
Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

# Genomic analysis of the spatial organization of the genome and its effect on cell type-specific p53 transcriptional responses

Thesis submitted in partial fulfillment of graduate requirements for

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

School of Computer Science

By

Hadar Amira Haham

Prepared under the supervision of

Prof. Ron Shamir

Prof. Ran Elkon

Feb 2023

## 1. Acknowledgments

I would like to extend my sincere thanks to the people who helped me make this thesis become a reality.

First and foremost, I want to thank from the bottom of my heart my outstanding supervisor Prof. Ron Shamir for supporting me on this journey. I feel fortunate to have been mentored during my work by an extraordinary researcher and rare person like Ron. From you I learned how to perform excellent science thoroughly, professionally and with never-ending dedication to the profession. I would like to thank Prof. Rani Elkon, for showing me how deep and wonderful the world of biology is, and for his creativity and broadening the horizons of thinking, planting peace and calm in the difficult moments of research.

Thank you, Ron and Rani, for believing in me and giving me the opportunity to learn from you in recent years.

Secondly, special thanks to Gony Shanel, who collaborated with me in this research and helped me to make it happen. You are great.

Thirdly, I would like to thank my collaborator Dr. Tsung-Han Stanley Hsieh from Prof. Xavier Darzacq lab, Department of Molecular and Cell Biology, University of California, Berkeley for creating and providing us with the unique datasets on which this study is based.

A great thanks to Naama, Tom and Hagai for the fruitful scientific discussions and for being amazing friends and always supportive. it wouldn't have looked the same without you.

I would like to deeply thank my friends from the ACGT group: Dr. Lianrong Pu, David Pellow, Tom Hait, Nimrod Rappoport, Dan Coster, Hagai Levi, Naama Kadosh, Omer Noy, Yonatan Itai,

Dan Flomin, Eran Shpigelman, Ron Saad and Maya Metzger, for the helpful discussions and for being great friends.

Additionally, I owe special thanks to Gilit Zohar-Oren for the administrative help that was always done with a smile and kindness. You are the best.

I deeply thank for the financial support I was granted during my studies: the Edmond J. Safra Center for Bioinformatics at Tel Aviv University, Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics, The Israel Science Foundation (grant 1339/18 and grant 3165/19 within the Israel Precision Medicine Partnership program) and German-Israeli Project Cooperation DFG-DIP RE 4193/1–1.

I wish to deeply thank to my parents Mazal and Moshe Amira who always supported and believed in me, instilled in me the instinct of curiosity from childhood and always encouraged me to keep learning.

Last but not least, my precious love Gil for your endless understanding and patience, for long nights of debugging and resolving problems with creative ideas, and for being my partner for life. I wouldn't have done it without you all.

## 2. Abstract

Many studies have observed that transcriptional responses to multiple stresses are highly cell type-specific. However, the mechanisms that underlie this tissue specificity remain largely elusive. In our study, we focus, as a model system, on the transcriptional networks activated by p53, and examine possible associations between cell type-specific genome 3D organization and cell type-specific transcriptional responses.

p53, known as the "guardian of the genome", is the major tumor suppressor gene in our genome, and it serves as a pivotal defense mechanism against cancer transformation. p53 activation in different cell types and tissues results in induction of very different transcriptional networks, alongside the activation of a universal p53 core response. The main goal of our research was to characterize cell type-specific responses to p53 activation and examine possible links with 3D genome organization. Our research utilized three layers of omics techniques: RNA-seq, ChIP-seq and Micro-C (an improved version of Hi-C, with enhanced resolution), that were applied to ten different cell lines. For each cell line, measurements were taken both in control conditions and after treatment by Nutlin-3a, a potent p53 activator.

In the analysis of these extensive datasets, we identified (1) dozens of cell type-specific p53-chromatin binding events; (2) cell type-specific p53 cofactors; (3) cell type-specific p53 binding events correlated with cell type-specific p53-induced gene expression, and (4) cell type-specific enhancer-promoter physical interactions. We specifically tested correlations between cell type-specific p53-induced responses and cell type-specific features of the spatial organization of the genome. Interestingly, we found that in contrast to differential expression between cells of different tissues of origin, which are strongly associated with difference in the spatial

organization of the genome, transcriptional changes in response to p53 activation do not show a strong link with corresponding spatial genomic alterations.

## 3. Table of Contents

## 4. Introduction

In order to understand the functioning of the human genome, it is not enough to consider the primary linear DNA sequence. Rather, a full understanding of genome function requires investigation and understanding of the three-dimensional (3D) folding and spatial organization of chromosomes in the nucleus [1].

The entire genome appears in the nucleus of every cell in our body, packed in the form of 23 pairs of chromosomes: 22 pairs of autosomes and a pair of sex chromosomes. Each pair of chromosomes includes a chromosome that originates from each parent. The total length of the human genome is about 3 billion bases (nucleotides) [2]. The number of protein-coding genes in the human genome is about 20,000 [3]. In 2021, a complete version of the human genome was published, covering the whole genome without any deficiencies. The name of the full genome is called T2T-CHM13. This version replaces the current genome GRCh38[4], on which our study was conducted.

In this thesis we sought to further investigate the relationship between the three-dimensional spatial structure of the genome and its responsiveness to states associated with p53 activation.

Our aim was to compare two contrasting models:

1. Induction of stress causes major changes in the spatial organization of the genome, among them – rearrangements of enhancer-promoter loops that precede the transcriptional induction of certain stress-induced genes.

2. Induction of stress does not result in a major change in the spatial genome organization. That is, the spatial structure largely remains as it was under basal conditions. That is, transcriptional changes in response to stress are not accompanied by gross 3D changes.

To investigate the relationship between genome organization and gene expression we analyzed a very large-scale data set containing ten cell lines, each profiled before and after Nutlin-3a (hereafter referred to as Nutlin) treatment, using three different techniques:

1. RNA-seq, which measures the expression of genes in the sample.

2. ChIP-seq, a method used to profile protein interactions with chromatin.

3. Micro-C, an improved method of Hi-C with enhanced resolution. It is used to analyze physical interaction between any two genomic loci, and thus, infer the 3D organization of the genome.

# 5. Background

This chapter provides the background and terminology required for the thesis. First, we present the relevant biological basis that includes concepts in gene regulation, high throughput sequencing methods, the relationship between structure and function, and the p53 gene and its biological importance. Next, we describe the high throughput methods used in this thesis, including a discussion of biases in these methods and the way they are treated. In addition, a computational background is given on the various computational methods we used, such as unsupervised learning methods, including clustering and PCA, a comprehensive explanation of the Hi-C method, as well as the statistical tests used in this thesis.

## 5.1 Biological Background

In this section we introduce biological concepts and definitions that are needed for understanding the motivation of this thesis, and the computational problems that we deal with.

### 5.1.1 Biological Concepts

#### 5.1.1.1 Structure-Function Relationship

In biology, a key idea is that structure determines function. The way in which a biological unit is arranged in space allows it to perform a specific task. We see this at all levels in the hierarchy of biological organization from atoms up to the biosphere.

#### 5.1.1.2 Chromatin organization

Certain proteins compact chromosomal DNA into the microscopic space of the eukaryotic nucleus. These proteins are called histones, and the resulting DNA-protein complex is called chromatin. Within the nucleus, histones provide the energy (mainly in the form of electrostatic interactions) to fold DNA. As a result, chromatin can be effectively packed into a very small volume.

Histones are a family of small, positively charged proteins termed H1, H2A, H2B, H3, and H4 (Van Holde, 1988). DNA is negatively charged, due to the phosphate groups in its phosphate-sugar backbone, so histones bind with DNA very tightly. The basic unit of organization of chromatin is the nucleosome, a structure of DNA and histone proteins that repeats itself throughout an organism's genetic material. Approximately 150 bp of DNA wrap around this protein structure almost twice to make a nucleosome core particle. With linker histone (e.g., histone H1) and linker DNA, this is called the nucleosome. The linker DNA can vary in length, usually between 10 to 90 bp, depending on the species, gene activity, developmental stage, and other factors [5]. High levels of DNA packing enable the final dense structure of the chromosome (Figure 1).

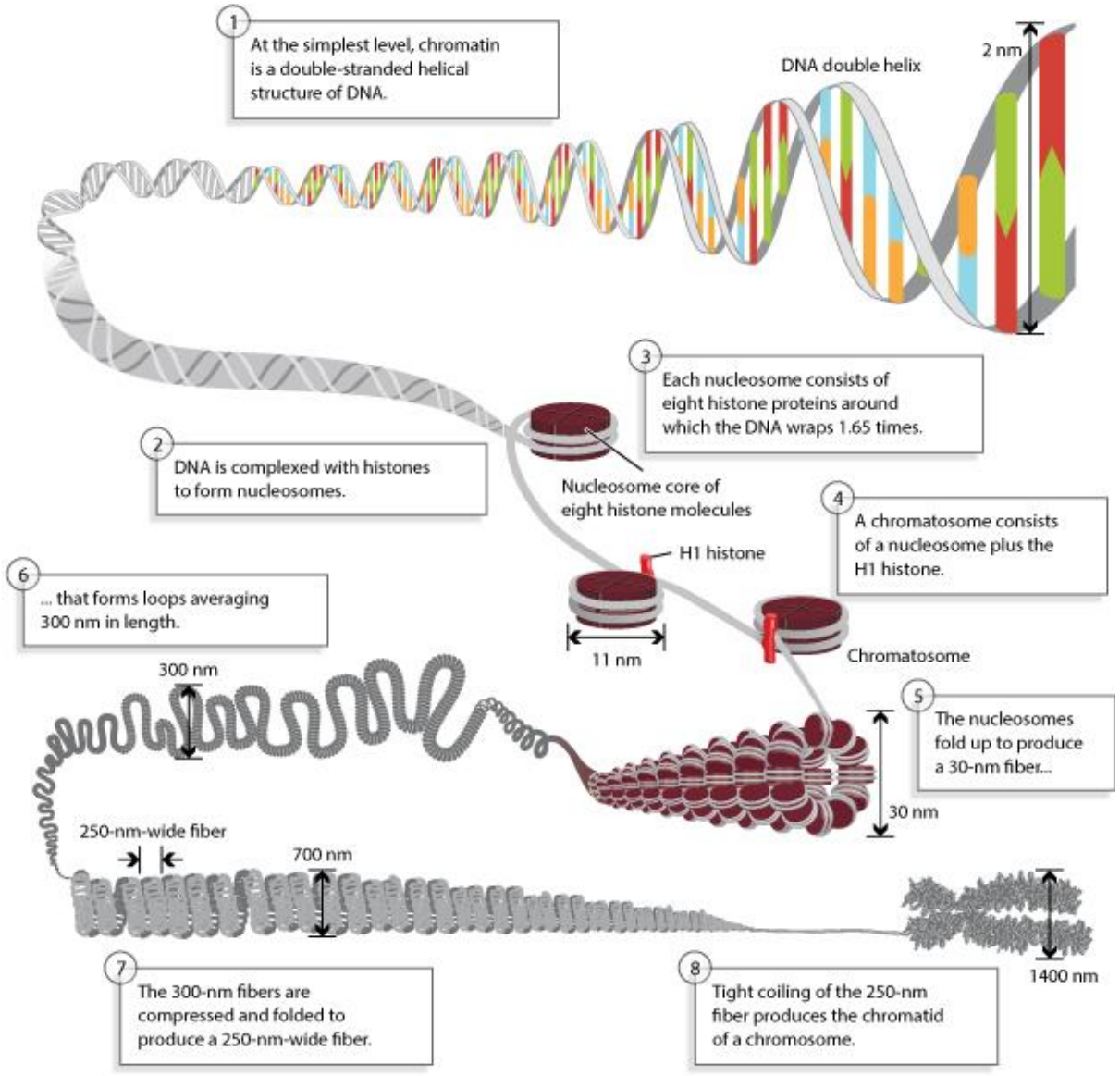


Figure 1. | Chromosomes are composed of DNA tightly wound around histones. Chromosomal DNA is packaged inside microscopic nuclei with the help of histones. These are positively-charged proteins that strongly adhere to negatively-charged DNA and form complexes called nucleosomes. Each nucleosome is composed of DNA wound 1.65 times around eight histone proteins. Nucleosomes fold up to form a 30-nanometer chromatin fiber, which forms loops averaging 300 nanometers in length. The 300 nm fibers are compressed and folded to produce a 250 nm-wide fiber, which is tightly coiled into the chromatid of a chromosome. Source: Nature Education Adapted from Pierce, Benjamin. Genetics: A Conceptual Approach, 2nd ed. 2013. All rights reserved.

#### 5.1.1.3 Gene regulation

Gene regulation is the process used to control when, where and to what level genes are expressed. The process can be complicated and is carried out by a variety of mechanisms, including regulatory proteins and chemical modification of DNA. Gene regulation is key to the ability of an organism to respond to environmental changes.

Gene regulation is one of the fundamental processes that a cell carries out in order to produce the transcripts that will be translated into proteins. A lot of the cell's energy is devoted to fine-tune its gene regulation in the context of development, response to stress or other conditions.

