SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Comprehensive analysis of human transcription factor binding sites with ChIP-seq and topological arrangements of transcription factor complexes on the DNA

by Erik Czipa

Supervisor: Dr Barta Endre, PhD

UNIVERSITY OF DEBRECEN

DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2019

Comprehensive analysis of human transcription factor binding sites with ChIP-seq and topological arrangements of transcription factor complexes on the DNA

By Erik Czipa
Molecular Biology Master's Degree
Supervisor: Dr. Endre Barta

Doctoral School of Molecular Cell and Immun Biology, University of Debrecen

Head of the **Examination Committee**: Prof. Dr. Sándor Bíró, DSc
Members of the Examination Committee: Dr. György Vámosi, PhD
Dr. Zoltán Gáspári, PhD

The Examination takes place at Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen; at 10:00, 22th of March, 2018.

Head of the Defense Committee: Prof. Dr. Fésüs László, MD, PhD, DSc, MHAS
Reviewers: Dr. Gáspári Zoltán, PhD
Dr. Penyige András, PhD

Members of the Defense Committee:
Prof. Dr. Fésüs László, MD, PhD, DSc, MHAS
Dr. Gáspári Zoltán, PhD
Dr. Penyige András, PhD
Dr. Benkő Szilvia, PhD
Dr. Sebestyén Endre, PhD

The PhD Defense takes place at the Lecture Hall of Building A of the Department of InternalMedicine, Faculty of Medicine, University of Debrecen, at 1 p.m., 15th of January, 2020.

1. Introduction

The DNA, as a blueprint, contains the genetic code to build an organism. In a multicellular organism, every single cell contains almost the same genetic information. In contrast, the cells that cooperate in the organism show large morphological and physiological differences relative to each other. The cellular differentiation is a complex multi-stage process, which starts with organization of the active and inactive DNA territories. The effects on the transcriptional activities of genes influence the proteome of the cell.

Transcriptional regulation and transcription factors

Transcription is the central event of gene expression, which is preceded by many biochemical processes. Transcription starts with the assembly of the transcription pre-initiation complex (PIC), which consists of general transcription factors (GTF). These factors gather around the promoter before the 5' end of the gene, which is the transcription start site. The promoter has a specific sequence, which can be recognized by general transcription factors. Key DNA sequences, named core promoter elements, have frequently occurring patterns. These include the TATA-box, downstream promoter elements, initiators, TCT, B recognition element, and motif ten element. The general transcription factors attach to the structurally and spatially distinct promoter elements. These general transcription factors include TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which do not function universally on all core promoters.

Core promoters and general transcription factors are essential for direct transcription initiation but generally have low basal activity, which can be further suppressed by chromatin or activated by more remote regulatory elements, such as enhancers. These regions and the promoters are occupied by other transcription factors, which can recognize specific sequences as well. Although these TFs are not part of the PIC, they play a significant role in its assembly and maintain the conditions for unperturbed transcription. TF binding occurs in nucleosome-depleted regions, which generally encompass regions with lengths similar to those protected by nucleosomes.

Transcription factors

In addition to the general transcription factors, which show DNA sequence preferences, several other proteins influence transcription to regulate gene expression. An assessment of the scientific literature resulted in the identification of 3230 "transcription factors". TFs are diverse, not only in function and cell-type specificity, but also in structure. However, all TFs have DNA-

binding domains (DBD), which recognize specific sequences in the genome. These short genomic regions are called transcription factor binding sites or responsive elements.

Varying combinations of transcription factors have different molecular functions. Functions include the degree of activity and the composition of recruited co-regulator proteins. Most of the cofactors are enzymes that mediate post-translational modification of target proteins. This group of enzymes is extremely diverse in structure and enzymatic activity. Most of these enzymes are transfer proteins, which enact the transfer of specific functional groups, methylation, phosphorylation, sumoylation, ubiquitylation, ADP ribosylation, butyrylation, citrullination, crotonylation, formylation, proline isomerization, and propionylation) to histones or other regulator proteins.

Topologically associated domains

The unfolded DNA from one single human cell would measure 2 meters end to end. This extensive length of DNA needs to be packed inside the nucleus, which is approximately 5 μm in diameter. The major features of chromosome architecture are barely known. The first level of chromosome packing is the previously described nucleosome. Nucleosomes wrap about 146 base pairs around a histone octamer. A nucleosome is about 11 nm wide. To make the chromatin structure more compact, approximately 30 nm “zigzag” chromatin fibers are formed. These fibers need further packing; meanwhile, functionally active and inactive DNA needs to be separated.

Chromosome conformation capture techniques clarified the presence of chromatin domains or topologically associated domains. TADs represent hundreds of kilobases to several million bases in length. These domains are evolutionarily conserved and stable for many cell divisions. Their role is not completely understood, but there is convincing experimental evidence to support a simultaneous “insulator” and “co-regulator” feature. TADs may create autonomous gene regulatory domains, where genes share coordinated gene expression profiles within the same TAD. TADs also block the spread of activity between neighboring TADs. Smaller functional units can be observed within TADs, called sub-TADs or loops. Sub-TADs are regions that display both self-associative and insulating properties, similar to TADs. CTCF and cohesin complexes mediate chromatin looping, a physical process that brings two distal DNA regions close to each.

CTCF and cohesin

High-throughput 3C methods revealed the significance of the CTCF-mediated subdivision of chromatin into TADs. CTCF binding sites demarcate the individual TAD boundaries and other chromatin loop borders. Together with cohesin subunits (RAD21, SMC1/2 and STAG1), CTCF can coordinate interactions between enhancers and their corresponding promoters by composing loops, while protecting interactions between sequences located inside and outside the loops in interphase nucleus.

