO:0033108)	NDUFS3, SDHAF3, ACAD9, OXA1L, TTC19, NDUFAF2, NDUFB5, SDHAF2, UQCC3, NDUFB2, NDUFS8, SDHAF1, NDUFB9, SMIM20, FOXRED
DNA metabolic process (GO:0006259)	
Reactome pathway	**Genes**
Processing Of Capped Intronless Pre-mRNA R-HSA-75067	ZNF473, WDR33, CPSF2, CSTF1, SNRPD3, CPSF3, PCF11, NCBP1, and SNRPF
Major Pathway Of rRNA Processing In Nucleolus And Cytosol R-HSA-6791226	RPS28, DDX49, FCF1, FBL, RPP38, EXOSC9, WDR3, RPL37A, EXOSC7, RPL38, RPS7, RPL35A, RPS27A, TBL3, EMG1, SENP3, BUD23, RPS19, MPHOSPH10, UTP15, RIOK1, RPS8, NOC4L, EXOSC8, RPL27A, MTREX, UTP18, UTP4, RPS21, PDCD11, and NIP7
rRNA Modification In Nucleus And Cytosol R-HSA-6790901	TBL3, MPHOSPH10, UTP4, NOC4L, UTP15, BUD23, DDX49, EMG1, FBL, FCF1, UTP18, WDR3, PDCD11, and RPS7
Metabolism Of RNA R-HSA-8953854	PRCC, MPHOSPH10, MRM3, CWC27, NXF1, RPS19, RPS27A, GTF2H3, CDC5L, GTF2F2, WDR3, TBL3, TRMT13, EXOSC7, PPWD1, TRMT44, CWC25, HSPB1, RPS7, TRMT10C, SRRM2, SRSF11, UBB, PUF60, NCBP1, PSMD3, WTAP, RPS8, TYW3, SENP3, RPL27A, RNASE7, EXOSC9, PSMC5, NUP50, UTP18, PCF11, BUD31, YBX1, SNRNP25, UTP15, TYW1, RAN, SNRPA, RPS21, RIOK1, SF3B3,

	MTREX, SNRPF, CDC40, RPL35A, WDR33, NUP93, CSTF1, EXOSC8, NIP7, PDCD11, LSM8, DDX20, RNMT, RPL38, FBL, UPF3A, PSMD4, ZRSR2, RPP38, PAN3, NUP214, GLE1, ZNF473, EMG1, QTRT2, BUD23, SNRPD3, RPL37A, CPSF2, PRPF38A, SART1, PTBP1, MAPKAPK2, CPSF3, DDX49, RPS28, FYTTD1, SRRT, LSM3, CTU2, FCF1, GTF2H5, EIF4A1, UTP4, CNOT2, PRPF4, NOC4L, SMG7, SEH1L, NUP85, and CHTOP
Processing Of Capped Intron-Containing Pre-mRNA R-HSA-72203	NCBP1, SNRPF, SNRNP25, PRPF38A, SRRM2, PTBP1, SRRT, FYTTD1, LSM3, MTREX, CDC40, PUF60, PRCC, LSM8, RNASE7, SRSF11, CSTF1, GLE1, SNRPA, NUP50, SEH1L, SF3B3, GTF2F2, NXF1, PPWD1, CWC25, CDC5L, WTAP, CWC27, NUP214, ZRSR2, CHTOP, BUD31, PRPF4, SART1, WDR33, CPSF2, CPSF3, NUP85, YBX1, NUP93, SNRPD3, and PCF11
mRNA Splicing R-HSA-72172	SART1, GTF2F2, CDC5L, MTREX, CDC40, CWC25, PRCC, BUD31, SRRM2, CPSF3, WDR33, LSM8, SNRPD3, PPWD1, SRRT, CWC27, PRPF4, SNRNP25, PUF60, PTBP1, NCBP1, CPSF2, PCF11, PRPF38A, SNRPA, ZRSR2, SF3B3, CSTF1, SRSF11, SNRPF, YBX1, and LSM3
KEGG Pathway	**Genes**
Oxidative phosphorylation	UQCR10, NDUFB2, SDHA, NDUFB5, COX4I1, NDUFS6, NDUFS3, NDUFB3, ATP6V0A1, ATP6V0E2, UQCRQ, NDUFB9, ATP5MG, ATP6V0A2, ATP5F1C, NDUFS8, COX6B1, SDHD, COX10, SDHC, and ATP5PB
Citrate cycle (TCA cycle)	SDHC, MDH2, PCK2, SDHD, DLST, SDHA, and DLAT