#### 5.1.1.4 Transcription factors

Transcription factors are proteins involved in the process of converting, or transcribing, DNA into RNA. A generic component of transcription is RNA polymerase, which initiates and performs the transcription of genes. One distinct feature of transcription factors is that they have DNA-binding domains that give them the ability to bind to specific sequences of DNA called transcription factor binding sites (TFBSs). Some transcription factors mainly bind to promoter regions proximal to the transcription start site (TSS) and help form the transcription initiation complex. Other transcription factors mainly bind to distal regulatory sequences, such as enhancer sequences, and can either stimulate or repress transcription of the target gene. These regulatory sequences can be many thousands of base pairs upstream or downstream from the gene they control. Regulation of transcription is the principal layer of gene control. The action of transcription factors allows for unique expression of each gene in different cell types and during development [6].

## 5.2 Next-Generation Sequencing

Next-generation sequencing (NGS) is a general name for new sequencing techniques developed over the last two decades. NGS performs deep high-throughput sequencing in a short time that can provide hundreds of millions of short sequences (e.g., 150 bases paired end). NGS has revolutionized genomic research in terms of time and cost needed to generate sequence data compared to the previous Sanger sequencing technology [7] used in the original Human Genome Project.

### 5.2.1 RNA-seq

RNA sequencing (RNA-seq) is a technique that uses NGS to quantify the expression level of all transcripts in a biological sample at a given time point, analyzing the modulation of the cellular transcriptome. Specifically, RNA-seq facilitates the ability to look at changes in gene expression over time, or differences in gene expression among different groups or treatments. The RNA-seq workflow is described below in Figure 2.

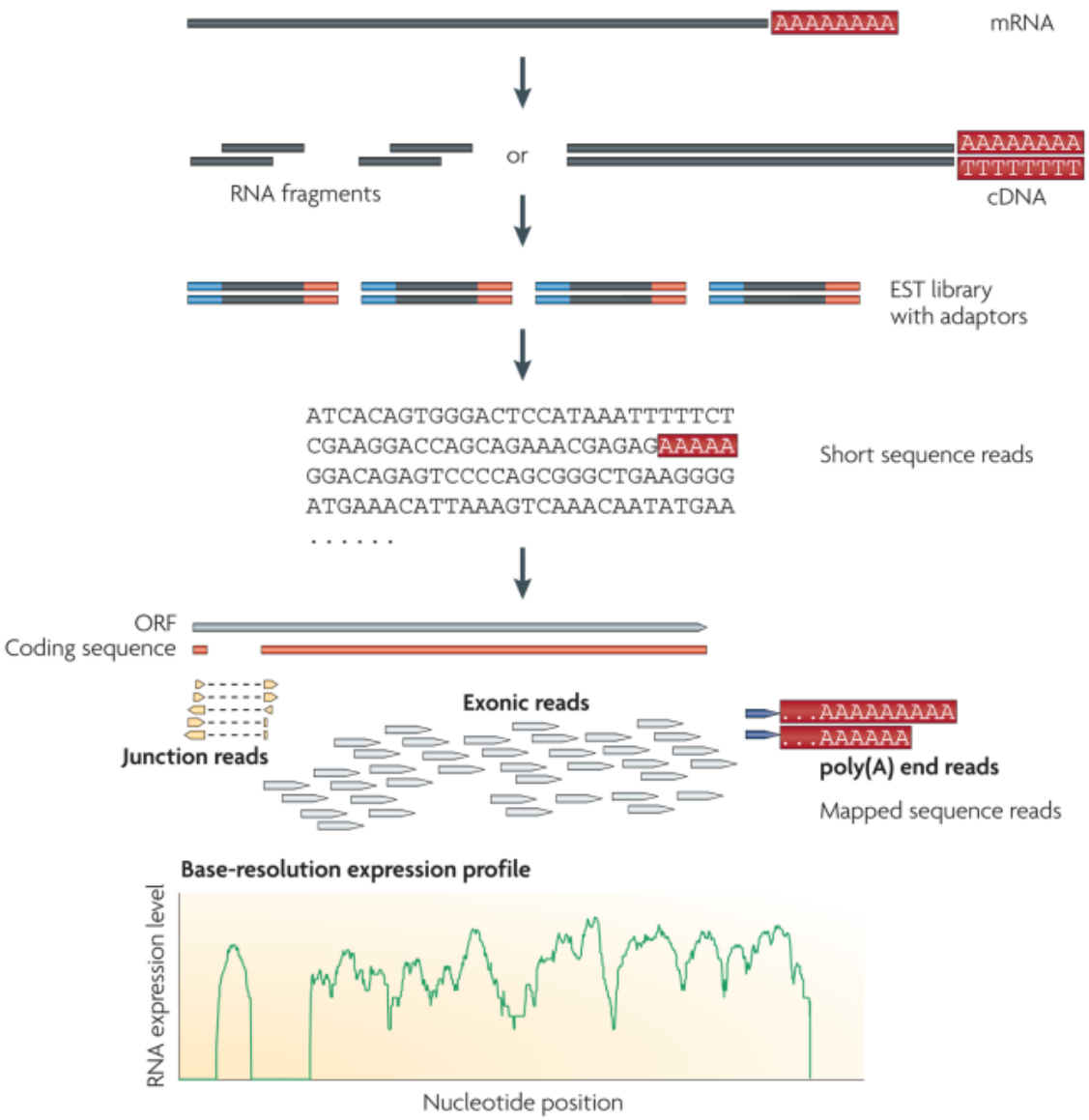


Figure 2. | RNA-seq workflow: long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned to the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown. Source: [8].

### 5.2.2 ChIP-seq

ChIP-sequencing (ChIP-seq) is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. The ChIP-seq workflow is described below in Figure 3.

ChIP-seq data can also identify histone markers that characterize different chromatin states. Histone modifications are roughly divided into two groups, which characterize open (transcriptionally active) and closed (transcriptionally repressed) chromatin states.

Identification of DNA-protein binding sites from ChIP-seq reads count data requires computational tools that perform peak calling (Figure 4). The most popular method at present is MACS [9], [10], which finds genomic intervals that are statistically enriched for reads, compared to the background read coverage in their local genomic neighborhood.

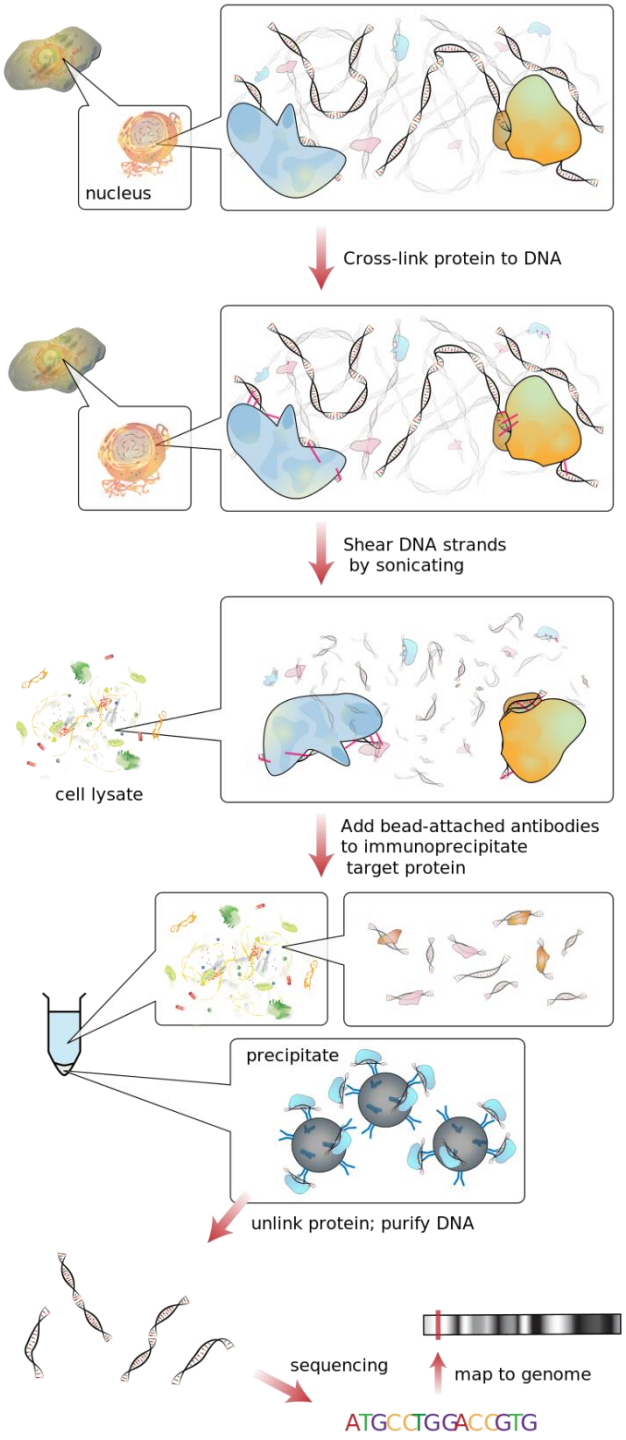


Figure 3. | ChIP-seq workflow: First, the DNA is extracted from the nucleus and cross linked to the protein to prevent detaching during the sonication process. Second, the DNA is sheared and fragmented by sonication. Third, a protein antibody is attached to the protein of interest. Forth, the antibody is precipitated and selects only those DNA fragments attached to the protein of interest. Finally, the proteins are removed from the DNA segments, and the segments are then sequenced and mapped to a reference genome. Source: https://en.wikipedia.org/wiki/ChIP_sequencing.

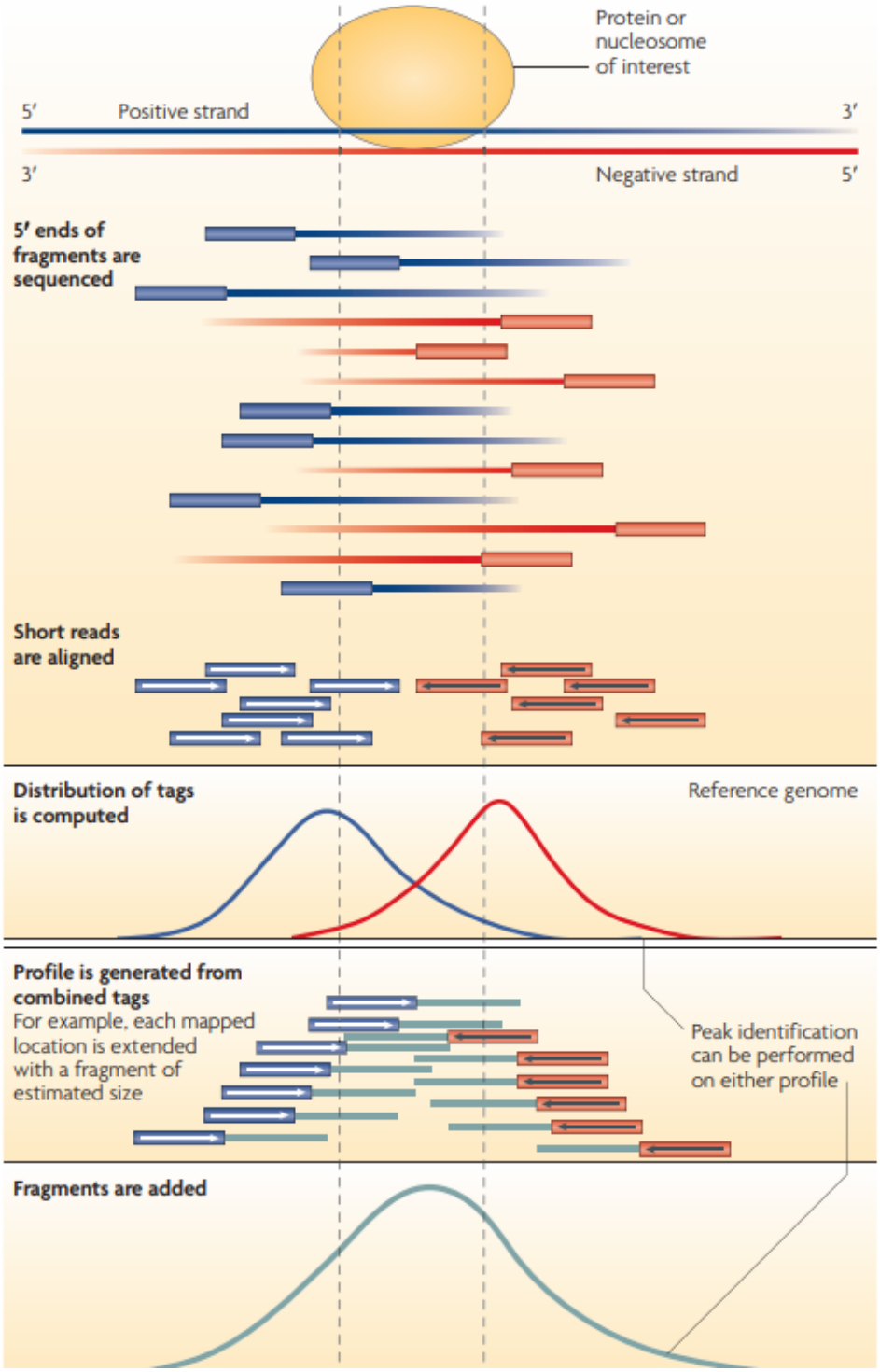


Figure 4. | ChIP-seq peak calling: DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5′ end. Therefore, the alignment of these tags to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest. This strand-specific pattern can be used for the optimal detection of enriched regions. To create an approximate distribution of all fragments, each tag location can be extended by an estimated fragment size in the appropriate orientation and the number of fragments can be counted at each position. Source: [10].

