Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

3-D genomic interactions and their relation to gene expression

Thesis submitted in partial fulfillment of graduate requirements for

The degree "Master of Sciences" in Tel-Aviv University

School of Computer Science

By

Idan Nurick

Prepared under the supervision of

Prof. Ron Shamir

Dr. Ran Elkon

August 2017

Acknowledgements

I would like to use this short note to express my gratitude to a group of people that helped me conduct this research.

First, I would like to thank Prof. Ron Shamir for setting an example of a true scientist, both hard working and brilliant, who was incredibly patient with me throughout the past years. We knew our ups and downs, and I'm truly grateful you lead us to this point. I would like to thank Dr. Rani Elkon for joining us seven months ago with an amazing spirit and wisdom, showing me the strength of innovation and curiosity.

I would like to thank Michal Ozery-Flato and Liat Ein-Dor for collaborating with us for more than a year, sharing with me their exceptional knowledge and experience.

I also like to thank my lab members in the past years – Tom Hait, Ron Zeira, Roye Rozov, Yaron Orenstein, David Pellow, Kobi Perl and Gal Dinstag. Two members deserve a special thank you – David Amar for the inspiration and the passion for science and Dvir Netanely for being a mentor in his own special way. I would also like to thank the Edmond J. Safra Foundation for the support over the past months.

Last but definitely not least, I would like to thank my parents for everything. It is always a great honor to get an opportunity to thank them.

Abstract

The 3-D organization of the genome in the nucleus has an important role in many aspects of cellular life, including gene expression control. Recently, several high-throughput methods have been developed that allow insights into the chromosomal architecture and chromatin interactions at unprecedented resolution.

One of these techniques, Hi-C, is based on chromosome conformation capture. Using Hi-C, several genome structures were characterized. At low resolution, there are two different types of compartments in every chromosome – type A (gene rich) and type B (gene poor). Each compartment can then be divided into sub-compartments using epigenomic features. At higher resolution, topological associated domains (TADs) were identified. TADs are chromosomal segments that tend to span most interactions within them. Another technique, ChIA-PET, is based on dynamic conformation capture, allowing the capture of even higher resolution of chromatin interactions inside TADs.

Gene expression profiles that were recorded in response to a multitude of stresses in different cell types showed that a large portion of the transcriptional response to stress is cell-type specific and only a small minority is universal. Our goal is to examine to what extent the spectrum of genes induced by stress in each cell type is determined by structure constraints that exist before stress was applied.

In order to do so, we performed a wide examination of 13 different cell types under different treatments and analysis (RNA-Seq, GRO-Seq, Chip-Seq, etc.), and checked whether measured response correlates with pre-defined chromatin organization. Our results imply that there is a significant correlation between cell-type specific response to stress and structure constrains in the untreated cell.

Contents

Abstract 5
1. Introduction 8
2. Background 10
2.1. Biological background 10
2.1.1. Biological concepts 10
2.1.1.1. Gene regulation 10
2.1.1.2. Chromatin organization 12
2.1.2. Next Generation Sequencing 16
2.1.2.1. ChIP-Seq 16
2.1.2.2. RNA sequencing 17
2.1.2.3. Global Run-On (GRO-Seq) 18
2.2. High Throughput methods for detecting chromatin 3D interactions 20
2.2.1. Hi-C 20
2.2.2. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) 24
2.3. Relationship between 3D organization and gene expression 26
2.3.1. Defining 3D structures from Hi-C data 26
2.3.2. Correlation of 3D structures to transcription levels 30
2.3.3. 3D structures and response to stress 32
2.4. Computational background 34
2.4.1. Hi-C contact matrix normalization 34
2.4.2. Principal Component Analysis (PCA) 35
2.4.3. Data Analysis 36
2.4.3.1. Chi-square two sampled test 37
2.4.3.2. Nonparametric statistical tests 37
3. Results 39
3.1 Chromosome compartmentalization 39
3.1.1. Gene expression correlations with A/B compartmentalization 41
3.1.2. Correlation between A/B compartmentalization and TF binding. 43
3.2. Correlation between gene expression level and extent of promoter interactions 48
3.2.1. Promoter-chromatin interactions are correlated with gene expression 48
3.2.2. Changes in promoter interactions across cell lines correlate with changes in gene expression levels 52

3.3. Correlation between basal chromatin interactions and gene induction upon treatment 55
3.3.1. Chromatin compartmentalization and induction of TF binding upon treatment 55
3.3.2. Chromatin compartmentalization and induction of TF binding upon treatment; Comparison between cell lines 57
3.3.3. Stress-induced genes are enriched for A compartment 58
3.3.4. Promoters of induced genes have more basal chromatin interactions prior to treatment 59
3.3.5. Correlation between chromatin compartmentalization and gene induction upon treatment between cell lines 60
4. Discussion 62
References 64
Supplementary Tables 68

1. Introduction

The human genome is highly organized inside the nucleus. The genome is divided into 23 pairs of chromosomes, folded and compressed by various mechanisms, which are known to play an important role in the regulation of gene expression. Several novel methods made it feasible to measure the genome structure itself in addition to understanding the folding mechanisms. This allows us for the first time to systematically explore different functions of the 3-D organization of the genome.

The 3-D structure of the genome has been studied in different resolutions, each reveals another layer of gene expression regulation. The lowest resolution divides the chromosomes into open and closed regions, defined by dense and sparse interactions [1]. These two types of compartments, A and B, were shown to correlate with euchromatin and heterochromatin properties respectively – A compartment is gene rich and genes located in A are more highly expressed compared to B. The median size of a contiguous A and B segment along the genome is 500Kbp. A compartments cover about 45% of the genome, and B compartments cover 48%. The A compartment is enriched with epigenetic markers known to be prevalent in euchromatin areas (such as H3k9ac), while B compartment is enriched with epigenetic markers known to be prevalent in heterochromatin areas (such as H3k27me3)[1], [2]. Analyzing Hi-C data also revealed that segments of the same type, even in distal linear locations, tend to physically cluster together, forming A and B regions in the nucleus [1]. This suggests that Hi-C data can be useful to obtain a close approximation of euchromatin/heterochromatin areas in the genome.

At higher resolution, each A or B segment is composed of smaller chromatin structures called topological associated domains (TAD) (median size around 200Kbp) [2]. Most of the intrachromosomal interactions revealed by Hi-C occur within TADs. TADs have a significant effect on gene expression - genes within TADs have lower expression than genes at the borders of TADs [2]. Combined analysis of high-resolution Hi-C data (up to 1KB) and ChIA-PET data demonstrated that the borders of TADs are demarcated by loops strongly correlating with enhancer-promoter interactions [3].

When comparing compartmentalization between cell lines, compartment patterns are similar but there are still many discordant loci [1], suggesting that A/B partition is cell line specific. TADs behave differently from the larger A and B compartments in this manner - analysis of different cell lines suggests that the majority of TADs are consistent across tissues [2]–[4]. TADs can be active or inactive in terms of gene expression, and adjacent TADs are not necessarily of opposite chromatin states, suggesting that TADs are conserved chromatin features, and groups of adjacent TADs form A and B compartments [5].

Multiple gene expression studies demonstrated that much of the transcriptional response to stress is cell type specific [6]. Yet, the regulatory mechanisms that dictate the set of target genes that are induced in each cell type in response to stress are poorly understood. Previous studies also indicated that enhancer-promoter interactions already exist in cells before the binding of stress-induced transcription factors, suggesting that transcriptional activation in response to stress triggers relatively few changes in the 3-D organization [7].

In this thesis, we aimed to further explore, on a genomic scale, the relationship between the 3D organization of the genome and gene expression in different cell types, and examine whether the

spectrum of genes induced by stress in different cell types is also determined by 3D structures and enhancer-promoter interactions existing before stress was applied.

Our goal was to test two main alternative hypotheses:

1. Preexisting partition into A/B compartments and pre-existing enhancer-promoter interactions in the untreated cell have a major role in defining “poised genes” that will be induced by different types of stress.
2. In response to stress, multiple changes occur in the spatial organization of the genome, resulting in changes in genes expression.

To explore the relationship between genome organization and gene expression we analyzed all the publicly available Hi-C data that were recorded to date in human cells (13 cell lines) and carried out the following analysis: (1) We defined A/B compartments for each cell type. (2) We validated that known features of A/B compartments indeed correlate with the output of our compartmentalization. (3) We tested for correlation between promoter interactions as measured in Hi-C\ChIA-PET and gene expression. (4) We examined publicly available transcription factor binding sites and gene expression data recorded on cells with 3D organization data after various treatments, and checked for the relationship between gene response and A/B compartmentalization. (5) After observing significant correlation, we checked whether the same correlation exists when comparing gene expression data to promoter interactions of untreated cells, defined by significant Hi-C interactions and ChIA-PET data for RNA pol 2.

2. Background

This chapter lays out the background and terminology required for the thesis. We first introduce basic biological definitions and motivation for our research. Next, we present the high-throughput methods and data types that were used in this thesis, including discussion of inherent biases in these methods and the way they are handled. Finally, we give background for the computational and statistical tests used in the thesis.

2.1. Biological background

2.1.1. Biological concepts

In this section, we present basic concepts in biology and review previous studies relevant to our work.

2.1.1.1. Gene regulation

One of the most basic questions in molecular biology is how one genome sequence can give rise to a variety of tissues. The answer to this question lies, at least in part, in the ability of distinct cell types to express genes, and the proteins they encode, at different levels and combinations, and to react differently in response to changes in their environment. The process of generating products from genes encoded in the DNA is mainly composed of two steps, also known as the central dogma of molecular biology, where each step is regulated by various mechanisms.

The first step, **transcription**, copies the data from the DNA to an RNA molecule, a temporary copy of the gene data. The next step, **translation**, decodes the data in the RNA molecule in order to produce a protein (**Figure 1**).

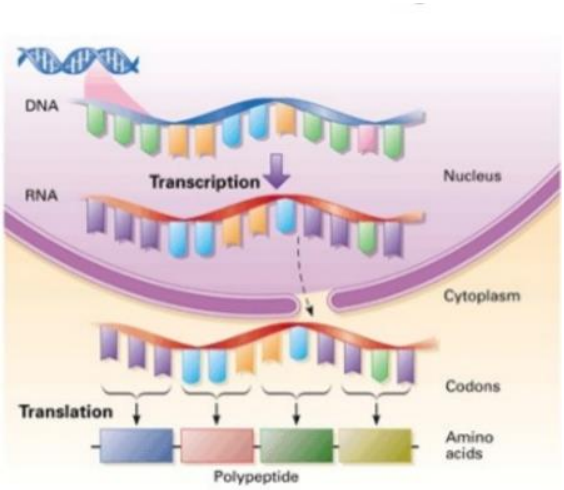


Figure 1 *The central dogma of molecular biology. First DNA is being copied to a temporary RNA molecule called mRNA, inside the nucleus. mRNA molecules cross to the cytoplasm of the cell where they are being translated to proteins.* [8]

Regulation of transcription controls the amount of RNA products available for protein translation. It is considered the primary regulatory mechanism since there is a strong (although not perfect) correlation between RNA amount and protein amount [9]–[12]. The imperfect correlation suggests that features of the genome beyond its primary nucleotide sequence must contribute to the cell specific gene regulation that underlies cellular identity. Many proteins are involved in transcription regulation – RNA polymerase, transcription factors, histone and scaffold proteins, chromatin modulators and many more [13], [14].

Transcription factors are proteins that bind to the DNA in areas called transcription factors binding sites (TFBS) and contribute to the efficiency by which RNA polymerase is recruited to the **promoter** of a gene. Transcription factors and other DNA binding proteins, such as histones, also determine the chromatin structure and packaging, influencing genes accessibility. **Figure 2** demonstrates how histone side chains (also called histone modifications/markers) mark different chromatin states.

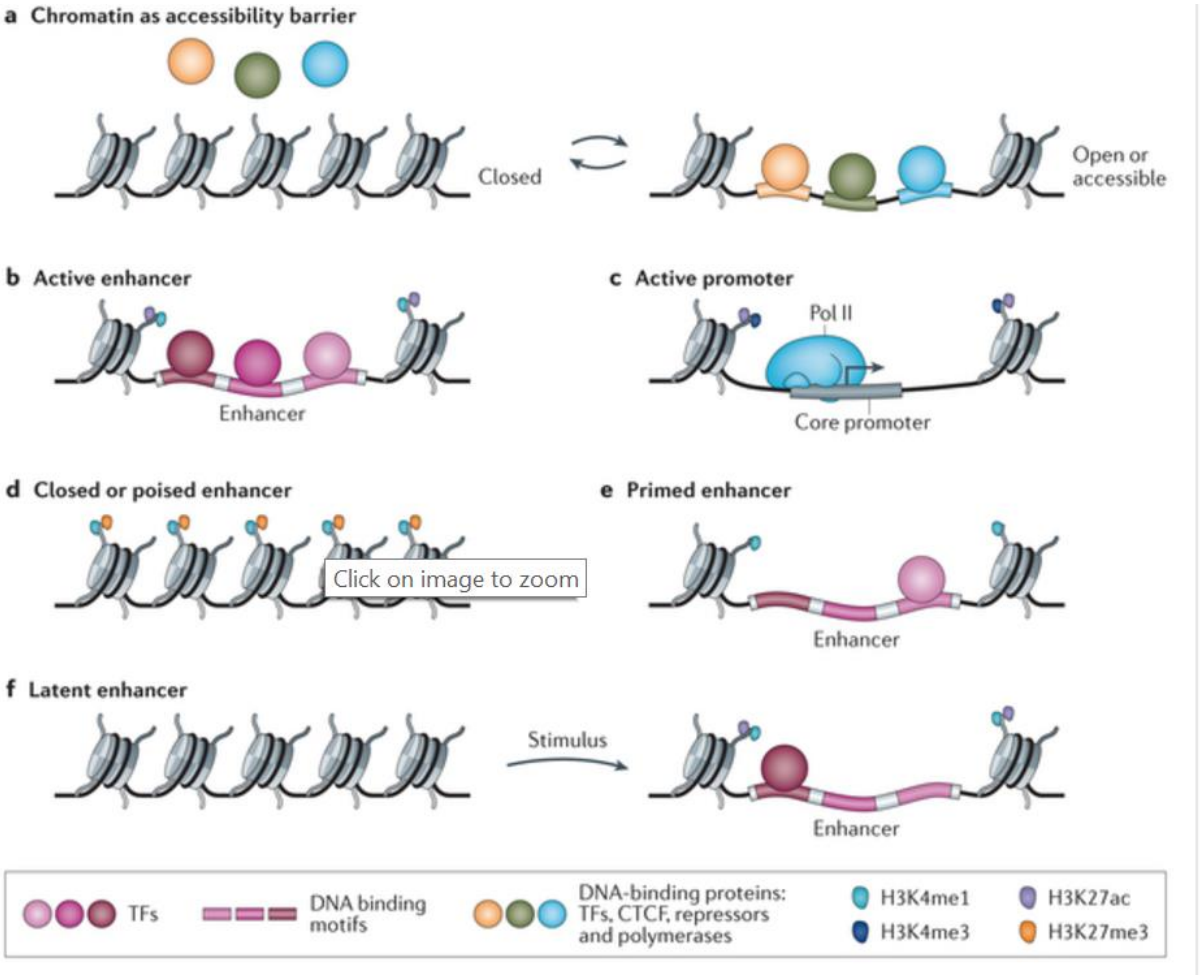


Figure 2 *Chromatin states that allow or restrict access of transcription factors, RNA polymerase II and other proteins. (a) Transcription factors can bind open and accessible regions. The transition between different chromatin states is mediated by histone modifications and pioneer transcription factors. (b),(c) Histone markers H3K27ac, H3K4me3 and H3K4me1 are enriched in active enhancers and promoters, correlating with active transcription (d) H3K27me3 is a histone marker enriched for repressed regions* [15].

Translation regulation controls the amount of protein produced from mRNA using ribosomes. Translation is less understood since until recently there were few high-throughput methods to measure protein levels. Ribo-seq, a novel method to measure the amount of translated RNA, is beginning to shed light on this process. [16]

2.1.1.2. Chromatin organization

Chromatin is a collection of macromolecules in the nucleus of the cell, composed of DNA and DNA-binding proteins [17]. In order to "package" a large amount of DNA in a compact way that fits the cell's nucleus, chromatin is folded in several levels. DNA segments bind and fold around histones, and form complex molecules that eventually can be seen as chromosomes under the microscope (**Figure 3**).

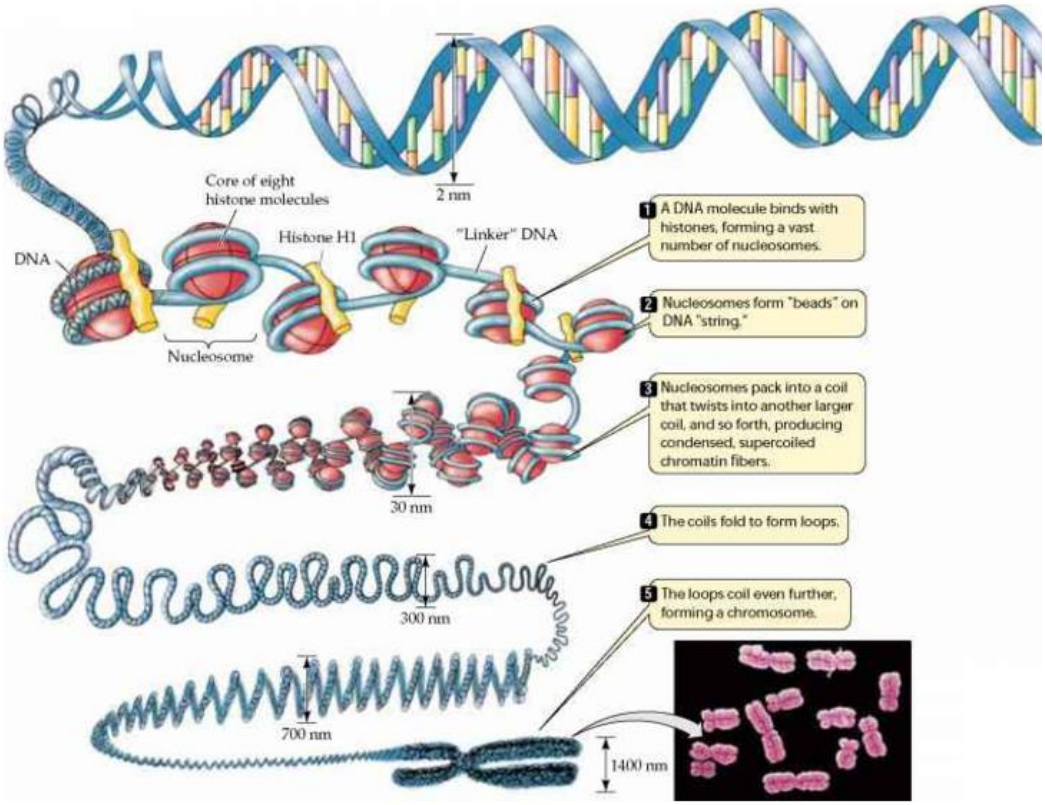


Figure 3 *Chromosome organization. Chromosomes are highly complex macromolecules, constructed from DNA molecules packed around histone proteins forming nucleosomes. The structures hierarchy is described in the text boxes.* [18]

At the lowest resolution, chromosomes are divided to euchromatin and heterochromatin. These are sub-chromosomal structures that can be observed under optical and electron microscope and are characterized experimentally as a dense (heterochromatin) and loose (euchromatin) areas in the chromosome (**Figure 4**). Chromatin stain and epigenetic markers are widely used to distinguish between these two states. One main group of epigenetic markers are histone side chains, which biochemically determining if a certain area in the chromatin will be highly dense and less accessible or less dense and therefore more accessible [19].

In addition to having similar properties, these structures tend to cluster together in the same areas in the nucleus – heterochromatin in areas near the lamina (nucleus envelope), euchromatin near the center of the nucleus. (**Figure 4**)

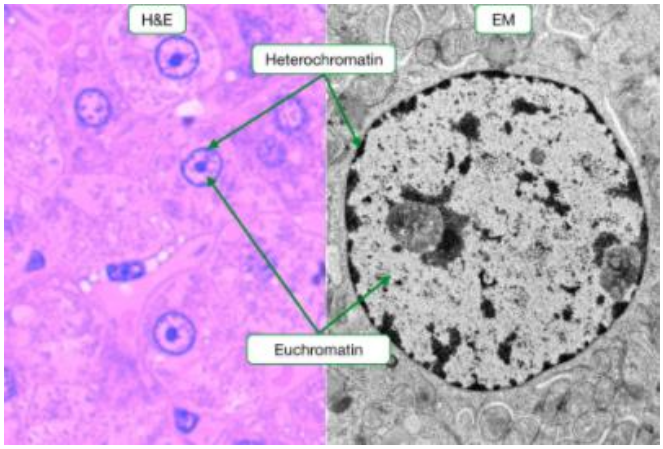


Figure 4 *Euchromatin and heterochromatin viewed by electron microscope (right image) and by H&E staining (left image). The main cytological differences can be easily observed experimentally – heterochromatin regions are mainly located near the lamina, while euchromatin regions are located at the center of the nucleus. Heterochromatin is dense so it looks darker and stains better than euchromatin.*[20]

In higher resolution, each compartment can be sub-divided into topological associated domains (TADs), chromatin structures with a median length of approximately 200KB (**Figure 5**). The human genome is composed of thousands of TADs covering together more than 90% of the entire genome. These structures constrain chromatin interaction such that most of the intrachromosomal interactions occur within TADs and very few occur across TAD boundaries[1]–[3], [21]. TADs are relatively stable between cell types but can change their compartment, typically as a whole unit [4]. Remarkably, TAD positions also highly conserved between mouse and human[3], [4]. Thus, TADs have been shown to play an important role in the regulation of gene expression. [2], [22]

Figure 5 *Illustration of topological associated domains. Each of the two "bundles" is a TAD. Most chromatin interactions are spanned within TADs, inter-TAD interactions are relatively rare.*[23]

In even higher resolution, many studies suggest that TADs themselves are also subdivided to sub-TAD structures, named chromatin loops (**Figure 6, 7**). These structures, similarly to compartments and in contrast to TADs, seem to have cell-type specific properties [3], [7]. Chromatin loops, having a relatively small size and highly dynamic nature, can be better evidenced only in higher-resolution assays [3], [24].

A large fraction of TADs have chromatin loops in their borders. Moreover, the appearance of a loop is usually (in 65% of cases) associated with the appearance of a TAD demarcated by the loop, also referred as loop domains (**Figure 6**). Combined with the fact that highly expressed genes tend to cluster in the boundaries of TADs (see details in the next sections), these loops are suggested to have a regulatory role.

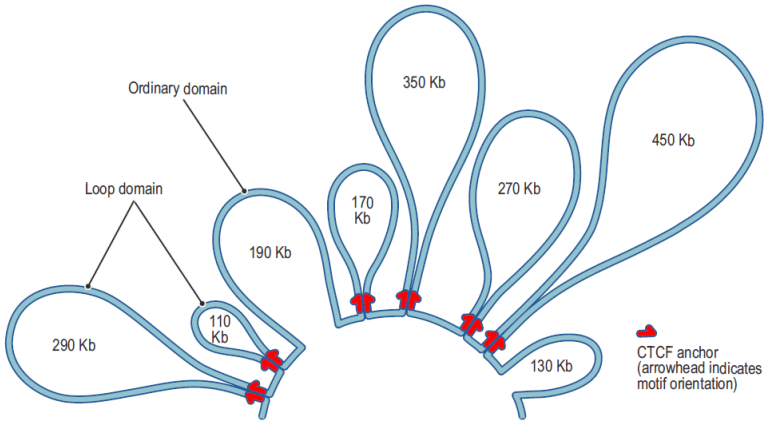


Figure 6 *TADs and loops. An example of 2.1 Mb region on chromosome 20. The figure demonstrates the relation of the CTCF protein marks to loops boundaries and the types of TADs (ordinary and loop) [3]*

The mechanisms involved in TAD establishment are not yet fully understood. Some novel studies [25]–[27] suggest a model in which a complex, including the proteins CCCTC-binding factor (CTCF) and cohesin, mediates the formation of loops by a process of extrusion (**Figure 7**). TADs form as a byproduct of this process. The model was tested on high-resolution spatial proximity maps and showed high consistency, using only information about the locations at which CTCF is bound. Disruption of TAD borders (by impairment of binding of associated factors)[27] seems to produce different architectural and functional effects, indicating the importance of this complex in chromatin organization.

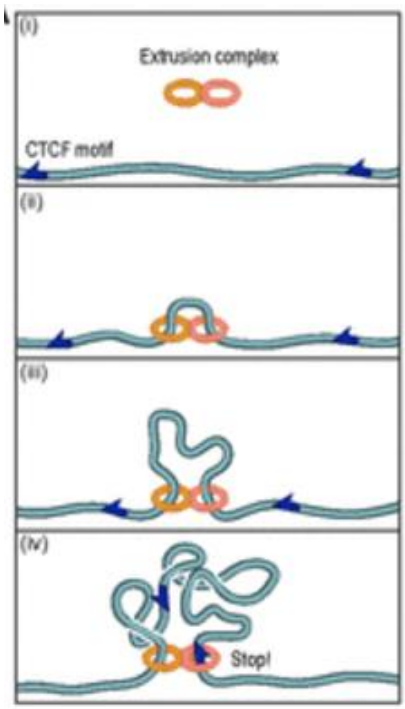


Figure 7 *Extrusion model for loops formation using CTCF binding sites. Chromatin is wrapped by the extrusion complex and the loop grows until it reaches converging CTCF binding sites. The result is a chromatin loop with CTCF binding sites in its boundaries. According to Rao et al., 2014* [3] *converging CTCF binding sites exist in more than 90 percent of loops boundaries.*

Figure 8 summarizes sub-chromosomal structures and their hierarchy.