Profile 5	
GO Biological Process	**Genes**
rRNA processing (GO:0006364)	RPL15, NAT10, LSM6, IMP4, RPS24, RPL14, UTP20, NSA2, SRFBP1, RPS15, WDR55, NOC4L, MRTO4, NOP53, FCF1, RIOK1, MAK16, EXOSC8, UTP11, RPL11, RPS27, UTP3, BMS1, UTP4, MRPL1, RRP1, RIOK3, DDX49, RPL36, RPL5, WDR43, DDX27, RPLP1, RPL4, PWP2, DCAF13, RPL18A, EXOSC7, RRP1B, RPL35A, POP4, RPL21, WDR75, UTP15, RRP15, NOL11, RPL7A, RPL12, RPS9, XRN2, DDX10, RPL41, NOLC1, WDR36, and EXOSC2

Ribosome biogenesis (GO:0042254)	MALSU1, MRTO4, RPL15, NOLC1, RPS24, UTP4, DDX49, ABCE1, FCF1, XRN2, UTP11, RPL14, WDR75, WDR43, RRP1B, WDR55, RPS9, RPL35A, NOL11, WDR36, RPL7A, MRPL22, UTP3, DCAF13, RPL12, RPL41, NOC4L, UTP15, RPL4, METTL17, RRP1, RPLP1, RIOK1, RPL21, DDX10, RPS27, EIF2A, LTV1, RPL5, BMS1, EXOSC7, EXOSC2, EXOSC8, RPS15, ZNF658, RPL11, PWP2, RPL18A, RPL36, POP4, UTP20, NOP53, IMP4, DDX27, and RRS1
Cellular macromolecule biosynthetic process (GO:0034645)	TICRR, RPL7A, RNASEH2A, CDK1, RPL35A, HARS2, IGHMBP2, RPL12, MCM8, TOP1, RPS26, GAL3ST2, GRWD1, HARS1, RPL14, CDC6, RPL5, FAM111B, POLR2I, TSFM, ORC3, EXO1, SARS1, FUT9, MLF1, MRPS16, RPS15, MCM5, DBF4, RAD50, MCM10, ALAS1, RPS3A, INHA, DRG1, GCNT1, MRPS18C, RPA1, HELB, CREBL2, RFC5, POLR2B, RPLP1, EEF2K, PWP1, ATM, RPL4, RPL18A, RPS6KB2, RPS24, ST3GAL4, MRPL52, EIF4EBP2, RPS27, POLR2H, MRPL32, TYMS, MRPS21, CDC7, RFC3, MCM2, MRPS36, MRPS18B, ORC6, MRPL42, POLA1, RPS9, NFKBIB, KAT5, RPL36, RPL41, RAD9B, RPS3, MRPL18, RPL15, EXT2, RPL21, POLR2G, MRRF, POLR2K, RPL11, and POLR2F
Positive regulation of viral process (GO:0048524)	POLR2K, MDFIC, NELFCD, VAPB, POLR2G, PPIA, PPARA, RAD23A, SUPT5H, PPIE, POLR2B, CDK9, PKN2, POLR2F, IFIT1, POLR2H, PPIH, POLR2I, DHX9, ADARB1, DDX3X, CD28, VPS4A, and KPNA2
DNA metabolic process (GO:0006259)	PRIMPOL, UNG, HELB, GTF2H3, NEIL1, GEN1, POLE4, CDT1, MCM8, CDK1, TYMS, RFC5, MCM2, TERF2IP, RPA1, GMNN, POLA1, CHD1L, ADPRS, SMARCAL1, RAD51B, RAD9B, DCLRE1B, SUPV3L1, ATM, KMT5C, RAD50, MCM10, POLK, CDC7, XRCC5, PWWP3A, OGG1, THAP9, NEIL3, ORC3, NUDT15, KAT5, PURA, MSH3, RIF1, EXO1, TSN, MSH6, MNAT1, RAD51, MCM5, ORC6, GTF2H1, POLH, NOP53, NABP2, RFC3, XRCC6, TOP1, KPNA2, EXD2, PARP3, UBE2U, NUDT1, DDX21, PARP2, HINFP, IGHMBP2, SMARCAD1, TICRR, ALKBH3, RNASEH2A, POLB, DBF4, MUS81, CDC6, and GTF2H4
Reactome pathway	**Genes**
Cell Cycle Checkpoints R-HSA-69620	MAPRE1, PSMD9, EXO1, ANAPC11, ORC6, RNF8, PSMB6, ATM, NDC80, DBF4, H2BC9, H2BC3, RAD50, CCNB2, ORC3, CDCA8, CENPQ, INCENP, RFC5, CDC16, BUB1B, GTSE1, MCM10, CDC20, PSME3, PSMD13, CENPA, BABAM1, DYNC1I2, H2BC5, H3-4, PSME1, SGO2, CENPO, YWHAH, H2BC15, CDK1, CDC7, NSD2,