## 5.3 The Hi-C Technique and Chromosome Conformation

In this section we will describe methods used to study how DNA is organized within the nucleus. First, we describe Hi-C, a method for capturing chromosome conformation by using high-throughput sequencing, developed by Lieberman-Aiden et al [11]. Next, we describe the Micro-C method [12], introduced by Hsieh et al., our collaborators in this current project, which is an improvement of the Hi-C method with enhanced resolution. Micro-C allows the detection of chromosomal interactions at the nucleosome level. Last, we present some key concepts in the field of Hi-C.

### 5.3.1 Hi-C

Hi-C gives information on the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. Hi-C allows unbiased identification of chromatin interactions across the entire genome.

Briefly (Figure 5), cells are crosslinked with formaldehyde; DNA is digested with a restriction enzyme that leaves a 5′ overhang; the 5′ overhang is filled, including a biotinylated residue; and the resulting blunt-end fragments are ligated under dilute conditions that favor ligation events between the cross-linked DNA fragments. The resulting DNA sample contains ligation products consisting of fragments that were originally in close spatial proximity in the nucleus, marked with biotin at the junction. A Hi-C library is created by shearing the DNA and selecting the biotin-containing fragments with streptavidin beads. The library is then analyzed by using massively parallel DNA sequencing, producing a catalog of interacting fragments.

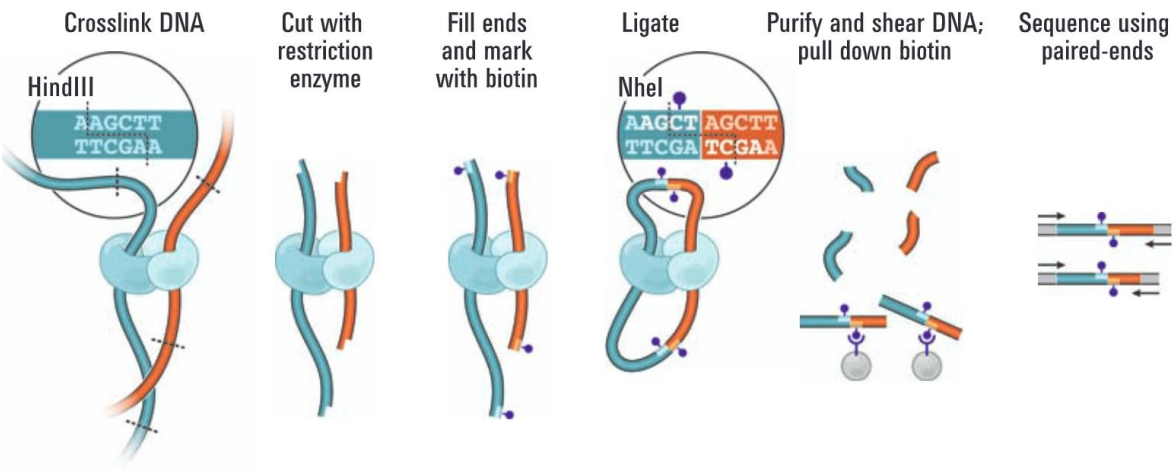


Figure 5. | Hi-C workflow. Source: [11].

After alignment of the sequence reads to the reference genome, a genome-wide contact matrix $M$ is constructed by dividing the genome into, typically, 1-Mb regions ("loci"); the matrix entry $M_{i,j}$ is the number of ligation products between locus $i$ and locus $j$. This matrix reflects an ensemble average of the interactions present in the original pool of cells; it can be visually represented as a heatmap, with intensity indicating contact frequency (Figure 6).

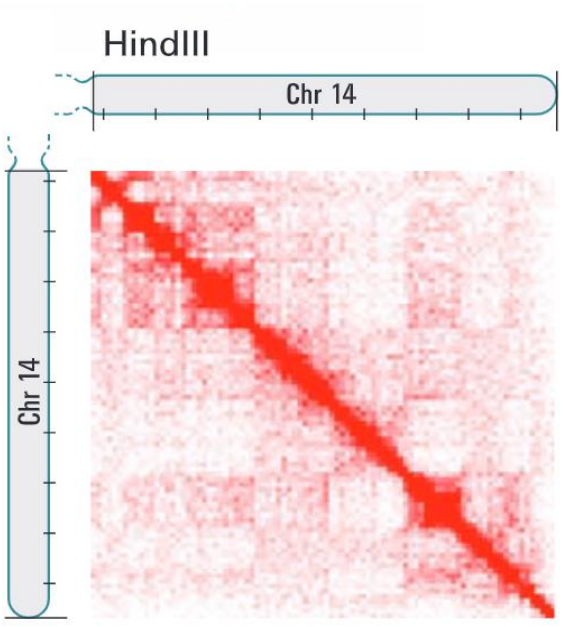


Figure 6. | Hi-C contact matrix. This is the result of Hi-C workflow for chromosome 14. Each pixel represents the total number of interactions observed between two 1-Mb chromosomal intervals. Color intensity corresponds to the total number of interactions. Tick marks appear every 10 Mb. Source: [11].

The theoretical resolution limit of Hi-C is determined by the restriction enzyme used by the protocol (~2-3k bps), although the practical resolution is determined by the sequencing depth and is typically much poorer (~100k – 1M bps).

The contact matrix should be normalized into a new matrix $\boldsymbol{M}^*$. This can be done, for example, by dividing each entry in the contact matrix by the genome-wide average number of contacts for loci at that genomic distance. More advanced normalization methods include:

- Iterative correction and eigenvector decomposition (ICE) [13]. This algorithm works by iteratively correcting for various sources of bias in the Hi-C data. In the first step, it corrects for systematic biases such as GC content and mappability. In the second step, it estimates the contact frequencies between all pairs of genomic bins, and then computes the eigenvalues and eigenvectors of this matrix of contact frequencies. By decomposing the matrix of contact frequencies into eigenvalues and eigenvectors, the ICE algorithm is able to identify patterns in the Hi-C data that are not explained by systematic biases. Finally, this algorithm uses these patterns to normalize the Hi-C data and adjust for any remaining noise. The final output of the normalization process is a matrix of corrected contact frequencies between different regions of the genome, which can be used for downstream analysis. [13]
- The Knight-Ruiz (KR) method [14] is a fast method for balancing the row and column sums of the matrix of contact frequencies, i.e., reweighting rows and columns so that the sum of each row and column equals one. The KR method works by iteratively scaling the rows or columns of the matrix using the conjugate gradient method. This method was demonstrated to handle various sources of bias and noise in HiC data. [14]

In the next step, a correlation matrix $\boldsymbol{C}$ is calculated, in which $\boldsymbol{C}_{i,j}$ is the Pearson correlation between the $I^{th}$ and $J^{th}$ columns of $M^*$. Last, PCA is performed on the correlation matrix $\boldsymbol{C}$. PCA seeks to

maximize the variance captured by reduced dimensions. In Hi-C data, the first PC usually represents the division of the genome into two highest-level types of structural units, called A/B compartments, that correspond to genomic loci that are overall transcriptionally active and inactive, respectively. Figure 7 summarizes the consequential process.

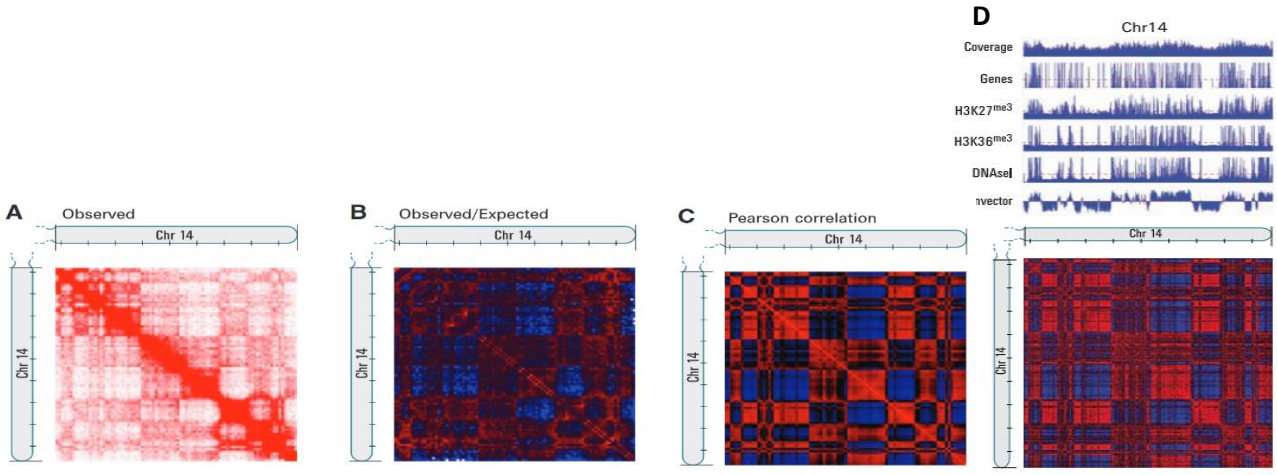


Figure 7. | Hi-C contact matrix of chromosome 14 in different stages of Hi-C analysis workflow. A. Contact matrix. Map of chromosome 14 at a resolution of 1 Mb, Tick marks appear every 10 Mb. B. Normalized contact matrix. The observed/expected matrix shows loci with either more (red) or less (blue) interactions than would be expected, given their genomic distance. C. Correlation matrix calculated on B, showing the correlation [range from –1 (blue) to +1 (red)] between the profiles of every pair of 1-Mb loci along chromosome 14. The plaid pattern indicates the presence of two compartments within the chromosome. D. Correlation map of chromosome 14 at a resolution of 100 kb. The tracks above show, from top to bottom, the read coverage, gene locations, three epigenetic tracks and the first eigenvector of the correlation matrix, corresponding to A/B compartments in the correlation figure. Source: [11].

## 5.4 Micro-C

Micro-C is a Hi-C-based method, in which micrococcal nuclease (MNase) is used instead of restriction enzymes to fragment the chromatin, thereby enabling nucleosome resolution (~150 bps) chromosome folding maps.

This protocol is based on the Hi-C protocol [11], with key alterations being the MNase digestion step, subsequent mononucleosomal end repair, and a modified two-step method for specifically purifying ligation products. After purification of ligation products between mononucleosomes, paired-end deep sequencing is used to characterize the ligation products. Figure 8 shows the workflow of the method and Figure 9 shows the resulting interaction matrix.

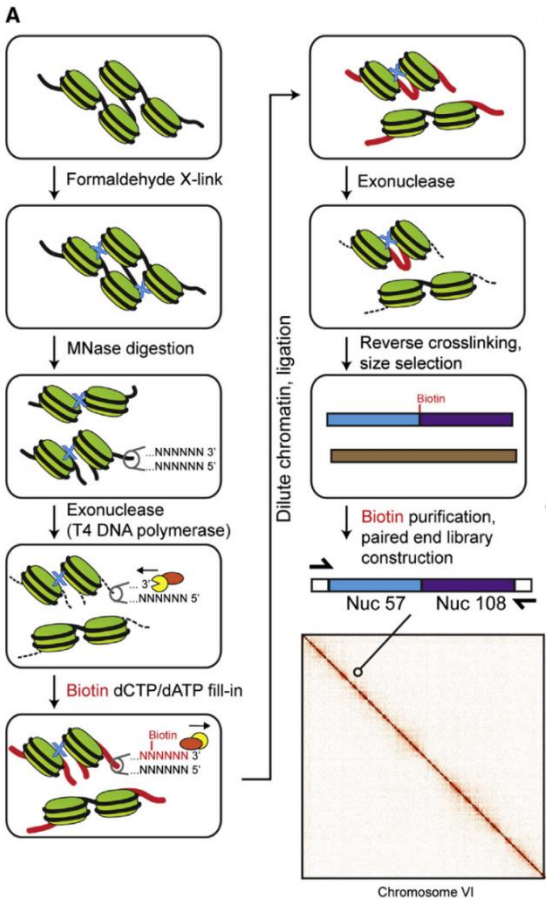


Figure 8. | Micro-C Workflow. Source: [12].

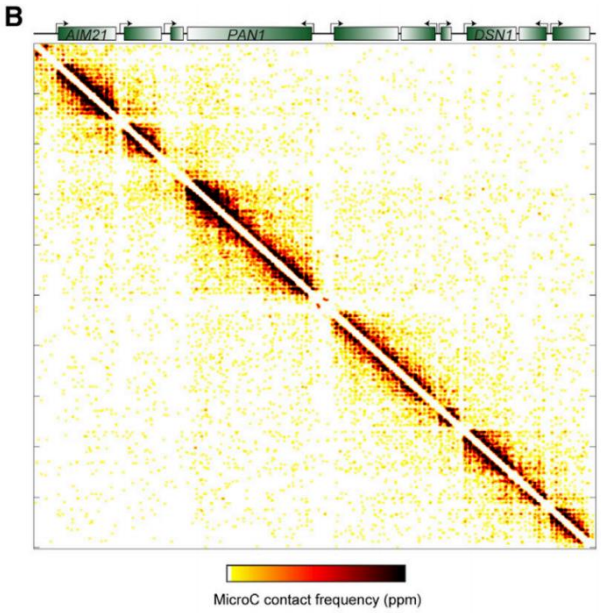


Figure 9. | Nucleosome-nucleosome interaction matrix. Zoom-in on a 20 kb X 20 kb submatrix from chromosome 9 (360,001–380,000), with Micro-C interactions represented in white-yellow-red-black heat map showing the interaction intensity between pairs of loci. Source: [12].