Investigation of transcriptional regulation with High-Throughput Sequencing Technologies

Research tools in functional genomics and molecular biology have developed by leaps and bounds over the last few decades. The development of high-throughput sequencing technologies enabled relatively easy and rapid parallel DNA sequencing at a reasonable price. The combined techniques advanced genomic research on a global level, including gene expression profiling, chromosome counting, DNA-protein interactions, and detection of epigenetic changes. In this study, I am focusing on the processing and analysis of chromatin immunoprecipitation techniques, like ChIP-seq and ChIA-PET.

The ChIP-seq technique involves the specific antibody treatment to “fish out” the DNA-protein complex of interest, after cross-linking, followed by a random fragmentation step.

2. Aims of the study

During the analysis of CTCF and cohesin ChIP-seq data, we noticed that there is a visible shift between the summit positions of these proteins. Since the CTCF is the only member which have known DNA binding domain in this complex, we wanted to identify why the summits are not located on the same genomic localization and if there is any measurable system behind this shift. We hypnotized that the juxtaposing summits referring to topological position of cohesin proteins. We assumed that there is a structural feature of complex which holds the cohesin subunits in close proximity of DNA. So thus, the non-DNA binding proteins are crosslinked with DNA during the formaldehyde treatment of chromatin immunoprecipitation.

We downloaded and processed several CTCF and cohesin subunit ChIP-seq data from human and mouse (see in Material and methods) to answer the following questions:

- Can the observed shift be seen in all analyzed samples?
- Does the shift show any strand specificity?

- Does the shift follow the orientation of CTCF motif?
- Is there any order between the shifts of difference subunits?
- Is there any linkage between the protein positions and the known topology of CTCF-cohesin complex?
- Is there any correlation between the shift orientation and CTCF mediated chromatin looping?

After the publication of results, we extended our focus to other available human ChIP-seq data, which were analyzed using the summit-based topology analysis. We decided to collect as many human ChIP-seq data as we can and process them with our method. For unique data processing and comperability we faced with the following tasks:

- Development of a pipeline, which is able to automatically extract topological and network information from large amount of data.
- Create a database which can be used not only for data storing but for protein position analysis too.
- Discover unknown protein-protein interactions and complexes and create new topological models.
- Create a web interface which provides access to our result for other researchers.

3. Material and methods

Primary analysis

Data from 4068 ChIP-seq experiments, covering a wide range of proteins and cell types, were collected from the NCBI SRA and ENCODE databases. The naming and automatic download of experiments were performed using a homemade script. Processing of the downloaded raw data was carried out with the previously mentioned in-house developed ChIP-seq analysis pipeline. We used PeakSplitter for summit predictions and, thus, more accurate identification of local maxima

Peak filtering

Identifying peaks with well-defined maxima was crucial at the early stage of data processing because false positive peaks could result in the false prediction of the protein's position. The peak summits show the highest coverage for the peak region and coincide reasonably with the center of the corresponding DNA elements bound by transcription factors.

Therefore, the identification of regions suitable for the clear determination of the summit position was required. Current software packages use different strategies, such as the evaluation of peak prediction reproducibility or false discovery rates, for peak prediction, which dramatically decrease the false positive rates. Unfortunately, using these methods necessitated configuring the filtering algorithms differently for each experiment, making automation of the processing of large datasets more difficult. For better filtering, we have developed a pipeline, which reduces the false positive discovery rate even further.

To avoid false positive results, we filtered out duplicated reads using a step in the ChIP-seq analysis pipeline and developed a Perl script that classified and filtered the subpeaks based on their size and shape. In the script, two parameters are responsible for the detection of the previously mentioned large signal intensity increase.

JASPAR CORE motif and ChIP-seq data pairing

Identification of the exact positions of TF binding sites is the basis of ChIPSummitDB. These motif positions are not only a collection of regulatory regions, but the motif centers are also used as reference points for summit position analysis. Our primary goal was to create consensus binding site sets for as many transcription factors as possible. To do this, we used the JASPAR CORE database, which is a "curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes" and incorporates 579 non-redundant motifs. We were able to allocate only 338 motifs to at least one ChIP-seq experiment, because no sequence and HTS data were available for the rest of motifs in human. The motifs were manually curated and the most similar motifs were paired with the corresponding antibodies.

Numerous tools can be used to find the occurrences of individual motifs. Instead of choosing one single tool, we combined 3 popular methods: HOMER, FIMO, and MAST. The positions, which were identified for certain motifs by at least two programs, were filtered in the first step.

Summit distance calculation

The identified consensus sequence locations are not only used to show the genome-wide distribution of transcription binding sites, but are also used as reference points for landscaping of possible co-bindings and measuring motif-protein or protein-protein distances. All motif occurrences obtained from every set were screened to identify ChIP-seq experiments containing

peak summits in the +/- 50 bp vicinity of the motif center. The distances between motif centers and summit positions were calculated

4. Results

ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA

Our first observation was that there is a visible shift between CTCF and cohesin subunit ChIP-seq peaks. Due to possible protein-protein cross-linking events, components of a protein complex that are not directly involved in specific DNA binding can produce ChIP-seq peaks that overlap with the peaks of TFs, which anchor them to DNA, and their corresponding summit positions approximately coincide.

Since CTCF has the only DNA binding domain among the components of the CTCF/cohesin complex, we expected that the corresponding ChIP-seq peaks will point to the same position with respect to CTCF binding site. To measure the occurrence and extent of the shift, we compared summit positions relative to a reference point. The reference point needs to be fixed in the genome. The center of the CTCF motif seemed reasonable as a comparison site. At this point, since the CTCF binding site is a non-palindromic element, we could measure the distance distribution strand specifically. Depending on the location of the proteins and the motif orientation, the distance value can be positive or negative. The distance was measured in base pairs. We extended the analysis to 237 human experiments and 183 mouse samples. The overall plotting of protein positions validated the observed shift. Strikingly, the average position of different proteins showed a characteristic separation around the binding sites. The plots within a cluster had the same protein target, which suggested that proteins have a position preference relative to each other and the binding site. The serial order of peak summit positions was invariably CTCF –> SMC1/3 –> RAD21, STAG1/2, irrespective of whether the average positions were calculated for a cell type or for the entire dataset.