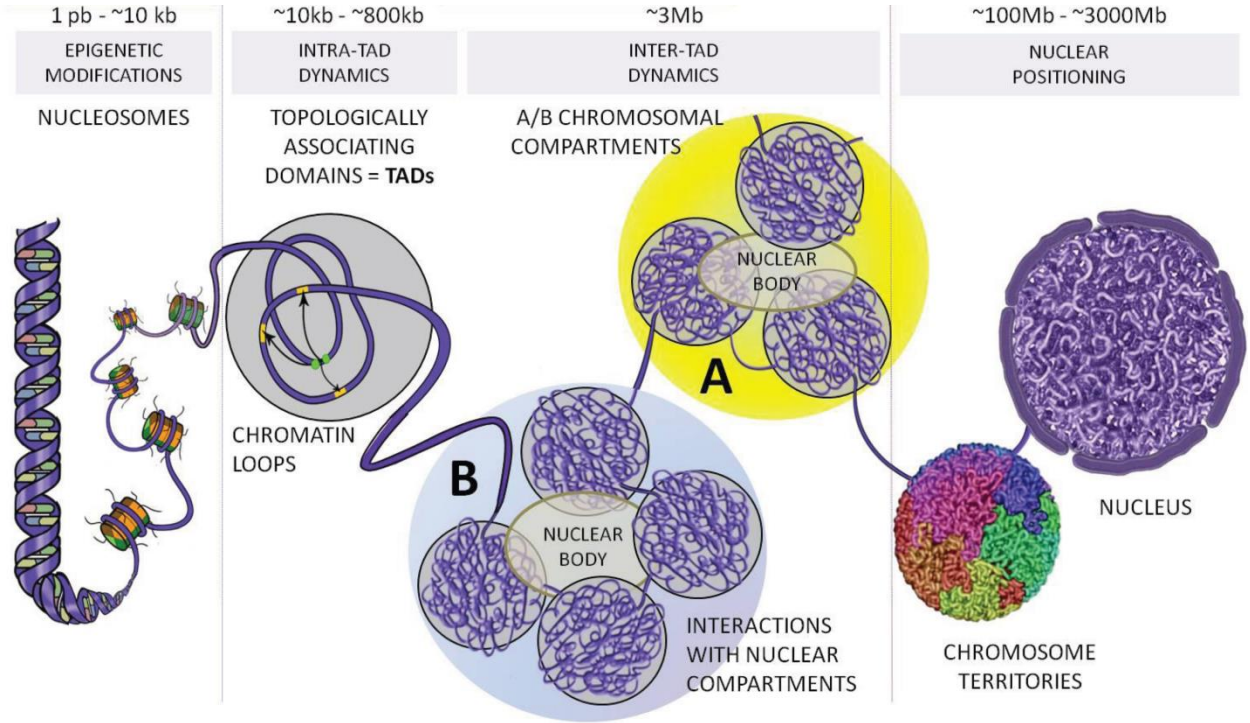


Figure 8 *Summary of sub-chromosomal 3D structures.* [21]

2.1.2. Next Generation Sequencing

DNA sequencing is the process of determining the sequence of nucleotides within a DNA molecule. Next generation sequencing is a general name for several new high throughput methods that were implemented in commercial DNA sequencers over the last decade. These methods revolutionized genomic research and made DNA and RNA sequencing much faster and cheaper, and therefore a widespread and popular tool.

The large quantities of data produced by these methods made it necessary to develop proper tools and programs to handle and analyze it, giving rise and changing the field of bioinformatics. This section introduces the relevant techniques used in our work. It also describes the pipeline we used to process the data and explore it.

2.1.2.1. ChIP-Seq

ChIP-Seq is a method for analyzing protein interactions with the DNA. ChIP-Seq combines chromatin immunoprecipitation (ChIP) with DNA deep sequencing to identify binding sites of DNA-associated proteins [REFs]. ChIP-Seq data can identify histone markers that characterize different chromatin states. Histone modifications are roughly divided into two groups which characterize open and closed chromatin states. We used these markers to validate that our partition to A/B compartments indeed correlated with corresponding chromatin states.

Specifically, ChIP-Seq works as follows:

1. Chromatin immunoprecipitation – (ChIP) is a method to selectively enrich for DNA sequences bound by a specific protein. The process enriches for specific cross-linked DNA-protein complexes by precipitation with an antibody against the protein of interest. After removing the antibodies, oligonucleotides adaptors are ligated to the DNA molecules to enable PCR.
2. Sequencing – The resulting fragments are sequenced using NGS. Sequenced fragments are aligned to the reference genome.
3. Peak Calling - The last step is to computationally identify areas in the genome that have been enriched with aligned reads, indicating protein binding sites. In this work, we used MACS [28], which will be further discussed in Data Analysis section.

Additional details and illustration are found in **Figure 9**.

Peaks detected by ChIP-Seq do not necessarily imply functional DNA-protein interactions. In fact, most binding sites discovered by ChIP-Seq are not functional. [29]

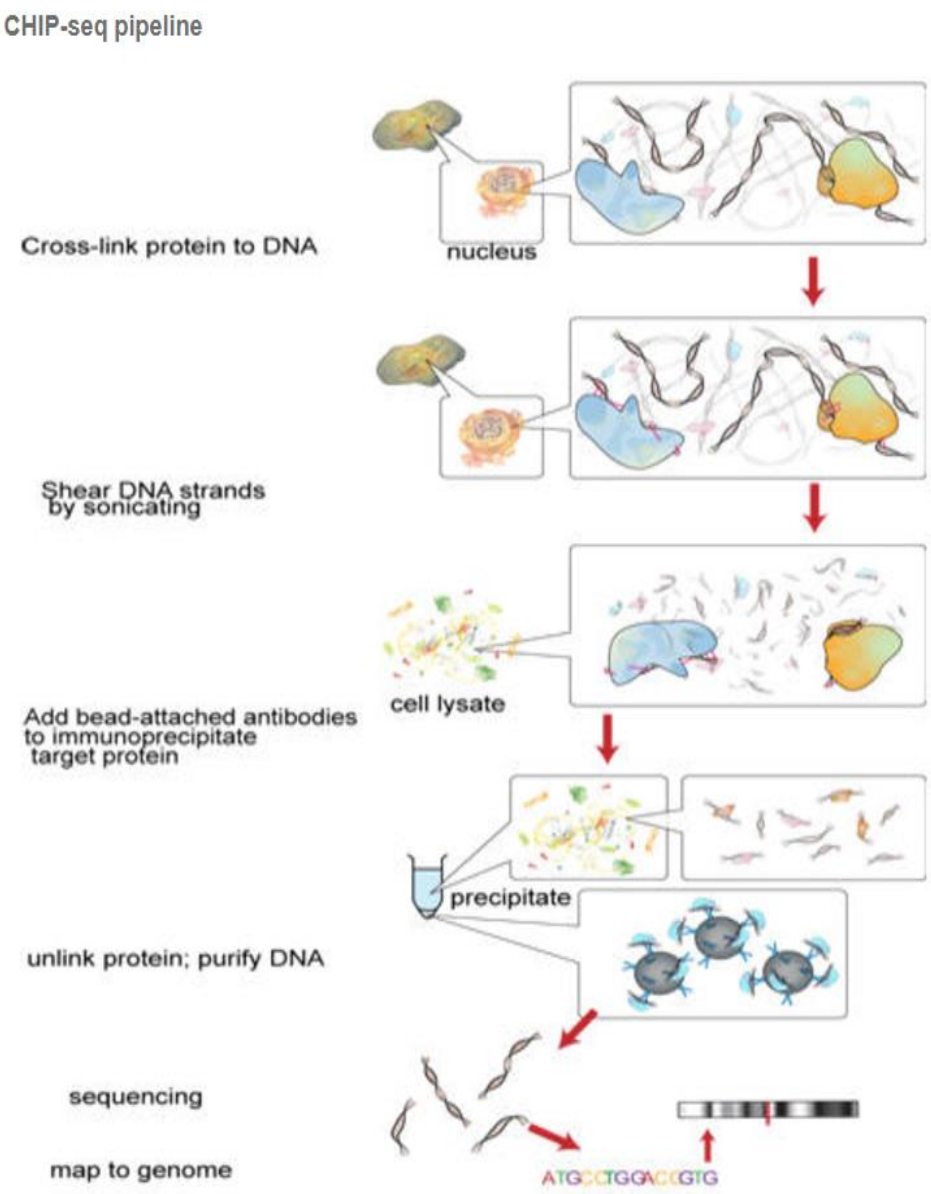


Figure 9 *ChIP-Seq overview. In order to find specific protein binding sites on the genome, a cross-linking agent is injected to the nucleus resulting in DNA-protein complexes. Sonication is used to shear the DNA. Marked antibodies for the protein of interest are added in order to pull down the DNA-protein complexes. DNA is separated from the protein, sequenced and aligned.* [30]

2.1.2.2. RNA sequencing

RNA sequencing (RNA-Seq) is a technique that uses NGS to measure RNA expression levels in sample. In a typical RNA-Seq protocol, cellular RNA is filtered to enrich for transcripts with 3' polyA tails, which characterizes the vast majority of mRNAs. Next, reverse transcription is applied to form cDNA libraries which are then sequenced (**Figure 10**).

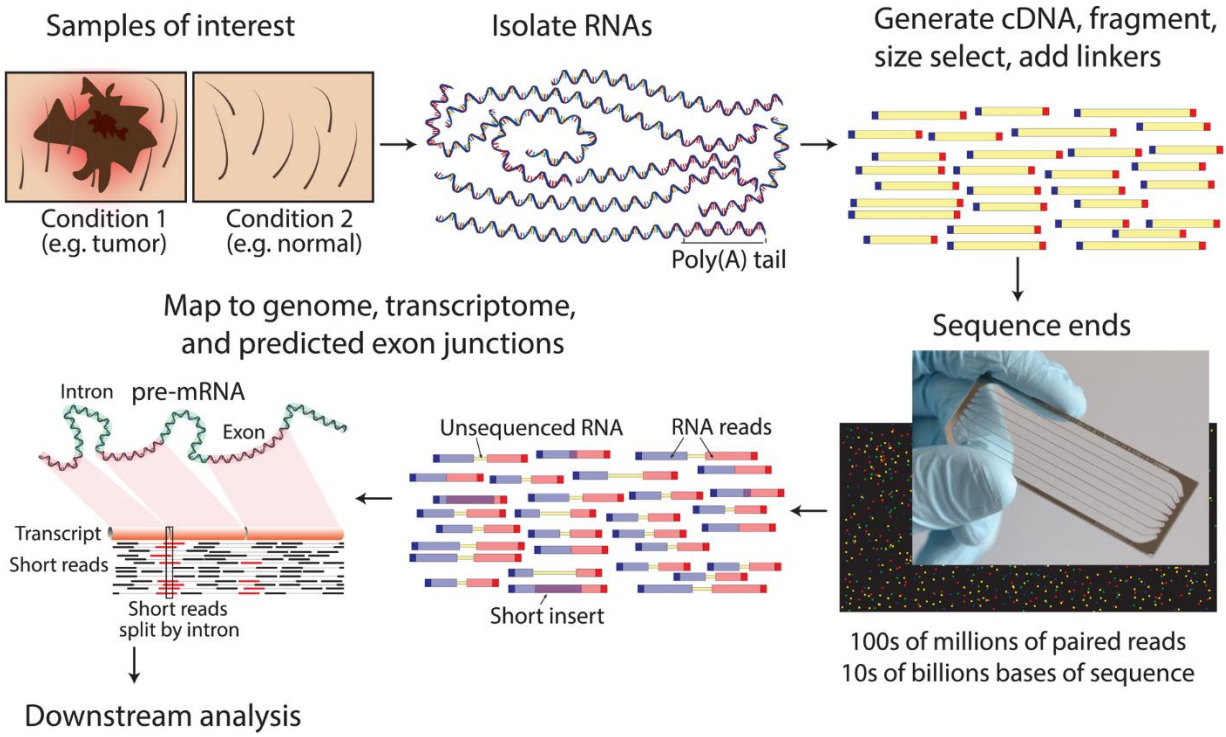


Figure 10 *RNA-Seq overview - A typical RNA-Seq experimental workflow involves the isolation of RNA from samples of interest, generation of sequencing libraries, use of a high-throughput sequencer to produce hundreds of millions of short paired-end reads, alignment of reads against a reference genome or transcriptome, and downstream analysis for expression estimation, differential expression, transcript isoform discovery, and other applications* [31]

2.1.2.3. Global Run-On (GRO-Seq)

RNA-Seq measures mRNA steady state levels, determined by the balance between production and degradation rates (that is, by the rate of transcription and transcript stability). The GRO-Seq technique focuses on nascent transcription and provides estimates of mRNA production rates on a genomic scale.

This measurement is achieved by inhibiting RNA-polymerase and simultaneously labeling all RNA molecules that are actively transcribed. The labeled RNAs are then isolated, sequenced and aligned. The output of the workflow is a list of actively transcribed regions in the genome at the time point when the experiment was conducted (**Figure 11**)

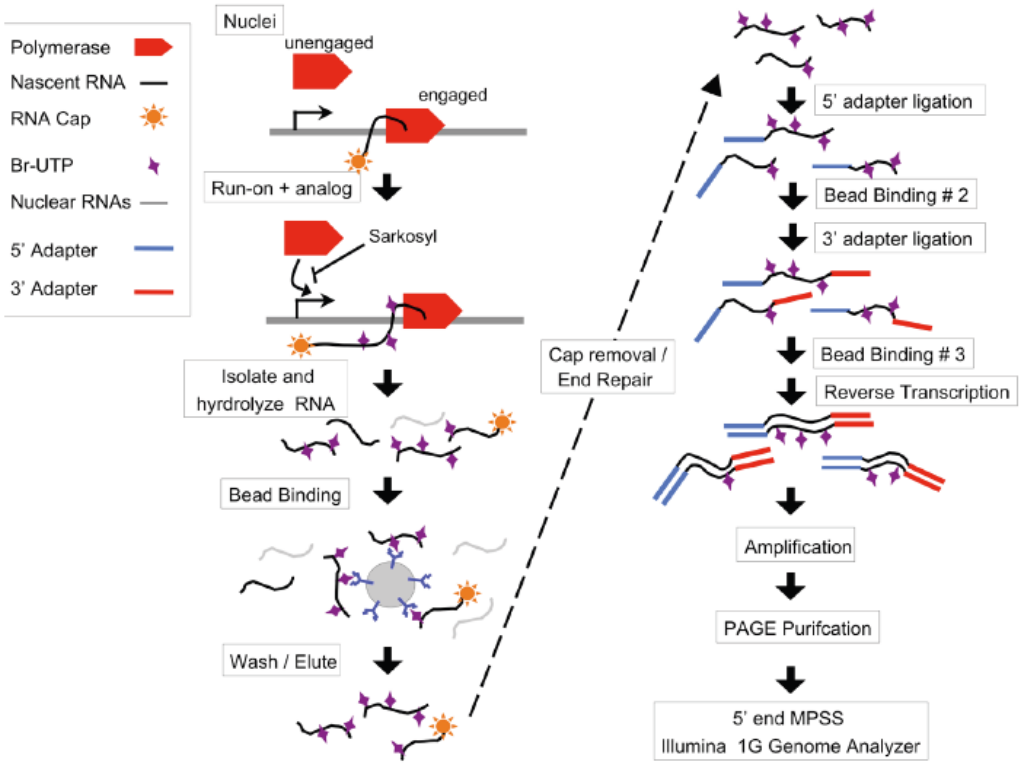


Figure 11 *GRO-Seq overview. The goal of GRO-Seq is to measure active transcription products. In order to do so, Sarkosyl is added, inhibiting RNA polymerase from binding to the DNA. Nuclei are incubated with Br-UTP which marks all newly synthesized transcripts. When isolating all transcripts that contain Br-UTP, the results are products of RNA polymerases that bound to the DNA before Sarkosyl was attached. The next steps are similar to ChIP-Seq and RNA-Seq - amplification and reverse transcription of the resulting transcripts, generation of sequencing libraries and alignment to reference genome* [32]

2.2. High Throughput methods for detecting chromatin 3D interactions

In this section we will describe novel high-throughput methods used to understand how the DNA is organized inside the nucleus. First, we will describe Hi-C, chromosome conformation capture combined with high-throughput sequencing method, presented by Lieberman-Aiden et al. Next we describe a method used to detect dynamic interactions mediated by proteins, ChIA-PET, which identified 3D genome-wide functional interactions.

2.2.1. Hi-C

Hi-C is a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with high-throughput sequencing [1]. It evolved from a series of assays based on chromosome conformation capture – 3C, 4C and 5C [1], [33].

In order to capture chromatin conformation, formaldehyde is injected into the nucleus. As a result, proximal segments are cross-linked. DNA is digested with a restriction enzyme that leaves 5' prime overhang. The 5' overhang is filled with biotinylated residue which is ligated under dilute conditions that favor ligation between the cross-linked DNA fragments. The resulting DNA sample contains ligation products of fragments that were in close spatial proximity when the cell was sampled. These products are marked with biotin at the junction. The methods differ in the last step when sequencing of the ligation products is done **(Figure 12)**.

3C and 4C are designed to for studying an **individual locus** of interest, a gene promoter for example, thus generating single interaction profile. In 3C, also called one-vs-one method, we test a genomic element of interest versus surrounding chromatin. In order to do so, the interaction between a specific pair of loci is measured using PCR with known primers.

The extreme complexity of the 3C library and the low relative abundance of each specific ligation product made it necessary to find less specific methods for large-scale analysis. 4C generates a genome-wide interaction profile for a single locus (one-vs-all), by performing another step of digestion with a different restriction enzyme and self-circularization ligation. Inverse PCR is then performed with primers for both ends of the locus of interest, amplifying the sequence of the unknown paired locus [33].

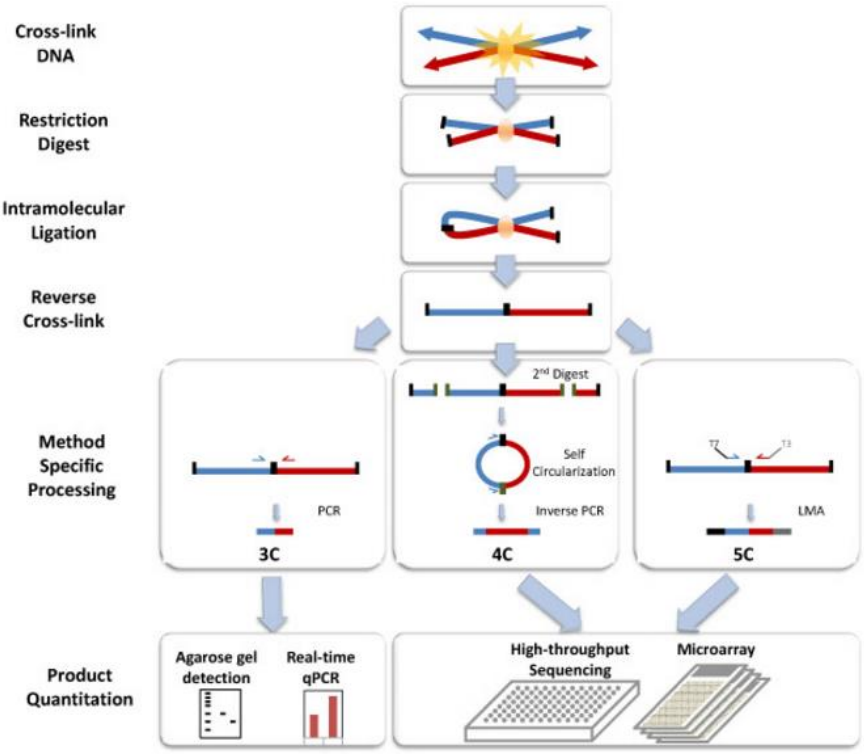


Figure 12 *Principles of 3C-derived methods - All the methods derived from the Chromosome Conformation Capture (3C) protocol start similarly. First, nuclei are incubated with formaldehyde cross-linking chromatin segments in close spatial proximity. Next, a restriction enzyme is used to produce pairs of short cross-linked fragments and a ligation step connects sequence ends that remained in proximity. Each method proceeds differentially to generate genomic libraries: secondary digestion with known primers for 3C, circularization and inverse PCR in 4C, "carbon copy" amplification in 5C.* [34]

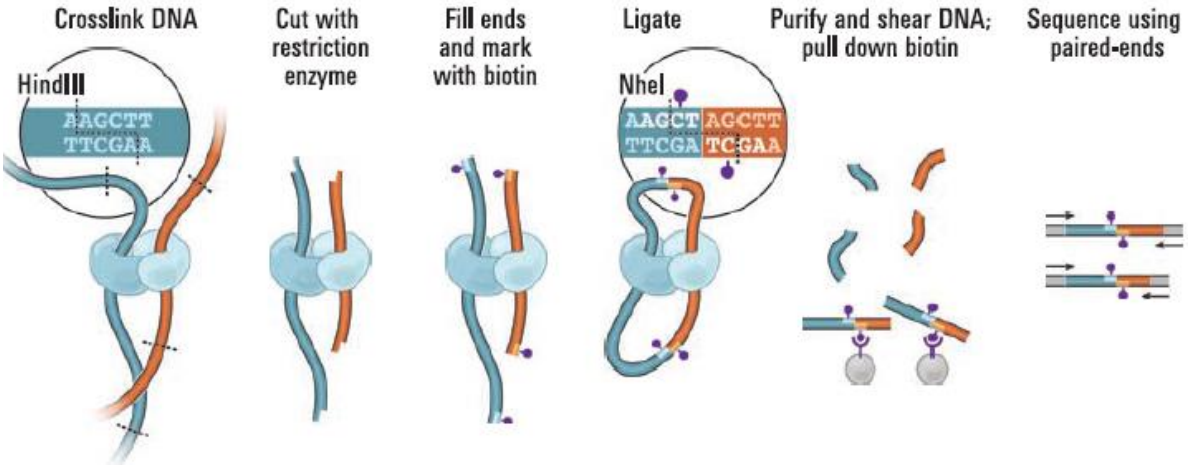


Figure 13 *Overview of Hi-C - As in 3C-derived methods described above, cells are cross-linked with formaldehyde, resulting in links between segments with high linear distance but close spatial proximity. Chromatin is digested with a restriction enzyme and the resulting sticky ends are filled in with nucleotides marked with biotin (purple dot). Ligation is performed under extremely dilute conditions to create DNA circles composed of two chimeric molecules and DNA binding protein. DNA is sheared and separated from the protein, resulting in multiple DNA fragments, where only a subset of them are chimeric products that indicate a 3D interaction. Streptavidin binds biotinylated nucleotides and isolates the fragments of interest that are next sequenced and aligned. [1]*

Chromosome conformation capture carbon copy (5C) measures the frequency of interactions between all fragments defined by a specific restriction enzyme, within a given region no greater than a mega-base (hence this method is called many-vs-many). 5C uses highly multiplexed ligation-mediated amplification (LMA) to first copy and then amplify parts of the 3C library. In the next step, 5C performs another ligation of constant ends to ligation products and amplifies them using universal products (instead of locus-specific primers used in 3C, 4C). This approach is useful for identifying multiple chromatin interactions in a specific area but unsuitable for genome-wide interactions.

In Hi-C, the library is created by identifying the biotin containing fragments, which indicates they are ligation products, with Streptavidin bead (also known as all-vs-all). The library is then analyzed by using high-throughput DNA sequencing methods, producing pairs of interacting fragments (**Figure 14)**.

The output of the Hi-C method is a genome contact matrix M. The genome is divided into bins, where bin size varies from 1MB to 1KB depending on sequencing depth. The entry M_{ij} is the number of interacting fragments between bin i and bin j. This matrix reflects the probability of two loci to interact, as the measurement is based on a pool of many cells in different stages of cellular life. In the last eight years, significant efforts have been made to obtain Hi-C maps at ever increasing resolutions [3], [4], [7].

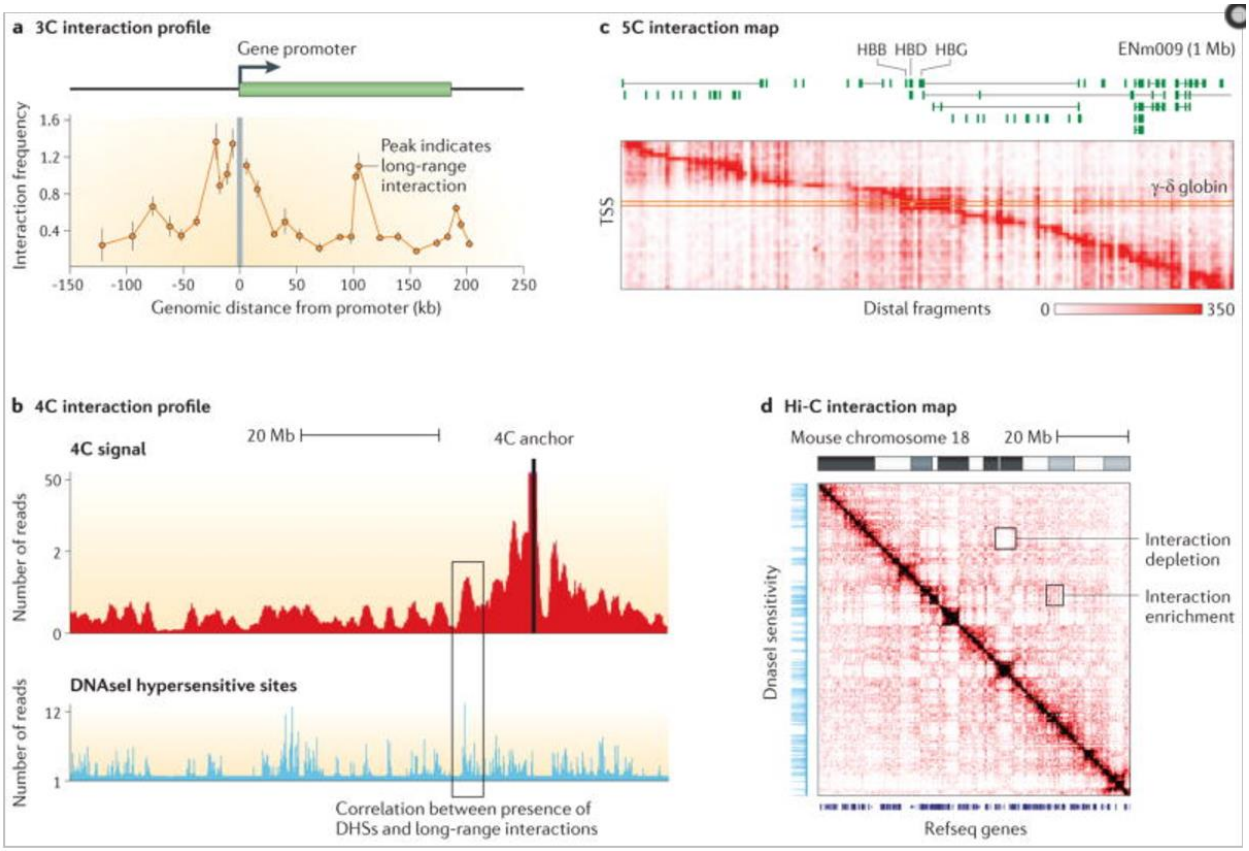


Figure 14 *Visualizing 3C-based results. 3C and 4C compare one locus to other regions, so they are usually presented as a one dimension graph. 5C and Hi-C are presented as two-dimensional heatmaps* [33]

A main drawback of 3C-derived methods is the fact that they report on the frequency of interaction between two loci in the cell population, but they do not distinguish functional from non-functional associations, nor do they reveal the mechanisms that led to the co-localization. Functional associations are specific contacts between two regions, mediated by proteins that bind them, such as enhancer-promoter interactions. Non-functional interactions can result from indirect co-localization of two regions at the same dense packed area, such as nuclear lamina (in the case of heterochromatin) and transcription factories (in the case of euchromatin). In addition, they can be a side effect of specific long-range interactions involving nearby fragments or other chromatin constraints. Finally, the proportion between the length of the long fiber-like chromosomes and the nucleus size makes random collisions an abundant phenomena. When analyzing Hi-C data, various methods are used to try and extract functional interactions (these methods will be further discussed in the following sections).

A summary of the type of results of the different 3C methods is shown in **Figure 15**. Hi-C data reveal 3-D structures across the genome and significant long-range interactions. In this thesis, we analyzed correlation of this structure and interactions with gene expression.