### 5.4.1 Key Concepts in Hi-C Data

**Hi-C resolution:**

The average size of DNA fragments that are created by Hi-C when chromosomes are cut by restriction enzyme.

**A/B compartments:**

The highest-level organization of the genome, observable even under the microscope, is into two types of compartments, heterochromatin and euchromatin. Heterochromatin is typically highly condensed, gene-poor, and transcriptionally silent, whereas euchromatin is less condensed, gene-rich, and more accessible to transcription. The typical size of a contiguous compartment is a few mega base pairs. In Hi-C, the terms compartment A and B correspond to euchromatin and heterochromatin, respectively. Compartments tend to vary among cell types

and tissues. A/B compartmentalization is outlined by PCA analysis of the Hi-C correlation matrix (usually, by the main PC – PC1, See Figure 7D).

## Topologically Associated Domains:

Topologically Associated Domains (TADs) are self-interacting genomic regions. DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD (Figure 10). TADs are also known as CID – Chromosomal Interaction Domains in other species (not human). The typical size of TADs is of ~0.5 Mb, containing 1-5 genes. TADs are largely conserved between cell types and species [12], [15].

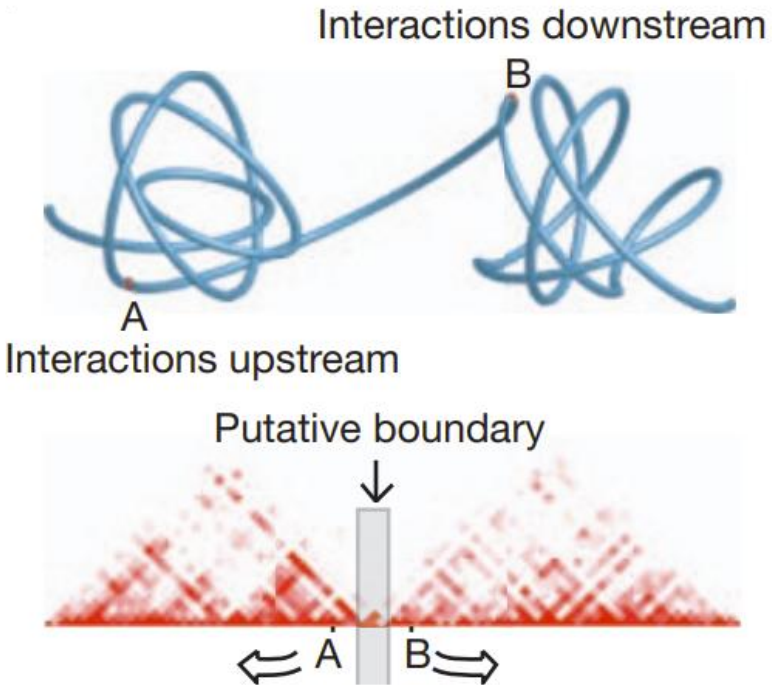


Figure 10. | Schematic illustrations of TADs. The top figure shows two adjacent DNA segments that are compacted into two separate dense parts, denoted A and B. As a result, there are many interactions among segments in A, and similarly in B, but few interactions between A and B. The intensity of interactions is shown at the bottom, where the upper diagonal of the corresponding square of the contact matrix is rotated 45 degrees. The red triangles correspond to TADs A and B with the putative boundary (gray rectangle) between them. There are very few A-B interactions. Source: [15].

Loops:

Chromatin loops (Figure 11) are defined as pairs of genomic sites that lie far apart along the linear genome but are brought into spatial proximity through chromatin folding maintained by a cohesin unit. The two ends of the loops are called anchors and they delimit the loop.

One of the loop's anchors commonly contain a gene's promoter whose activity is induced when the loop is created. Two common types of regulatory loops are promoter-promoter loops and enhancer-promoter loops (which increase the expression of a gene).

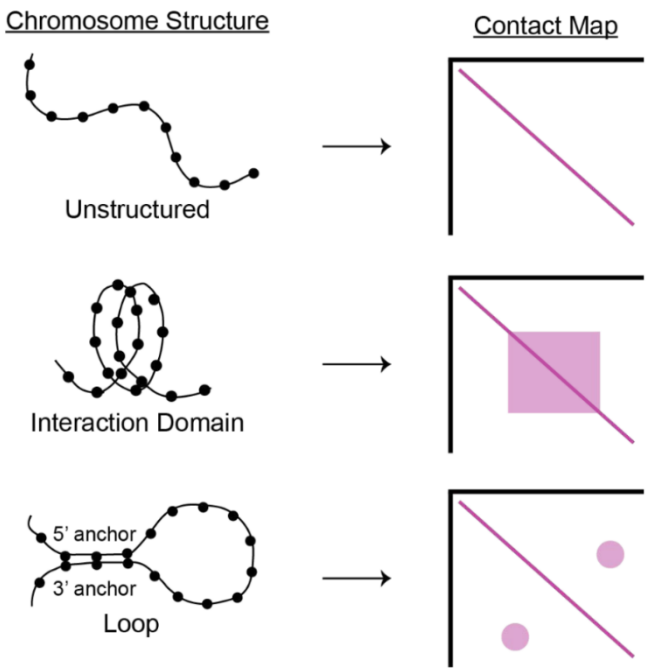


Figure 11. | Illustration of different chromosomal structures and their corresponding signal on the contact map. Unstructured chromosomes: only neighboring nucleosomes will be crosslinked producing a linear signal along the diagonal. Interaction Domain: the nucleosomes within a domain will be also crosslinked forming a square along the diagonal. Loop: the nucleosomes at the base of the loop (5' and 3' anchors) will be crosslinked forming a spot away from the diagonal. These different structures can form concomitantly on chromosomes producing contact maps with squares and spots along the diagonal. Source: [16].

### 5.4.2 The p53 Gene

p53 is a tumor suppressor protein that promotes apoptosis, DNA repair or cell-cycle arrest in response to DNA damage and other cellular stresses. It thus serves as a cell defense mechanism. The protein encoded by this gene is a transcription factor that regulates the expression of dozens of target genes [17]. It is the most frequently mutated gene in human cancer [18]. P53 is usually found in the cytoplasm in an inactive state, inhibited by the MDM2 protein, but in response to genotoxic stress, the p53 protein is activated by several kinases, such as the ATM protein kinase. After its activation, p53 enters the cell's nucleus where it induces the transcription of numerous targets, key among them is the gene encoding the CDKN1A (p21) protein [19].

Nutlin-3 is a molecule that occupies p53 binding site of MDM2 and effectively disrupts the p53–MDM2 interaction, which leads to activation of the p53 pathway in p53 wild-type cells. Inhibiting the interaction between MDM2 and p53 stabilizes p53 and is thought to selectively induce a growth-inhibiting state called senescence in cancer cells. Nutlin is thus used as an effective p53 activator. Nutlin-3 has been shown to affect the production of p53 within minutes. It is the compound that is used for p53 activation in our study.

# 6. Materials and Methods

## 6.1 Data sets

The data analyzed in my research consists of 10 human cell lines with a p53 status as shown in Table 1.

For each cell line, measurements were taken in control cells as well as in cells treated with Nutlin, a potent p53 activator, with two replicates per condition.

To characterize the change in the spatial structure and transcriptome associated with p53 activation, three different omics techniques were applied to each treated and untreated cell line: RNA-seq, ChIP-seq, and Micro-C. All data were generated by our collaborator Dr. Tsung-Han S. Hsieh from the lab of Prof. Darzacq at the Department of Molecular and Cell Biology, University of California, Berkeley.

| Cell Line | A549 | GM12878 | HCT116 | HEK293 | HeLa | HepG2 | MCF7 | IMR90 | SKNSH | U2OS |
|---|---|---|---|---|---|---|---|---|---|---|
| Tissue | lung | blood | colon | kidney | cervix | liver | breast | lung | brain | bone |
| Type | cancer | normal | cancer | immortal | cancer | cancer | cancer | normal | cancer | cancer |
| P53 status | WT | WT | WT | Mutated | Absent | WT | WT | WT | WT | WT |

Table 1. | Properties of the different cell types used in this study. Most of cell types are p53 wild type profile, while HeLa and HEK293 cells are with non-functional p53.

## 6.2 Methods for analysis of Hi-C data

For the analysis of Hi-C data we used the following methods and tools:

### 6.2.1 Juicer

Juicer is a unified pipeline for processing tera-base scale Hi-C datasets. It enables processing raw fastq files to create Hi-C maps binned at many resolutions, and automatically annotates loops and contact domains [20]. We mainly used this tool to annotate TADs in 10 kbp resolution.

### 6.2.2 Mustache

Chromatin loops are defined as pairs of genomic sites that lie far apart along the linear genome but are brought into spatial proximity by a mechanism called loop extrusion [21]. Mustache is a local enrichment-based method for high-resolution Hi-C and Micro-C data. It uses scale-space representation from computer vision to identify "blob-shaped" objects in the contact map and this way identifies chromatin loops at multiple resolutions. We used this tool mainly to annotate chromatin loops in 5 kbp resolution.

### 6.2.3 HiCDC+

HiC-DC+ estimates significant changes in interactions between two conditions measured by Hi-C or HiChIP experiments. It works on the raw contact matrix and analyses interactions for each chromosome up to a specified genomic distance, binned by uniform genomic intervals or restriction enzyme fragments. It trains a background model to account for random polymer ligation and systematic sources of read count variation [22].

This tool was mainly used for identifying significant changes in chromatin loop interactions between two conditions in 5 kbp resolution.

### 6.2.4 FANC

FAN-C is a command-line tool and Python API for matrix generation, analysis, and visualization on Hi-C data. FAN-C also includes a basic genome browser utility that allows for interactive exploration of Hi-C and additional genomic datasets. These include various visualizations of Hi-C matrices: square; triangular; mirrored, in which two triangular Hi-C matrices are shown above and below a horizontal dividing line; and split, where the diagonal separates two different matrices in a square plot. All of the above matrix plots can also be used to display difference and fold-change maps. The latter was our main interest in using this tool [23].

### 6.2.5 PCA

Principal component analysis (PCA) is a statistical method for reducing the dimensionality of the data while retaining most of the variation in the data set. It accomplishes this reduction by iteratively identifying directions, called principal components, along which the residual variation in the data is maximal. By using the few first components, each sample can be represented by a small number of features instead of by values for the (potentially, thousands of) original variables. Samples can be plotted by projecting their top components in 2D or 3D, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped [24].

The input of PCA is a matrix $G_{mXn}$ 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), has the greatest possible variance out of all the linear combinations over the variables, the second principal component (PC2) has the second greatest residual variance given the first, and so on. The number of features can be up to the number of variables, but typically, low variance components are discarded, leading to low dimension representation of the data.

Formally, let $G$ be centered, so that the mean of each observation is zero. The projection of a vector $v$ is given by $G_v$. The normalized variance of the projection is $\frac{1}{n-1} G_v^T \cdot G_v = v^T \cdot \left(\frac{1}{n-1} G^T G\right) \cdot v = v^T C v = D$ where $C$ is the covariance matrix of $G$, and D is the diagonal matrix of eigenvalues of C. We would like to find $v$ that maximizes $v^T C v$. Since $C$ is symmetric, it can be diagonalized by its eigenvector basis denoted by $\{z_i\}$. The diagonal matrix D contains the eigenvalues that correspond to the eigenvectors in C. 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 that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, we can represent the correspond eigenvectors accordingly and get $PC1, PC2, \ldots, PC_n$, while $PC1$ fits the $\lambda_1$ and will give the first feature with the maximum variance.

PCA is used as part of the pipeline of Hi-C data analysis, for calculation of the genome compartments. The first PC of the correlation matrix of the Hi-C data usually reflects the A/B chromosomal compartments. Regions with negative and positive values of PC1 represent this partition. The determination of whether positive or negative values of PC1 correspond to A or B compartments in the genome is based on gene density, which is markedly higher in the A compartment. In Figure 12, heterochromatin compartment (label 'B') is represented in blue, whereas euchromatin (label 'A') is represented in red.