As previously mentioned, the CTCF-cohesin subunit interaction order has the following sequence: CTCF-> STAG1/2-> RAD21-> SMC1/3. To understand the identified serial order between ChIP-seq signals, we converted the positional distances into approximate 3D spatial constraints. For this, we took all average positions from all human and mouse cells and chose the human median positions for CTCF, SMC1/3 and RAD21, and for STAG1.

This pattern was then mapped on the surface of a B-DNA model that we built using a sequence dependent modeling procedure. Normally, 10/11 base pairs are involved in one turn

of B-DNA. The summit distance gap between CTCF and RAD21-STAG1/2 is approximately equal to one turn of the DNA. This makes the peak summits of CTCF, STAG1/2, and RAD21 map to one face of the double helix, while the intermediate contact sites of SMC1/3 are located on the opposite face.

Combining the published structural models and our position data, we could give a possible explanation of the protein positions and create a hypothetical model for CTCF mediated chromatin looping. The double embrace model was integrated into our topological model that involves two cohesin rings. This model can help explain the unusual positioning of ChIP-seq summits of SMC proteins. The SMC hinge domain has nonspecific DNA binding capability, which stabilizes the chromatin loop formation. The hinge and head domains are separated by a relatively long rod like structure, which embraces the chromatin. The head connects directly to other members of the cohesin ring and indirectly to the CTCF protein. On the opposite side of the rod, the hinge domain can form a non-specific bond with the distal DNA region, which comes close during loop formation. This means that the detected SMC signals belong to the other anchor region's cohesin ring. This explains the opposite localization of SMC on B-DNA relative to CTCF-RAD21 and STAG1. Our results clearly suggest that RAD21 and STAG are in contact with each other and also either one or both proteins are in close contact with DNA, around the 3' end of CTCF binding site. Their position preference is unspecific and may be the consequence of physical proximity to DNA and the cross-linking procedure during chromatin immunoprecipitation. The double embrace arrangement provides a testable hypothesis that may help to clarify several, seemingly contradictory features, of loop closure.

CTCF-binding site orientation shapes the chromatin loops

In light of the previous result, we wanted to investigate the correlation between topological data and chromatin looping. The 3C techniques were developed to analyze the spatial organization of chromatin. The 3C techniques, in combination with High-Throughput Sequencing, enable the identification of DNA loops on the genome level. We can rely on HiC data or chromatin immunoprecipitation for TAD mapping. We used publicly available ChIA-PET data to investigate the orientation of CTCF motifs, which are involved in chromatin looping. First, we downloaded prepared MCF7 CTCF ChIA-PET data from the ENCODE database, because it has biological replicates. After the complex processing of this data, interaction tables were generated to store information about interacting distal DNA regions. To simplify the prepared data, the identified interaction can be divided into 3 parts: i) first anchor, ii) interior loop region, and iii) second anchor. The anchor regions frame the chromatin interaction and serve as the binding platform for the loop mediating proteins. In terms of 3C

data processing, the anchor regions represent regions with variable width that contain the possible interaction points. Most available Hi-C datasets have relatively low resolution, between 25 to 40 kb, and the most advanced procedures produce 5kb data. The resolution could be improved to 1 kb in the case of ChIA-PET. The MCF7 CTCF ChIP-seq have an ~850 average anchor length. This provides more accurate identification of the anchoring of the CTCF binding site.

To find the exact interaction point, we created a pipeline that scans anchor regions. The pipeline requires CTCF ChIP-seq data for experimental validation and motif enrichment analyses. The enriched de novo CTCF motif was remapped to the MCF7 CTCF peak regions to find the motif instances. Then, a home-made scanner program found the most proximal CTCF motifs relative to the midpoint of the anchor region. This allowed us to distinguish different types of loops. Since one loop has two anchor regions and we paired every anchor with a CTCF, we can cluster the loops depending on the CTCF binding site orientation as follows :

- Convergent: the motifs face each other and are directed inside the loop
- Divergent: the motifs are oriented in opposite directions towards the outside of the loop
- Same direction on the strand: both motifs are on the same strand
- Convergent with unidentified pair: One anchor's motif is not identified, but the other motif anchor faces the inside of the loop
- Divergent with unidentified pair: One anchor's motif is not identified, but the other motif anchor faces the outside of the loop

The frequency analysis showed a clear enrichment of convergent motif orientations. We compared this data with ChIA-PET results from other cell types. First, we created a consensus MCF7 loop set, considering the interactions with overlapping CTCF binding sites on both anchor regions. Then, we processed the consensus loop set and the other downloaded ChIA-PET data with the previously described procedure. The results were congruent.

Histone modifications in CTCF mediated chromatin looping

As mentioned previously, substructures within TAD are associated with cohesin and form functional units like enhancer-promoter loops. The substructures within TAD are also associated with CTCF, whose function is difficult to define. CTCF sites facilitate gene activation, while other CTCF sites function as insulators.

To identify correlations between transcriptional regulation and chromatin looping, we investigated histone modifications in the vicinity of the loop anchor regions. We used the previously defined consensus CTCF interaction set from the MCF7 cell line and seven publicly available MCF7 histone ChIP-seq datasets. To visually check with the genome browser, we downloaded the GRO-seq data.