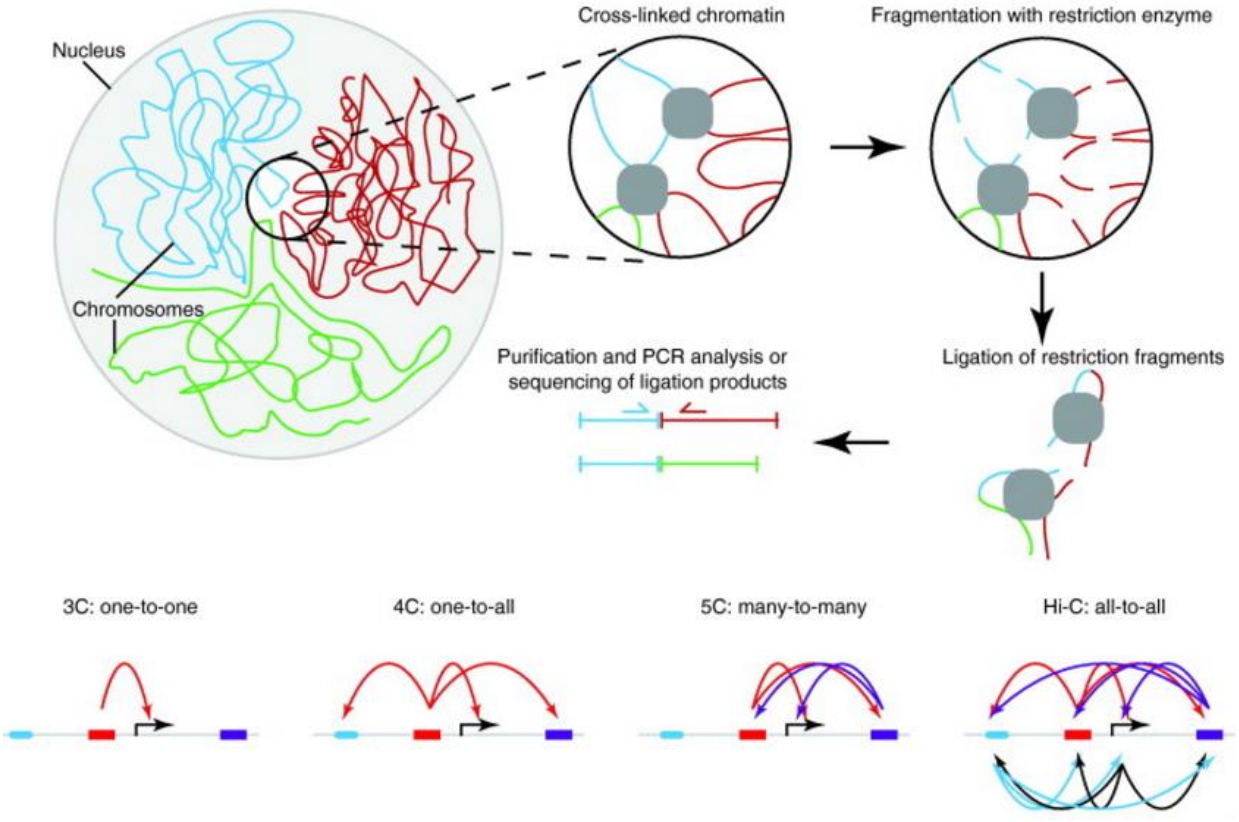


Figure 15 *– Chromosome conformation capture summary. 3C finds interaction degree of one locus against another locus, 4C of one locus against all regions, 5C captures all interactions in a limited region, and Hi-C measures genome-wide associations.* [35]

2.2.2. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)

ChIA-PET is a technique that incorporates ChIP-based enrichment (as described above), chromatin proximity ligation, Paired-End Tags, and high-throughput sequencing to determine genome-wide long-range chromatin interactions related to a protein of interest.

ChIA-PET is used to identify gene regulation made by interactions between the promoters (or gene coding regions) and regions far from them, such as transcription factors binding sites and enhancers. The technique can identify functional chromatin interactions that involve the target protein, suggesting it mediates the interaction. The combination of ChIP-Seq and 3C helps identify specific interactions, removing most of the background noise other 3C based methods suffer from. Instead of mapping all long range chromatin interactions, like in Hi-C, DNA-protein complexes are enriched for the protein of interest. Next, DNA fragments attached to the protein are ligated to form a chimeric ligation product. These fragments are identified, sequenced and aligned (**Figure 16**). The results are pairs of DNA fragments with a genomic span bigger than 3KB (less than 3KB is considered a result of self-ligation). When ChIA-PET is done with immunoprecipitation (IP) of a certain TF, chromatin interactions mediated by that TF are detected. To obtain a more global view of chromatin interaction the cell, RNA polymerase II can be used as the protein.

In our study, we used ChIA-PET data for RNA POL 2 along with Hi-C data to identify enhancer-promoter interactions in basal conditions and examined its correlation with genes response to treatment.

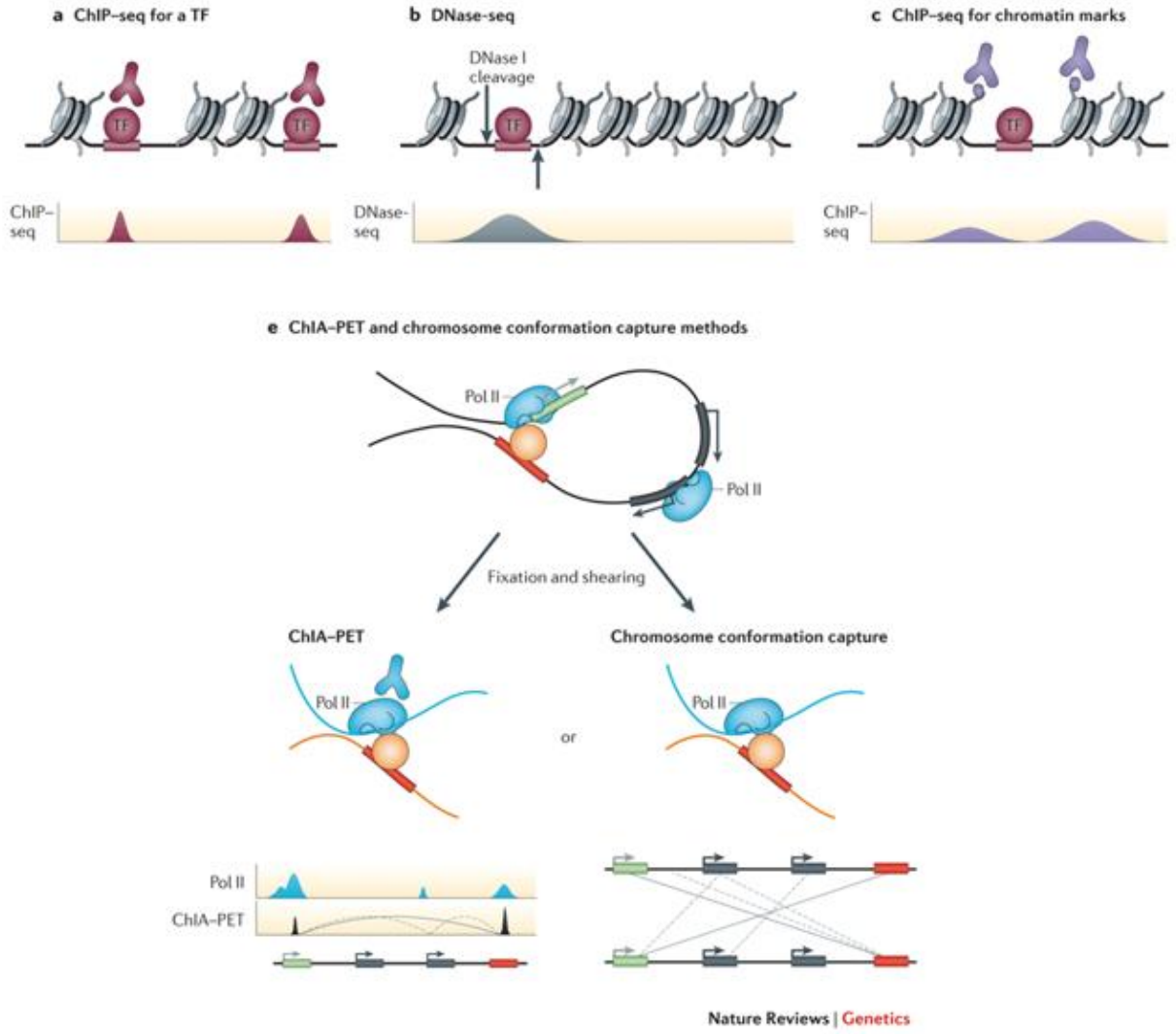


Figure 16 *ChIA-PET overview. (a) ChIP-Seq for the protein of interest (in our case RNA POL 2) is the first step. (b) DNase I is used to remove positions not occupied by TFs. (c) Marked antibodies against the protein of interest are used to pull down DNA-protein complexes (e) ChIA-PET and chromosome conformation capture based methods can be both used to identify chromatin interactions mediated by a protein of interest. All 3C-based methods (described above) can be used, depending on the purpose of the experiment.* [15]

2.3. Relationship between 3D organization and gene expression

Previous studies have shown a relationship between the 3D organization of the chromatin and gene expression.

2.3.1. Defining 3D structures from Hi-C data

Methods, including 5C and Hi-C, that map all interactions in a genomic region of interest or in complete genomes in an unbiased fashion, can be analyzed in various ways to identify structural features of chromosomes. One of the first Hi-C experiments, conducted by Lieberman-Aiden et al. [1], generated 1MB resolution contact matrix of the human genome. These maps, even at low resolution compared to more recent experiments, showed computationally that the nucleus segregates into two compartments denoted A and B:

Normalized intrachromosomal contact matrices present a plaid pattern (**Figure 19)**. This pattern suggests that the chromosome is segmented into sub-chromosomal regions, where each segment belongs to one of two distinct compartments (A/B), such that most intrachromosomal chromatin interactions occur within compartments. We can consider each bin in the contact matrix as both an observation (row) and a feature (column) and try to find feature combination that divides the observations set into two compartments, by using PCA (see details in the following sections).

Lieberman-Aiden et al., showed that the first principal component (PC1) corresponded to the plaid pattern, giving one compartment a positive sign and negative to the other. In a small fraction of the cases, PC1 divided the chromosome to the two chromosome arms. In this cases, PC2 corresponded to A/B partition. The compartments demonstrated the previously known features of the two chromatin states, euchromatin (arbitrarily labeled A) and heterochromatin (arbitrarily labeled B), such as gene density and histone markers (**Figure 19**) [1].

Comparing A and B to known eu/hetero features

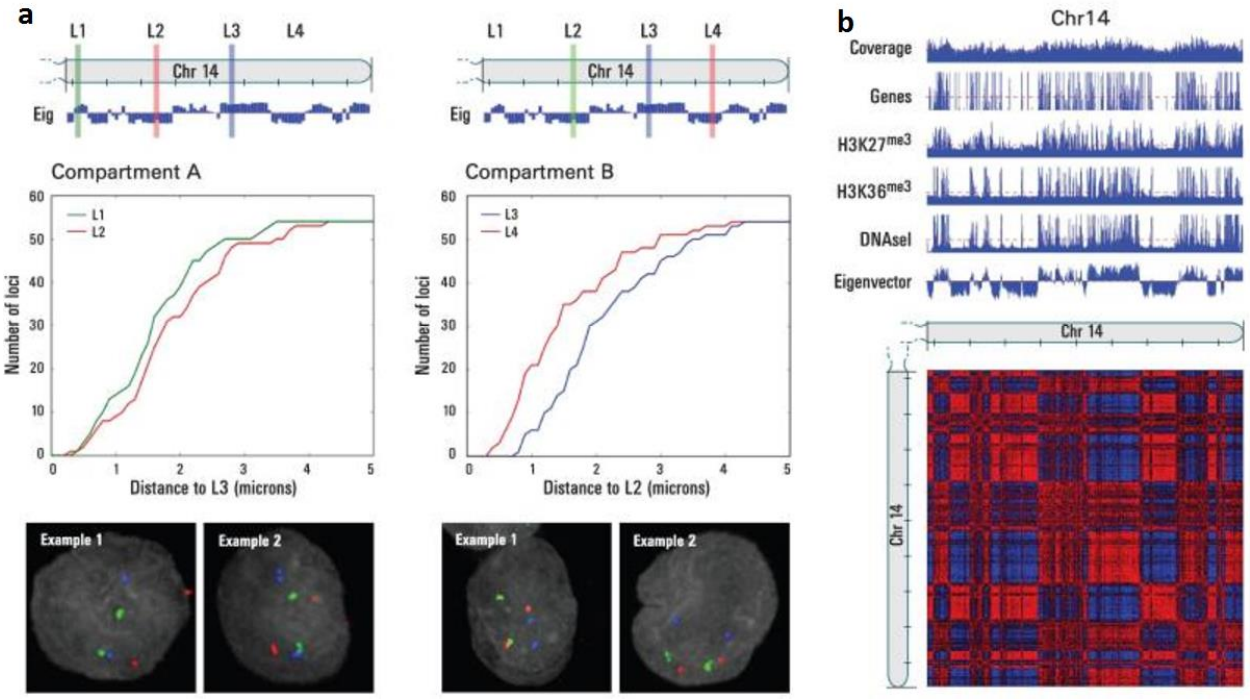


***Figure 19**. PCA and cellular compartments. (a) Compartments cluster together in the nucleus - PCA on chromosome 14 assigned markers L1, L3 to A compartment and L2, L4 to B compartment. FISH assay shows that L1 and L3 (green and blue on left examples) are closer to each other than to L2 (red). Same holds for L2 and L4 (green and red on right examples) compared to L3 (blue). (b) A/B compartments defined by the value of PC1 correlate with eu/hetero features such as (top down) – gene density, histone markers and chromatin density measured by DNAse|.* [1]

Further analysis of chromatin interactions revealed sub-compartments called topological associated domain (TAD)[2]. Most of the intrachromosomal interactions take place within TADs, while inter-TAD interactions are rare [2], [3]. This feature enabled researchers to identify TADs throughout mammalian chromatin by analyzing genome-wide Hi-C contact matrices using Hidden Markov Model approach.

To identify systematically all such topological domains in the genome, the following analysis was made[2]:

1. A directionality Index (DI) was defined for each genomic region as a function of the ratio between upstream interactions and downstream interactions involving that region. High DI regions have more upstream interactions, while low DI regions have more downstream interactions. Hence, at TAD boundary a change in the DI value from negative to positive is expected (**Figure 20**).
2. Hidden Markov model (HMM) was applied using the DI of adjacent regions, with hidden states used to find TADs boundaries. Boundaries between TADs were defined in loci where a sudden inverse in interaction bias was identified (**Figure 21**). The domains defined by HMM were reproducible between replicates, showing that the results are robust.

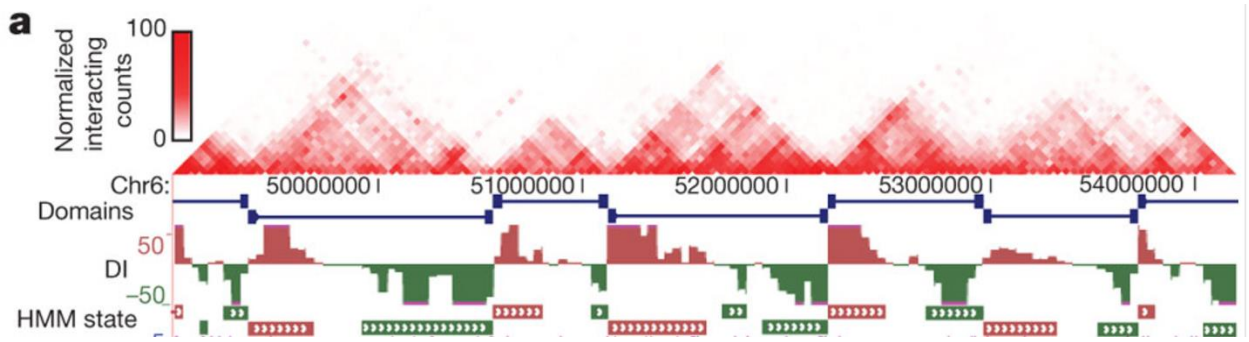


***Figure 20** Identifying TAD borders using Hi-C data and HMM – Normalised Hi-C data is presented above (a short region in chromosome 6). Under it domains are marked with blue lines, square marks identify TADs borders. Directionality Index (DI) and HMM state are demonstrated, green stands for upstream interaction bias, red for downstream interactions bias.*[2]

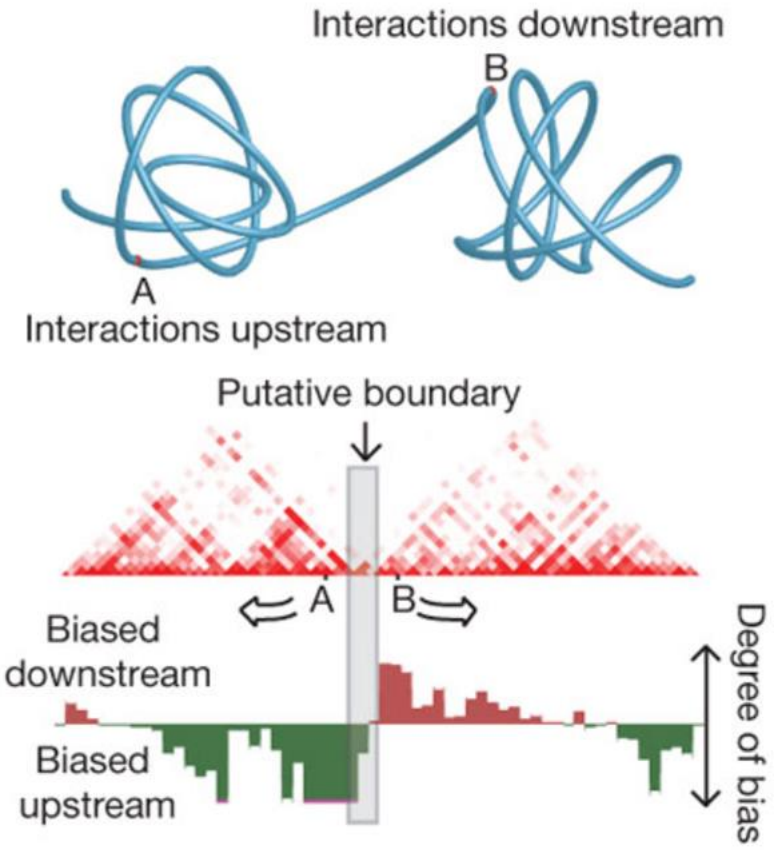


Figure 21 *Illustration of TADs revealed by directionality index calculated on Hi-C data [2]*

Different cell lines show similar but not an identical division into compartments. In contrast, TADs are much more conserved across cell lines. Broad analysis of Hi-C data demonstrates this difference (**Figure 22**)

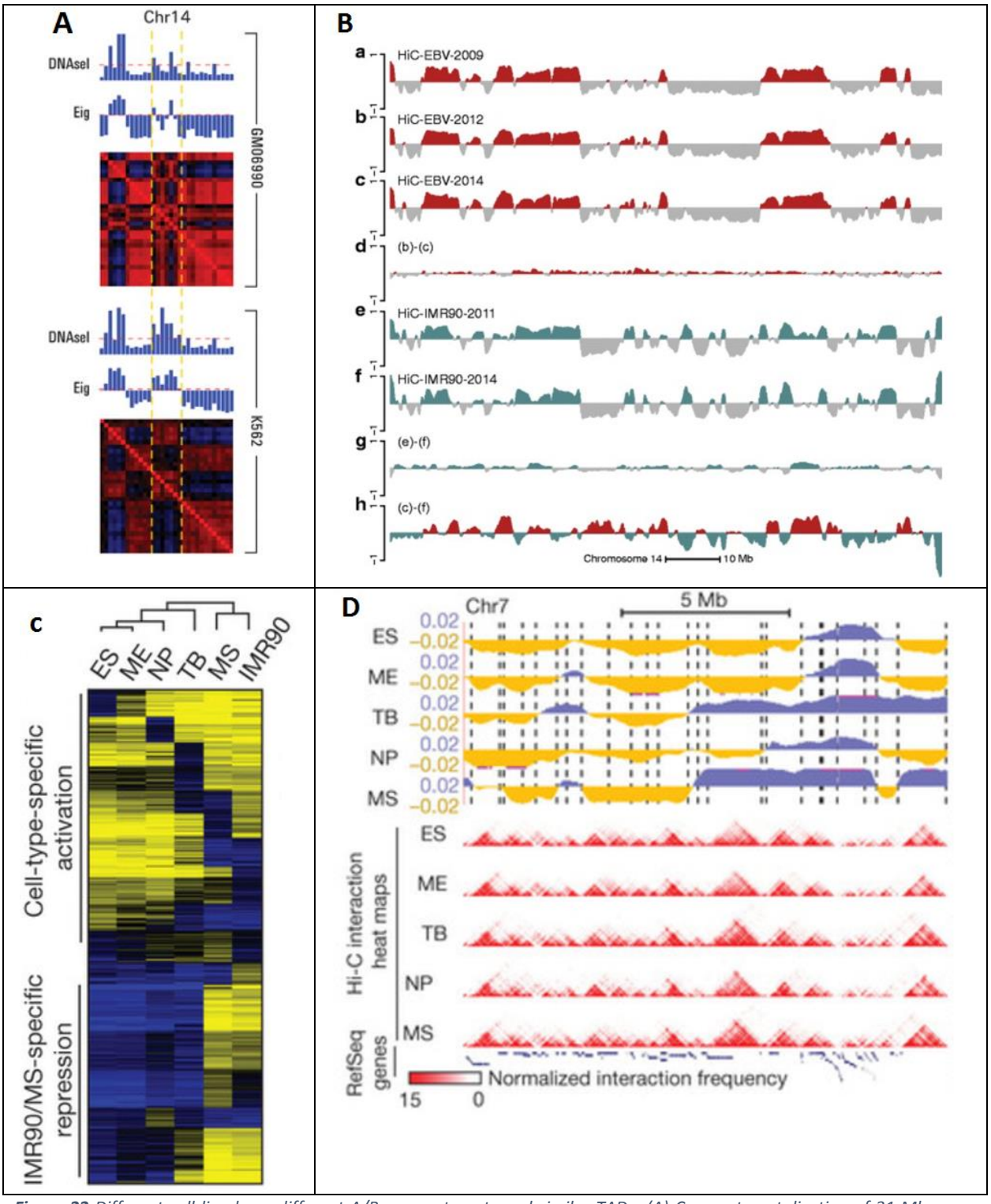


Figure 22 *Different cell line have different A/B compartments and similar TADs. (A) Compartmentalization of 31-Mb segment from chromosome 14 in cell lines GM06690 and K562, based on 1MB Hi-C data. For each cell line Hi-C contact heatmap, PC1 values and DNAse1 result is shown. The indicated region (yellow dashes) demonstrated the difference between an alternating area in GM06690, corresponding to a stable area in K562. Compartmentalization corresponds to open/closed areas measured by DNAse1* [1]. ***(B)*** *Analyses of different Hi-C data for the same cell line, (a-c) for EBV and (e-f) for IMR90 gives the same compartmentalization output as seen by their difference, (d) for EBV and (g) for IMR90. Comparing different Cell lines gives much more significant differences in compartmentalization, implying that A/B partition is cell type specific (h)* [36]. ***(C)*** *Another*

example for the difference between A/B compartments in different cell lines. K-means clustering (K=20) of PC1 values for Hi-C data binned to 40KB resolution. [4] ***(D)*** *TADs are much more preserved across cell lines than A/B compartments. Compartment shifting between cell lines mostly occurs within TAD boundaries, suggesting that TADs are regulatory units.* [4]

2.3.2. Correlation of 3D structures to transcription levels

In a prominent study, Dixon et al. [4] induced the differentiation of embryonic stem cells (ESCs) into different cell types and examined dynamic changes in genome organization and gene expression that occurred during the differentiation process. Their analysis showed that genes that changed their compartment status (A in ES and B in the differentiated cell or vice versa) had different distribution of fold-change of expression: genes whose compartment was changed from 'B to A' were generally up-regulated during differentiation compared to genes that moved from 'A to B' that were generally down-regulated (**Figure 23**).

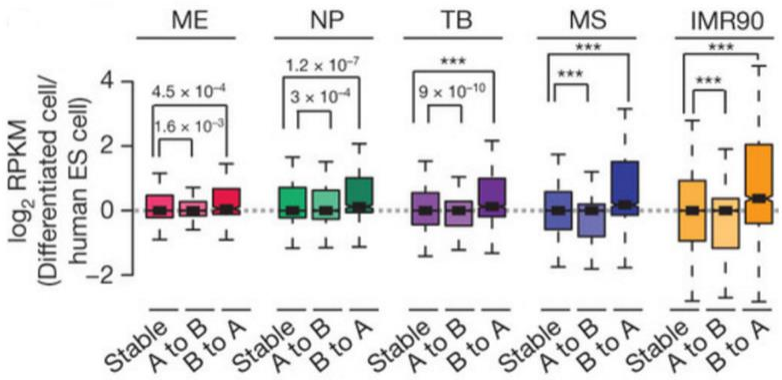


Figure 23 *Compartment - expression correlation. Genes are divided into three groups according to compartmet status in hESC and in the titled cell. The distribution of fold-change in gene expression, measured by RNA-Seq, is presented by boxplots. Transition from B to A resulted in upregulation of gene expression while transition from B to A resultd in downregulation in gene expression. Genes that remain the same ("Stable") seem to distribute around mean FC = 1. ((***)P < 2.2 × 10−16, P values calculated by Wilcoxon test; whiskers correspond to interquartile range). [4]*

This study also stated that these transitions only affect a subset of genes, enriched for lineage-specific genes. It implies that compartments switching underlies a mechanism for cell-type specific gene expression profile.

TAD role in gene expression has been investigated in many directions. In mammals, TADs borders are enriched with Transcriptional Start Sites (TSS). The epigenetic features of TADs can help understand some aspects of their role in gene regulation. Boundaries of TADs are enriched with CTCF, the insulator TF, along with activating TFs and histone markers of highly expressed areas (H3K4me3 and H3K36me3). Intra-TAD areas are more enriched with gene repressing markers (**Figure 24**).

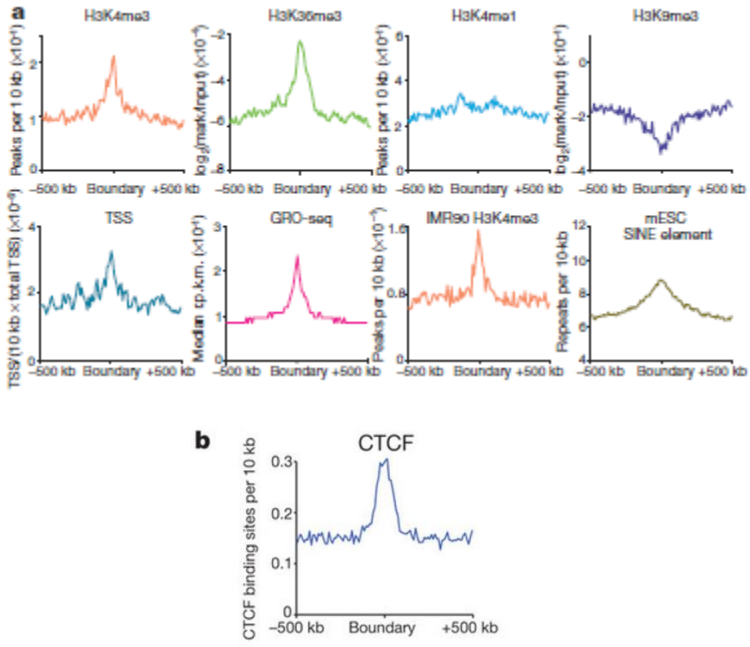


Figure 24 *(a) Active histone markers, TSS, GRO-Seq and SINE elements are enriched in boundary regions in mouse ES cells or IMR90 cell while closed histone markers are enriched in the center or spread uniformly. (b) Enrichment of CTCF at boundary regions. .[2]*

It was also observed that genes with high expression tend to be located at the boundaries of TADs rather than at the middle (**Figure 25**). Combined with the findings that TADs borders are mainly demarcated by long-range looping interactions anchored with CTCF transcription factor, these loops are suggested to be enhancer – promoter contacts, concentrated at the boundaries of TADs [3].