	SKA1, PPP2R5D, PSMD7, CDKN1B, CCNB1, CENPF, MCM5, MIS12, NUP107, UBE2D1, RPS27, RFC3, MCM8, BUB3, NUDC, H2BC13, REXO2, PSMA8, PSMD8, PIAS4, DYNC1I1, KAT5, PSMC1, POLR1B, PSMA3, PSMC2, RAD9B, ANAPC4, SPDL1, ZWILCH, PPP2R5A, PPP2R1A, PSMB4, and CDC6
Cell Cycle R-HSA-1640170	CHMP2B, CENPQ, ORC3, MAU2, DYNC1I2, HMMR, RAD50, GTSE1, MSH4, H2BC13, BLZF1, MNAT1, CDC7, ANAPC11, NSD2, PSMB4, DBF4, PSMC2, CNEP1R1, H2BC3, CDK1, PSMD7, HAUS1, CENPF, NDC80, ESPL1, TYMS, CDC16, GORASP2, TUBB3, PPME1, H2AC6, BANF1, AKAP9, NEK7, CEP78, PRKAR2B, CDC20, REXO2, POLA1, KIF23, CDCA8, CCNB1, PCM1, CUL1, CHMP4B, TUBB4B, TUBG1, ATM, SKA1, MLH1, MCM5, LCMT1, GINS1, SGO2, TUBA8, SPDL1, DYNC1I1, CENPA, BABAM1, NCAPG, H2BC15, POLR2H, CDC6, PSMC1, CEP135, PIAS4, H2BC9, CHTF18, E2F4, RAB1B, POLE4, PPP2R5A, PSME1, POLR1B, TEX12, NUDC, ZWILCH, RAB8A, RNF8, TUBB6, PSMA8, KPNB1, RPS27, UBE2D1, BUB3, HAUS2, HAUS5, RBBP4, PPP2R5D, MCM8, H3-4, RCC1, PSME3, POLR2B, CCNB2, H2BC5, NUP42, POLR2F, RFC3, H2AC19, RFC5, KAT5, RBL2, PRDM9, RBX1, CEP43, CEP57, BORA, PSMD13, MIS12, CENPO, CDC25B, NUP58, PSMB6, REC8, RAD9B, MND1, PPP1R12B, SKP2, POLR2K, PSMA3, CDKN1B, SMC3, RB1, TMPO, LMNB1, CDT1, ESCO2, EXO1, CEP72, H2AZ1, LEMD3, MASTL, RAD51, VPS4A, PSMD8, YWHAH, HDAC1, POLR2I, SPAST, CENPW, NUP107, PPP2R1A, ANAPC4, INCENP, GAR1, LPIN3, TERF2IP, CCND1, RUVBL2, ORC6, MCM10, POLR2G, RAB2A, GMNN, LIN52, SFI1, PSMD9, MAPRE1, NUP205, BUB1B, and MYBL2
Cell Cycle, Mitotic R-HSA-69278	SPAST, RFC5, H2BC13, PSME1, NUP42, MCM5, PCM1, GORASP2, DBF4, RBBP4, CDC25B, LEMD3, SKP2, KIF23, PSMD9, ANAPC11, KPNB1, PPP2R1A, CDKN1B, NDC80, LPIN3, MASTL, BUB1B, NEK7, ESCO2, TMPO, RBX1, PSMA3, MAPRE1, HMMR, PSMD7, HAUS1, CNEP1R1, PSMA8, CDC20, CUL1, ANAPC4, RB1, CHMP4B, VPS4A, ORC6, CENPO, H2BC5, NUP58, H2AC19, ESPL1, GMNN, PSMB6, GTSE1, SFI1, E2F4, H2BC9, MCM8, CDC16, H3-4, CDT1, SKA1, RCC1, UBE2D1, PPP2R5D, SMC3, ORC3, PRKAR2B, BORA, RAB8A, GINS1, H2AC6, MNAT1, MIS12, PSMD8, H2BC3, BANF1, PSME3, PSMC2, CENPA, CCNB1, TUBA8, HAUS5, AKAP9, CEP57, SPDL1, RAB2A, LIN52, HDAC1, POLA1, CCND1, TUBB6, CEP43, NUP205,