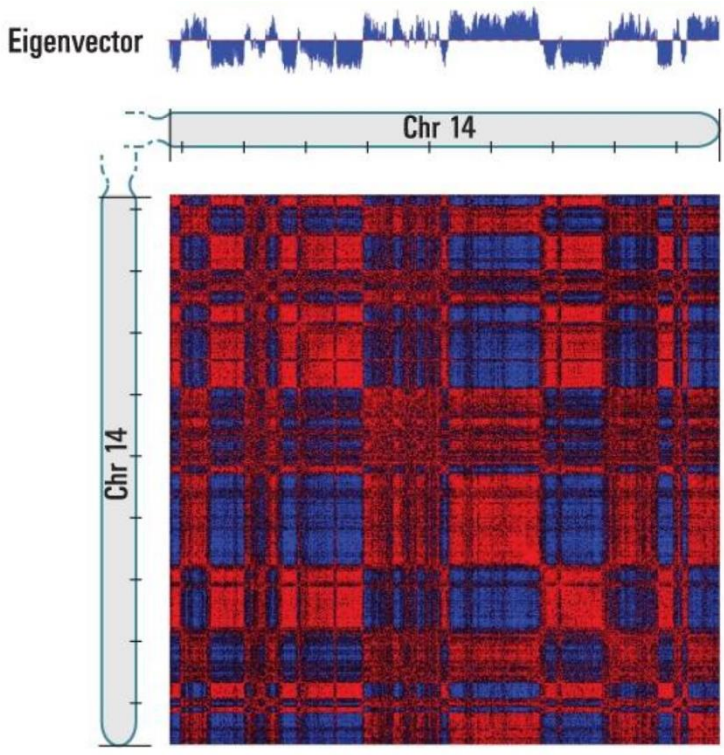


Figure 12. | Correlation map of chromosome 14 at a resolution of 100kb. The principal component (eigenvector) correlates with features of open chromatin. The matrix shows loci with either more (red) or less (blue) interactions than would be expected, given their genomic distance. Source: [11].

### 6.2.6 The hypergeometric test

This test is used in enrichment analysis. Formally, let $B$ be a background set of genes of size $N$, let $S \subseteq B$ be a set of $n$ target genes, and let $D \subseteq B$ be a fixed set of $K$ apriori defined genes (see Figure 13). This can be a set of genes defining a certain biological process, a pathway, or the targets of some regulatory factor. Suppose $|S \cap D| = k$. We wish to compute the probability of obtaining such overlap size $k$ given the null hypothesis that the genes in the target set were selected randomly without replacement from the background group. Under that assumption the probability of intersection $k$ is given by:

$$\mathrm{p_X}(k) = \Pr(k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}$$

The hypergeometric p-value for enrichment is calculated as the probability of obtaining overlap of at least $k$ when making $n$ draws in total, i.e., $\sum_{i=k}^{\min(n,K)} \Pr(i)$.

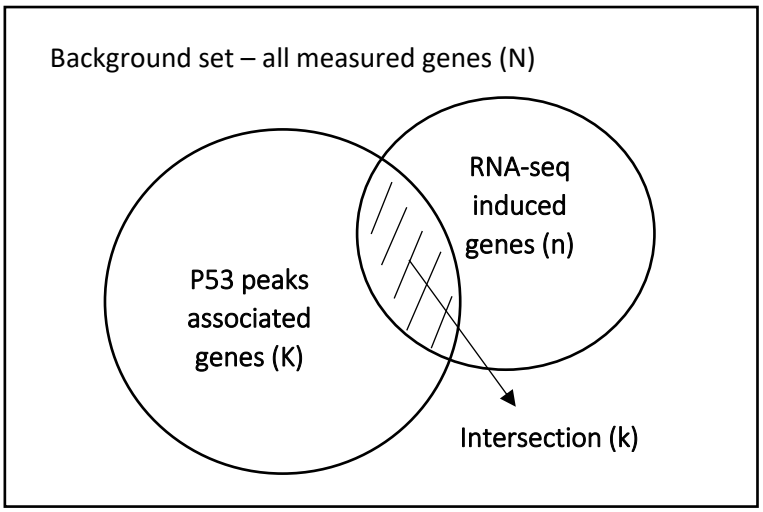


Figure 13. | Schematic of the hyper geometric test. In this example the test is for enrichment of genes associated with P53 chip-seq peaks among induced genes.

In addition, for each test, we also calculated its **enrichment factor** as follows:

$$EF = \frac{intersection \cdot Background\ set}{p53\ peaks\ nearest\ genes \cdot RNA-seq\ induced\ genes} = \frac{k \cdot N}{K \cdot n}$$

### 6.2.7 Non-parametric statistical tests

Parametric tests are used when data is assumed to follow a particular distribution (e.g., a normal distribution). Nonparametric tests are used when a particular distribution cannot be assumed; they are based on ranking the values rather than taking the actual values into account. Parametric tests generally have higher statistical power.

# 7. Results

## 6.1 Gene expression analysis of the response to p53 activation

Gene expression count data can be represented as an integer matrix $M \in \mathbb{R}^{nxm}$, where n is the number of genes in the data and m is the number of conditions. Each row in the matrix contains the expression level of a specific gene, and each column represents the biological condition (cell line, pre/post-Nutlin treatment) of a sample. The entry $M_{i,j}$ in the matrix is the number of reads of gene $i$ under a certain probed condition $j$. Values are normalized to counts per million (CPM). This calculation is done in two steps:

1. Count the total reads in the sample (column) and divide that number by 1,000,000 – this is our "per million" scaling factor.
2. Divide the actual read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving counts per million (CPM).

We filtered out all the genes that were not robustly detected in any condition in our dataset. Specifically, we filtered out the genes whose CPM value did not reach 1 CPM in both replicates of at least one cell line. The remaining gene set contained 21,651 genes, of which 15,459 were protein-coding genes.

Next, to examine the quality of our expression data, we applied PCA analysis to the transposed expression matrices of control and treatment, without combining replicates. PCA plots (Figure 14) indeed show that replicate samples of the same cell line and biological condition are located close to each other.

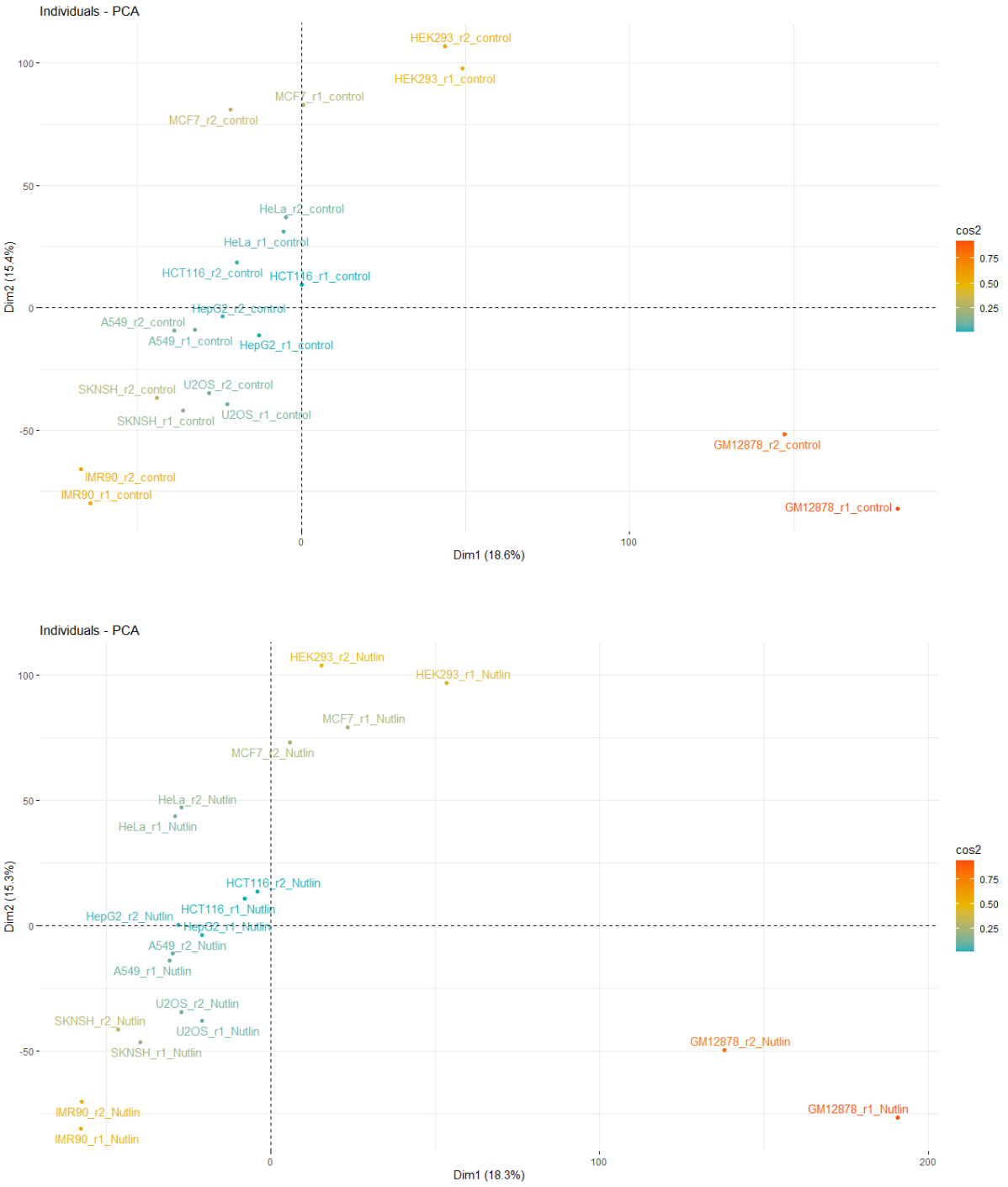


Figure 14. | PCA plots of the expression data. Top: basal conditions. Bottom: treatment conditions. It can be seen that replicates in each condition tend to be close to each other and that there is a difference between basal and after treatment with Nutlin.

We then identified the genes that significantly responded to Nutlin treatment in our dataset.

We applied DESeq2 on each cell line separately to identify differentially expressed genes (DEGs; either induced or repressed) upon Nutlin treatment. Each DESeq2 run included four

samples: 2 replicates of the control condition and 2 replicates of the Nutlin-treated condition. The genes with q-value < 0.05 and fold-change > 1.5 (up or down) were defined as DEGs. In all cell lines combined; we detected a set of 583 unique DEGs, of which 451 were protein-coding genes. Table 2 and Figure 15 summarize the number of DEGs detected per cell line.

| Cell Line | No. of Induced Genes | No. of Repressed Genes |
|---|---|---|
| A549 | 260 | 74 |
| GM12878 | 66 | 0 |
| HCT116 | 103 | 13 |
| HEK293 | 1 | 0 |
| HeLa | 21 | 7 |
| HepG2 | 181 | 29 |
| MCF7 | 288 | 39 |
| IMR90 | 60 | 7 |
| SKNSH | 72 | 2 |
| U2OS | 156 | 36 |

Table 2. | Summary of differentially expressed genes per cell line.

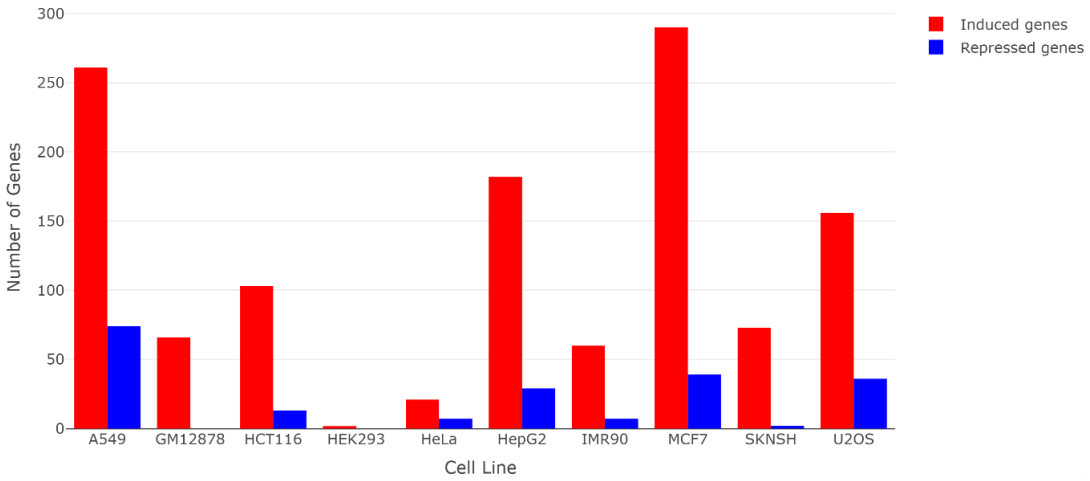


Figure 15. | Bar plot representing the number of differentially expressed genes.

Reassuringly, the two cell lines that do not carry a functional p53 (HEK293 and Hela) showed the lowest number of DEGs.

For each cell line, we created a volcano plot for visualizing the responsive genes and marked on it some canonical target genes of p53. Figure 16 presents the plot for the A549 cell line; the other volcano plots are presented in Figure S1.