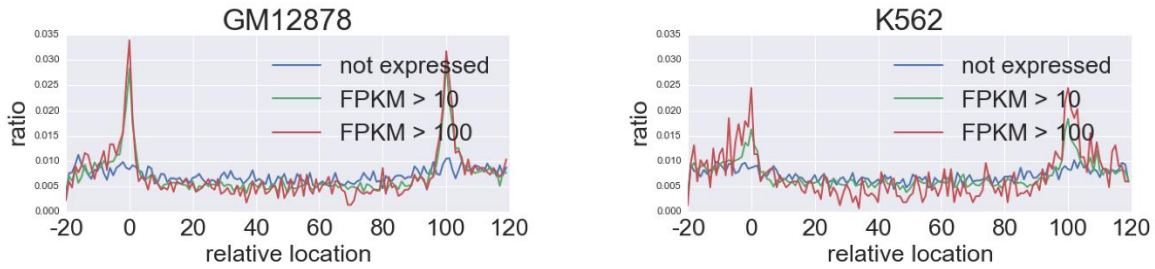


Figure 25 *Gene expression and location within TADs. In order to unify data from TADs of different sizes, we calculated the relative position for each gene in the TAD it is located in where TADs are normalized to (0,100). Genes were divided according to their FPKM, and relative location plots are presented for each group. These plots show that expressed genes tend to cluster at the boundaries of TADs while genes that are not expressed equally distributed (p<0.0001).*

In addition, when a TAD changes its compartment state from B to A, it has more intra-TAD interactions and genes within it are generally up-regulated, while a compartment's change from B to A correlates with a decrease in intra-TAD interactions and its genes are generally down-regulated [22].

2.3.3. 3D structures and response to stress

How the high-level 3D organization of the genome is changed in response to stress or treatment is largely unknown. One study applied Hi-C to T47D cells before and after progestin and estrogen treatment [22]. This study detected changes in intra-TAD interactions, maintaining TADs borders, akin to intra-TAD changes that were recorded when cells differentiate from ES. Furthermore, it was noted that genes within the same TAD tend to respond in the same direction to stress. This observation suggests that TADs form regulatory units. The study suggested that chromatin remodeling as a response to stress is an abundant and significant mechanism. **Figure 26** shows the strong correlation found between hormone-induced changes in intra-TAD interactions and response of gene expression in T47D breast cancer cell line treated with progesterone analog.

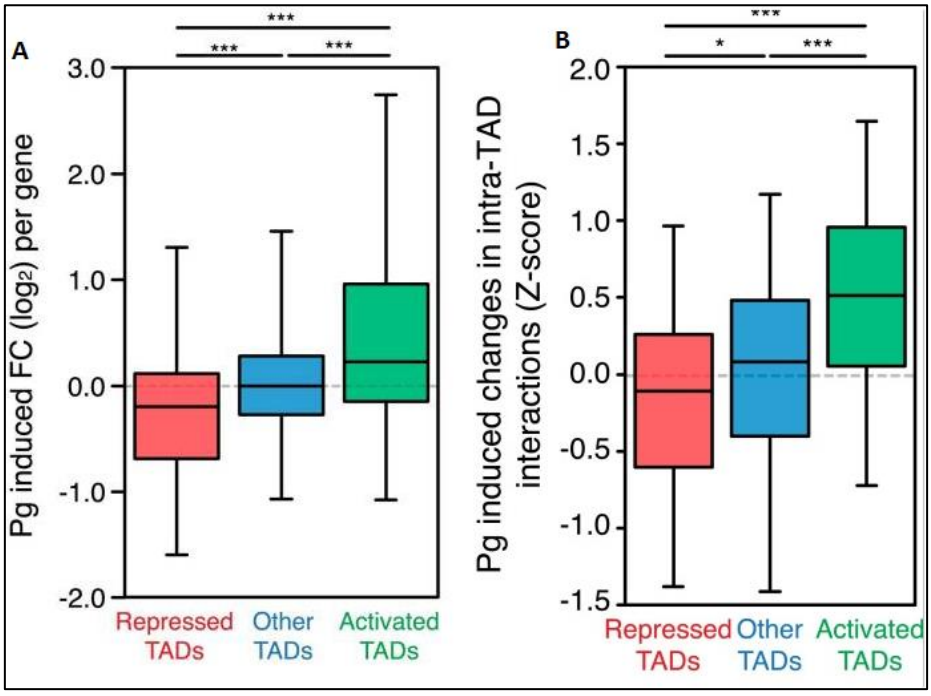


Figure 26 *Expression and interaction within TADs. (A) For each TAD, the average fold change (FC) of its genes expression in response to 6h Pg treatment is calculated. Based on this score, the top 100 TADs are defined as Activated TADs, the bottom 100 TADs are defined as repressed. The boxplots show the difference in fold change distribution between TADs groups.* ***(B)*** *Changes in expression correlate with changes in intra-TAD chromatin organization. The boxplots show the Z-score distribution of the number of changes in intra-TADs interactions for the groups defined in (A). (***) P < 0.001; (**) P < 0.01; (*) P < 0.05 (Bonferroni-corrected Mann-Whitney test)* [22]

However, another seminal study applied Hi-C to examine 3D changes in IMR90 cells after treatment with TNF-α. Importantly, this study observed that enhancer-promoter interactions of induced genes were already in place in the untreated cell line [7].. More specific 3C experiments indeed confirmed that genes that were induced by TNF-α had already enhancer-promoter contacts in the untreated IMR90 samples (**Figure 27**). Changes in looping frequency of induced genes upon treatment were minor, unlike changes in looping frequency of ES cell-specific genes when comparing ES and IMR90.

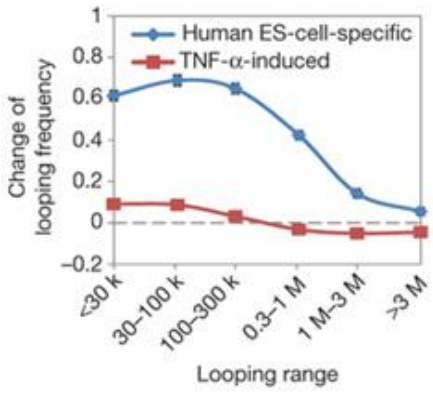


Figure 27 *Topological changes associated with enhancer activation. The blue plot shows the significant topological difference associated with enhancers of hESC-specific genes compared to untreated IMR90 cell line. The red plot shows the topological difference associated with enhancers of induced genes in IMR90 under TNF-a treatment, which are relatively mild. [7]*

These two studies illuminate different mechanisms that play a role in cell-type specific response to treatment or stress: (1) A/B compartmentalization and (2) intra-TAD E-P interactions that are present in the cell also in basal (unstimulated) condition. A key open question is the relative importance of these two mechanisms: 1) TADs conserved preexisting structure and boundaries, which allow intra-TADs modifications to take place only within them, and therefore regulate sets of genes in the same TAD in the same direction, and 2) cell specific preexisting enhancer-promoter interactions across TADs that lead to cell specific response.

2.4. Computational background

In this chapter, we lay out the computational background of this thesis. Each section deals with a different type of computational problem. More details on the computational problems addressed are given in the references in each section.

2.4.1. Hi-C contact matrix normalization

The results of Hi-C data are pairs of reads, aligned to the genome. The genome is partitioned to bins of some size (between 1KB to 1MB), and a contact frequency matrix is computed as described above. Like all biological experimental methods, Hi-C has systematic biases that we want to take into consideration when calculating the contact probability matrix[37].

The main experimental biases are distance related. First, proximal bins along the DNA are overrepresented due to incomplete digestion and to a fragment re-ligated to itself

Additional biases include (**Figure 17):**

1. Non-specific fragments – DNA tends to randomly shear non-specifically regardless of restriction enzyme.
2. Fragment lengths – different fragment lengths affect cross-linking and ligation efficiency.
3. GC content – high/low GC content changes DNA molecule strength and condensation, affecting ligation product processing.
4. Mappability – repetitive regions in the genome are harder to map to, which affects data reliability for certain fragments. For example, repetitive or non-sequenced areas (i.e., centromeres and telomeres) will not have reliable data, while unique sequences will.

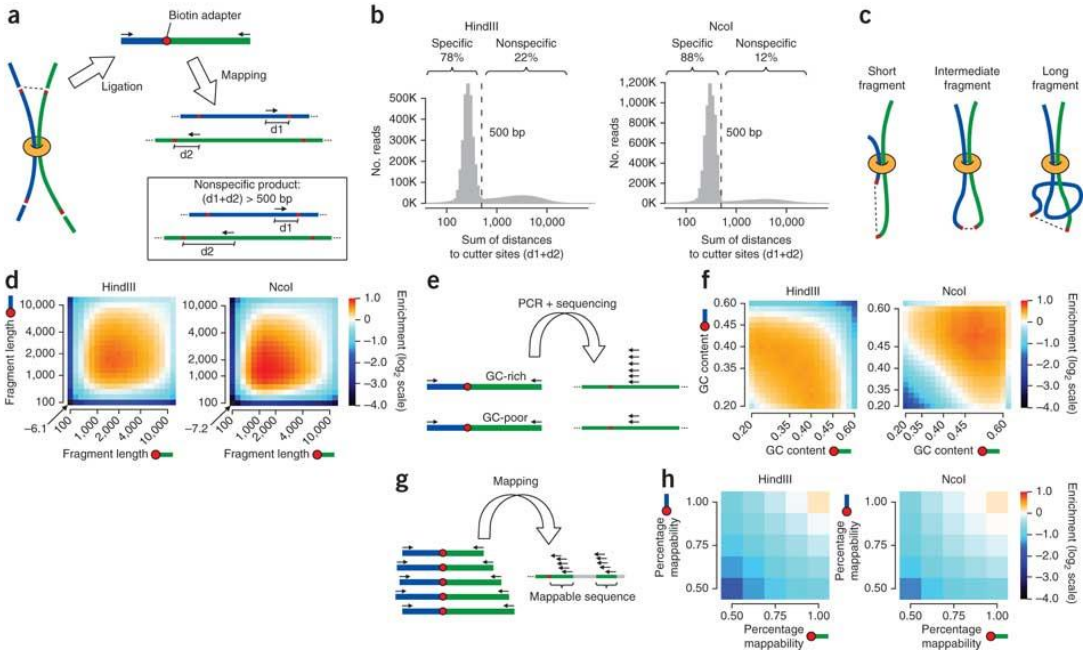


Figure 28 *Systematic biases in Hi-C as described by Yaffe et al.* [37] *(c,d) Fragment lengths are determined by the restriction sites located on the genome. (e,f) Local GC content varies between chimeric products of the ligation step (g,h) Mappability, as measured by the uniqueness of the sequence, is determined by the read length and whether it is located in a repeat locus. The biases might affect the results processing.*

There are different methods for normalizing the matrix according to known biases. These are the ones used on the data analyzed in our work:

1. Matrix balancing methods such as Knight and Ruiz [38]. This method was used in order to normalize most of the contact matrices we used [3].
2. HiCNorm – a method which uses a Poisson regression in order to remove the biases [39].
3. Naïve approach that calculates the expected probability for two loci to interact, taking into consideration sequence depth and distance [40]
4. Calculating the likelihood of any two fragments using a model that takes into consideration the biases above, using numerical optimization[37].
5. LOWESS normalization with different parameters[41].

In order to account for distance related biases, normalizing methods calculate an expected value for each diagonal (i.e., all pairs with same 1D distance) separately.

2.4.2. Principal Component Analysis (PCA)

PCA is a useful statistical technique for finding patterns in data of high dimension, which has found application in many fields. The main purpose of the method is extracting high variance features from the data, in order to classify observations with a simple, low dimensional vector.

The input of PCA is a matrix $G_{m\times n}$, where m is the number of observations and n is the number of variables. The output is a linear transformation that transforms the data to a new coordinate system. The first coordinate, named first principal component (PC1) or first feature, has the greatest possible variance out of all the linear combinations over the variables, the second coordinate has the second greatest and so on. The number of features can be up to the number of variables, but in vast the majority of the cases, low variance components are discarded, leading to lower dimension representation of the data.

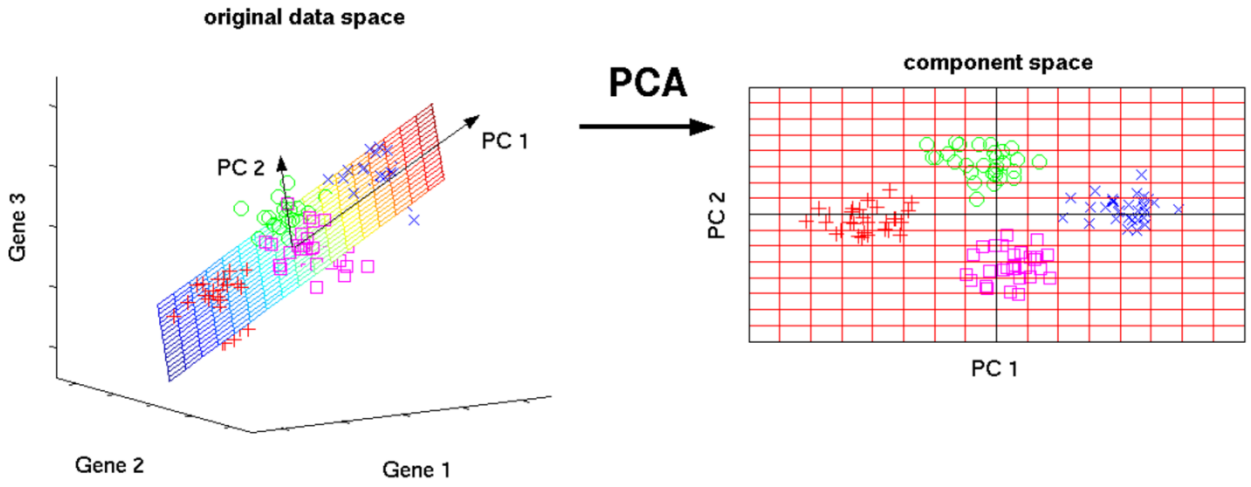


Figure 29 *PCA illustration.* [42]

Finding the feature with the highest variance also minimizes projection error from the new coordinate represented by the feature value and the real value, since the sum of the variance and the projection

error is constant and equals to the squared distance between the origin of the axis and the original coordinate, by Pythagoras theorem.

We can also define the problem with the following formula:

Assume that G, which is centered G (i.e., the mean of each observation is 0). The projection of a vector v is given by Gv. The variance of the projection is $\frac{1}{n-1}(Gv)^T \cdot Gv = v^T\left(\frac{1}{n-1}G^TG\right)v = v^TCv$, where C is the covariance matrix of G. So, we want to find v that maximizes v^TCv.

According to *spectral theorem*- since C is symmetric, it can be diagonalized by its eigenvector basis denoted by $\{z_i\}$. We can represent each vector v by a linear combination of the eigenvector basis vectors:

$$v = \sum w_i z_i$$

and calculate its variance as $\sum \lambda_i w_i^2$.

Given $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, assigning $w_1 = 1$, $w_i = 0\ for\ i > 1$, resulting in $v = z_1$, will give us the feature with maximum variance.

Lieberman-Aiden et al. (2009) performed PCA on the Hi-C matrix and used PC1 to compute the A and B compartments division described before.

In summary, PCA can be computed as follows:

1. Center each variable in $G_{m\times n}$ so it mean in each observation will equal 0.
2. Calculate the symmetric covariance matrix $C_{n\times n} - C_{i,j}$ is the covariance between variable i and variable j.
3. Compute the matrix of eigenvectors which diagonalizes the covariance matrix C:

$$V^{-1}CV = D$$

4. The D diagonal contains the eigenvalues of the corresponding eigenvectors in V. When sorting the eigenvalues in decreasing order, and sorting the eigenvectors accordingly, we get $PC1..PCn$.
5. The product of observation matrix G and the vector PC_i $(1 \leq i \leq n)$ is a linear combination on each variable, represented by a row in the matrix, defining a new property called feature i that has the i^{th} highest variance out of all linear combinations on the matrix rows.

2.4.3. Data Analysis

We used a variety of data samples in our work, in order to find genomic features that are common to many different cell types under many different treatments. We used several statistical methods in order to determine the significance of our findings, described in the following section.

2.4.3.1. Chi-square two sampled test

Chi square two sample test is used to check if two data samples, usually binned data, come from the same distribution, especially when the common distribution in unknown. The null hypothesis is that the two samples come from a common distribution.

Since we used it for comparing data divided into 2 bins, we calculated the statistic in the following way:

Given the following table-

	Bin 1	Bin 2
Sample 1	O_{11}	O_{12}
Sample 2	O_{21}	O_{22}

For each cell O_{ij}, E_{ij} is the product of the sum of the row i and column j, divided by the sum of the table. For example:

$$E_{11} = \frac{(O_{11} + O_{12}) * (O_{11} + O_{21})}{O_{11} + O_{12} + O_{21} + O_{22}}$$

The statistic is evaluated by one sample chi square test, with the calculated expectation:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

P-value is derived from Chi square table with DF = 1.

2.4.3.2. Nonparametric statistical tests

Nonparametric statistics tests are used when we do not want to rely on any assumptions regarding the probability distribution of our data. We encountered such situation when using gene expression and chromatin interaction data.

2.4.3.2.1 Wilcoxon test

We used Wilcoxon test when checking whether two sets of genes, G_1, G_2 , have the same mean gene expression. Let us denote the distributions of gene expressions for the sets as D_1, D_2. The null hypothesis is that when sampling $d_1 \sim D_1, d_2 \sim D_2 \rightarrow$ the probability of $d1 > d2$ equals the probability that $d_2 > d_1$ ($P(d_1 > d_2) = P(d_2 > d_1)$). The null hypothesis will be rejected when the mean of the two distributions is significantly different.

The test involves the calculation of a statistic usually called U, whose distribution under the null hypothesis is known. U is calculated in the following way:

Join both observation sets to one set. Sort this set and give each observation its ranking in the group (starting with 1 for the minimal value). For each set we get:

$$U_i = R_i - \frac{|G_i|(|G_i| + 1)}{2}$$

R_i is the sum of ranks of observations in G_i. We can choose U_1 or U_2 since they completely derived from each other. In order to calculate the significance of U, we assume that for a large set (>20), the distribution of U under the null hypothesis approximates a normal distribution.

2.4.3.2.2. Permutation test

We used a permutation test when checking whether a group of genes G, has a significant mean and median number of chromatin interactions at their transcription start site (TSS). We performed this test both for ChIA-PET data and significant chromatin interactions from Hi-C data.

Given a set of genes G, we performed the following:

1. Calculate the mean and median, G_{mean}, G_{median} of the number of chromatin interactions at the genes TSS.
2. Generate n random sets of $|G|$ genes (we used $n = 10000$). We maintained that the number of genes taken from each chromosome will be the same as in G.
3. The empirical P-value for G_{mean} is the fraction of sets with higher or equal mean than G.
4. The empirical P-value for G_{median} is the fraction of sets with higher or equal median than G.

3. Results

Our analyses were based on public datasets of two high-throughput methods for profiling the 3D organization of the genome:

1. Hi-C, a method based on chromosome conformation capture, measuring the frequency of interactions between DNA segments in the sampled cells. The frequency, as explained in the background section, represents the probability of interactions between any two chromatin segments. We mainly used intrachromosomal interactions in order to divide each chromosome to different compartments as described below. We used Hi-C data for 13 cell lines with bin size 40Kbp and 100Kbp. **Supplementary Table 1** reports the source study and resolution for each.

2. ChIA-PET, a method that measures dynamic interactions between DNA segments mediated by a specific protein. We mainly used ChIA-PET data for RNA Pol2 in order to get enhancer-promoter interactions. We used ChIA-PET data from the ENCODE project for three cell lines, with mean chromatin segment size 3Kbp. **Supplementary Table 2** reports the source study and mean resolution for each.

3.1 Chromosome compartmentalization

First, we defined A/B compartments for 13 human cell lines for which Hi-C data are available (**Supplementary Table 1**). We normalized each Hi-C matrix (as described in the background section) and performed principal component analysis (PCA) for each intrachromosomal matrix separately. The first principal component partitions the chromosome into two compartments, A and B, according to the sign of the elements. As seen in previous studies[1], the A compartment is gene rich and its chromatin is less dense (showing correlation with known areas and markers of euchromatin in the genome), while the B regions are gene poor and their chromatin is denser (correlating with known areas and markers of heterochromatin in the genome). For each chromosome separately, we determine whether positive or negative values of PC1 correspond to A or B based on genes richness – the compartment with higher gene density was labeled as A compartment. Centromeric regions were not included in the A/B partitions since no chromatin interactions are identified by Hi-C in such regions.

Table 1 shows summary statistics for the size and number of genes for the A and B compartments for each cell line. **Figure 30** shows the compartments in chromosome 1 for each of the cell lines. Lines are organized hierarchically using agglomeration single – linkage clustering. Profile similarity was computed using Jaccard coefficient. The mean similarity between profiles was 0.75.

Cell line	A Total size (Mbp)	Genes in A	B Total size (Mbp)	Genes in B
GM12878	1322	15184	1410	3958
K562	1376	15401	1356	3741
HUVEC	1382	15116	1350	4022
HMEC	1317	14593	1415	4543

NHEK	1433	14864	1300	4276
IMR90	1310	13569	1423	5577
T47D	1372	14114	1350	4979
MCF7	1384	15056	1450	4758
MCF10	1386	15090	1451	4772
LNCAP	1433	14112	1299	5021
PC3	1395	13341	1313	5692
KBM7	1301	14506	1431	4631
PrEC	1327	13994	1387	5041

Table 1 Compartment size and genes distribution in compartments in 13 human cell lines for which Hi-C data is available

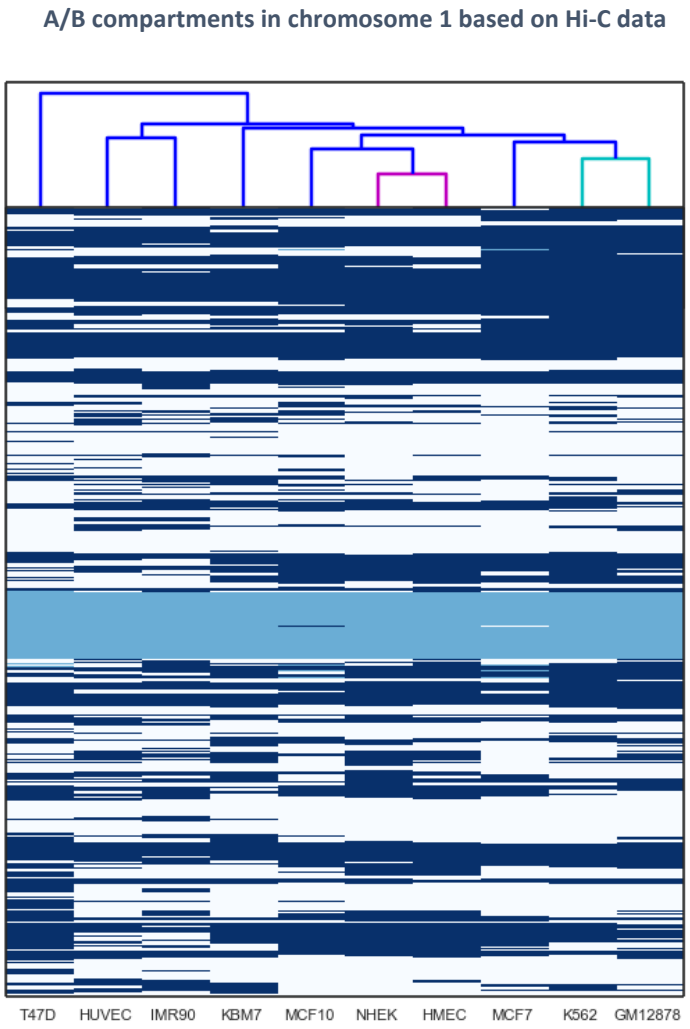


Figure 30 *– A/B partition of chromosome 1 for different cell lines based on Hi-C data in 100KB resolution. Dark blue indicates A compartments and white indicates B compartments. Light blue indicates areas which Hi-C could not measure interactions for, e.g., centromeres. The hierarchy graph represents the similarity between A/B compartments in different cell lines calculated by single-linkage clustering. The colored links indicate the most similar pairs.*

3.1.1. Gene expression correlations with A/B compartmentalization

A compartments are known to generally correlate with higher transcriptional activity. We therefore first performed several sanity checks on our A/B partitions, in order to confirm that they are consistent with this known feature. Using RNA-Seq data from the same cell lines for which Hi-C data were available, we verified that gene expression in A compartment is indeed significantly higher than in B compartment, for each cell line separately (**Figure 31**)

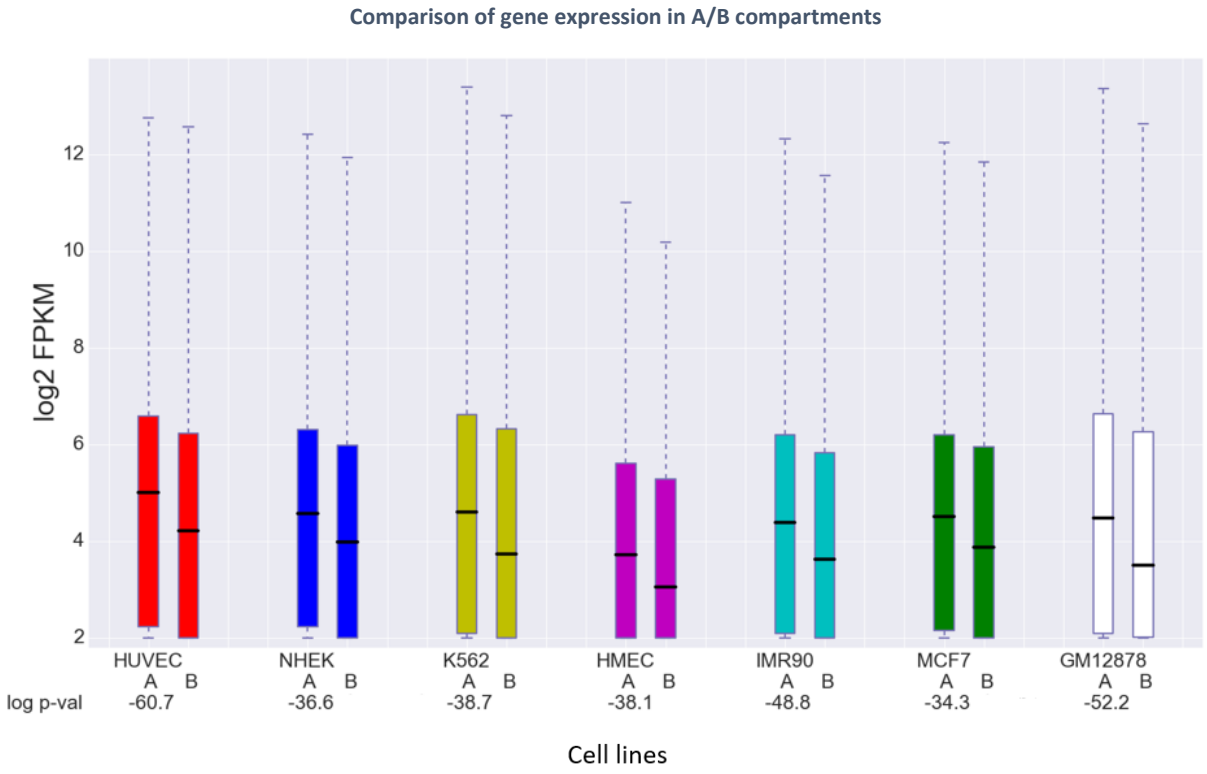


***Figure 31** – Comparison of gene expression in A and B compartments for each cell line. The number below each cell type is the log (base 10) p-value for the significance of the difference between the two means computed using Wilcoxon test. The plots show that genes in A compartment are significantly more expressed than genes in B compartment*

The previous analysis was done on each cell line separately. Next, we examined the correlation between differences in A/B compartmentalization and gene expression across different cell lines. Specifically, for each pair of cell lines, we examined whether genes located in A compartment in one cell line and in B compartment in the other cell line show higher expression in the former. For each pair of cell lines, we divided the genes into four groups – A in both cell lines (AA), B in both cell lines (BB), A in cell line 1 and B in cell line 2 (AB) and B in cell line 1 and A in cell line 2 (BA). Next, we calculated gene-expression ratios between cell line 1 and 2 and compared the distribution of these ratios between the four gene groups. As expected, genes in the group AB tend to be more highly expressed in cell line 1, genes in the group BA tend to be more expressed in cell line 2 and genes in AA or BB have a mean ratio close to 1 (**Figure 32a**). All pair-wise tests gave significant p-value (<0.0001) for comparison between AB and BA gene sets (**Figure 32b**).