	PPP2R5A, CHMP2B, MAU2, RAB1B, DYNC1I1, HAUS2, CDC7, RBL2, ZWILCH, CEP72, CEP78, PSMB4, NUP107, BUB3, TUBB4B, LCMT1, POLE4, TYMS, CEP135, TUBG1, POLR1B, CENPQ, CCNB2, INCENP, MYBL2, LMNB1, NCAPG, BLZF1, H2AZ1, CDC6, MCM10, PPME1, DYNC1I2, CDCA8, H2BC15, NUDC, SGO2, PSMC1, RPS27, TUBB3, PPP1R12B, CENPF, CDK1, RFC3, and PSMD13
M Phase R-HSA-68886	TUBG1, RAB1B, UBE2D1, CDCA8, PPP2R5D, DYNC1I1, MAU2, PSMD7, CDC16, ANAPC4, H2AZ1, PPP2R5A, PSMB4, PSMD9, CDK1, H2BC3, SPDL1, ANAPC11, TUBB6, CEP43, LMNB1, TUBA8, CENPA, KIF23, HAUS5, CEP78, MASTL, HAUS1, H2BC9, PSMC2, MIS12, CDC20, NDC80, PSMC1, CEP135, PSME1, PSMA8, H2AC6, PRKAR2B, CCNB1, H2BC13, INCENP, GORASP2, BUB1B, LEMD3, PSMD8, ZWILCH, CHMP4B, NUDC, RB1, CENPF, ESPL1, PSMD13, TUBB3, CEP72, BLZF1, NCAPG, H2AC19, TMPO, NUP42, SPAST, H3-4, PSME3, NUP58, NUP205, DYNC1I2, RCC1, CEP57, TUBB4B, RAB2A, NUP107, CHMP2B, SFI1, PSMA3, SMC3, PPP2R1A, H2BC5, NEK7, CENPQ, RPS27, AKAP9, MAPRE1, SKA1, VPS4A, CENPO, PCM1, HAUS2, PSMB6, H2BC15, BUB3, SGO2, KPNB1, BANF1, CNEP1R1, LPIN3, and CCNB2
Metabolism Of RNA R-HSA-8953854	RPS15, PNO1, MRM3, METTL14, CNOT6, RPL21, NAT10, RPL7A, GTF2H3, LSM11, NUP107, PSMD8, GEMIN2, SNRNP48, APOBEC2, TRMT13, EXOSC7, RPS3, PPWD1, PCBP2, SRSF5, PRPF3, RRP1B, SRRM1, TRMT44, U2SURP, PRORP, WBP4, RPL11, LTV1, TRMT112, SNRNP27, SRSF11, WDR36, HNRNPUL1, ADARB1, DDX1, PSMB6, ZC3H11A, EDC4, HNRNPA1, RPL15, SRSF3, PSMA3, DUS2, NCBP1, PSMD9, EXOSC2, LSM6, GTF2H4, DCAF13, PPP2R1A, RIOK3, TSEN2, RPL41, PCBP1, DDX42, PSME3, WDR43, RPS27, RPL4, PPIE, TRNT1, SNRNP70, RPS3A, POLR2B, DHX38, NUP58, GTPBP3, SNRPB2, APOBEC3C, HNRNPA3, SNRNP25, PUS1, UTP15, NUP205, RBM17, RIOK1, PSMD13, DDX46, RBM42, GTF2H1, CDC40, RPL36, RPL35A, UTP25, CSTF1, EXOSC8, POLR2H, RPS24, CDKAL1, PYDC1, BMS1, IMP4, MNAT1, WDR75, TRMT12, NXT1, RBM5, UTP3, XRN2, RPL12, RPP38, PSMC2, QTRT1, PSMC1, POLR2F, TRMT61A, POLR2K, QTRT2, POLR2G, BUD23, SNRPD3, PRPF6, PSMB4, TSEN34, CPSF2, PRPF38A, PSMA8, NOL11, GAR1, PSME1, RPL5, SYMPK, PPIH, MAPKAPK2, RPS9, RPL14, DDX49, THG1L, FUS, SMG8, CTU2, FCF1, GTF2H5,