We investigated the histone ChIP-Fragment Coverage, which indicates the density of aligned tags relative to CTCF anchor regions. The average tag coverage of different histone experiments was calculated within 1000 bp frame relative to CTCF centers. The overall meta profile of histone occupancy followed the previously described peak-valley-peak pattern, where the center valley contains the cis-regulatory element, the CTCF motif in our case. Remarkably, signal intensities for H3K4 methylations were higher compared to other histone modifications.

Thus, the analysis was focused on the H3K4 methylation data. MCF7 consensus interactions were used to profile the histone ChIP-seq fragment coverage of every CTCF anchor position. Anchor regions were occupied by at least one H3K4 methylation type in the 4188 interactions. Heatmaps, centered on the anchor CTCF motif centers from both sides of the loop, were paired next to each other. This approach facilitated the simultaneous investigation of histone coverage on both anchor regions. We clustered the profile with k-Means clustering. The most prominent phenomenon was the characteristic asymmetry of histone signals on opposite sides of the loops. The signals were high for only one anchor region in each cluster. In the first round, we distinguished 6 clusters, which were reduced to 3 with side ordering. The "downstream and upstream side of loops" are artificial constructs and there are large similarities between cluster 2-4, 1-3, and 5-6; thus, we ordered the anchors with "strong signal" into one side of the heatmap, resulting in 3 clusters as follows:

- Cluster 1: strong H3K4me2 and H3K4me3 signals with relatively low H3K4me1
- Cluster 2: High H3K4me1, medium H3K4me2, and low H3K4me3 signals
- Cluster 3: Strong H3K4me2 signal and medium H3K4me1 and H3K4me3 signals

Cluster 1 is mostly involved in promoter specific interactions, while clusters 2 and 3 interact with bridge enhancer regions for introns or other enhancer regions.

An article was published with results consistent with our studies. In this study, ESC ChIA-PET data were processed. The investigators introduced the definition of polycomb domains, which are characterized by a particularly high presence of the polycomb proteins, like EZH2 and SUZ12, in association with H3K27me3 histone modification. This complex structure represses lineage-specifying developmental regulators. Their meta-analysis had a similar result as ours. Taken together these studies indicate that the CTCF loops have structural roles in both

gene activation and repression, which enable the physical proximity between enhancer and promoter regions.

ChIPSummitDB

The main goal of analyzing ChIP-seq experiments is to identify regions in the genome where we find more sequencing reads than we would expect to see by chance. These regions are called peak regions due to the appearance of the visualized distribution of mapped tags. Our goal was to create a global database based on combining the location of identified transcriptional regulatory elements with the positional information of the co-bound regulatory proteins. By investigating a global picture of different transcription factors and cofactors, we can identify previously unknown transcriptional regulatory networks. Using the database, we can browse co-bound proteins on TREs and acquire information about their positioning relative to each other and the bound transcription factor motif.

The comparison between experiments requires uniform processing of data. Several databases contain pre-processed ChIP-seq data. They differ in the stage and the approach to data processing. Their downloadable content makes the gene regulatory research work easier by providing information about identified transcription factor binding sites or motif enrichment.

After the developmental phase, we extended the data collection and processed it with the custom made pipeline. We collected 4052 human ChIP-seq experiments and successfully analyzed 3782. The remaining 270 experiments lacked identified peak regions due to the low quality of the data. Overall, more than 93.4 million peaks were used in the database creation, which covered more than 1 Giga base pairs in the human genome. A total of 2659 ChIP-seq targets were classified as transcription factors the others were cofactors. A total of 2496 experiments belonged to transcription factors that have described motifs in the JASPAR CORE database. These experiments were used for motif optimization and binding site prediction. The JASPAR CORE database stores 579 non-redundant motifs. From the 579 non-redundant motifs, 338 PWMs could be paired to at least one ChIP-seq experiment from the 2496 experiments with described motifs. We could find motif instances for 280 JASPAR CORE motifs. This kind of reduction is not striking, because many motifs are theoretical and cannot be linked to a ChIP-seq experiment.

Finally, more than 5 million transcription factor binding sites were identified, which cover 40.8 Megabase pairs in the genome. The identified binding sites represent valuable information in themselves. However, a comparison of all motif data with complete ChIP-seq dataset provides information about the protein complexes and transcription factor network in

correlation with specific binding sites. We calculated the protein position preferences with respect to the corresponding motif centers and compared the results to identify topological relationships between proteins. The complete dataset is stored in the database, which is publicly available on the ChIPSummitDB's web interface.

Database and web interface

To reach the viewable results from raw sequence data requires a large investment in time and computing resources. Currently, transcriptional regulation related studies are efficient due to the ChIP-seq technique. This requires unified data processing for comprehensive analysis. The comparison of two or more samples can provide large scale biological correlation, but this data is still insufficient for regulatory network mapping. We aimed to collect data about transcription factor/cofactor occupancy on different types of transcription factor binding sites. We established ChIPSummitDB, a web interface to browse processed ChIP-seq data and identified transcription factor binding sites in a global manner.

The website provides information about:

- Transcription factor binding site profiles: JASPAR CORE motifs are optimized with HOMER analysis of ChIP-seq experiments to extract accurate PWMs. The motifs are carefully paired with available ChIP-seq experiments. The matrices represent frequently presented sequence motifs in the peak regions, which resemble the original JASPAR motif.
- Topological data: Spatial organization of DNA bound protein complexes is based on summit analysis.
- The overlap between ChIP-seq peaks: The juxtaposing ChIP-seq signal in correlation with a given motif can be examined.
- Regulatory SNP: Using dbSNP and ClinVar databases, an SNP finder has been integrated. The identified transcription factor binding sites can be scanned for regulatory SNPs.
- Genomic map: All data are viewable in a genome browser format.

The six display modes are provided to visualize the data during different approaches.