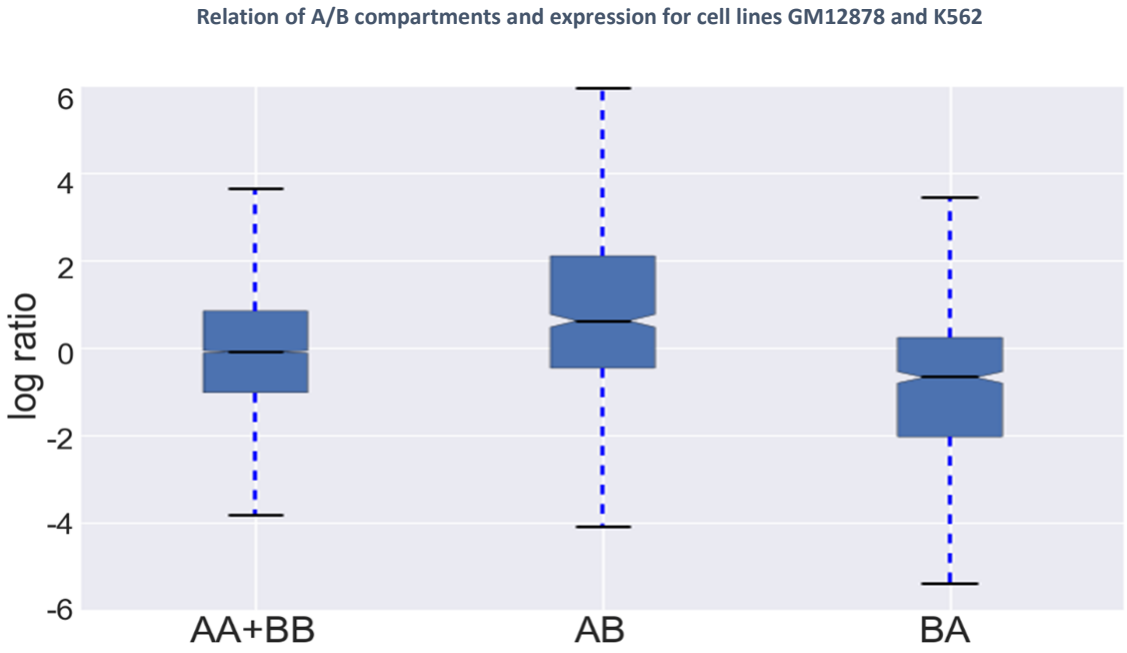


Figure 32a *- Compartmentalization-expression relation between cell lines GM12878 and K562. The genes are divided into four groups.In the figure AA and BB are united. P-value (Wilcoxon) between gene expression ratio distributions in AB and BA is* 10^{-40}

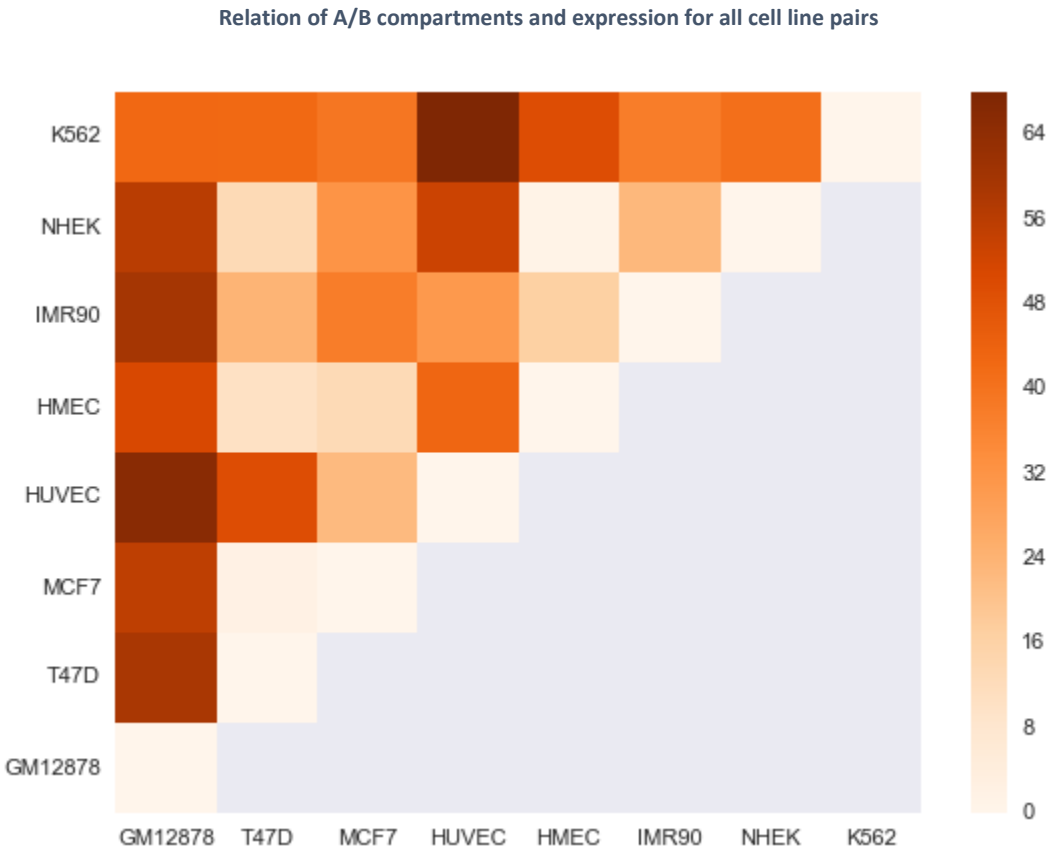


Figure 32b *– Compartmentalization-expression relation between all cell lines. For each pair of cell lines, we calculated the significance of the difference between expression ratio for genes in AB and mean expression ratio for genes in BA using Wilcoxon rank-sum test. For all pairs, the p-values are highly significant ($p < 0.0001$) besides HMEC and NHEK.*

3.1.2. Correlation between A/B compartmentalization and TF binding.

As mentioned before, epigenetic markers such as transcription factors binding sites and histone modifications related to euchromatin are enriched for A compartments, while heterochromatin epigenetic markers are more prevalent in B compartments. We next examined the correlation between A/B compartmentalization and TF binding. In this analysis, we included 122 TFs that had ChIP-Seq data recorded by the ENCODE project for cell lines with Hi-C data (**Supplementary Table 3**).

First, we tested TFBS enrichment for A compartments. For each transcription factor and cell line we computed *density factor*, D, defined as follows: Let the number of observed binding sites in region S be $O(S)$ and number of expected binding sites in region S be $E(S)$:

$$D = \frac{O(A)/O(B)}{E(A)/E(B)}$$

$D > 1$ implies that binding sites are enriched for A compartment and $D < 1$ implies that binding sites are enriched for B compartment. For TF binding sites, $E(A)/E(B)$ is the size ratio between the compartments.

For example, we compared the number of CTCF binding sites (**Table 2**) in A and in B to their total size (**Table 1**) and calculated the p-value using chi-squared test:

Cell line	Observed A	Observed B	total	Expected A	Expected B	D	log 10 p-value
K562	48141	15524	63665	32056	31599	3.06	<-300
HUVEC	35329	12544	47873	24216	23656	2.75	<-300
NHEK	45665	16701	62366	32700	29665	2.48	<-300
HMEC	36469	18135	54604	26322	28281	2.16	<-300
IMR90	29184	14880	44064	21121	22942	2.13	<-300
GM12878	42295	17575	59870	28970	30899	2.57	<-300

Table 2 *CTCF BS in A and B compartments in six cell lines. The columns Expected A/B gives the expected number of binding sites based on the relative sizes of A and B compartments.*

Indeed, CTCF BS are highly enriched for A compartment in all the cell lines we tested (p-value<10^{-300}), even though compartments sizes are almost equal.

Next, we wanted to see if A-B transitions between cell lines are also reflected in TF binding sites. For each pair of cell lines, we segmented the genome into four regions according to A/B assignment in the two cell lines as described above. For a given TF, we divided the TF binding sites into three groups: binding sites common to cell line 1 and 2, binding sites detected only in cell line 1 and binding sites detected only in cell line 2. We then tested for a relationship between the two divisions. Specifically, we tested whether TFBSs that are specific to a cell line tend to occur more often in genomic regions assigned to A compartment in that cell line and to B in the other cell line. We also computed the *occupancy enrichment ratio* R, defined as follows: Let the number of cell line i only BSs in region S be $n(i,S)$. Then

$$R = \frac{n(1,AB) + n(2,BA)}{n(1,BA) + n(2,AB)}$$

(The ratio of the sum of numbers in blue to the sum of numbers in orange in Table 3). **Table 3**, as an example, the results obtained for CTCF binding sites in HMEC and HUVEC cell lines.

HMEC HUVEC	AA	AB	BA	BB	total	R	p-val
HMEC_only_BSs	7241	**1655**	947	4775	14618	2.34	10^{-168}
HUVEC_only_BSs	4516	264	**1180**	1524	7484		
Common_BSs	24986	2587	4647	9750	41970		

Table 3 *CTCF binding sites are divided into 3 groups and each group is divided according to A/B assignment in HUVEC and HMEC. R is the occupancy enrichment ratio (see text). P-value is calculated based only on AB/BA division.*

As expected, we observed that CTCF BSs specific to HMEC (HUVEC) were significantly enriched in AB (BA) genomic regions. R for this comparison was 2.34 (p-value= 10^{-168}; chi-square test between n(1,AB),n(1,BA) and n(2,AB), n(2,BA)).

As CTCF ChIP-Seq data is available for 6 cell lines with Hi-C data, we could carry out a wide comparison for this factor. For all comparisons, we got a significant association ($p<0.001$) between CTCF binding and A (**Figure 33a**).

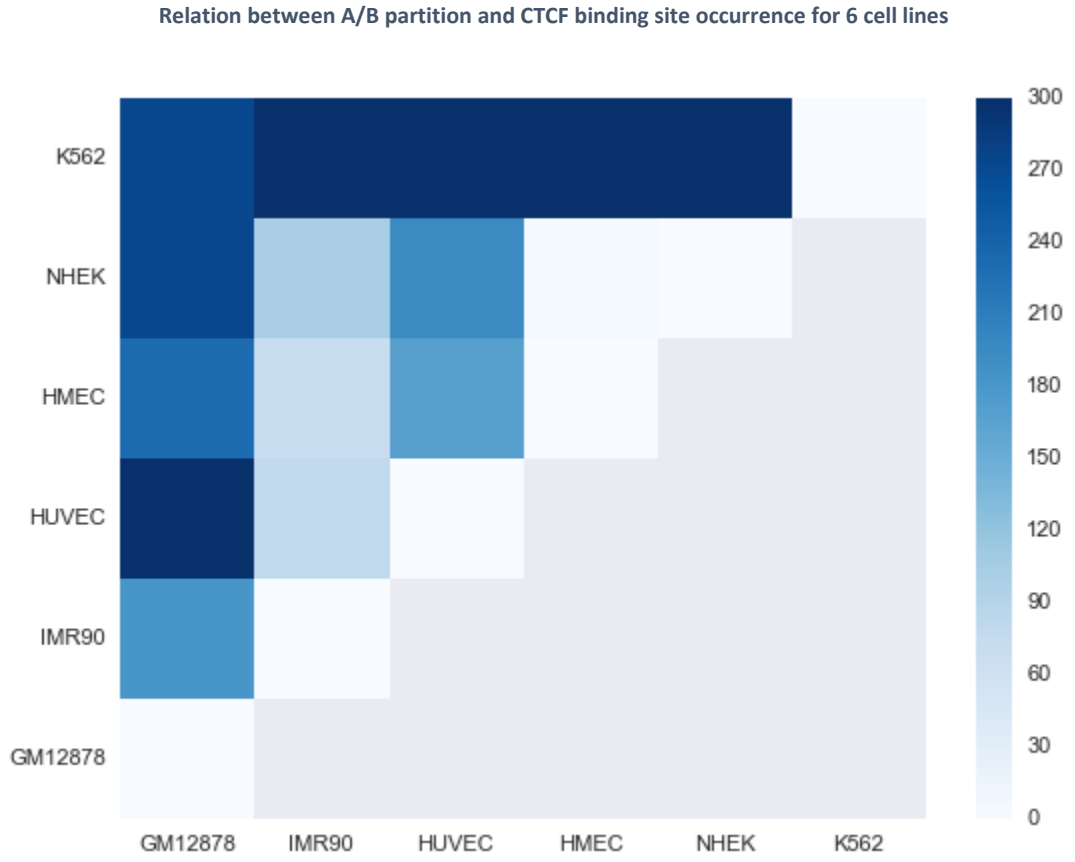


Figure 33a *– Relation between A/B partition and CTCT binding site occurrence for 6 cell lines. Each cell contains the p-value of chi-square test between (AB, BA) count of cell line 1 only binding sites and (AB, BA) count for cell line 2 only binding sites. All p-values are significant (<0.001).*

To study the relation of binding sites and compartments across many TFs, we focused on GM12878 and K562, which have ChIP-Seq data for 49 common TFs. For each TF we calculated the p-value and the occupancy enrichment factor. Strong TFBS-compartment relationship was observed for the vast majority of TFs (**Figure 33b**).

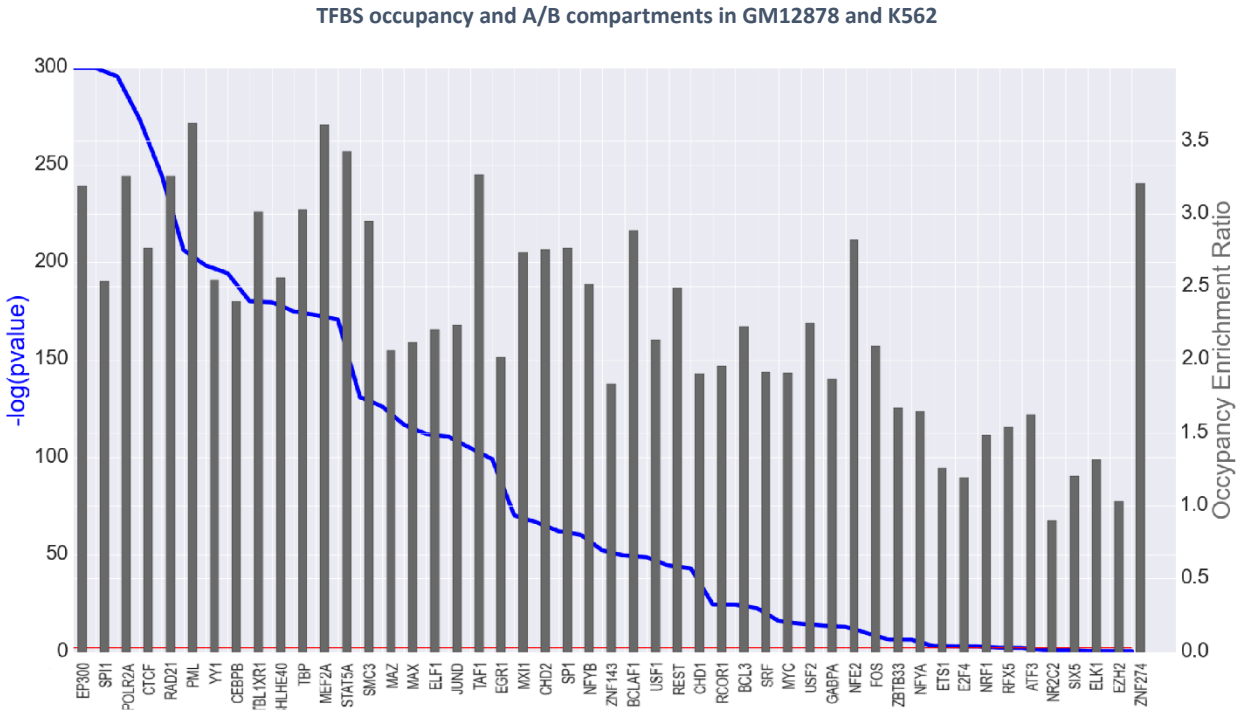


Figure 33b *- TFBS occupancy and A/B compartments in GM12878 and K562. 49 TFs are sorted by p-value calculated by chi-square test as described above (y-axis is -log10(p-value)). TF with p-value above the horizontal line (marked in red) have significant p-value (p<0.05). The occupancy enrichment ratio is calculated for each TF and represented by the grey bars.*

We obtained a significant enrichment for 44 out of 49 TFs (FDR < 0.05). The strongest effect was observed for EP300, a transcriptional activator that marks active enhancers. For three out of the five TFs with non-significant p-value, the reason for the non-significant result is that one of the groups is too small, a basic problem when using the chi-square test (**Supplementary Table 4**) shows an example of ZNF274).

We performed another test that calculates the ratio between cell line specific binding sites to common binding sites. This test was made to distinguish between cell line enrichment of binding sites to A compartment presented by D, to transition enrichment presented by R.

We computed *transition enrichment ratio* T, defined as follows: $n(i,S)$ is defined as described above, $i=3$ refers to common binding sites:

$$T_1 = \frac{n(1,AB)/n(1,BA)}{n(3,AB)/n(3,BA)}, T_2 = \frac{n(2,BA)/n(2,AB)}{n(3,BA)/n(3,AB)}$$

For the example in **Table 3**, we get $T_1 = 3.14$, p-value= 10^{-133}; chi-square test between n(1,AB),n(1,BA) and n(3,AB), n(3,BA)), $T_2 = 2.49$, p-value= 10^{-38}; chi-square test between n(2,BA),n(2,AB) and n(3,BA), n(3,AB)).

We obtained a significant enrichment for 42 out of 49 TFs (FDR < 0.05) for both T1 and T2.

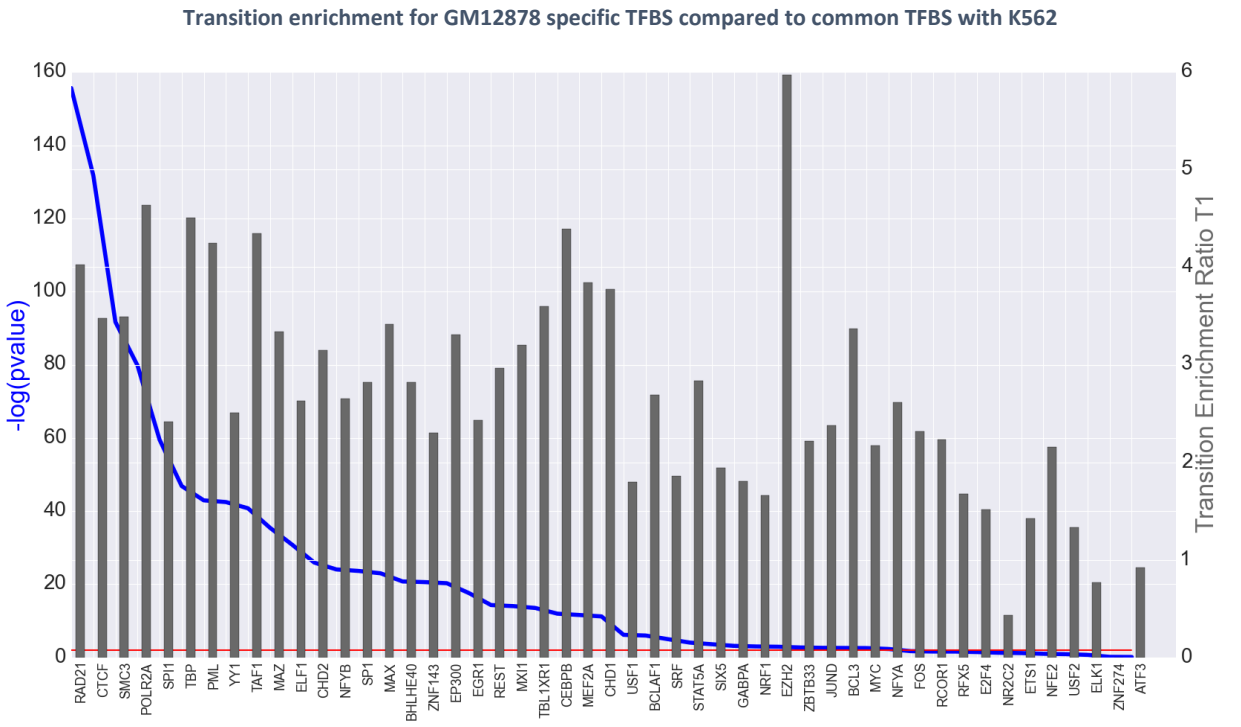


Figure 33c *- 49 TFs are sorted by p-value calculated by chi-square test as described above (y-axis is -log10(p-value)). TF with p-value above the horizontal line (marked in red) have significant p-value (p<0.05). The transition enrichment ratio for GM12878, T1, is calculated for each TF and represented by the grey bars.*

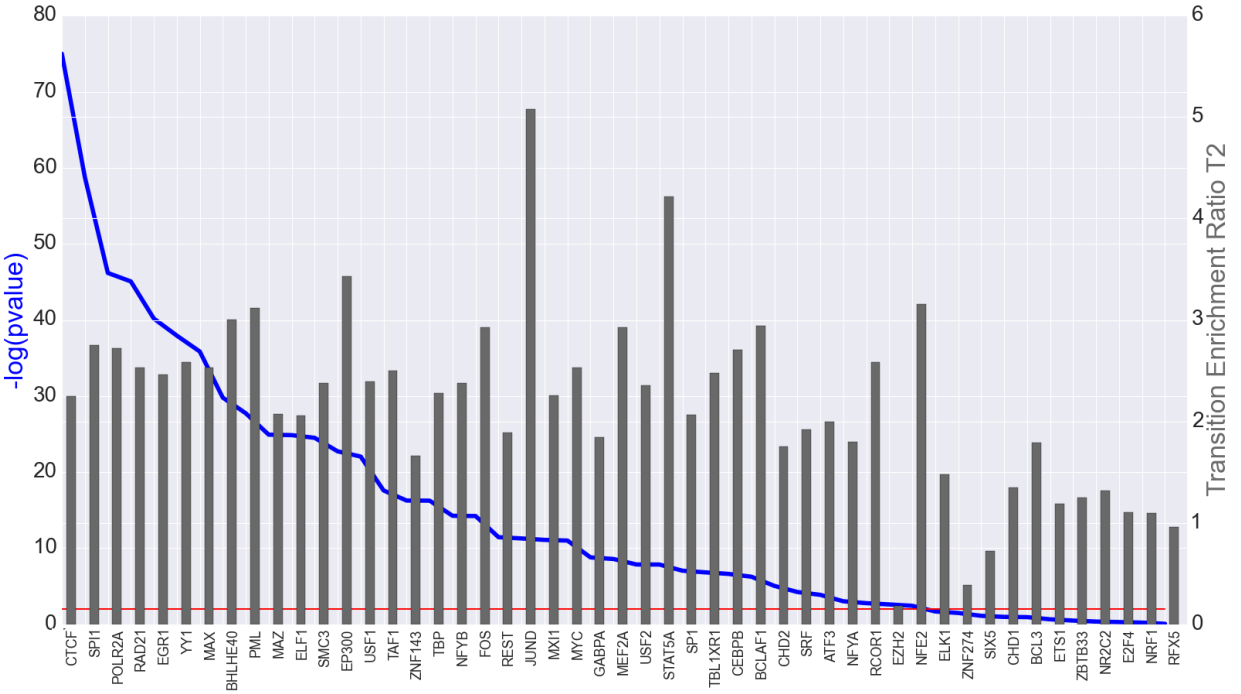


Figure 33d *- 49 TFs are sorted by p-value calculated by chi-square test as described above (y-axis is -log10(p-value)). TF with p-value above the horizontal line (marked in red) have significant p-value (p<0.05). The transition enrichment ratio for K562, T2, is calculated for each TF and represented by the grey bars.*

Next, we carried out similar tests for selected epigenetic marks. First, we examined H3k9ac, which marks active regions. Here too, as expected, we found a significant correlation between cell type specific H3K9ac signal and cell type specific compartmentalization. An example is shown in **Table 4** (p-value < 10^{-300}, chi-square test).