	NUP42, PSMD7, FAM98B, CPSF4, UTP4, POLR2I, TRMT11, UTP20, RPLP1, PRPF4, RPL18A, NOC4L, DHX9, PWP2, CCAR1, DCPS, SUPT5H, UTP11, CASC3, DCP2, PNRC2, U2AF2, SF3B5, and PHAX
KEGG Pathway	**Genes**
Huntington disease	NDUFB3, TUBB3, PSMC1, AP2M1, KIF5A, NDUFS5, PSMD8, COX6A2, POLR2I, TUBB4B, NDUFA7, CREB1, COX4I1, POLR2B, APAF1, NDUFV2, COX8C, TFAM, PLCB1, DNAH14, GRIN1, DNAH9, TUBB6, UQCRH, ATP5PB, NDUFB4, DNAI1, TAF4B, PSMB6, DCTN5, NDUFB5, POLR2H, WIPI2, TBP, AP2B1, DNAL1, NDUFA13, NDUFS3, TUBA8, SDHD, NDUFS6, SOD1, PSMC2, NDUFS4, ATP5F1C, POLR2K, PSMD13, NDUFV3, COX5B, ATG13, HDAC1, GPX5, SLC1A3, UQCR10, PSMD1, DNAI2, PIK3C3, POLR2G, PSMB4, GPX8, DNALI1, PSMA3, ATP5MC1, CLTA, AP2A2, NDUFS7, NDUFB9, PSMA8, PSMD9, VDAC3, NDUFA9, COX7C, BAX, KLC3, SDHA, PSMD7, POLR2F, and ACTR10
Amyotrophic lateral sclerosis	ATP5F1C, APAF1, DNAH9, COX7C, GPX8, RAB8A, SOD1, PSMD8, NDUFA7, COX8C, HNRNPA3, MAP2K3, DNAI2, MCU, NDUFB5, WIPI2, GABARAPL2, ATP5PB, GRIN1, UQCR10, SDHD, SQSTM1, NDUFS3, TUBB6, PSMC1, PSMD9, NDUFS7, NDUFV3, TOMM40, NDUFV2, COX6A2, ANG, FUS, NDUFB9, DCTN5, COX5B, TUBA8, PSMC2, NDUFS5, ATP5MC1, ACTR10, SMCR8, COX4I1, HNRNPA1L2, NDUFS6, NXT1, FIG4, DNAL1, CHMP2B, TUBB4B, RAB5A, VAPB, BAX, PSMB4, KIF5A, NDUFS4, PSMA8, NUP58, PSMA3, NDUFB4, NUP205, TARDBP, DNAH14, DNALI1, NDUFA9, SRSF3, NCBP1, PIK3C3, GPX5, SDHA, MAP2K6, NDUFB3, UQCRH, PSMD13, DNAI1, HNRNPA1, NUP107, ATG13, TUBB3, PSMD7, PSMB6, NDUFA13, KLC3, CAT, NRG3, and PSMD1
Citrate cycle (TCA cycle)	MDH2, PCK1, PDHA2, PCK2, OGDH, SDHD, ACO1, SDHA, ACO2, IDH3A, PDHA1, MDH1, and DLD
RNA polymerase	POLR2F, POLR1C, POLR2I, POLR3B, POLR3E, POLR3F, POLR2H, POLR2B, POLR3G, POLR2G, POLR2K, and POLR1B
Glyoxylate and dicarboxylate metabolism	CAT, SHMT2, ACO2, AFMID, AGXT, GRHPR, PCCA, MDH1, DLD, ACO1, MDH2, and PCCB

Sources of the data are publicly available data sets:

1. Genomic coordinates of the regulatory elements in human genome (http://www.ensembl.org/biomart/martview/c299972c97f2b6ce2952cdfb87ba729f)

2. Encode project(https://www.encodeproject.org/)

3. Experimental nucleosome mapping data (https://bigd.big.ac.cn/nucmap/ ,https://generegulation.org/nucleosome-positioning-database/)