MotifView

In this viewing mode, the average distances between the read peaks from the ChIP-seq experiments is obtained and the given consensus motifs are visualized on a scatterplot. Each scatter represents an experiment. Circles represent transcription factors with defined binding sites, while triangles represent co-factors and other indirectly bound proteins. Different colors indicate the antibodies used in the immunoprecipitation. The X axis shows the average distance

between peak summits and the center of the binding sites overlapped by the peaks. The Y-axis shows either the number of the peaks overlapping the center of the binding sites or, in the default mode, the standard deviation of the shift values between the peak summits and motif centers. Such a scatterplot representation is available for every consensus binding motif set. The displayed data can be filtered for the number of the peaks or for the standard deviation. Data can also be displayed based on the ChIP antibody or cell type. The average data obtained by the same antibody in different experiments can also be calculated and shown.

If browsing the standard deviation scatterplots, the dots show a visible clustering among factors. Since the dots are colored according to the experimental antibody, different color groups are distinguishable around the preferred position and standard deviation. The standard deviation showed unexpected correlation with factor-DNA proximity. Apparently, the factor that is responsible for the motif binding has a significantly lower standard deviation than other associated proteins. The fixation of DNA binding proteins to DNA limits the variability of summit positions relative to the bond sequence. In contrast, the spatial distance between DNA and indirectly attached factors causes higher mobility, which is restricted only by the structural characteristics of protein-protein interactions. Approximate position preferences of cofactors can be tracked and their connectivity order can be distinguished.

Pair Shift View

The pair shift view shows the summit distance distributions of the selected ChIP-seq data related to the motif as a histogram. The X axis represents the distance from the middle of the given motif, which is marked as the "0" point. The numbering of the axis is consistent with the position weight matrix below the diagram. The Y axis shows the frequency of summit occurrences at the positions relative to the motif center. In the case of a well-defined protein topology, high overlap frequency, and close DNA localization, the curve has a bell-like pattern. According to our observations, the narrowness of the curve is inversely proportional to the protein's physical distance from the DNA. This relationship can be detected when looking at the standard deviations as well. Factors with low overlap frequency and no position preference show plateau distribution. Setting the parameters on the drop down boxes, we can investigate the summit distribution of 1 to 3 experiments around an adjusted motif. The minimum and maximum values of the axes are configurable as well, in the text boxes below the diagram. A rolling mean with a 5 bp frame was applied to smooth the frequency curves. There is a possibility to select an experiment in this view and see it in the ExperimentView.

Experiment view

At the early stage of our work, we collected 4068 human ChIP-seq data from public databases. From this data, 3782 experiments were successfully processed and used in the following steps of the analysis. The basic information from this data is at least as crucial as the final results. As previously mentioned, we tried to use a wide variety of ChIP-seq data considering both the origins and the target proteins. To track the source of the data, we created an "Experiment view", which is a more manageable and readable way to browse essential information about the distinct experiments by putting all of the data into a simple table.

VennView

The diagrams of the motif view cumulatively represent the statistical data of all occupying ChIP-seq experiments on all instances of an adjusted motif type. The co-occurrence frequency of distinct ChIP-seq summits from different experiments is not taken into account here. To fill this gap, we created a Venn diagram view. The Venn diagram displays all possible logical relationships between different sets. In our case, the sets are the motifs, which overlap with the peaks of a chosen ChIP-seq experiment, and the relationship is the number of common motifs that are simultaneously occupied in these experiments.

CTCF binding sites in genome regulation and gene expression

Using the motif view of ChIPSummitDB provides not only protein positioning information around the adjusted motif type, but also an extensive picture of the occupied protein network. The number of summit-motif co-occurrences can be tracked and visualized on a scatterplot. Using the co-occurrence frequency as the Y value, we can browse the most frequently occurring ChIP-seq experiments around a motif type. This highlights the members of commonly assembled protein complexes.

As we were familiar with the CTCF and cohesin complex, we started to analyze their related network. The identified CTCF motifs showed frequent ChIP-seq occupancy with various transcription factors. Strikingly, in addition to cohesin signals, other factors were also located downstream of the CTCF element. In the order of co-occurrence frequency, the YY1 and ZNF143 signals were the most enriched factors at the CTCF motifs next to the cohesin ChIP-seq signals. The list and frequency of juxtaposing factors were variable, while the abundant presence of YY1 and ZNF143 was relatively constant between different cell types. Interestingly, the global hierarchical clustering analysis of CTCF elements revealed that other

peaks could be observed in close proximity to CTCF binding sites only in the presence of YY1 and/or ZNF143. We performed hierarchical clustering with Manhattan distance on CTCF binding sites in a cell type specific manner. ChIP-seq experiments were collected from the same cell line but with different antibodies. The signal from these experiments was compared on experimentally validated CTCF binding sites. Four cell lines were used in the analysis. The results were consistent in all cell lines; details are shown for GM12878 only.

We investigated the complete peak sets of CTCF, YY1, and ZNF143. All factors have a remarkable peak number in the GM12878 cell line. We investigated the overlap between factors. The analysis distributed the peaks into subsets, which represent large populations. We investigated the presence of CTSs in the subsets. The results suggest that all of the peak sets overlap with CTSs up to 55 %. In the case of CTCF, this means that the CTCF ChIP-seq signal appears not only on CTSs binding sites, but several phantom peaks can also be observed in the genome. The phantom peaks may represent genomic regions, which connect to CTCF indirectly, through other factors. The large overlap indicates a common co-occurrence between ZNF143, YY1, and CTCF; however, the low CTS presence under ZNF143 and YY1 peaks highlights that these two factors have several other interaction sites in the genome. Thus, this relationship between factors is not mutually exclusive between ZNF143-YY1 and CTCF. However, the ratio of CTS was higher in common binding sites.