GM12878 NHEK	AA	AB	BA	BB	total	R	p-value
GM12878_only_BSs	14695	**3111**	596	1708	20110	4.54	$<10^{-300}$
NHEK_only_BSs	19997	1401	**5949**	4154	31501		
Common_BSs	21594	2036	1078	2911	27619		

Table 4 – H3k9ac sites in A and B compartments for cell lines GM12878 and NHEK

The opposite trend was observed for epigenetic repressive marks. As an example, we tested H3k27me3 and found that cell type specific signal is enriched for cell type-specific B compartments. An example is shown in **Table 5** (p-value < 10^{-300}))

MCF7 GM12878	AA	AB	BA	BB	total	R	p-value
MCF7_only_BSs	5213	1751	**3637**	9712	20313	0.54	10^{-122}
GM12878_only_BSs	7176	**1765**	1145	3608	13694		
Common_BSs	318	95	118	317	848		

Table 5 – H3k27me3 sites in A and B compartments for cell lines GM12878 and NHEK

3.2. Correlation between gene expression level and extent of promoter interactions

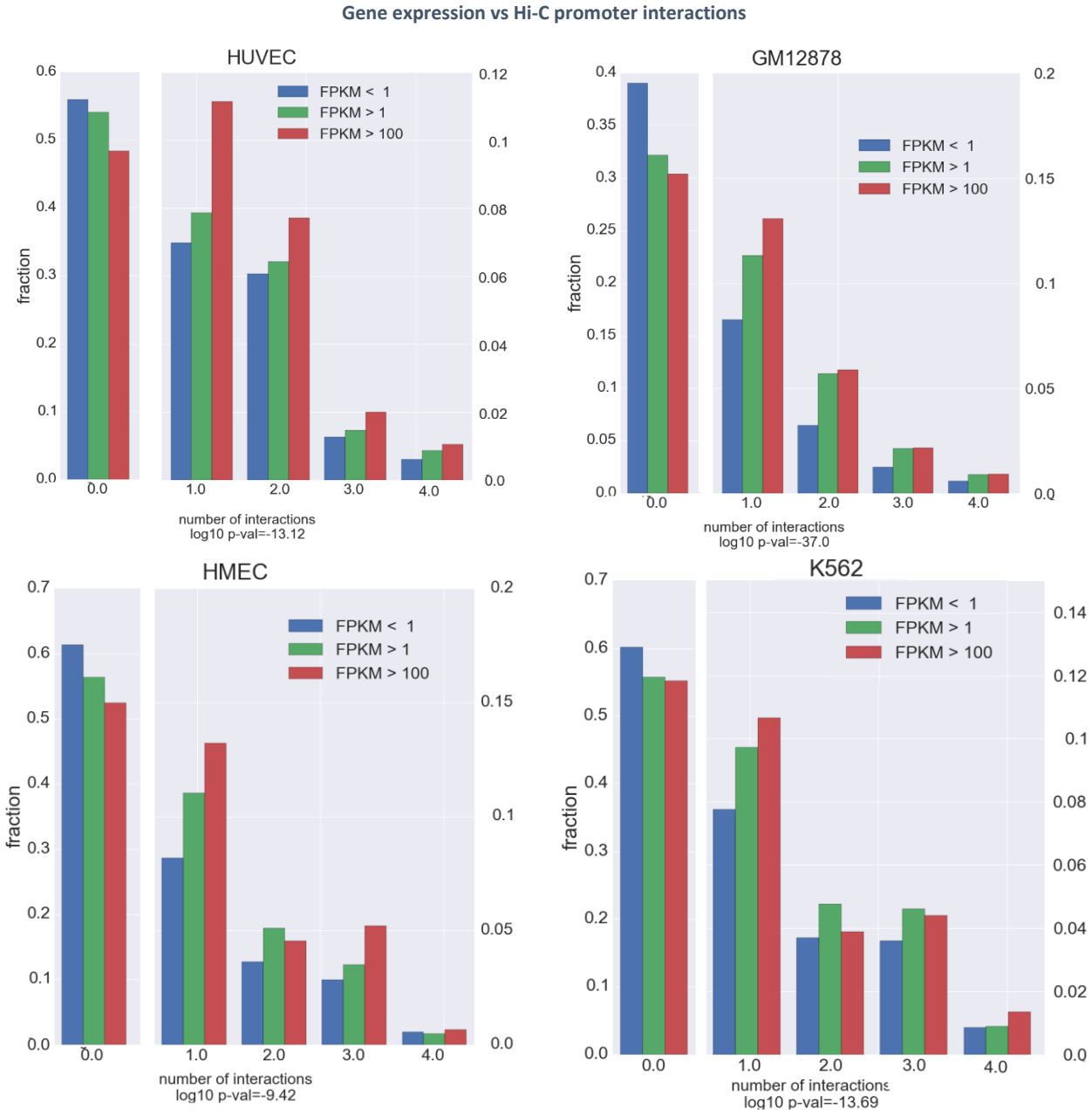
We next examined the relationship between gene expression level and chromatin interactions involving gene promoters. We hypothesized that promoters of highly expressed genes are more likely to be engaged in chromatin interactions than lowly expressed genes. To test this hypothesis, we analyzed only genes located within the A compartment.

3.2.1. Promoter-chromatin interactions are correlated with gene expression

The A compartment is generally characterized by high transcriptional activity. Yet, genes within this compartment show high expression variability and many of them are not expressed at detectable levels. We used promoter interactions inferred from both Hi-C and ChIA-PET data to test whether gene expression level is correlated with promoter-chromatin interactions. We reasoned that most promoter-chromatin interactions presumably reflects promoter-enhancer interactions, and thus are expected to correlate with higher expression of the involved gene.

For Hi-C, we used PSYCHIC [43], a tool that removes intra-TAD signal biases in Hi-C data and finds significant contacts between promoters and other chromatin segments. In accordance with our

hypothesis, we indeed observed that the number of interactions in which a gene promoter is involved is positively correlated with the gene's expression level. **Figure 34a** shows the distribution of the expression levels for five groups of genes in A compartment distinguished by their number of interactions. **Figure 34b** displays the same data but dividing the promoters into only two groups: those with no interactions and those with one or more interactions. Here too, the expression level of genes in A compartment whose promoter was not engaged in any chromatin interaction was significantly lower than the expression of those who were engaged.



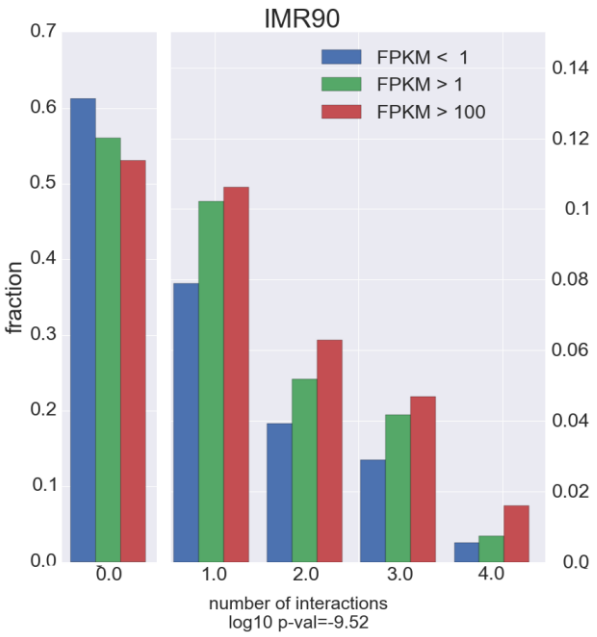


Figure 34a *- Gene expression and promoter interactions. Genes in A compartment were partitioned into three groups according to their expression levels. For each group, the distribution of the number of genes with 0,1 … 4 interactions in their promoter is shown. Interactions are extracted from Hi-C data using PSYCHIC. P-value is calculated using Wilcoxon test comparing the distributions in the least and most abundant expression groups.*

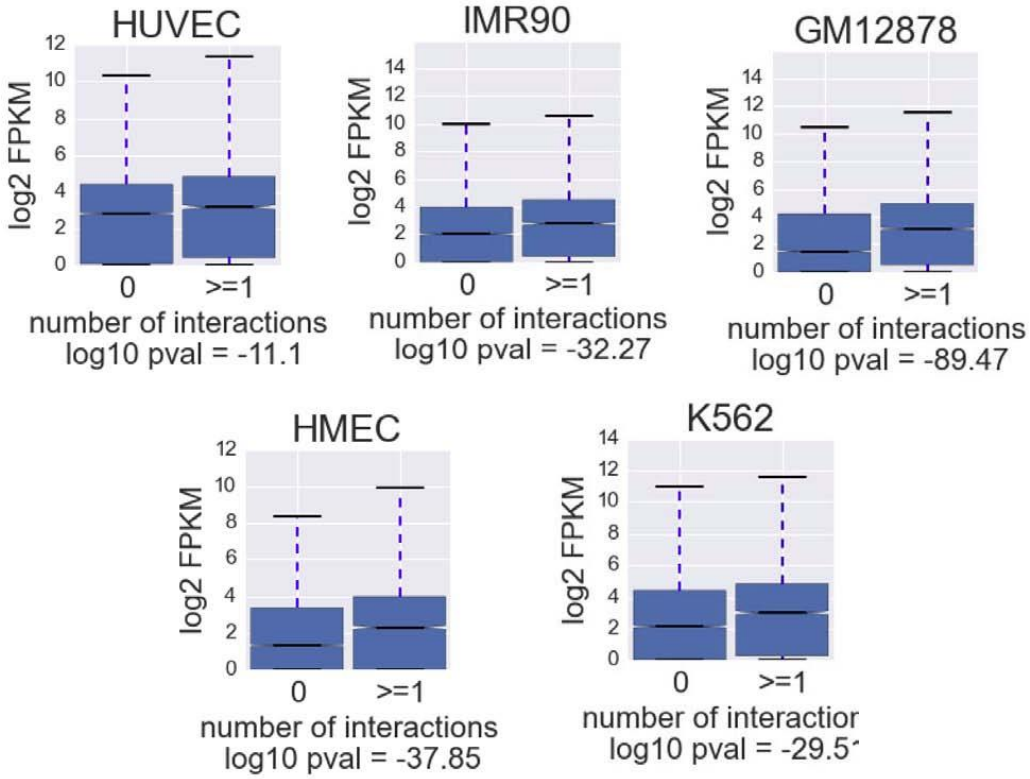


Figure 34b *- Genes in A compartment were partitioned into two groups according to number of interactions of their promoters. For each group, the distribution of the gene expression levels measured by RNA-Seq is shown. P-value is calculated using Wilcoxon test.*

Next, we applied a similar test, using ENCODE's ChIA-PET data for RNA POL2. Here too, we found the same correlation, albeit with stronger significance (**Figure 34c, 34d**). The reason for the higher significance is that while Hi-C data measures all chromatin interactions, ChIA-PET measures interactions mediated by RNA POL2, so it emphasizes transcription related interactions.

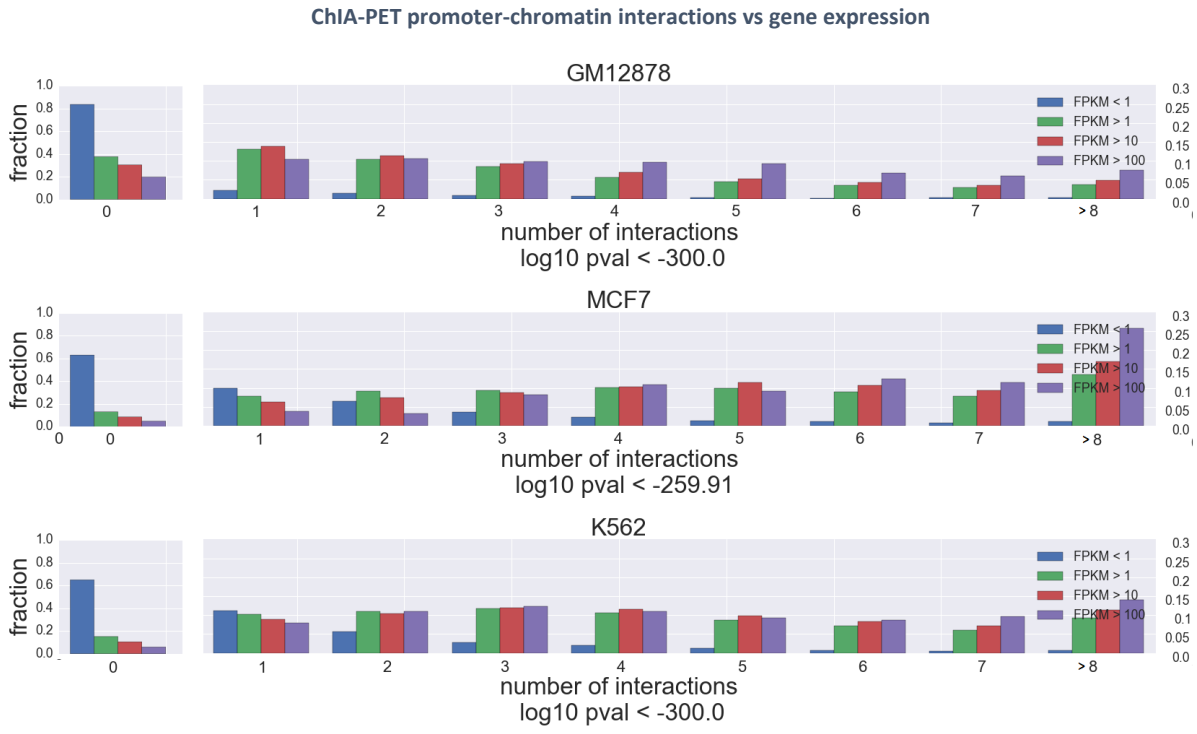


Figure 34c *– ChIA-PET promoter-chromatin interactions vs. gene expression. Genes in A compartment were partitioned into four groups according to their expression levels. For each group, the distribution of the number of genes with 0,1 … 8 interactions in their promoter is shown. Interactions are extracted from ChIA-PET RNA POL2 data. P-value is calculated using Wilcoxon test comparing the distributions in the least and most abundant expression groups.*

Gene expression vs ChIA-PET promoter-chromatin interactions

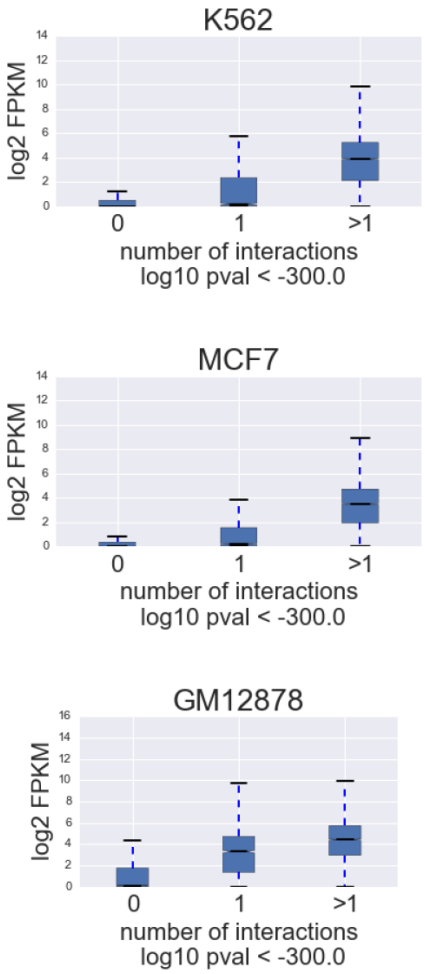


Figure 34d *– Gene expression vs ChIA-PET promoter-chromatin interactions. Genes in A compartment were partitioned into three groups according to number of promoter-chromatin interactions measured by ChIA-PET RNA POL2. For each group, the distribution of the gene expression levels measured by RNA-Seq is shown. P-value is calculated using Wilcoxon test comparing the distributions of genes with no interactions to genes with at least one interaction.*

3.2.2. Changes in promoter interactions across cell lines correlate with changes in gene expression levels

Having observed correlation between promoter contacts and expression for each cell line separately, we turned to examine differences between cell lines. We asked if the difference in expression of a gene in different cell lines is associated with difference in the number of interactions in which the gene's promoter is involved in these cell lines. This analysis too was confined to genes located within A compartment in both cell lines (AA genes). For each pair of cell lines, we divided the genes into four groups: no promoter interactions in both cells ("00" group; promoter interactions only in cell line 1 ("01"); promoter interactions only in cell line 2 ("10") and promoter interactions in both ("11"). We used Pol2 ChIA-PET data for this analysis. **Figure 35** shows the results of this analysis applied to MCF7 vs.

K562. The analysis strongly demonstrates that genes were more highly expressed in the cell line in which their promoter was more highly engaged in chromatin interactions (**Figure 35**).

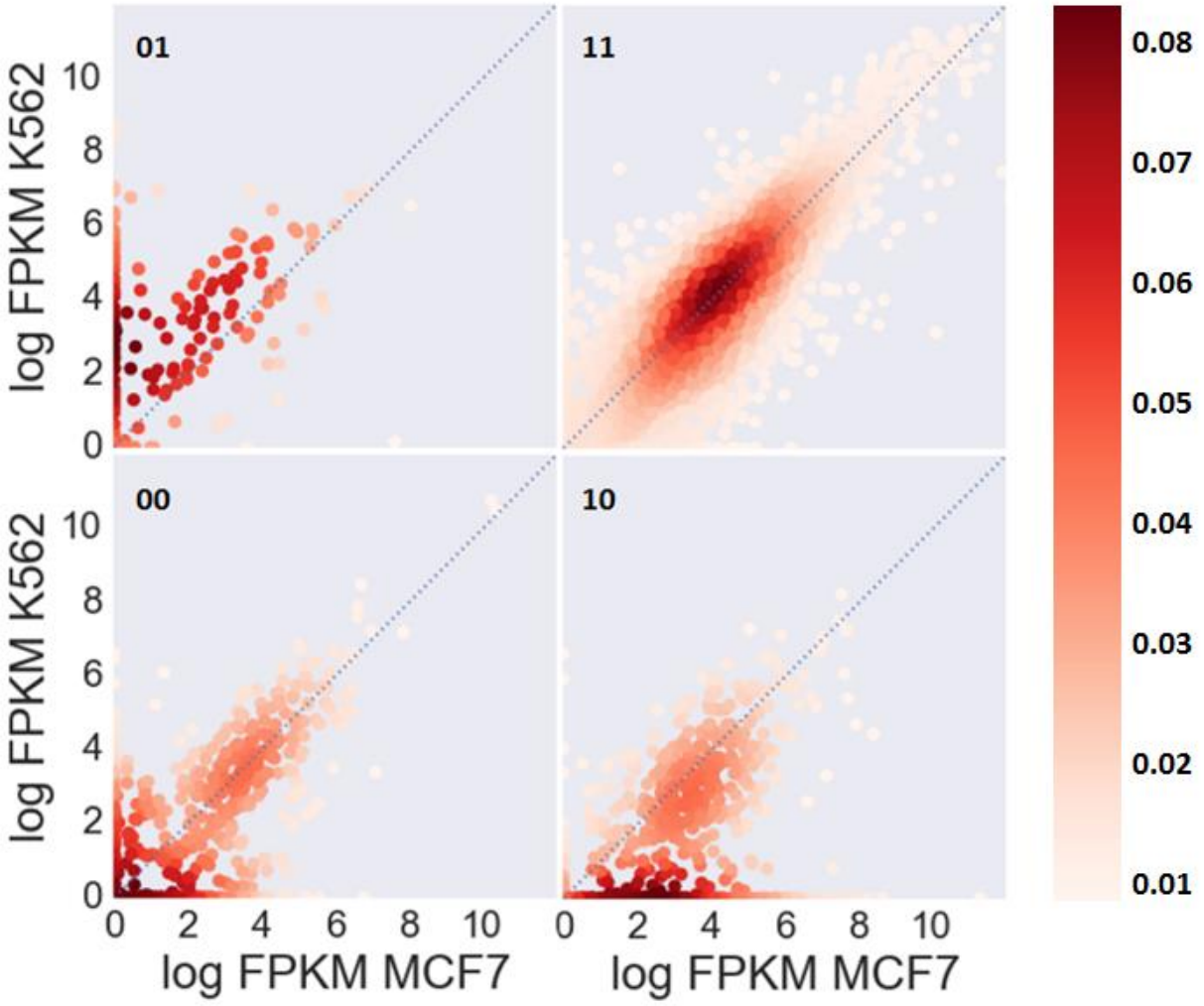


Figure 35 *Relation between Gene expression and ChIA-PET promoter-chromatin interactions between cell lines. AA genes were divided into four groups according to the involvement of their promoters in chromatin interactions in each of the cell lines in the examined pair (00, 01, 10, 11. 00) and expression levels were compared between the two cell lines. Genes were placed on the 2D plot according to their expression in both cell lines, clustered (with overlaps) using KDE. The color of a cluster stands for its gene density. The results show that there is a correlation between different gene expression in the examined cell lines and the number of promoter's chromatin contacts.*

We next calculated fold-change in gene expression between any pair of cell lines and, for each pair, compared the distribution of these ratios between the four groups of genes. Difference in expression level was significantly associated with difference in promoter involvement in chromatin interactions (**Figure 36**).

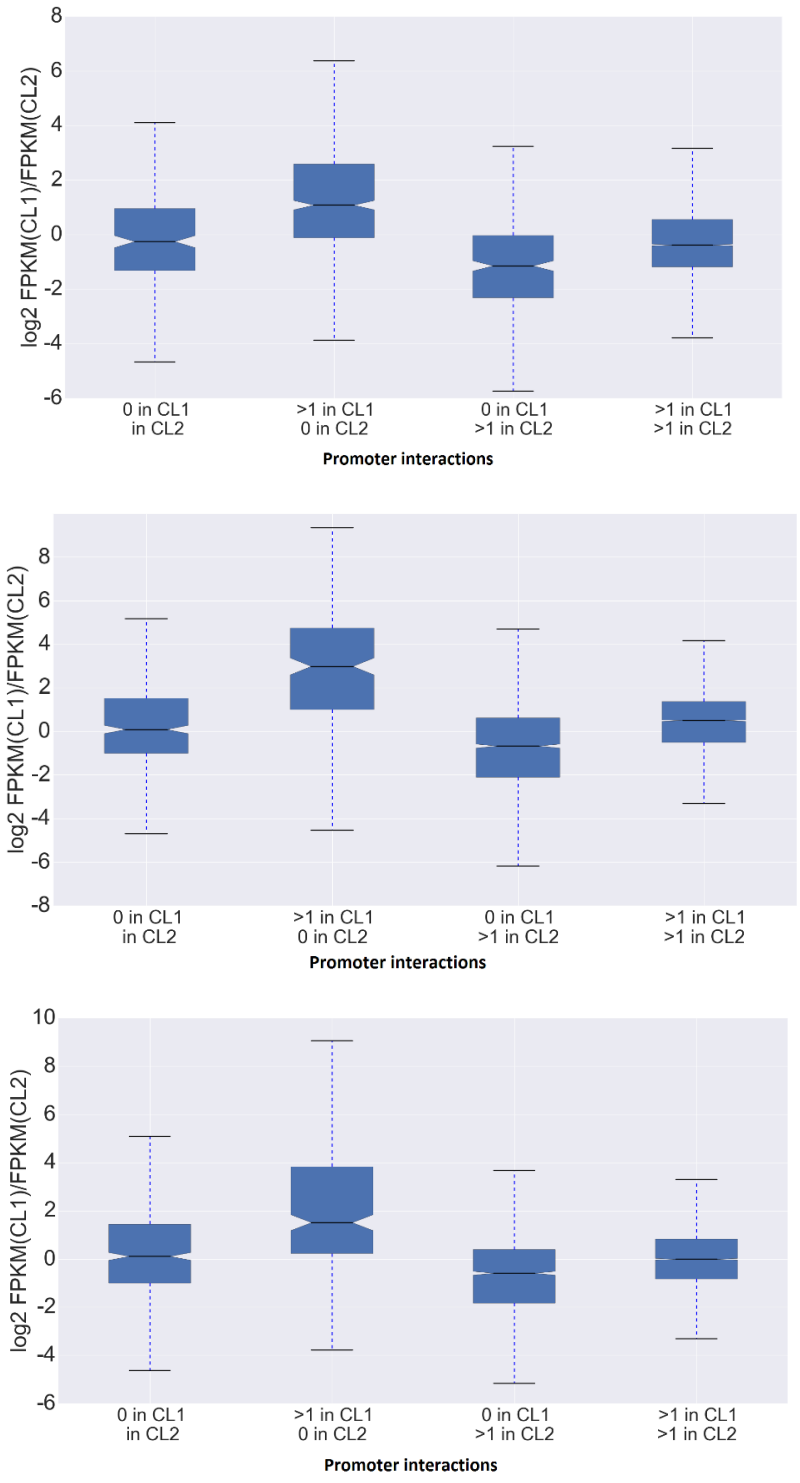


Figure 36 *Gene expression ratio between cell lines partitioned into groups depending on promoter interactions. The groups are as described in Figure 35. Top:* GM12878 and K562, middle: MCF7 and K562, bottom: GM12878 and MCF7. In all tests, *p-value* $< 10^{-300}$ *between group 10 and group 01.*

3.3. Correlation between basal chromatin interactions and gene induction upon treatment

Many transcriptomic studies observed that a large portion of the transcriptional response to various challenges is cell-type specific. We next sought to examine the role of chromatin interactions in the cell-type response to treatment.

In this section, we examine if chromatin interactions that are already in place in the cells prior to the treatment affect the set of genes that respond to the treatment. Specifically, we test the relationship between preexisting, basal chromatin interactions and cell's transcriptional response to treatment. To allow us to draw some general conclusions, we analyzed a variety of cell lines and multiple treatments covering diverse biological processes.

3.3.1. Chromatin compartmentalization and induction of TF binding upon treatment

In this analysis, we analyzed 110 publicly available Chip-Seq datasets from GEO that recorded TF binding profiles in cells (for which Hi-C data are available) before and after the application of treatment/stress. Overall, we analyzed 21 TFs in 7 cell lines in response to 22 treatments. (**Supplementary Table 5** summarizes the analyzed conditions). To ensure analysis uniformity, we downloaded raw sequence reads and detected TF peaks ourselves. Briefly, for each ChIP-Seq experiment, reads were aligned to the human genome (hg19) using bowtie2. Control and treated samples were then compared using MACS2 [44] to identify TF peaks that were induced or repressed upon treatment.