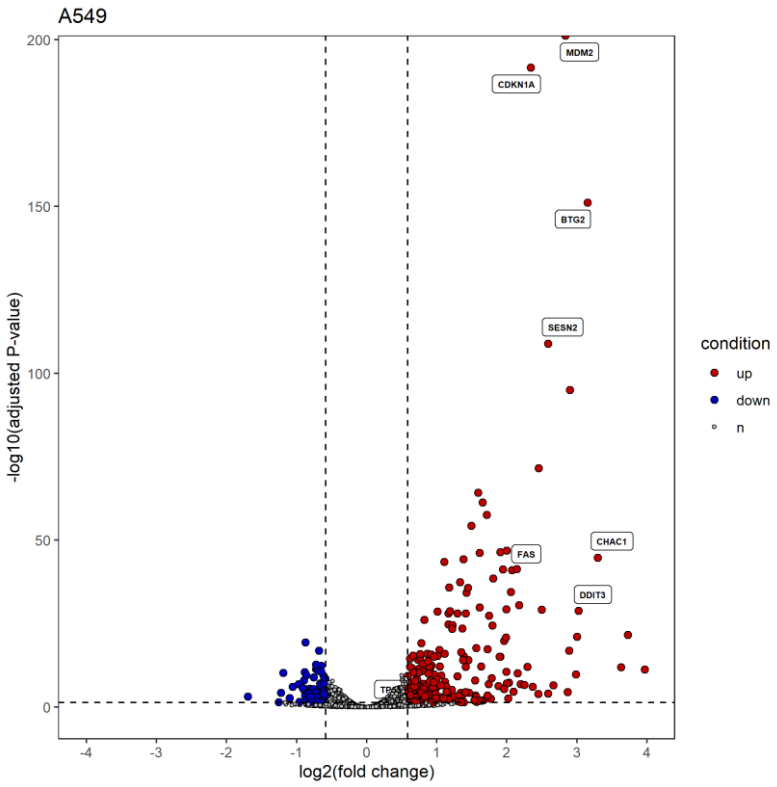


Figure 16. | Volcano plot of A549 cell line: Each dot represents a differentially expressed gene. Red dots represent up-regulated genes, blue dots represent down-regulated genes, and genes in grey are not significantly differentially expressed. Labeled genes are well known targets of p53.

Most of the transcriptional response upon p53 activation was cell type-specific: 338 genes were induced in only one cell line, while 45 were induced in at least six, see Figures 17,18.

We refer to the set of genes that were induced in at least 6 cell lines (out of 8 cell lines with functional p53) as the “p53 core/canonical responsive genes”. This gene set is highly enriched for the known p53 network [25] (enrichment factor = 37.77; p-value = 1.27e-37). A description of each protein-coding p53 core gene is given in Table 3. There was one gene that was induced in 9 out of 10 cell lines, and it is SESN2, a well-established p53 target gene [26].

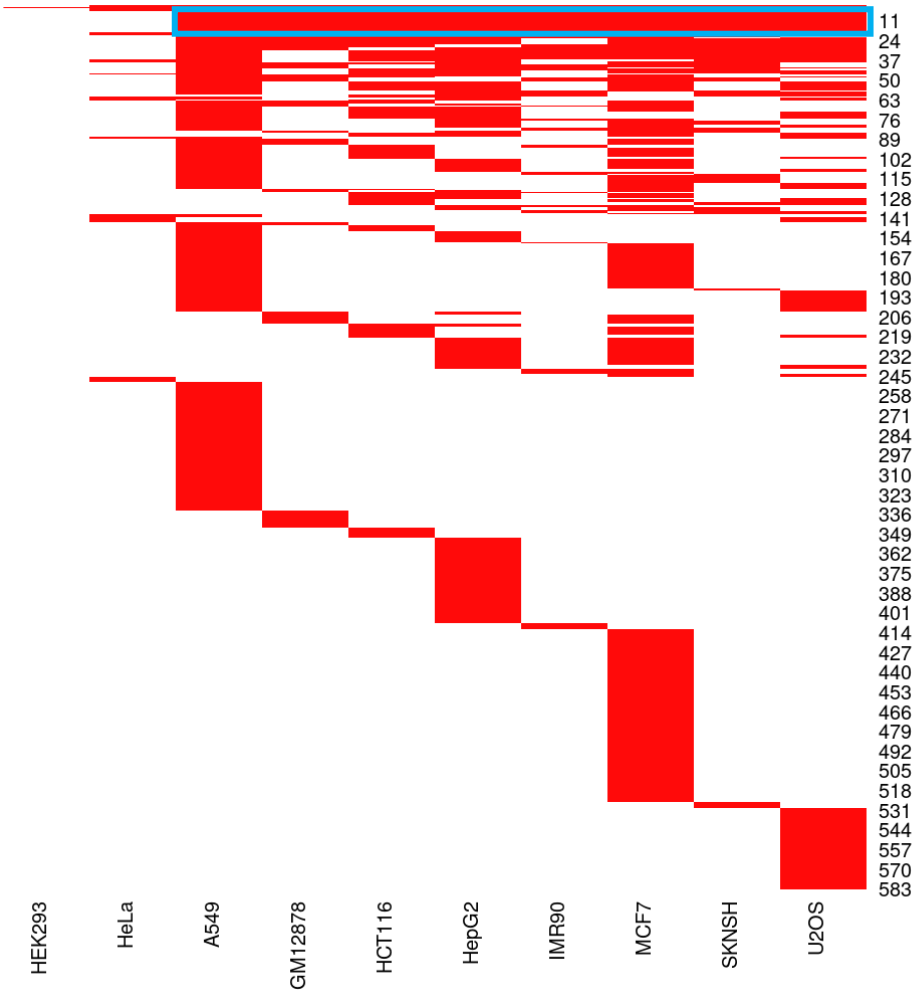


Figure 17. | Binary DEGs heatmap. red: up-regulated gene. The blue rectangle at the top indicates the set of genes that were DEGs in eight of the cell lines.

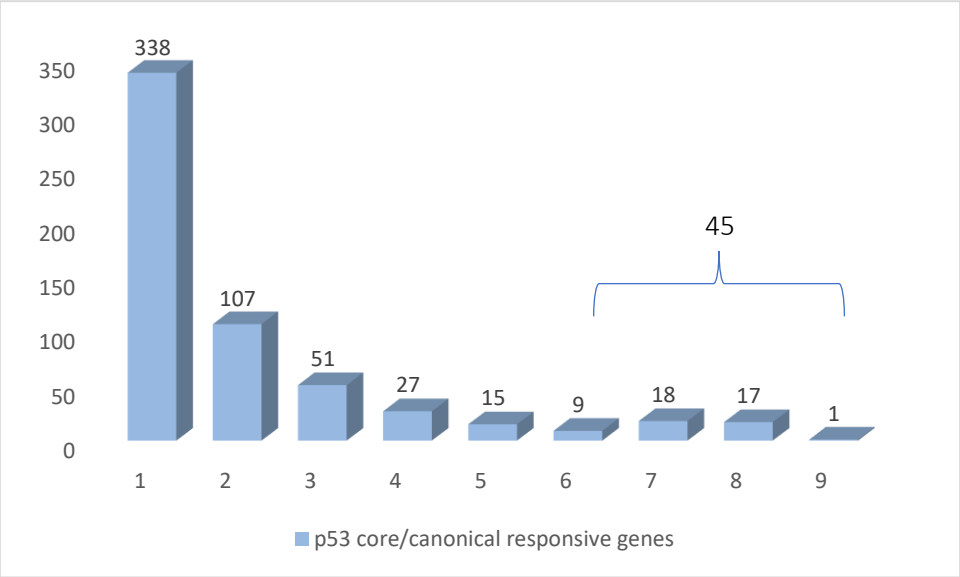


Figure 18. | The number of cell lines in which genes were induced. The histogram shows how many genes were induced in $i$ cell lines for $i$ =1, ..., 9, cell lines. 45 genes were induced in at least 6 cell lines.

| Gene ID | Gene name | Number of Cell Lines | Gene description |
|---|---|---|---|
| ENSG00000130766 | SESN2 | 9 | sestrin 2 [Source:HGNC Symbol;Acc:HGNC:20746] |
| ENSG00000128965 | CHAC1 | 8 | ChaC glutathione specific gamma-glutamylcyclotransferase 1 [Source:HGNC Symbol;Acc:HGNC:28680] |
| ENSG00000130513 | GDF15 | 8 | growth differentiation factor 15 [Source:HGNC Symbol;Acc:HGNC:30142] |
| ENSG00000128165 | ADM2 | 8 | adrenomedullin 2 [Source:HGNC Symbol;Acc:HGNC:28898] |
| ENSG00000078237 | TIGAR | 8 | TP53 induced glycolysis regulatory phosphatase [Source:HGNC Symbol;Acc:HGNC:1185] |
| ENSG00000080546 | SESN1 | 8 | sestrin 1 [Source:HGNC Symbol;Acc:HGNC:21595] |
| ENSG00000105327 | BBC3 | 8 | BCL2 binding component 3 [Source:HGNC Symbol;Acc:HGNC:17868] |
| ENSG00000124762 | CDKN1A | 8 | cyclin dependent kinase inhibitor 1A [Source:HGNC Symbol;Acc:HGNC:1784] |
| ENSG00000135679 | MDM2 | 8 | MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973] |
| ENSG00000143217 | NECTIN4 | 8 | nectin cell adhesion molecule 4 [Source:HGNC Symbol;Acc:HGNC:19688] |
| ENSG00000154767 | XPC | 8 | XPC complex subunit, DNA damage recognition and repair factor [Source:HGNC Symbol;Acc:HGNC:12816] |
| ENSG00000164938 | TP53INP1 | 8 | tumor protein p53 inducible nuclear protein 1 [Source:HGNC Symbol;Acc:HGNC:18022] |
| ENSG00000170734 | POLH | 8 | DNA polymerase eta [Source:HGNC Symbol;Acc:HGNC:9181] |
| ENSG00000170836 | PPM1D | 8 | protein phosphatase, Mg2+/Mn2+ dependent 1D [Source:HGNC Symbol;Acc:HGNC:9277] |
| ENSG00000170855 | TRIAP1 | 8 | TP53 regulated inhibitor of apoptosis 1 [Source:HGNC Symbol;Acc:HGNC:26937] |
| ENSG00000177076 | ACER2 | 8 | alkaline ceramidase 2 [Source:HGNC Symbol;Acc:HGNC:23675] |
| ENSG00000196152 | ZNF79 | 8 | zinc finger protein 79 [Source:HGNC Symbol;Acc:HGNC:13153] |
| ENSG00000197852 | INKA2 | 8 | inka box actin regulator 2 [Source:HGNC Symbol;Acc:HGNC:28045] |
| ENSG00000168209 | DDIT4 | 7 | DNA damage inducible transcript 4 [Source:HGNC Symbol;Acc:HGNC:24944] |
| ENSG00000162772 | ATF3 | 7 | activating transcription factor 3 [Source:HGNC Symbol;Acc:HGNC:785] |
| ENSG00000048392 | RRM2B | 7 | ribonucleotide reductase regulatory TP53 inducible subunit M2B [Source:HGNC Symbol;Acc:HGNC:17296] |
| ENSG00000116717 | GADD45A | 7 | growth arrest and DNA damage inducible alpha [Source:HGNC Symbol;Acc:HGNC:4095] |
| ENSG00000120889 | TNFRSF10B | 7 | TNF receptor superfamily member 10b [Source:HGNC Symbol;Acc:HGNC:11905] |
| ENSG00000164331 | ANKRA2 | 7 | ankyrin repeat family A member 2 [Source:HGNC Symbol;Acc:HGNC:13208] |
| ENSG00000167196 | FBXO22 | 7 | F-box protein 22 [Source:HGNC Symbol;Acc:HGNC:13593] |
| ENSG00000131080 | EDA2R | 7 | ectodysplasin A2 receptor [Source:HGNC Symbol;Acc:HGNC:17756] |
| ENSG00000177595 | PIDD1 | 7 | p53-induced death domain protein 1 [Source:HGNC Symbol;Acc:HGNC:16491] |
| ENSG00000026103 | FAS | 7 | Fas cell surface death receptor [Source:HGNC Symbol;Acc:HGNC:11920] |
| ENSG00000159388 | BTG2 | 7 | BTG anti-proliferation factor 2 [Source:HGNC Symbol;Acc:HGNC:1131] |
| ENSG00000168918 | INPP5D | 7 | inositol polyphosphate-5-phosphatase D [Source:HGNC Symbol;Acc:HGNC:6079] |
| ENSG00000173846 | PLK3 | 7 | polo like kinase 3 [Source:HGNC Symbol;Acc:HGNC:2154] |
| ENSG00000175197 | DDIT3 | 7 | DNA damage inducible transcript 3 [Source:HGNC Symbol;Acc:HGNC:2726] |
| ENSG00000196072 | BLOC1S2 | 7 | biogenesis of lysosomal organelles complex 1 subunit 2 [Source:HGNC Symbol;Acc:HGNC:20984] |
| ENSG00000051108 | HERPUD1 | 6 | homocysteine inducible ER protein with ubiquitin like domain 1 [Source:HGNC Symbol;Acc:HGNC:13744] |
| ENSG00000176046 | NUPR1 | 6 | nuclear protein 1, transcriptional regulator [Source:HGNC Symbol;Acc:HGNC:29990] |
| ENSG00000115129 | TP53I3 | 6 | tumor protein p53 inducible protein 3 [Source:HGNC Symbol;Acc:HGNC:19373] |
| ENSG00000164237 | CMBL | 6 | carboxymethylenebutenolidase homolog [Source:HGNC Symbol;Acc:HGNC:25090] |
| ENSG00000166592 | RRAD | 6 | RRAD, Ras related glycolysis inhibitor and calcium channel regulator [Source:HGNC Symbol;Acc:HGNC:10446] |
| ENSG00000161513 | FDXR | 6 | ferredoxin reductase [Source:HGNC Symbol;Acc:HGNC:3642] |
| ENSG00000100647 | SUSD6 | 6 | sushi domain containing 6 [Source:HGNC Symbol;Acc:HGNC:19956] |
| ENSG00000172831 | CES2 | 6 | carboxylesterase 2 [Source:HGNC Symbol;Acc:HGNC:1864] |

Table 3. | p53 core/canonical responsive genes. The 41 protein-coding genes induced in six or more tissues, with a short description of the function of each.