We separated the CTSs which had CTCF ChIP-seq signals in GM12878 into two populations. In the first population, YY1 ChIP-seq signals were also present, while the second population lacked YY1 ChIP-seq signals. The results showed that CTSs are highly enriched in other transcription factor ChIP-seq signals in the presence of YY1. This enrichment almost vanished in the absence of YY1.

Other factor peaks usually show relatively lower intensities than their instances in promoter or enhancer regions. We can link these regions to each other: CTCF anchor regions with the presence YY1 and low factor signal and transcription factor binding sites with strong ChIP-seq signals and motif presences. The RUNX3 binding sites are located within 100 kbp of the CTCF binding site.

These observations lead us to conclude that ZNF143 and YY1 not only co-occupy binding sites with CTCF and cohesin, but they also establish a connection between cohesin and transcription factors on regulatory regions of nearby genes. It is worth mentioning that there are differences between the YY1 and ZNF143 overlap with CTCF. ZNF143 generally shows more frequent juxtaposition with CTCF than YY1. The occupied ZNF143 summit displays a narrow distance distribution curve with a lower standard deviation relative to the CTCF motif

center and CTCF summit positions. In contrast, YY1 displays a broad summit distance distribution curve with higher SD. The maxima of both factors are co-located. This position coincides with the predicted SMC1/3 positions.

Investigation of GATA1 and TAL1 binding events with summit analysis

As mentioned previously, two or more closely situated binding sites form a composite element. CEs are well studied and their PWMs are represented in motif databases, e.g. TAL1:TCF3, MAX:MYC, POU5F1:SOX2, RXR:VDR FOS:JUN, etc.. The collaboration between GATA1 and TAL1 in erythroid development and differentiation from multiprogenitor cells into red blood cells is well studied, and their CE is represented in the JASPAR database. If we separately investigate GATA proteins in the vicinity of GATA1 binding sites, we can observe a large population of summits, which are situated 7 base pairs upstream from the GATAA sequence's guanine nucleotide. The summit position enrichment is located approximately 9-10 base pairs downstream, relative to the GATA1 consensus motif center. This observation is valid for all investigated GATA proteins.

In the case of TAL1, the summit positions cannot sharply delineated. Due to the well-known heterodimerization between TAL1 and TCF3 proteins, we can investigate their CE. Both proteins bind to the CAG DNA sequence in convergent orientations, which makes the CE palindromic. Unfortunately, the palindromic nature of the CAGCTG sequence affects the definition of protein positions despite the elimination of redundancy. The maxima of the distribution curve are located around position 1-2, which represents strong TG base pairs. Interestingly, the shoulders of the distribution curves are at -20 and +20 base pair away from the motif center. These represent a summit population with a position away from the core motif. Because of the disturbance of the palindrome sequence, the opposite shoulders can be considered one population. Their remote location remains unclear, but we hypothesize that this is caused by the co-binding events of other factors. We investigated the TCF3 motif also, which is quite similar to the TAL1::TCF3 CE. The summit distances of TAL1 and TCF3 proteins show a congruent distribution on both motifs.

Since GATA1 and TAL1 bind non-palindromic sequences, they are useful for further study and validation of our technique and database. The scatterplot for the GATA1:TAL1 composite element indicates discernible segregation between TAL1 and GATA1 proteins. Both individual signals are located in the proximity of their binding site with a relatively low standard deviation. However, we can observe the same position preference for GATA1 summits as in the case of GATA1 motif, which is not situated directly on their binding site, with an average

signal around 5-6 base pairs downstream of the GATA1:TAL1 motif center. In contrast, the TAL1 signal is overlapping with the TAL1 motif at position 8-10, which does not correlate with the observed GATA-like shift. However, the TAL1 signal can be compared to a non-palindromic motif; the summit distance distribution shows a broad enrichment around the mentioned position. The maxima of distance distribution have similar locations as observed in the case of the TAL1::TCF3 motifs. The TAL1 summits, which are juxtaposing with GATA1:TAL1 motifs or TAL1:TCF3 motifs, represent two different clusters. The two CEs are barely overlapping the same summits from the same experiment. The overlap ratio is less than 20 % of the TAL1 summits from bone marrow samples.

To approach the phenomenon from a structural perspective, we investigated the published GATA and TAL X-Ray Diffraction results. A common chart including both GATA1 and TAL1 data was not available. Therefore, we combined separate models. The complete protein structures were not determined, only the DNA recognition element has been crystallized. Approximately 15% of GATA and 27.5 % of TAL1 protein are known. Both identified regions are closer to the N-terminus of the proteins, but the vast majority of the GATA N-terminus is unknown. The uncharted region may be related to the shift in ChIP-seq summits. Mapping the results on a B-DNA model allows us to present the structural background of the phenomenon. The GATA1 binds the DNA in the major groove. Accordingly, the model shows how the zinc finger is interlocking into the major groove. The bHLH motif of TAL1 acts in a similar manner. The GATA motif can be found between -9 - -4 relative to the CE center. The maximum of summit distance distribution is at 1 bp. In the B-DNA model, this position is in the adjacent minor groove, exactly where the uncharted N-terminus of GATA is facing. Results from an investigation into the distance distribution of GATA and TAL in different cell types are congruent.