First, per experiment, we divided the induced sites into A/B compartments and examined their enrichment towards A. P-value was calculated using chi-squared test, comparing the counts of induced sites in A and in B to the number of base pairs in A and in B. We get very significant p-values in the vast majority of experiments, indicating that the preexisting A/B compartmentalization within a cell line constrains the TF-chromatin interactions induced in response to stress. The results are summarized in **Figure 37** and the full results are shown in **Supplementary Table 5.**

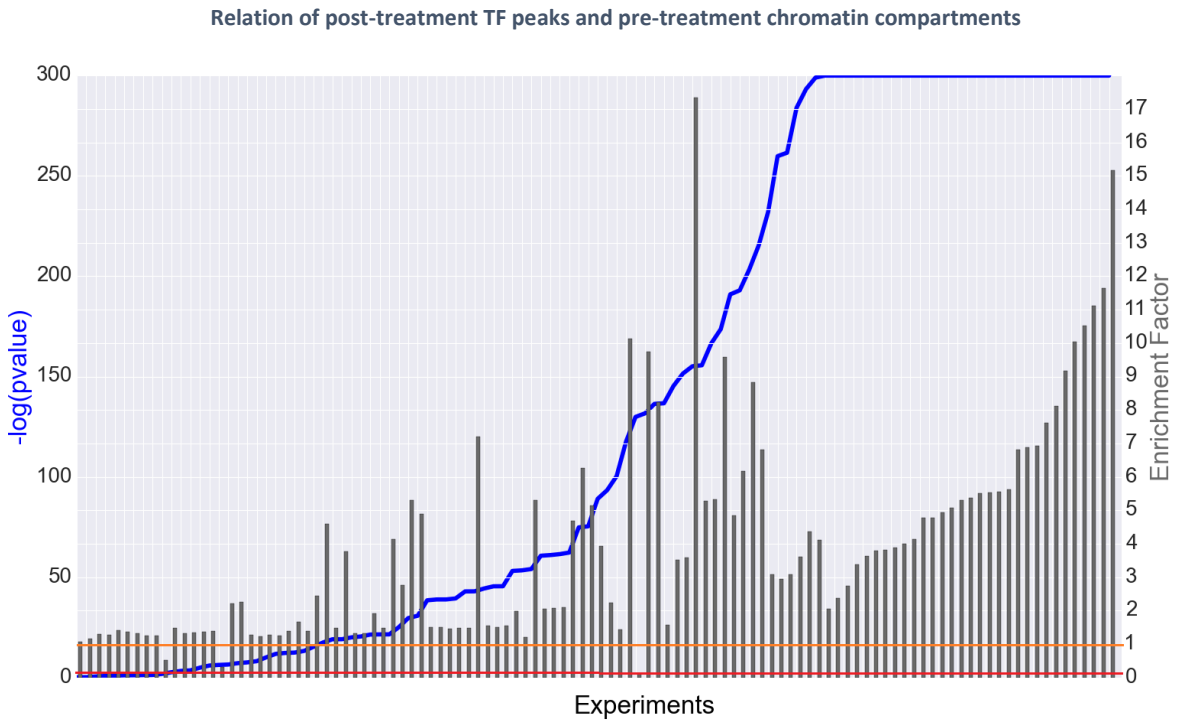


Figure 37 *–. Enrichment of TF BSs induced by various treatments to the A compartment (as exist in the cells before treatment) Experiments are sorted by p-value, occupancy enrichment ratios are represented by bars. Red line stands for p-value = 0.01 and orange line stands for occupancy enrichment ratio = 1.*

3.3.2. Chromatin compartmentalization and induction of TF binding upon treatment; Comparison between cell lines

Next, we sought to further examine the relationship between cell-type specific chromatin organization and response to treatment. To carry out such analysis, we searched for ChIP-Seq datasets that profiled chromatin binding of the same TF upon the same treatment in two different cell lines (for which Hi-C/ChIA-PET data are also available). For each such pair, we again divided the induced TF binding sites into three groups: binding sites induced upon treatment only in cell line 1, binding sites induced only in cell line 2 and binding sites induced in both. Induced TFs in each group were then divided into four categories – AA, AB, BA, BB as before. Our hypothesis was that TFBSs induced only in cell line 1 (2) would be enriched in region AB (BA). We found a significant relation for all comparisons (**Table 6**).

	AA	AB	BA	BB	total	Enrichment	R	p-value
MCF7 T47D Treatment - 17ÃŸ-estradiol Antibody - ER cocktail: Ab-10 Thermo Scientific Lab Vision, HC-20 sc-543 Santa Cruz								
Replicate 1:								
Cell1_only_BSs	2354	522	230	534	3640	2.33	1.99	7.06E-23
Cell2_only_BSs	1177	149	233	295	1854	1.63		
Common_BSs	1344	158	103	144	1749			
Replicate 2:								
Cell1_only_BSs	7302	1834	1229	2751	13116	1.56	1.5	6.41E-13
Cell2_only_BSs	854	108	177	180	1319	1.63		
Common_BSs	1794	248	227	317	2586			
Two rep combined								
Cell1_only_BSs	9656	2356	1459	3285	16756	1.61	1.61	3.29E-29
Cell2_only_BSs	2031	257	410	475	3173	0.63		
Common_BSs	3138	406	330	461	4335			
LNCAP MCF7 Treatment - TNFa Antibody - p65								
Cell1_only_BSs	74	28	16	43	161	1.7	1.59	0.001421
Cell2_only_BSs	1194	166	262	271	1893	1.56		
Common_BSs	67	11	12	12	102			
HUVEC MCF7 Treatment - TNFa Antibody - p65								
Cell1_only_BSs	12443	2568	321	842	16174	8	6.47	4.70E-99
Cell2_only_BSs	671	100	154	227	1152	1.44		
Common_BSs	690	83	34	54	861			
HUVEC IMR90 Treatment - TNFa Antibody - pol2								
Cell1_only_BSs	507	132	7	42	688	19	0.59	3.01E-12
Cell2_only_BSs	6726	1428	718	437	9309	0.53		
Common_BSs	208	28	1	6	243			
LNCAP HUVEC Treatment - TNFa Antibody - p65								
Cell1_only_BSs	75	39	28	39	181	1.47	8.73	1.62E-37
Cell2_only_BSs	12237	374	3470	866	16947	10		
Common_BSs	60	7	12	4	83			

Table 6 description of 5 comparisons of cell specific induced TFBSs and their relationship to A/B division. P-value is calculated by chi-square test for AB/BA in first 2 rows (cell 1/2 only BS)

Notably, we observed that despite the significant association between cell-type specific induced TFBSs and chromatin organization (preexisting in the cells prior to the application of the challenge), most of the cell-type specific induced TFBSs were located in AA regions (Table 6), indicating that other factors besides A/B compartmentalization underlie most of the cell-type specific transcriptional response.

3.3.3. Stress-induced genes are enriched for A compartment

At this point, we returned to our main question: to what extent the spectrum of genes induced by stress in each cell type is determined by the A/B structure constraints that exist in the cells before stress was applied? Here, we analyzed 36 gene expression datasets (**Supplementary Table 6**). Briefly, we downloaded raw sequence reads from GEO/SRA DB, mapped them to the human genome (hg19) using tophat2, counted the number of reads that mapped to each annotated gene using HTSeq-counts and GENCODE annotations and normalized gene expression estimates to RPKM. We compared expression profiles between treated and control samples and defined the genes whose expression was changed by at least 1.5-fold as differential genes. (To avoid inflation of lowly expressed genes among the called genes we used a floor level of 1.0 RPKM.) Then, for each cell line and treatment, we tested whether the set of induced genes was over-represented in the A compartment. In most conditions that we tested, the induced genes were enriched for the A compartment (**Figure 38**), meaning occupancy enrichment ratio was above 1 with p-value < 0.01 (chi-square test).

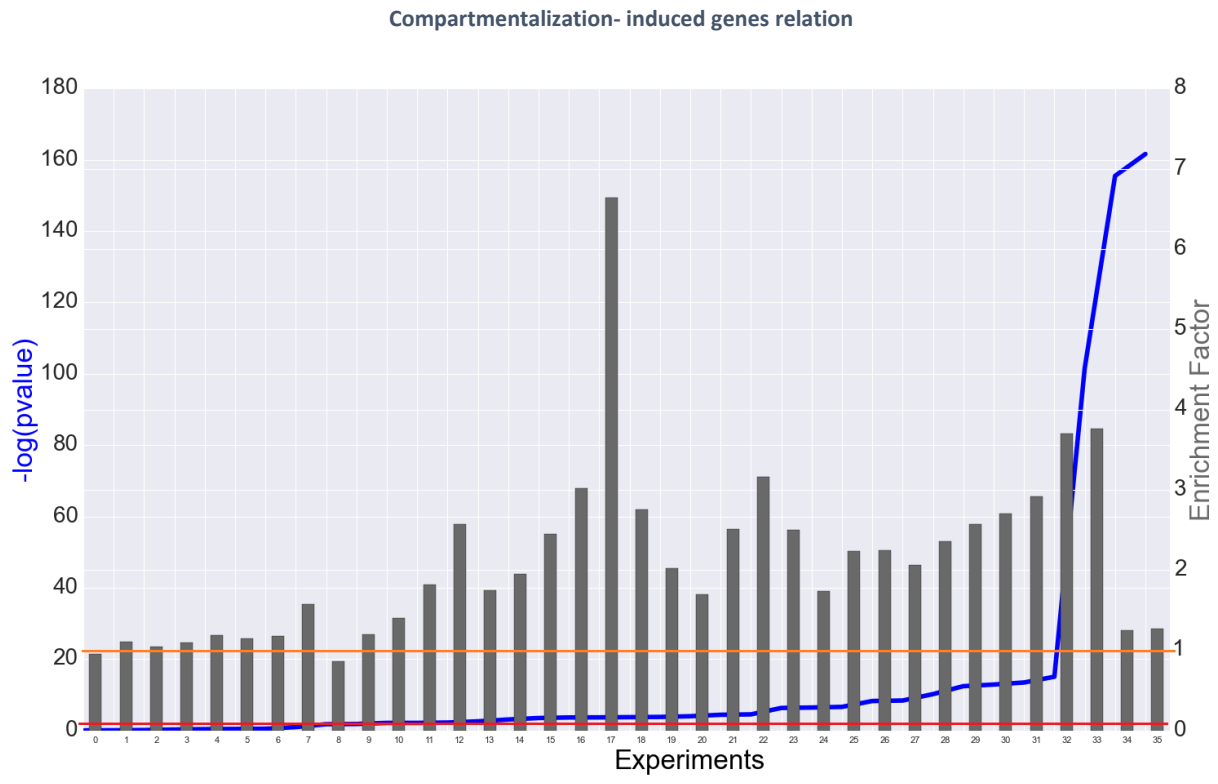


Figure 38 *- Summary plot of post-treatment RNA-Seq data (supplementary table 6) - experiments are sorted by p-value, enrichment factor is represented by bars. Red line stands for p-value = 0.01 and orange line stands for occupancy enrichment ratio = 1.*

This analysis resulted in less significant results than TFBS analysis because while there are thousands of induced TFBS measured by ChIP-Seq, the number of responsive genes measured by RNA-Seq is much lower (less than 10 in some cases). Nevertheless, 28 out of 36 experiments had a significant p-value, FDR < 0.05, and the fact that 34 out of 36 experiments have enrichment factor larger than 1 indicates this is not a random feature (p-val < $3 * 10^{-5}$).

3.3.4. Promoters of induced genes have more basal chromatin interactions prior to treatment

In this section, we tested if promoters of induced genes have a higher number of chromatin interactions prior to stress than non-induced ones. If this holds, it would suggest that the induced genes in each cell type are to a certain extent predetermined by the structural organization of the chromatin prior to the application of the treatment.

We analyzed the gene expression experiments described in section 3.3.3 and examined if the promoters of the induced genes are engaged in a significantly higher number of chromatin interactions. We estimated significance using permutation test with 10000 iterations, in each iteration selecting a random set of genes of the same size as the induced genes set (**Figure 39**). We obtained significant p-values (p<0.05) for all conditions but one (**Table 7**). That particular experiment (MCF7 E2 + ICI) had very few induced genes in A compartment (**Supplementary Table 6**).

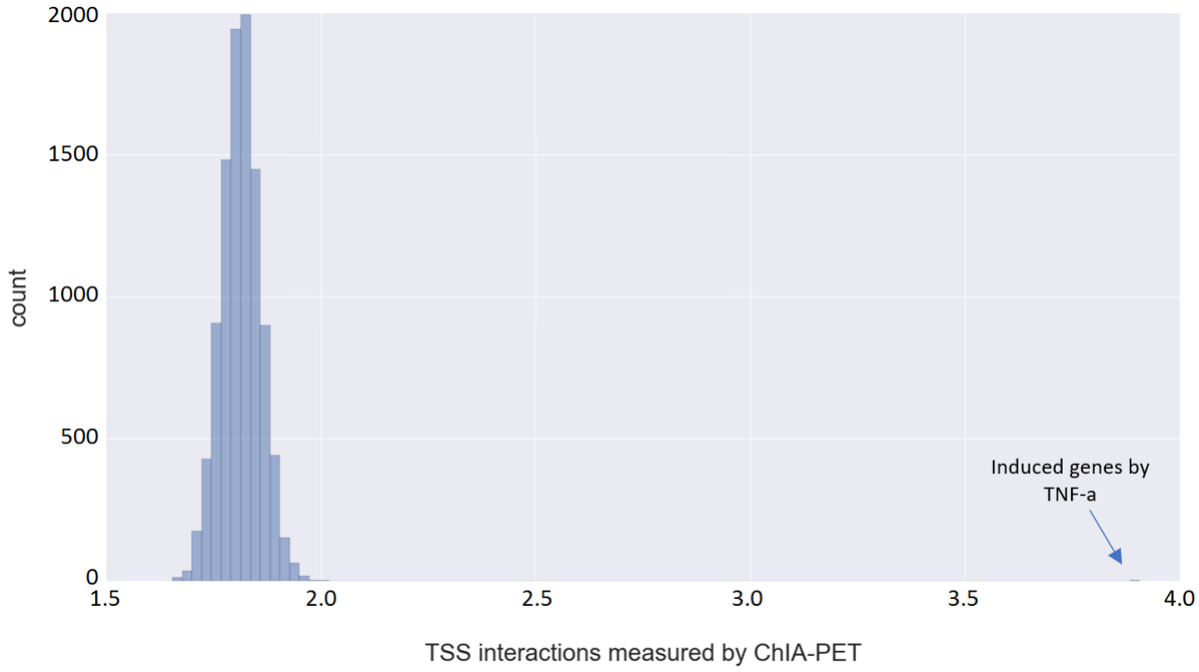


***Figure 39**- Evaluating the significance of the number of interactions of induces genes using permutation tests. The histogram shows the mean number of chromatin interactions in 10000 randomizations for GM12878 cell line treated with TNF-α. The arrow shows the number for the real set of induced genes.*

Cell line	Description	Data type	Mean number of promoter interactions	p-val
HUVEC	IFN	Hi-C	1.46	<10E-04
HUVEC	TNFa	Hi-C	1.57	<10E-04
K562	SAHA	ChIA-PET	5.44	<10E-04
K562	SAHA	Hi-C	1.12	<10E-04
K562	NaBut	ChIA-PET	5.27	<10E-04
K562	NaBut	Hi-C	1.03	<10E-04
HMEC	TNFa	Hi-C	1.4	<10E-04
MCF7	IL1B	ChIA-PET	7.91	<10E-04
MCF7	E2 + ICI	ChIA-PET	4.75	0.41
MCF7	E2+TOT+TNFa	ChIA-PET	7.16	<10E-04
MCF7	E2+TOT	ChIA-PET	4.94	0.04
MCF7	E2+TOT+IL1b	ChIA-PET	7.51	<10E-04
MCF7	E2	ChIA-PET	8.16	<10E-04
MCF7	IL1b+ICI	ChIA-PET	7.67	<10E-04
MCF7	TNFa	ChIA-PET	7.27	<10E-04
MCF7	estradiol	ChIA-PET	10.63	<10E-04
MCF7	TNFa+ICI	ChIA-PET	6.89	<10E-04
IMR90	Nutlin-3a	Hi-C	1.21	<10E-04
IMR90	TNFa	Hi-C	1.25	<10E-04
IMR90	TNFa+cycloheximide	Hi-C	1.24	<10E-04
GM12878	TNFa	ChIA-PET	3.91	<10E-04
GM12878	TNFa	Hi-C	1.77	<10E-04

Table 7 - Significance of the relation between the number of chromatin interactions and gene expression. Empirical p-values were calculated by permutation test.

3.3.5. Correlation between chromatin compartmentalization and gene induction upon treatment between cell lines

Finally, we examined if cell-specific gene induction correlates with pre-existing chromatin compartmentalization. In order to do so, we repeated the steps in section 3.3.2 with the set of induced genes instead of transcription factors binding sites for five cell lines treated with TNF-α. Here too, for all cases, we found that genes induced specifically in cell line 1 (2) were significantly enriched for AB (BA)

regions (**Table 8**). Yet, here too, the majority of cell type specific responsive genes were located in AA regions, again indicating that other factors (e.g., cell-type specific basal chromatin interactions within the A region) play critical roles in determining the specific spectrum of genes that respond to a challenge in each cell type.

	AA%	AB%	BA%	BB%	Enrichment	R	p-value
HMEC GM12878							
only induced in cell line 1	0.54	0.26	0.11	0.09	2.36	3.4	7.44E-08
only induced in cell line 2	0.81	0.03	0.09	0.08	3		
induced in both	0.83	0	0.14	0.02			
GM12878 IMR90							
only induced in cell line 1	0.76	0.13	0.03	0.07	4.33	3.2	1.49E-06
only induced in cell line 2	0.71	0.14	0.1	0.05	0.71		
induced in both	0.77	0.19	0	0.04			
IMR90 MCF7							
only induced in cell line 1	0.68	0.13	0.14	0.06	0.93	1.5	1.00E-04
only induced in cell line 2	0.62	0.06	0.22	0.1	3.67		
induced in both	0.7	0.08	0.15	0.08			
GM12878 MCF7							
only induced in cell line 1	0.79	0.11	0.06	0.04	1.83	1.7	3.00E-03
only induced in cell line 2	0.74	0.08	0.1	0.09	1.25		
induced in both	0.81	0.14	0.04	0			
HMEC MCF7							
only induced in cell line 1	0.45	0.31	0.09	0.14	3.44	2	9.00E-04
only induced in cell line 2	0.75	0.06	0.1	0.09	1.67		
induced in both	0.79	0.09	0.02	0.09			
HUVEC MCF7							
only induced in cell line 1	0.54	0.38	0.04	0.04	9.5	2	7.34E-05
only induced in cell line 2	0.75	0.07	0.1	0.08	1.43		
induced in both	0.77	0.09	0.06	0.09			

Table 8 Correlation between cell-type specific A/B compartmentalization and response to TNFa

4. Discussion

In this thesis we examined the relationship between pre-defined genome structures in the nucleus and changes in the levels of gene expression in response to stress. Our goal was to perform a broad analysis in order to obtain insight into how pre-stress structure affects gene regulation in the cell in response to stress.

First, we collected Hi-C data of 13 cell lines from six different studies. We normalized these data sets and performed compartmentalization of the genome in each cell line, partitioning it into two compartments, A and B, which correlate with euchromatin and heterochromatin, respectively. We saw that different cell lines have similar but far from identical A/B partition (**Figure 30**). We validated what previous studies have shown on A/B compartments using ChIP-Seq and RNA-Seq data from hundreds of studies. RNA-Seq analysis demonstrated that genes in A compartments are highly expressed compared to genes in B compartments in all cell lines. In addition, ChIP-Seq data for transcription factor binding sites validated that the vast majority of TFBS are enriched for A compartment (**Table 2**). Following this results, we checked whether differences in gene expression and cell type-specific binding sites correlate with differences in A/B compartments. We showed that transition from A to B or B to A between cell lines has a significant correlation with changes in gene expression and in TFBS density (**Figure 33-34)**.

Next, we used Hi-C and ChIA-PET data of three cell types from two studies for estimating the number of gene promoter interactions with distal chromatin segments. Combining these data with RNA-Seq, we showed that high promoter-chromatin interactions correlate with high expression levels, suggesting that the data reflect enhancer-promoter interactions. When comparing between cell lines, we also demonstrated that differences in promoter-chromatin interactions are accompanied by differences in gene expression between cell lines (**Figure 35**).

After analyzing untreated cells data, we collected data (ChIP-Seq, RNA-Seq and GRO-Seq) on cell response to different treatments. These data spanned 85 experiments performed on the set of cell lines that we analyzed. In order to make the analysis of data from diverse sources as coherent as possible, we processed the reads from the experiments using our own pipeline and avoided using binding sites and expression levels as computed and reported in each study. In order to check the relationship between cell response and A/B compartmentalization, we first calculated the enrichment of induced TFBSs to A compartment and revealed a significant enrichment in 44 out of 49 ChIP-Seq data sets (**Figure 37**). Following this result, we made five comparisons between induced binding sites of the same transcription factor under the same treatment between different cell lines. The output of this comparisons revealed that predefined compartment state transitions correlate with cell specific induced binding sites (**Table 6**). Following induced TFBSs, we checked for enrichment in A compartment for genes with induced transcript level as measured by RNA-Seq and GRO-Seq. We revealed a significant enrichment in 28 out of 36 experiments (**Figure 38**). When comparing different cell lines under the treatment of TNFα, we detected that transition in compartment state shows a significant correlation with cell specific induced genes (**Table 8**).

Having observed these correlations, our last step compared gene expression data after treatment to promoter interactions of the untreated cells, extracted from Hi-C and ChIA-PET data. We showed that

for 20 out of 22 data sets, the mean number of promoter-chromatin interactions in the untreated cells for induced genes after treatment was highly significant (**Table 7**).

The results described above suggest that basal genome structures affect the cell-specific response to stress. In low resolution, A/B partition varies between cell lines and our analysis shows that these transitions correlate with induced protein-DNA complexes and with the induced level of gene transcription. At a much higher resolution, we show the enrichment of higher resolution promoter-chromatin interactions in induced genes, suggesting that some genes are "poised" well in advance of the stress, making them readier to respond to stress. Previous studies have shown that the compartments can be divided into topological associated domains (TADs) that are much more conserved between cell lines. Recent work demonstrated that TADs function as a regulatory structure, changing all of its gene expression levels in the same direction as a response to stress. We have not explored the role of these domains in cell response to stress, but the combination of our work with these insights suggests that poised genes tend to be in the same TAD. This might explain one of the differences between expressed genes and poised genes.

Our work emphasizes the assumption implied by previous studies, that a critical stage in the genome organization takes place during differentiation. The lack of Hi-C data and ChIA-PET data of cells after treatment did not allow us to perform a broad comparison of 3D organization in basal and induced cells. Such analysis might shed light on the importance of the predefined structural constraints and on the ability of the cell to additionally modify it in response to stress.

Throughout our research we noticed that the same features of induces genes exist for repressed genes as well. Since most of the expressed genes that were repressed in response to stress are in A compartment and have a high number of promoter-chromatin interactions, we did not focus on this enrichment in the thesis. Understanding whether there is a pre-existing long range interactions profile that characterizes repressed genes will add support to the hypothesis that the basal spatial organization has an important role in gene regulation.

An interesting future direction of future research can be using different methods in order to modify cells 3D organization, and testing whether it affects cell response to stress. Such study will help prove that the correlation that we see between the 3D organization and gene expression is indeed a mechanism that has a significant role in gene regulation.

References

[1] E. Lieberman-Aiden *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science (80-.).*, vol. 326, no. 5950, 2009.

[2] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions.," *Nature*, vol. 485, no. 7398, pp. 376–80, Apr. 2012.

[3] S. S. P. Rao *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014.

[4] J. R. Dixon *et al.*, "Chromatin architecture reorganization during stem cell differentiation.," *Nature*, vol. 518, no. 7539, pp. 331–6, Feb. 2015.

[5] J. H. Gibcus and J. Dekker, "The Hierarchy of the 3D Genome," *Mol. Cell*, vol. 49, no. 5, pp. 773–782, Mar. 2013.

[6] E. de Nadal, G. Ammerer, and F. Posas, "Controlling gene expression in response to stress," *Nat. Rev. Genet.*, vol. 12, no. 12, p. 833, Nov. 2011.

[7] F. Jin *et al.*, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, vol. 503, no. 7475, pp. 290–4, Oct. 2013.

[8] "Hasan H. Otu From Sequence to Function to Network: Analysis Issues in Bioinformatics BIDMC Genomics CenterHarvard Medical School. - ppt download." [Online]. Available: http://slideplayer.com/slide/10711347/. [Accessed: 24-Jul-2017].

[9] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein, "Comparing protein abundance and mRNA expression levels on a genomic scale.," *Genome Biol.*, vol. 4, no. 9, p. 117, 2003.

[10] D. Greenbaum, R. Jansen, and M. Gerstein, "Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts," *Bioinformatics*, vol. 18, no. 4, pp. 585–596, Apr. 2002.

[11] M. P. Washburn *et al.*, "Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae," *Proc. Natl. Acad. Sci.*, vol. 100, no. 6, pp. 3107–3112, Mar. 2003.

[12] D. U. Gorkin, D. Leung, and B. Ren, "The 3D genome in transcriptional regulation and pluripotency," *Cell Stem Cell*, vol. 14, pp. 762–775, 2014.

[13] N. D. Heintzman *et al.*, "Histone modifications at human enhancers reflect global cell-type-specific gene expression," *Nature*, vol. 459, no. 7243, pp. 108–112, May 2009.

[14] D. T. Odom *et al.*, "Control of Pancreas and Liver Gene Expression by HNF Transcription Factors," *Science (80-.).*, vol. 303, no. 5662, 2004.

[15] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: from properties to genome-wide predictions," *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 272–286, Mar. 2014.

[16] L. Calviello *et al.*, "Detecting actively translated open reading frames in ribosome profiling data,"

Nat. Methods, vol. 13, no. 2, pp. 165–170, Dec. 2015.

[17] A. T. Annunziato, “DNA Packaging: Nucleosomes and Chromatin,” *Scitable*.

[18] “Niveles de empaquetamiento del ADN | Biologia 508_07 | Pinterest | Búsqueda.” [Online]. Available: https://es.pinterest.com/pin/497647827560940648/. [Accessed: 24-Jul-2017].

[19] A. J. Bannister and T. Kouzarides, “Regulation of chromatin by histone modifications.,” *Cell Res.*, vol. 21, no. 3, pp. 381–95, Mar. 2011.

[20] “profobr.club - Chromatin In Mitosis.” [Online]. Available: http://profobr.club/jpgcpng-chromatin-in-mitosis.html. [Accessed: 24-Jul-2017].

[21] V. Ea, M.-O. Baudement, A. Lesne, and T. Forné, “Contribution of Topological Domains and Loop Formation to 3D Chromatin Organization,” *Genes (Basel).*, vol. 6, no. 3, pp. 734–750, Jul. 2015.

[22] F. Le Dily *et al.*, “Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation.,” *Genes Dev.*, vol. 28, no. 19, pp. 2151–62, Oct. 2014.

[23] “Genomics: Think Global, Act Local,” *Cell*, vol. 149, no. 7, pp. 1413–1415, Jun. 2012.

[24] Z. Tang *et al.*, “CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription,” *Cell*, vol. 163, no. 7, pp. 1611–1627, Dec. 2015.

[25] S. Sofueva *et al.*, “Cohesin-mediated interactions organize chromosomal domain architecture,” *EMBO J.*, vol. 32, no. 24, pp. 3119–3129, Dec. 2013.

[26] J. Zuin *et al.*, “Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 3, pp. 996–1001, Jan. 2014.

[27] A. L. Sanborn *et al.*, “Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 47, pp. E6456-65, Nov. 2015.

[28] J. Feng, T. Liu, B. Qin, Y. Zhang, and X. S. Liu, “Identifying ChIP-seq enrichment using MACS,” *Nat. Protoc.*, vol. 7, no. 9, pp. 1728–1740, Aug. 2012.

[29] S. G. Landt *et al.*, “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.,” *Genome Res.*, vol. 22, no. 9, pp. 1813–31, Sep. 2012.

[30] “Wikipedia - ChIP-Seq.” .

[31] M. Griffith *et al.*, “Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud,” *PLOS Comput. Biol.*, vol. 11, no. 8, p. e1004393, Aug. 2015.

[32] L. J. Core, J. J. Waterfall, and J. T. Lis, “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters,” *Science (80-.).*, vol. 322, no. 5909, pp. 1845–1848, Dec. 2008.

[33] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, “Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.,” *Nat. Rev. Genet.*, vol. 14, no. 6, pp. 390–403, Jun. 2013.