## 6.2 p53 Chip-seq analysis

We performed p53 ChIP-seq analysis to identify p53 binding sites in the ten cell lines analyzed in our study. In order to identify significantly enriched genomic regions ('peaks') in the p53 ChIP-seq data we used the MACS2 algorithm. For each cell line after Nutlin treatment, p53 ChIP-seq reads were mapped to the reference genome in order to call 'p53 peaks'. The number of peaks detected in each cell line is reported in Table 4. We also applied p53 ChIP-seq analysis to the untreated cells and obtained a profile of p53 binding events under basal condition.

We next applied motif analysis to the set of p53 peaks detected in each cell line after Nutlin treatment, to identify known as well as de novo motifs, using Homer [27] and DREME [28] algorithms. In particular, as a quality control for the called peaks, we sought to confirm that the p53 motif was highly enriched. Reassuringly, in all cell lines, the strongest detected motif corresponded to the p53 motif. We wished to find enriched motifs of additional TFs, as these could represent cofactors of p53 that drive cell type-specific responses. Table 4 shows the number of significant motifs and the enrichment for the p53 motif in each cell line:

| Cell Line | # p53 Peaks | # Total Motifs (p-value ≤1e-5) | # Enriched Motifs (p-value ≤1e-20) | p53 Motif Enrichment (p-value) |
|---|---|---|---|---|
| GM12878 | 14,119 | 19 | 14 | 1e-1359 |
| A549 | 11,113 | 21 | 14 | 1e-1435 |
| MCF7 | 7,731 | 22 | 19 | 1e-1892 |
| HepG2 | 7,657 | 22 | 17 | 1e-1691 |
| HCT116 | 6,584 | 23 | 17 | 1e-1412 |
| HEK293 | 2,797 | 21 | 15 | 1e-828 |
| U2OS | 2,596 | 21 | 16 | 1e-843 |
| HeLa | 1,750 | 22 | 15 | 1e-550 |
| IMR90 | 8,076 | 22 | 13 | 1e-1571 |
| SKNSH | 8,934 | 22 | 16 | 1e-1927 |

Table 4. | The number p53 peaks found in each treated cell line and the number of significantly enriched motifs found in them. "Total motifs" is the number of motifs found for $p - value \leq 0.05$. The last column shows enrichment p-value for the p53 motif. That motif was the most significant and strongest motif in all cell types.

We also sought de novo motifs using Homer and DREME. Figure 19 shows an example of such motif.

**a)**

**b)**

Figure 19. | De-novo motif found in p53 peaks of SKNSH cell line: a) The motif found de novo (P-value = 1e-1927 in HOMER de novo motif discovery tool.) b) a matching known motif.

The following heatmap summarizes the strongly enriched known motifs ($p<10^{-20}$) detected in our p53 ChIP-seq dataset, and the cell lines in which each was detected. As can be seen, p53 motif is the only one detected in all cell lines. This analysis suggests cell type-specific coactivators of p53, including AP-1 and NFkB in GM12878, PTX1 in HCT116, HOXA5 in HepG2, SOX3 and FOXL1 in MCF7 and TEAD4 in SKNSH. Some other factors seem to cooperate with p53 in multiple cell lines e.g., SHN and TBX21. The GATA1 motif was mildly enriched in five cell lines. It is known as a paralog of GATA2, which plays an essential role in regulating transcription of genes involved in the development and proliferation of hematopoietic and endocrine cell lineages. Another presentation that uses the p-value of each motif to create a heatmap is given in Figure S2.

SHN gene function is unknown for humans but known for Drosophila. In Drosophila it was demonstrated that schnurri (SHN) gene is required for cell differentiation in the dorsal ectoderm [29].

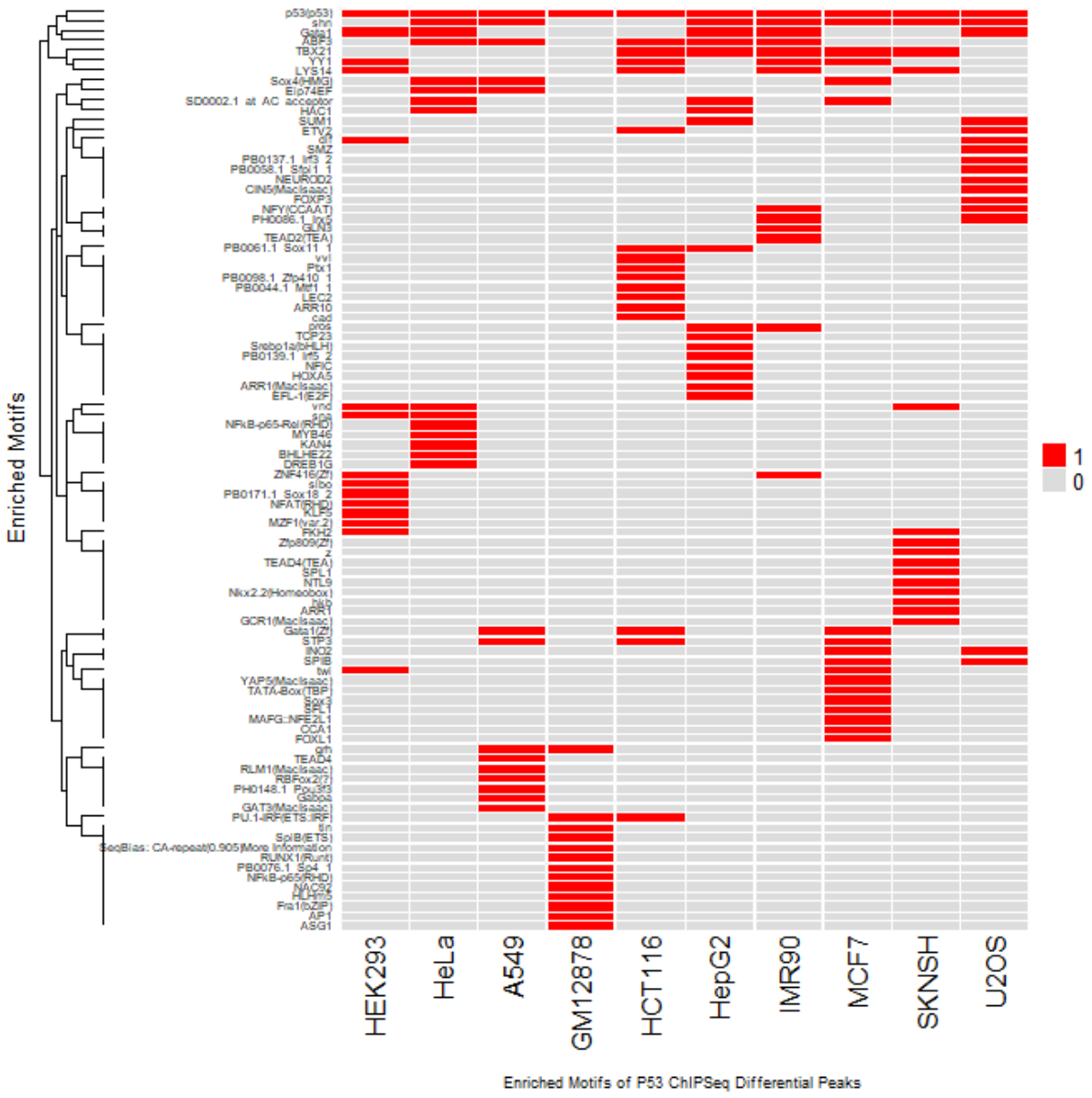


Figure 20. | Binary heatmap of enriched motifs in p53 peaks of treated cells. The motifs found de novo were matched to known motifs. A motif is considered enriched if it has $p - value \leq 1e - 20$. Red: enriched motif, grey: not enriched. P53 is enriched in all cell lines.

### 6.2.1 Integrated analysis of p53 ChIP-seq and gene expression data

Next, we carried out an integrative analysis of the transcriptomic and p53 ChIP-seq data. We wished to demonstrate that p53 binding to the chromatin is associated with induction of the target genes. For this task, we took a naïve approach and associated each p53 peak to its closest gene (the gene with the closest transcription start site (TSS)). In particular, we tested all p53 peak-nearest gene associations, as well as only those in which the TSS and the peak are within 50k and 20k bp. Table 5 shows the number of target genes that were associated with p53 peaks in each cell line:

| Cell Line | # p53 Peaks | # of Associated Genes | # of Associated Genes at Dis < 20K | # of Associated Genes at Dis < 50K |
|---|---|---|---|---|
| GM12878 | 14,119 | 13,852 | 6673 | 8854 |
| A549 | 11,113 | 10,860 | 5448 | 7135 |
| MCF7 | 7,731 | 7,555 | 2844 | 4341 |
| HepG2 | 7,657 | 7,464 | 2990 | 4484 |
| HCT116 | 6,584 | 6,411 | 2609 | 3828 |
| HEK293 | 2,797 | 2,707 | 974 | 1523 |
| U2OS | 2,596 | 2,482 | 774 | 1243 |
| HeLa | 1,750 | 1,698 | 637 | 991 |
| IMR90 | 8,076 | 7,814 | 3212 | 4662 |
| SKNSH | 8,934 | 8,682 | 3402 | 5010 |

Table 5. | The number of genes associated with p53 ChIP-seq peaks in each cell line. The two right most columns show the number of associated genes located within ±20,000 bp or ±50,000 bp.

We then examined the significance of the overlap between the closest genes to p53 peaks and the set of Nutlin-induced genes in each cell line, using the hyper geometric test (Figure 13) and the enrichment factor. We then tested the overlap between the set of Nutlin-induced genes and the set of genes that are closest to some p53 peak. Table 6 shows the results for all genes and for the two different distance cut-offs. Notably, in all these tests the overlap between the Nutlin-induced genes and p53 ChIP-seq closest genes is highly significant. Moreover, the

enrichment factor (EF) for those genes that lie within a distance of 20kbp from the p53 binding site is higher than the EF for genes within a distance of 50kbp, suggesting that the binding of p53 within the close region has a stronger effect on the expression of the target genes.

In many control samples, a higher significance was obtained compared to the treated samples. This can be explained as follows, the binding sites of p53 near the genes that undergo induction are the "strongest" binding sites (with the highest binding affinity). Therefore, ChIP-seq will detect a binding of p53 to these sites even when p53 levels are low (i.e., in control samples). In the treated cells, the level of the p53 protein increases significantly and therefore it also binds to hundreds/thousands of additional sites in the genome that are weaker and non-functional (in this sense, that do not induce the expression of a gene in their vicinity).