Regulatory SNP analysis in ChIPSummitDB

We integrated data from the human archive of Single Nucleotide Polymorphism Database, a broad collection of simple genetic polymorphisms containing more than 893 million submissions covering as many rSNPs as possible, and investigated their relationship with responsive elements. The overlapping motifs with a specific SNP are viewable in base pair resolution, which allows the examination of the modified nucleotide and its significance in DBD recognition. The most efficient way to simultaneously examine rSNPs and TFSBs is via the genome browser. We can use a dbSNP ID to investigate specific SNPs or we can examine a genomic region and the list of involved TFBSs and rSNPs. The binding sites are displayed as

PWMs, which facilitates the assessment of the effects of a specific nucleotide change. The dbSNP view provides a graphical interface, which displays the list the SNPs and their overlapping motifs with PWM scores. The dbSNP view also provides a list of ChIP-seq experiments, which have overlapping signals with the SNP. Therefore, we can collect data about the interacting proteins from different cell lines and highlight which cell lines have binding proteins at the SNP position. Thus, we can assess how the direct binding event is affected by the SNP. In loss of binding, other factors in the complex can vanish from this region. In the list, we can find information about the disturbed factors in different cell lines.

SNP rs2742624 is an A to G transition located in an intergenic region, approximately 4100 base pairs upstream of the UPK3A gene. UPK3A is expressed in the inner membrane of the urinary bladder and contributes to the strength of that membrane. The absence or loss of a functioning UPK3A protein leads to renal adyplasia. In recent studies, GATA1 was identified as a regulator of the UPK3A gene. A mutation in the GATA responsive element leads to decreased expression of the gene. According to their results, the presence of rs2742624 decreased GATA2 ChIP-qPCR and UPK3A mRNA expression in LNCaP cell lines. The mutation occurs at the last A nucleotide of the GATAA core motif, which leads to lower binding affinity. We also identified the GATA1 binding site in our database. The rSNP influences not only a GATA2 motif, but also affects predicted GATA1, GATA3, and GATA5 binding sites. The overlapping ChIP-seq peaks indicate that the SNP can disorientate GATA1 binding in pro-erythroblasts, GATA2 binding in HUVECs, and SHSY5Y and K562 and GATA3 binding in MCF7 cells.

5. Discussion

The data producing laboratories and authors are focusing on specific examinations, which support their projects. However, there is more information in HTS data, which has not been revealed yet. As the data requires large computing resources, especially in large scale comparisons, only a few laboratories undertake this challenge. Several processing steps are needed to extract the necessary information from raw data. Downstream analysis of raw HTS data is required for extracting meaningful information. The rapid maturation of HTS techniques is a result of the development of processing software and protocols. Numerous working groups are working simultaneously on the challenges of data processing, resulting in the appearance of distinct software, which combines already existing and newly developed algorithms. The programs often solve the same problem with slightly different methods. The selection of

processing programs is crucial in the construction of the pipeline. Since our investigations often require the comparison of several ChIP-seq experiments, we created a uniform data processing protocol. During our investigation, we observed that the ChIP-seq summit positions can reveal protein position information in complexes, which is demonstrated through the example of CTCF-cohesin. The visible shift and order between CTCF and cohesin subunit summit locations are related to the connection order and position preferences of the different proteins relative to each other and to a fixed genomic point. We used the center of the CTCF motifs and published protein structure data to understand the topology of CTCF-cohesin complexes. The extended analysis revealed the strand specific orientation of proteins, which follows a CTCF-SMC1/3-RAD21-STAG1 sequence. This locates the cohesin ring to the downstream of the CTCF motif. Further investigation with CTCF ChIA-PET data revealed that the CTCF motifs of chromatin loops within TADs face each other. Observation of the convergent orientation of two anchoring CTCF motifs was congruent with the results from other labs. Relying on this observation, we assumed that the cohesin ring has a proximal position in the DNA loops. We combined the appearance order of cohesin subunit ChIP-seq signals and the published protein structure data, which was plotted on a B-DNA model. Using the result, we created a hypothetical topology model of CTCF mediated chromatin looping that integrates the double embrace model. This explains the opposite position of SMC1/3 on the double helix relative to other subunits and supports the hypothesis of the hinge domain's nonspecific binding to DNA.

The cohesin ring, situated inside the loop, enables the physical proximity between inter loop enhancer-promoter regions. We created a "permanent" loop set from MCF7 ChIA-PET parallel replicas. About 60 % of the consensus loops could be linked to an active promoter/enhancer mark. The signal intensities showed high asymmetry between opposite sides of the loop, indicating that CTCF-cohesin mediated loops have a structural role in the formation of enhancer-promoter looping. Furthermore, a global analysis revealed transcription factor/ co-factor ChIP-seq signal enrichment in the vicinity of CTCF binding sites when ZNF143 and YY1 proteins are present. This observation indicates the complex role of CTCF loops in transcriptional regulation. The high transcription factor population completely vanishes when YY1 and ZNF143 proteins are not present. The summit positions and their standard deviations highlight some interesting correlation between ChIP-seq signal and protein topology. The low standard deviation of ZNF143 can be explained by direct binding between ZNF143 and cohesin. In contrast, binding seems to be looser as the distance from DNA increases, as for indirectly bound factors. The higher mobility of ChIP-seq summit positions increases the standard deviation. However, the maxima of distance distribution curves indicate the

approximate position of indirectly bound factors. In the case of YY1 and ZNF143, the maxima are overlapping with SMC proteins, suggesting that ZNF143 and YY1 interact with the hinge domain of SMC proteins and close enhancers through YY1, ZNF143, and other transcription factors at the cohesin ring. As a large number of correlations could be revealed with extended analyses of these factors, we decided to expand our focus to other regulator proteins and apply this technique. The large amount of data was provided by public databases.