[34] “Chromosome Conformation Capture, Contact Mapping Market Intelligence.” [Online]. Available:

https://citalytics.com/inside-contact-mapping-research-a-global-view-of-top-labs-and-authors/. [Accessed: 24-Jul-2017].

[35] "Chromatin And Chromosomes." [Online]. Available: http://h3.danieledance.com/index.php?q=chromatin-and-chromosomes. [Accessed: 24-Jul-2017].

[36] J.-P. Fortin and K. D. Hansen, "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data.," *Genome Biol.*, vol. 16, no. 1, p. 180, Aug. 2015.

[37] E. Yaffe and A. Tanay, "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture," *Nat. Genet.*, vol. 43, no. 11, pp. 1059–1065, Oct. 2011.

[38] P. A. Knight and D. Ruiz, "A fast algorithm for matrix balancing," *IMA J. Numer. Anal.*, vol. 33, no. 3, pp. 1029–1047, Jul. 2013.

[39] M. Hu, "HiCNorm: removing biases in Hi-C data via Poisson regression," *Bioinformatics*, vol. 28, pp. 3131–3133, 2012.

[40] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.," *Mol. Cell*, vol. 38, no. 4, pp. 576–89, May 2010.

[41] A. R. Barutcu *et al.*, "Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells," *Genome Biol.*, vol. 16, no. 1, p. 214, Dec. 2015.

[42] "Ph.D. thesis - Matthias Scholz - Max Planck Institute of Molecular Plant Physiology." [Online]. Available: http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/. [Accessed: 11-Jul-2017].

[43] G. Ron, D. Moran, and T. Kaplan, "Promoter-Enhancer Interactions Identified from Hi-C Data using Probabilistic Models and Hierarchical Topological Domains," *bioRxiv*, 2017.

[44] J. S. Carroll *et al.*, "Genome-wide analysis of estrogen receptor binding sites," *Nat. Genet.*, vol. 38, no. 11, pp. 1289–1297, Nov. 2006.

[45] P. C. Taberlay *et al.*, "Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations.," *Genome Res.*, vol. 26, no. 6, pp. 719–31, 2016.

[46] M. Malinen, E. A. Niskanen, M. U. Kaikkonen, and J. J. Palvimo, "Crosstalk between androgen and pro-inflammatory signaling remodels androgen receptor and NF-κB cistrome to reprogram the prostate cancer cell transcriptome.," *Nucleic Acids Res.*, vol. 45, no. 2, pp. 619–630, Jan. 2017.

[47] M. A. Sammons, J. Zhu, A. M. Drake, and S. L. Berger, "TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity," *Genome Res.*, vol. 25, no. 2, pp. 179–188, Feb. 2015.

[48] J. D. Stender *et al.*, "Structural and Molecular Mechanisms of Cytokine-Mediated Endocrine Resistance in Human Breast Cancer Cells," *Mol. Cell*, vol. 65, no. 6, p. 1122–1135.e5, Mar. 2017.

[49] E. Swinstead *et al.*, "Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin

Transitions," *Cell*, vol. 165, no. 3, pp. 593–605, Apr. 2016.

[50] ENCODE, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

Supplementary Tables

Cell line	Description	Hi-C data resolution (Kbp)	Source study
MCF10A	Non-tumorigenic epithelial breast cell line	40,250	[41]
MCF7	Breast cancer cell line with overexpression of estrogen receptor		
LNCAP	Androgen sensitive prostate adenocarcinoma cell line	40,100	[45]
PrEC	Prostate epithelial cell line	40	
PC3	Androgen insensitive prostate cancer cell line		
HUVEC	Human umbilical vein endothelial cell line	5,10,100,250,500,1000	[3]
NHEK	Primary normal human epidermal keratinocytes cell line		
IMR90	Fetal lung fibroblasts cell line		
K562	Myelogenous leukemia cell line		
KBM7	Chronic myelogenous leukemia cell line		
HMEC	Human mammary epithelial cell line		
GM12878	Lymphoblastoid cell line	1,5,10,100,250,500,1000	
T47D	Breast cancer cell line with mutated p53	100	[22]

Supplementary Table 1 – Hi-C datasets source study and resolution

Cell line	Description	ChIA-PET data resolution (Kbp)	Source study
MCF7	Breast cancer cell line with overexpression of estrogen receptor	3.2	ENCODE - GSE39495
K562	Myelogenous leukemia cell line	3.1	
GM12878	Lymphoblastoid cell line	1.2	[24]

Supplementary Table 2- ChIA-PET datasets source study and resolution

Transcription factor	Cell lines with available ChIP-Seq dataset in ENCODE
ZBTB33	K562, GM12878
CTCF	K562, HUVEC, NHEK, HMEC, IMR90, GM12878
EGR1	K562, GM12878
RUNX3	GM12878
MAZ	K562, GM12878
RAD21	K562, GM12878, IMR90
SMC3	K562, GM12878
MAFK	K562, IMR90
MAFF	K562
E2F6	K562
MAX	K562, GM12878, HUVEC
PAX5	GM12878
POLR2A	K562, HUVEC, GM12878, IMR90, NHEK
PHF8	K562
PML	K562, GM12878
YY1	K562, GM12878
TAF1	K562, GM12878
SIN3AK20	K562
GTF2F1	K562
ATF2	GM12878
MYC	K562, GM12878, HUVEC
MXI1	GM12878, K562
JUND	K562, GM12878
POU2F2	GM12878
KDM5B	K562

Transcription factor	Cell lines with available ChIP-Seq dataset in ENCODE
TBP	K562, GM12878
EP300	K562, GM12878
ELK1	GM12878, K562
RFX5	K562, GM12878
CHD2	K562, GM12878
ATF3	K562, GM12878
BRCA1	GM12878
NFYA	K562, GM12878
NFYB	K562, GM12878
JUN	K562, HUVEC
GABPA	K562, GM12878
E2F4	K562, GM12878
SP1	K562, GM12878
SRF	K562, GM12878
ELF1	K562, GM12878
USF1	K562, GM12878
ATF1	K562
SIX5	K562, GM12878
USF2	GM12878, K562
FOS	K562, HUVEC, GM12878
TBL1XR1	K562, GM12878
ZNF143	K562, GM12878
SP2	K562
EBF1	GM12878
CTCFL	K562

Transcription factor	Cell lines with available ChIP-Seq dataset in ENCODE
TEAD4	K562
THAP1	K562
ZEB1	GM12878
CEBPB	K562, IMR90, GM12878
PBX3	GM12878
UBTF	K562
CBX3	K562
BCLAF1	K562, GM12878
RBBP5	K562
RCOR1	K562, GM12878
FOSL1	K562
GATA2	K562, HUVEC
BHLHE40	K562, GM12878
TAL1	K562
BCL3	GM12878, K562
NFATC1	GM12878
MEF2A	GM12878, K562
MEF2C	GM12878
ZNF263	K562
CCNT2	K562
HDAC2	K562
TCF3	GM12878
TCF12	GM12878
ZNF274	K562, GM12878
STAT1	GM12878

Transcription factor	Cell lines with available ChIP-Seq dataset in ENCODE
BATF	GM12878
HMGN3	K562
SETDB1	K562
TAF7	K562
SPI1	K562, GM12878
ETS1	K562, GM12878
REST	K562, GM12878
ZBTB7A	K562
EZH2	NHEK, HMEC, K562, HUVEC, GM12878
JUNB	K562
NR2F2	K562
TRIM28	K562
GTF3C2	K562
SAP30	K562
CHD1	K562, GM12878
STAT5A	K562, GM12878
HDAC1	K562
NRF1	K562, GM12878
NR2C2	GM12878, K562
SIN3A	GM12878
GATA1	K562
NFIC	GM12878
IRF4	GM12878
BCL11A	GM12878
MTA3	GM12878

Transcription factor	Cell lines with available ChIP-Seq dataset in ENCODE
FOXM1	GM12878
RXRA	GM12878
KAP1	K562
BACH1	K562
HDAC8	K562
NFE2	K562, GM12878
ARID3A	K562
WRNIP1	GM12878
GTF2B	K562
HDAC6	K562
SMARCA4	K562
BRF2	K562
IKZF1	GM12878
SMARCB1	K562
STAT3	GM12878
BDP1	K562
RPC155	K562
SIRT6	K562
RDBP	K562
ZZZ3	GM12878
POLR3G	K562, GM12878
BRF1	K562

Supplementary Table 3 - 122 TFs that had ChIP-Seq data recorded by the ENCODE project for cell lines with Hi-C data

GM12878 K562	AA	AB	BA	BB	total
Cell1_only_BSs	5	0	1	11	17
Cell2_only_BSs	869	67	218	236	1390
Common_BSs	47	6	50	54	157

Supplementary Table 4 - Comparing GM12878 and K562 ZNF274 BS. Chi-square test yields p-value = 0.51 since there are only 17 specific BS for GM12878

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
IMR90	TNF-a (10ng/mL) 1hr	p300	312	113	-25.05	[7]
IMR90	TNF-a (10ng/mL) 1hr	H3K4me3	487	104	-62.27	
IMR90	TNF-a (10ng/mL) 1hr	H3K36me3 (Abcam ab9050)	170	45	-19.06	
IMR90	TNF-a (10ng/mL) 1hr	PolII (Santa Cruz sc-899)	7613	1837	-300	
IMR90	TNF-a (10ng/mL) 1hr	flavopiridol (1ÂµM, 1hr)	9154	1655	-300	
HUVEC	TNF-a (10ng/mL) 1hr	H3K27ac (Abcam, ab4729)	5711	490	-300	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-AR	4380	2816	-45.48	[46]
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-AR	4277	2751	-44.34	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-AR	4611	3084	-38.47	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-AR	4437	2943	-38.89	
LNCAP	DHT (100nM, 2h)	anti-AR	5301	3527	-45.45	
LNCAP	DHT (100nM, 2h)	anti-AR	5223	3550	-39.4	
LNCAP	DHT (100nM, 2h)	anti-AR	5475	3698	-42.89	
LNCAP	DHT (100nM, 2h)	anti-AR	5356	3662	-38.9	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-AR	0	0	-0.02	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-AR	0	0	-0.02	

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-AR	0	0	-0.02	[46]
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-AR	0	0	-0.02	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-FOXA1	4197	3270	-10.03	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-FOXA1	3623	2858	-7.52	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-FOXA1	6278	4755	-19.99	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-FOXA1	6011	4524	-20.43	
LNCAP	DHT (100nM, 2h)	anti-FOXA1	181	157	-0.16	
LNCAP	DHT (100nM, 2h)	anti-FOXA1	146	138	-0.14	
LNCAP	DHT (100nM, 2h)	anti-FOXA1	6789	5540	-8.14	
LNCAP	DHT (100nM, 2h)	anti-FOXA1	6760	5385	-11.76	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-FOXA1	222	168	-1.1	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-FOXA1	189	147	-0.78	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-FOXA1	506	404	-1.23	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-FOXA1	502	402	-1.18	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	2444	1751	-13.24	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	2419	1749	-12.23	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	832	622	-3.54	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	843	637	-3.27	

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
LNCAP	DHT (100nM, 2h)	anti-PIAS3+PIAS1+PIAS2	1061	763	-5.96	[46]
LNCAP	DHT (100nM, 2h)	anti-PIAS3+PIAS1+PIAS2	1022	751	-4.89	
LNCAP	DHT (100nM, 2h)	anti-PIAS3+PIAS1+PIAS2	2556	1713	-21.48	
LNCAP	DHT (100nM, 2h)	anti-PIAS3+PIAS1+PIAS2	2478	1686	-19.05	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	97	69	-0.89	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	86	66	-0.5	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	880	460	-21.4	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-PIAS3+PIAS1+PIAS2	810	485	-12.41	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-p65	21	6	-1.93	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-p65	21	7	-1.66	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-p65	180	80	-7.14	
LNCAP	TNF-alpha (1000 U/ml, 2h)	anti-p65	168	76	-6.47	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-p65	2075	1004	-61.01	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-p65	2015	961	-61.49	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-p65	2127	1041	-60.68	

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	anti-p65	2008	1011	-53.12	[46]
HUVEC	TNF-alpha (10 ng/ml, 30min)	anti-Pol3	867	50	-155.12	GSE34500
HUVEC	TNF-alpha (10 ng/ml, 30min)	anti-p65	15779	1040	-300	
IMR90	DMSO	H3 ChIP	0	0	-0.02	[47]
IMR90	Nutlin-3a	H3 ChIP	1	0	-0.29	
IMR90	DMSO	H3K4me3 ChIP	15429	4242	-300	
IMR90	Nutlin-3a	H3K4me3 ChIP	14800	3896	-300	
IMR90	DMSO	H3K4me1 ChIP	13573	2528	-300	
IMR90	Nutlin-3a	H3K4me1 ChIP	1967	286	-300	
IMR90	DMSO	H3K27ac ChIP	15455	3236	-300	
IMR90	Nutlin-3a	H3K27ac ChIP	13167	2666	-300	
IMR90	DMSO	H4K16ac ChIP	3098	550	-300	
IMR90	Nutlin-3a	H4K16ac ChIP	1303	211	-192.97	
IMR90	DMSO	RNAPII ChIP	13161	2757	-300	
IMR90	Nutlin-3a	RNAPII ChIP	9070	1713	-300	
IMR90	DMSO	p53 ChIP	97	65	-2.62	
IMR90	Nutlin-3a	p53 ChIP	1637	734	-93.36	
IMR90	DMSO	H3K4me2 ChIP	27136	7113	-300	
IMR90	Nutlin-3a	H3K4me2 ChIP	25274	6516	-300	
MCF7	E2 for 45m	ERa	20417	8637	-300	[48]
MCF7	E2 for 45m	ERa	4268	1069	-300	
MCF7	IL1b for 45m	ERa	2813	911	-232.08	
MCF7	IL1b for 45m	ERa	62	7	-10.75	
MCF7	TNFa for 45m	ERa	5196	1538	-300	
MCF7	TNFa for 45m	ERa	45	15	-4.19	
MCF7	IKK7	ERa	1280	240	-166.56	
MCF7	IL1b+IKK7	ERa	19	6	-2.06	
MCF7	IKK7	ERa	57	42	-1.07	
MCF7	IL1b+IKK7	ERa	24	46	-1.78	
MCF7	E2+ICI	ERa	3739	540	-300	
MCF7	IL1b+ICI	ERa	1564	435	-151.48	[48]
MCF7	E2+ICI	ERa	4	2	-0.36	
MCF7	IL1b+ICI	ERa	3328	1129	-259.85	
MCF7	E2+ICI	ERa	587	114	-75.34	

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
MCF7	IL1b+ICI	ERa	4520	819	-300	[48]
MCF7	E2+ICI	ERa	25	79	-6.26	
MCF7	IL1b+ICI	ERa	1203	228	-155.61	
MCF7	E2 for 45m	p65	190	46	-21.49	
MCF7	E2 for 45m	p65	2	0	-0.53	
MCF7	IL1b for 45m	p65	244	50	-30.78	
MCF7	IL1b for 45m	p65	1560	322	-190.96	
MCF7	TNFa for 45m	p65	1534	437	-145.27	
MCF7	TNFa for 45m	p65	3175	1029	-261.56	
MCF7	Dexamethasone	GR E-20X sc-1003 Santa Cruz	8474	4131	-300	[49]
MCF7	Dexamethasone	GR E-20X sc-1003 Santa Cruz	6377	4410	-100.34	
MCF7	17ÃŸ-estradiol	ER cocktail: Ab-10 Thermo Scientific Lab Vision, HC-20 sc-543 Santa Cruz	4375	861	-300	
MCF7	17ÃŸ-estradiol	ER cocktail: Ab-10 Thermo Scientific Lab Vision, HC-20 sc-543 Santa Cruz	11153	4080	-300	
MCF7	Dexamethasone	FoxA1	6315	4018	-136.68	
MCF7	Dexamethasone	FoxA1	238	97	-15.29	
MCF7	17ÃŸ-estradiol	FoxA1	9178	7561	-53.4	
MCF7	17ÃŸ-estradiol	FoxA1	845	215	-89.05	
MCF7	Dexamethasone and 17ÃŸ-estradiol	FoxA1	112	916	-129.88	
T47D	Dexamethasone	GR E-20X sc-1003 Santa Cruz	570	91	-74.81	
T47D	Dexamethasone	GR E-20X sc-1003 Santa Cruz	451	85	-54.06	[49]
T47D	17ÃŸ-estradiol	ER cocktail: Ab-10 Thermo Scientific Lab Vision, HC-20 sc-543 Santa Cruz	2854	654	-293.27	
T47D	17ÃŸ-estradiol	ER cocktail: Ab-10 Thermo Scientific Lab Vision, HC-20 sc-543 Santa Cruz	3052	741	-299	
T47D	Dexamethasone	FoxA1	244	46	-29.61	
T47D	Dexamethasone	FoxA1	156	34	-17.44	

Cell line	treatment	antibody	induced in A	induced in B	log p-value	source
T47D	17ÃŸ-estradiol	FoxA1	303	42	-42.94	[49]
T47D	17ÃŸ-estradiol	FoxA1	3266	903	-283.95	
T47D	Dexamethasone and 17ÃŸ-estradiol	FoxA1	12949	2332	-300	
K562	IFNa30	pol2	739	73	-117.62	[50]
K562	IFNa6h	pol2	838	86	-131.65	
K562	IFNg30	pol2	1342	152	-203.26	
K562	IFNg6h	pol2	1113	116	-173.65	
K562	IFNa30	cjun	2788	409	-300	
K562	IFNa6h	cjun	922	112	-136.38	
K562	IFNg30	cjun	2518	331	-300	
K562	IFNg6h	cjun	1583	232	-215.33	
K562	IFNa30	cmyc	3188	287	-300	
K562	IFNa6h	cmyc	5150	489	-300	
K562	IFNg30	cmyc	21370	2329	-300	
K562	IFNg6h	cmyc	11871	1181	-300	
GM12878	TNF	NFKB	4952	610	-300	

Supplementary Table 5 – Summary of A/B distribution of TF induced binding sites for different cell lines under different treatments, as measured by ChIP-Seq. Datasets were downloaded from GEO and analyzed as described above.

Cell line	treatment	induced in A	induced in B	Log p-value	enrichment	source
GM12878	TNF-a	3866	267	-101.57	3.76	PRJNA30709
IMR90	TNF-a (10ng/mL) 1hr	439	103	-6.43	1.74	[7]
IMR90	TNF-a (10ng/mL) 1hr	105	23	-2.12	1.82	
IMR90	cycloheximide (5Âµg/mL) pretreat 30min	254	46	-6.61	2.23	
IMR90	TNF-a (10ng/mL) 1hr; cycloheximide (5Âµg/mL) pretreat 30min	389	77	-8.39	2.06	
IMR90	TNF-a (10ng/mL) 1hr	80	27	-0.37	1.19	
HUVEC	IFN-? (50ng/mL) 2hr	124	10	-3.65	3.02	
MCF7	ÃŸ-estradiol (100nM) 160min	121	13	-3.73	2.75	
IMR90	GSM1418973: PolyA+ RNAseq (DMSO); Homo sapiens; RNA-Seq	1733	555	-155.48	1.25	
IMR90	GSM1418974: PolyA+ RNAseq (Nutlin-3a); Homo sapiens; RNA-Seq	1771	558	-161.62	1.27	
MCF7	E2 for 3h	131	16	-3.49	2.45	[48]
MCF7	E2 for 3h	129	12	-4.51	3.16	
MCF7	IL1b for 3h	166	20	-4.4	2.51	
MCF7	IL1b for 3h	172	27	-3.07	1.95	
MCF7	IL1b+ICI for 3h	204	31	-3.79	2.02	
MCF7	IL1b+ICI for 3h	182	32	-2.56	1.75	
MCF7	TNFa for 3h	338	62	-3.99	1.7	
MCF7	TNFa for 3h	348	78	-2.12	1.4	
MCF7	TNFa+ICI for 3h	895	235	-1.82	1.2	
MCF7	TNFa+ICI for 3h	1045	382	-1.76	0.86	
MCF7	E2	252	31	-6.31	2.5	
MCF7	E2	421	35	-15.06	3.7	
MCF7	E2+TOT	511	62	-12.42	2.57	

Cell line	treatment	induced in A	induced in B	log p-value	enrichment	
MCF7	E2+TOT	62	2	-3.67	6.64	[48]
MCF7	E2+TOT+IL1b	469	62	-10.19	2.36	
MCF7	E2+TOT+IL1b	495	57	-12.87	2.7	[48] [46]
MCF7	E2+TOT+TNFa	471	50	-13.45	2.92	
MCF7	E2+TOT+TNFa	404	56	-8.26	2.25	
LNCAP	TNF-alpha (1000 U/ml, 2h)	148	47	-0.26	1.1	
LNCAP	TNF-alpha (1000 U/ml, 2h)	158	47	-0.49	1.18	
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	161	49	-0.42	1.15	[46]
LNCAP	DHT (100nM, 2h) + TNF-alpha (1000U/ml, 2h)	175	59	-0.11	1.04	
LNCAP	DHT (100nM, 2h)	64	8	-2.22	2.57	
LNCAP	DHT (100nM, 2h)	65	14	-0.94	1.57	

Supplementary Table 6 - Gene expression data sets. P-value is calculated using chi-square test between induced genes in each compartment and number of genes in each compartment (considering A is gene rich and B is gene poor)

	AA	AB	BA	BB	total	R	log10 p-value
HMEC GM12878							
Cell1_only_BSs	17789	6309	1708	5087	30893	4.48	<-300
Cell2_only_BSs	15292	567	3880	1779	21518		
Common_BSs	21558	1201	2319	3613	28691		
HUVEC GM12878							
Cell1_only_BSs	16635	6747	651	3805	27838	7.64	<-300
Cell2_only_BSs	16046	612	2900	1844	21402		
Common_BSs	20606	1173	966	2614	25359		
HUVEC HMEC							
Cell1_only_BSs	14992	4292	468	3168	22920	4.16	<-300
Cell2_only_BSs	16370	1369	3341	4301	25381		
Common_BSs	27220	2629	984	4031	34864		
K562 GM12878							
Cell1_only_BSs	21793	4294	1019	3916	31022	4.56	<-300
Cell2_only_BSs	14158	580	2994	1849	19581		
Common_BSs	24463	1239	1350	2762	29814		
K562 HMEC							
Cell1_only_BSs	21915	5083	982	3672	31652	4.15	<-300
Cell2_only_BSs	17015	1667	5907	5110	29699		
Common_BSs	23268	2498	1285	3435	30486		
K562 HUVEC							
Cell1_only_BSs	22645	3551	1140	3770	31106	5.37	<-300
Cell2_only_BSs	15827	627	5944	3772	26170		
Common_BSs	22627	1193	1300	2594	27714		
NHEK GM12878							
Cell1_only_BSs	19997	5949	1401	4154	31501	4.54	<-300
Cell2_only_BSs	14695	596	3111	1708	20110		
Common_BSs	21362	1192	2042	3023	27619		
NHEK HMEC							
Cell1_only_BSs	11680	1082	471	2402	15635	1.52	-26
Cell2_only_BSs	7559	688	678	2879	11804		
Common_BSs	42741	2511	2267	7437	54956		
NHEK HUVEC							
Cell1_only_BSs	18223	3364	1152	3363	26102	3.91	<-300
Cell2_only_BSs	14273	584	3426	3056	21339		
Common_BSs	26132	1241	2646	3560	33579		
NHEK K562							
Cell1_only_BSs	18782	5447	1599	3903	29731	3.46	<-300
Cell2_only_BSs	21127	1132	3998	3561	29818		
Common_BSs	23818	1357	2117	3026	30318		

Supplementary table 7 - H3k9ac sites in A and B compartments for all cell lines with available ChIP-Seq data

	AA	AB	BA	BB	total	R	p-value
GM12878 NHEK							
Cell1_only_BSs	7244	1099	2136	3306	13785	0.73	1.16E-18
Cell2_only_BSs	4369	1026	1215	3272	9882		
Common_BSs	350	71	62	288	771		
GM12878 K562							
MCF7 NHEK							
Cell1_only_BSs	5682	1460	4114	9381	20637	0.45	8.56E-113
Cell2_only_BSs	4902	1066	904	3355	10227		
Common_BSs	129	64	55	168	416		
MCF7 GM12878							
Cell1_only_BSs	5213	1751	3637	9712	20313	0.54	2.04E-122
Cell2_only_BSs	7176	1765	1145	3608	13694		
Common_BSs	318	95	118	317	848		

Supplementary table 8 - H3k27me3 sites in A and B compartments for three cell lines

תקציר

למבנה המרחבי של הגנום בגרעין יש תפקיד חשוב באספקטים רבים של חיי התא, לרבות בקרה על ביטוי גנים. לאחרונה פותחו מספר שיטות בהיקף גדול המאפשרות לחקור בדיוק חסר תקדים את הארכיטקטורה הכרומוזומלית ואת האינטראקציות בין מקטעי כרומטין.

אחת מהשיטות, Hi-C, מבוססת על מדידת הקונפורמציה של הכרומוזומים (chromosome conformation capture). בעזרת שיטה זו אופיינו מספר מבנים בגנום. ברזולוציה נמוכה, כל כרומוזום מחולק למקטעים משני סוגים (compartments) – סוג A (עשיר בגנים) וסוג B (דל בגנים). ניתן לחלק כל מקטע לתת-מדורים על בסיס סמנים אפיגנטיים. ברזולוציה יותר גבוהה, זוהו תחומים בעלי קשר טופולוגי (topological associated domains, TADs). התחומים מוגדרים על ידי אזורים כרומוזומליים המכילים את רוב האינטראקציות המרחביות. שיטה נוספת, ChIA-PET, המבוססת על מדידת הקונפורמציה הדינמית של הכרומוזום, מאפשרת למדוד אינטראקציות ברזולוציה גבוהה יותר בתוך תחומים אלו.

מדידת רמת השעתוק של גנים בתגובה לתנאי עקה מגוונים בסוגי תאים שונים מדגימה שחלק משמעותי מתגובת התא אופייני ספציפית לסוג התא ורק חלק קטן ממנה הוא כללי. מטרתנו היא לבחון עד כמה התגובה האופיינית לעקה מושפעת מהארגון המרחבי של הגנום בסוג התא במצב הבזאלי.

לצורך כך, ביצענו אנליזה רחבה של 13 סוגי תאים שונים תחת טיפולים ומדידות שונות (RNA-Seq, GRO-Seq, ChIP-Seq וכו'), ובדקנו האם התגובה שנמדדה בניסויים קורלטיבית לארגון המרחבי הראשוני של הגנום בתא. התוצאות מצביעות על כך שישנו קשר מובהק בין התגובה האופיינית של תא לעקה לבין אילוצים מבניים בתא במצב הבזאלי.

אוניברסיטת תל אביב

הפקולטה למדעים מדויקים

ע"ש ריימונד ובברלי סאקלר

אינטראקציות מרחביות בגנום והקשר שלהן לביטוי גנים

חיבור זה הוגש כחלק מהדרישות לקבלת התואר

"מוסמך אוניברסיטה" – M.Sc. באוניברסיטת תל-אביב

ביה"ס למדעי המחשב

על ידי

עידן נוריק

העבודה הוכנה בהדרכתם של

פרופ' רון שמיר

דר' רן אלקון

אוגוסט 2017