Peak prediction gives us information about the approximate position of TFBS, but this provides only a blurry picture. Several factors can influence the width of predicted peaks, such as biological factors, like the cobinding of other proteins, or technical issues, like the selection of the peak prediction program. If investigating transcription factors with known DNA binding domains, the motif enrichment analysis can reveal the preferred sequence. A combination of known preferred sequences and summit positions can give us a more precise prediction of the concrete protein location. Large collections of transcription factor DNA-binding preferences are published, like JASPAR or HOCOMOCO. These databases contain curated sets of profiles, collected from literature data. Motif finding software frequently uses these databases to find the most similar profiles to the identified motifs. The profiles are stored as position weight matrices, which can be used to find the occurrences of the TFBSs in the genome. These motif centered databases usually store transcription factor profiles or motif enrichment reports of HTS data, but often lack positioning information for the TFBSs.

The ChIPSummitDb and the previously mentioned databases share common features:

- Large scale data collection
- Uniformly processed ChIP-seq data
- Transcription factor binding site prediction
- Comparable binding sites
- Combinable file formats
- Downloadable content
- Motif report to ChIP-seq experiments

But, only a few databases work with occurrences of a given motif. The difference between our database and other databases based on motif localization centered approach is that we use the JASPAR database as a source of motif matrices, and we attempted to identify their genome wide localization of motif instances. The collected ChIP-seq data provide experimental validation. The motif information is represented, not just as a motif enrichment report, but concrete genomic locations. We investigate the occupancy of different ChIP-seq experiment signals on the identified motifs. Thus, we can map the protein network, which is connected to

distinct regulatory sequences. The identified motifs serve as fixed reference points in the genome. This makes the positions of connecting proteins measurable relative to each other proteins or to other motifs.

The developed summit position based topology prediction was applied in our database. In addition to TFBS identification, we created a network map of the factors that show overlap with different types of binding sites, based on the downloaded ChIP-seq experiments and the presence of their signals in the vicinity of the binding sites. We can get a comprehensive view of the protein complexes and involved proteins for a motif type with the display modes of ChIPSummitDB. The provided information and features include:

- Occurrence of specific proteins on the adjusted motif
- The preferred position of proteins relative to the center of a motif
- Detailed histogram about the summit- motif center distance distribution for an adjusted experiment
- Overlap between ChIP-seq experiments in correlation with a given motif
- All produced data is viewable in a genome browser
- Downloadable content
- SNP scan on transcription factor binding sites
- Detailed information about processed ChIP-seq experiments, including origin, link to other databases, and motif enrichment

The GATA1:TAL1 motif was investigated as a composite element and the juxtaposition could be clearly tracked between GATA1 and TAL1. Thus, our results were congruent with the published structural data. The distance distribution maxima order followed the pattern in the CE. The TAL1 maxima coincided with the TAL1 core motif in the CE, while the GATA1 maxima signal showed a 7 bp shift relative to the start of the GATA1 motif. The shift can be explained with the structural characteristics of the GATA1 protein. However, the protein structure is barely charted with X-ray crystallography and the vast majority of missing structure is facing in the direction of the shift. The shift maxima are located almost one DNA turn away, relative to the core motif. This suggests that the detected ChIP-seq summit position preference may be a result of cross-linking between GATA1 protein non-DNA binding regions and the proximal DNA region during preparation of the ChIP.

6. Summary

The ChIP-seq technique can be used to extract topological information about protein complexes. We developed a summit position based technique, which was used to identify protein positioning relative to a fixed genomic point. To do this, we identified transcription factor binding sites genome-wide with ChIP-seq experimental validation. The identified TFBSs were used as reference points, to measure motif-protein and protein-protein distances. The technique was tested with a CTCF-cohesin complex analysis. The results revealed the cohesin subunit internal orientation in chromatin loops and its structural support in transcriptional regulation and insulation within TADs. The mediator function of YY1 and ZNF143 was also identified between cohesin and transcription factors.

The analysis was extended and we investigated several proteins with published ChIP-seq experiments. We created a database from the results, which contains more than 3702 processed ChIP-seq data. The results have been made publicly available through the http://summit.med.unideb.hu/summitdb/index.php domain. The web interface provides a surface to download and visualize data. Different display modes are provided to investigate transcription factor binding sites and their protein networks in detail. The database was tested with published structural data, including GATA1:TAL1.

Funding

This work was supported by the GINOP-2.3.2-15-2016-00044, the 2017-1.3.1-vke-2017-00026 and the FIKP_20428-3_2018_FELITSTRAT grants.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK376//2019.PL
Subject: PhD Publikációs Lista

Candidate: Erik Czipa
Neptun ID: SYFW19
Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

List of publications related to the dissertation

1. **Czipa, E.**, Schiller, M., Nagy, T., Kontra, L., Steiner, L., Koller, J., Pálné, S. O., Barta, E.: ChIPSummitDB: a ChIP-seq based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them.
   Database. [Epub ahead of print], 2019.
   DOI: http://dx.doi.org/10.1093/database/baz141
   IF: 3.683 (2018)

2. Nagy, G., **Czipa, E.**, Steiner, L., Nagy, T., Pongor, S., Nagy, L., Barta, E.: Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA.
   BMC Genomics. 17 (637), 1-9, 2016.
   DOI: http://dx.doi.org/10.1186/s12864-016-2940-7
   IF: 3.729

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

List of other publications

3. Simándi, Z., **Czipa, E.**, Horváth, A., Kőszeghy, Á., Bordás, C., Póliska, S., Juhász, I., Imre, L., Szabó, G., Dezső, B., Barta, E., Sauer, S., Károlyi, K., Kovács, I., Hutóczki, G., Bognár, L., Klekner, Á., Szűcs, P., Bálint, B. L., Nagy, L.: PRMT1 and PRMT8 regulate retinoic acid-dependent neuronal differentiation with implications to neuropathology.
Stem Cells. 33 (3), 726-741, 2015.
DOI: http://dx.doi.org/10.1002/stem.1894
IF: 5.902

Total IF of journals (all publications): 13,314
Total IF of journals (publications related to the dissertation): 7,412

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

20 November, 2019