| Cell Line | Condition (ChIPSeq:NUT/ Control, RNASeq:NUT) | Background set (M) | RNASeq induced genes (n) | P53 peaks nearest genes - ALL (N) | P53 peaks nearest genes - \|50kb\| (N) | P53 peaks nearest genes - \|20kb\| (N) | Intersection - ALL (k) | Intersection - \|50kb\| (k) | Intersection- \|20kb\| (k) | HG p-value - ALL | HG p-value - \|50kb\| | HG p-value - \|20kb\| | Enrichment Factor - ALL | Enrichment Factor - \|50kb\| | Enrichment Factor - \|20kb\| |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A549 | Case (N) | 21652 | 260 | 6370 | 5239 | 4306 | 145 | 133 | 126 | 5.0701E-19 | 3.76E-21 | 2.54E-25 | 1.9 | 2.1 | 2.4 |
| | Control | 21652 | 260 | 77 | 51 | 33 | 11 | 9 | 9 | 2.0114E-09 | 8.86E-09 | 1.36E-10 | 11.9 | 14.7 | 22.7 |
| GM12878 | Case (N) | 21652 | 66 | 7037 | 5881 | 4874 | 40 | 39 | 39 | 2.4774E-06 | 4.99E-08 | 1.59E-10 | 1.9 | 2.2 | 2.6 |
| | Control | 21652 | 66 | 460 | 282 | 180 | 17 | 16 | 16 | 2.6395E-14 | 2.14E-16 | 1.57E-19 | 12.1 | 18.6 | 29.2 |
| HCT116 | Case (N) | 21652 | 103 | 4431 | 3230 | 2327 | 52 | 49 | 47 | 1.1554E-11 | 3.74E-15 | 2.34E-19 | 2.5 | 3.2 | 4.2 |
| | Control | 21652 | 103 | 273 | 163 | 96 | 19 | 16 | 15 | 4.1543E-17 | 6.42E-17 | 4.78E-19 | 14.6 | 20.6 | 32.9 |
| HepG2 | Case (N) | 21652 | 181 | 4919 | 3630 | 2605 | 96 | 89 | 84 | 5.8747E-19 | 5.50E-24 | 1.28E-30 | 2.3 | 2.9 | 3.9 |
| | Control | 21652 | 181 | 537 | 325 | 217 | 35 | 33 | 30 | 2.2512E-21 | 2.92E-26 | 5.93E-28 | 7.8 | 12.1 | 16.5 |
| MCF7 | Case (N) | 21652 | 288 | 4877 | 3556 | 2497 | 138 | 129 | 124 | 1.4205E-21 | 5.27E-30 | 2.35E-42 | 2.1 | 2.7 | 3.7 |
| | Control | 21652 | 288 | 188 | 126 | 79 | 31 | 28 | 26 | 5.9885E-25 | 2.03E-26 | 1.47E-29 | 12.4 | 16.7 | 24.7 |
| IMR90 | Case (N) | 21652 | 60 | 4890 | 3710 | 2725 | 38 | 34 | 32 | 1.5606E-11 | 5.09E-12 | 3.79E-14 | 2.8 | 3.3 | 4.2 |
| | Control | 21652 | 60 | 438 | 262 | 166 | 14 | 11 | 11 | 1.1618E-11 | 1.34E-10 | 9.54E-13 | 11.5 | 15.2 | 23.9 |
| SKNSH | Case (N) | 21652 | 72 | 5357 | 4021 | 2926 | 51 | 48 | 48 | 2.3068E-16 | 4.32E-19 | 3.76E-25 | 2.9 | 3.6 | 4.9 |
| | Control | 21652 | 72 | 440 | 275 | 168 | 19 | 18 | 18 | 2.1129E-16 | 9.48E-19 | 1.18E-22 | 13 | 19.7 | 32.2 |
| U2OS | Case (N) | 21652 | 156 | 1932 | 1148 | 744 | 44 | 39 | 37 | 2.7471E-12 | 2.19E-16 | 5.84E-21 | 3.2 | 4.7 | 6.9 |
| | Control | 21652 | 156 | 521 | 278 | 173 | 22 | 19 | 18 | 2.6522E-11 | 1.58E-13 | 4.67E-16 | 5.9 | 9.5 | 14.4 |

Table 6. | Hypergeometric test results for the overlap between induced genes and p53 ChIP-seq peaks. Results are shown for all genes and for the two different distance cut-offs. ("Case" – Nutlin-treated cells; "Control" – untreated cells).

## 6.3 p53 Hi-C data: characterization of differential loops upon p53 activation

We first called TADs using Juicer [20] and loops using Mustache [30] in each cell line, both without and after Nutlin treatment. Tables 7 and 8 below show the number of loops and TADs in each condition for 5kb and 10kb resolution, respectively.

| Cell Line | # Basal | # Treatment |
|---|---|---|
| GM12878 | 24649 | 28863 |
| A549 | 15318 | 16438 |
| MCF7 | 25307 | 28954 |
| HepG2 | 24860 | 25147 |
| HCT116 | 19227 | 22189 |
| HEK293 | 30742 | 32716 |
| U2OS | 12418 | 14942 |
| HeLa | 20930 | 24207 |
| IMR90 | 17602 | 15730 |
| SKNSH | 17687 | 24178 |

Table 7. | Average number of loops between two replicates in 5kb resolution.

| Cell Line | # Basal | # Treatment |
|---|---|---|
| GM12878 | 4987 | 5772 |
| A549 | 4620 | 4395 |
| MCF7 | 5960 | 6935 |
| HepG2 | 4828 | 4780 |
| HCT116 | 4848 | 5032 |
| HEK293 | 6600 | 6884 |
| U2OS | 4049 | 4828 |
| HeLa | 3719 | 4294 |
| IMR90 | 4892 | 4510 |
| SKNSH | 3736 | 4999 |

Table 8. | Average number of TADs between two replicates in 10kb resolution.

The tables show that the number of loops detected in the basal and the treated condition is about the same in all cell lines, with a minor increase in the treated cells. A similar observation holds for TADs.

### 6.3.1 Analysis of differential loops

Next, we used the HiC-DC+ software to detect statistically significant differential loops between pairs of Micro-C samples [22]. We first used this tool to compare all pairs of cell types (45 pairwise comparisons between 10 different cell lines), under the basal condition. As these cell lines originate from very different tissues, we expected them to have many differential chromatin loops. Indeed, numerous loops passed this stringent statistical test employed by HiC-DC+ (Table 9).

| | A549 | GM12878 | HCT116 | HEK293 | HeLa | HepG2 | IMR90 | MCF7 | SKNSH | U2OS |
|---|---|---|---|---|---|---|---|---|---|---|
| **A549** | | 535 | 9 | 63 | 84 | 340 | 20 | 223 | 649 | 434 |
| **GM12878** | 393 | | 170 | 59 | 1000 | 302 | 66 | 214 | 66 | 426 |
| **HCT116** | 7 | 133 | | 21 | 122 | 127 | 42 | 63 | 12 | 179 |
| **HEK293** | 68 | 131 | 82 | | 266 | 120 | 74 | 118 | 678 | 212 |
| **HeLa** | 77 | 1684 | 169 | 306 | | 1295 | 353 | 728 | 1804 | 1143 |
| **HepG2** | 196 | 358 | 172 | 37 | 421 | | 9 | 94 | 833 | 1463 |
| **IMR90** | 1 | 11 | 11 | 3 | 0 | 0 | | 5 | 0 | 3 |
| **MCF7** | 229 | 352 | 156 | 120 | 531 | 178 | 72 | | 15 | 1004 |
| **SKNSH** | 36 | 10 | 0 | 0 | 102 | 51 | 0 | 0 | | 10 |
| **U2OS** | 552 | 281 | 405 | 60 | 1689 | 2733 | 15 | 1850 | 27 | |

Table 9. | Number of differential loops between basal cell types found by HiC-DC+ tool in chr12: Cell type measurements were taken in basal level only. Each entry in table represents the number of differential loops found by the tool after removal of diagonal loops and using the parameter padj<=0.1. For cell types $i, j$, entry $(i, j)$ is the number of differential loops that were stronger in $i$ and entry $(j, i)$ is the number of differential loops stronger in $j$. For example, for the pair: A549-GM12878 we get a total of 928 differential loops, 535 are stronger in A549 and 393 in GM12878.

Surprisingly, when we applied HiC-DC+ tests to compare Nutlin-treated to basal cells, *no differential loops were detected at all in any of the cell lines*. This finding suggests that chromatin loops that determine cell-identify transcriptional programs are markedly stronger than loops that are formed in response to stress (p53 activation in our research). This is in line with the fact that the scale of differential expression between different cell lines (originating from very different tissues) is an order of magnitude larger than the differential expression between Nutlin-treated and basal cells in the same tissue (Figure 21).

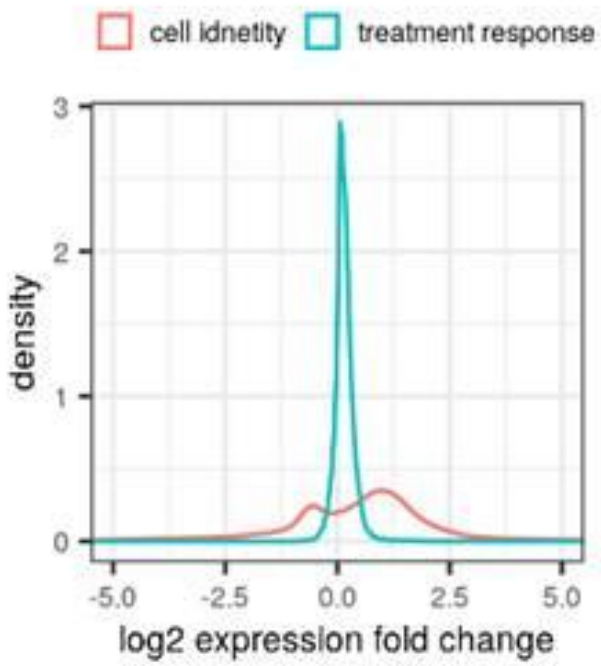


Figure 21. | Density plots of gene expression fold change between A549 and GM12787 (red) and between A549 Nutlin-treated and control sample (light blue). Fold change was calculated for all genes in RNA-seq (n=14,180).

Next, we examined associations between changes in promoter loop intensities and changes in expression levels of the associated genes. In this analysis too, when comparing different cell lines, we detected very strong associations: in basal cell lines where a gene had markedly higher expression, its promoter was associated with stronger loops (see Figure 22A for one

example. Similar strong associations were obtained for all 45 pair-wise comparisons between cell lines). In contrast, when we examined, for each cell line, association between changes in promoter loop intensities and changes in expression levels upon Nutlin treatment, no significant association was detected (Figure 22B).

Taken together, these results show that in our Micro-C dataset, we did not observe major changes in chromatin organization that corresponded to the observed modulation of gene expression upon p53 activation. This is in contrast to the strong association between genome organization and transcriptional programs that we observed in the comparisons between different basal cell lines.

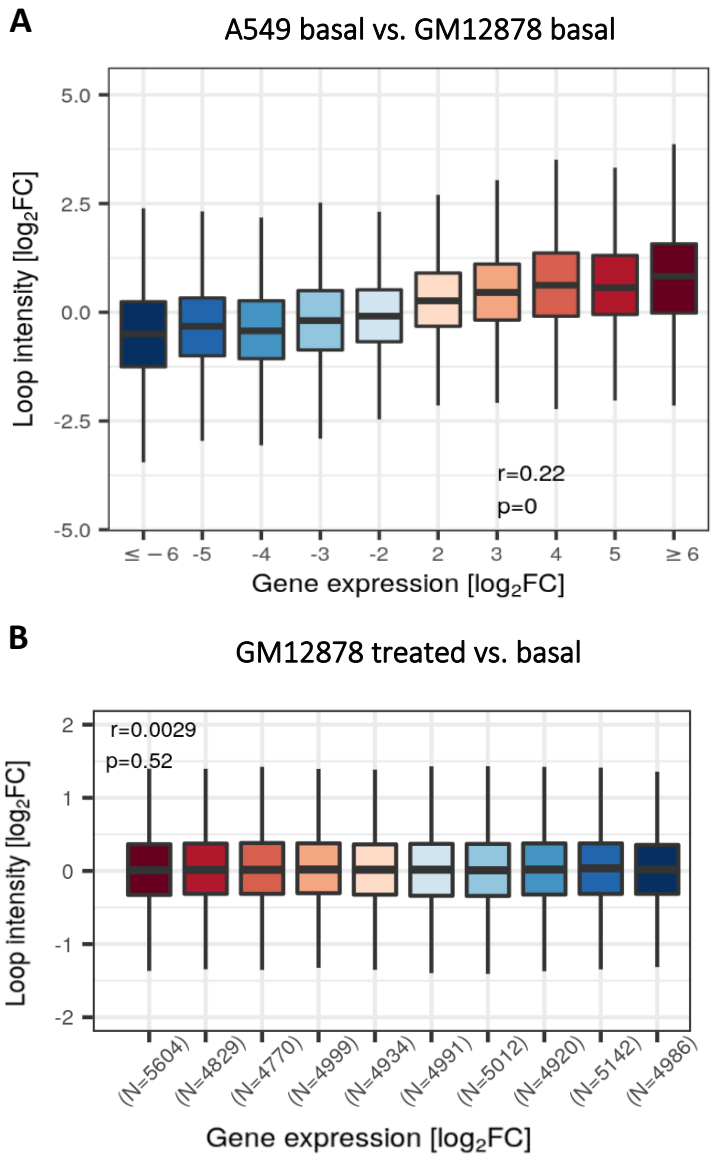


Figure 22. | Changes in gene expression levels vs. changes in interaction frequency of loops associated with their promoters, between different cell lines or conditions. X axis: log2 fold change in gene expression. Genes are binned according to log fold change between the two conditions. Y axis: distribution of log2 fold change of the loop intensity for the genes in each bin. Pearson's correlation value and p-value are indicated in each plot. A. A549 vs. GM12878 when both cell lines are basal. B. GM12878 treated with Nutlin vs. basal GM12878. In B, FC bins were defined as to include similar number of genes. Colors have no meaning.

In summary, we were not able to detect any statistically significant differential loops upon Nutlin treatment. This, most likely, points to the fact that changes that occur in chromatin interactions upon p53 activation were below the detection limit of our Micro-C analysis (despite very high sequencing depth). Additionally, we are limited by the resolution of the technique, and if most loops induced by p53 are shorter than 5k-10k, they are not detected due to the binning resolution we used (5k-10k bp). Yet, no differential loops were detected also when we used the resolution limit of our Micro-C protocol (0.5k bp).

### 6.3.2 Integrated analysis of Micro-C and gene expression data in response to p53 activation

Following the above observations, we tried to further examine the correlation between changes in gene expression levels and changes in the 3D organization of the chromatin. We chose to focus on six well-known target genes of p53 that were significantly induced in all the cell lines with functional p53 in our panel. Figure 23 shows the induction level of these genes across the ten cell lines.

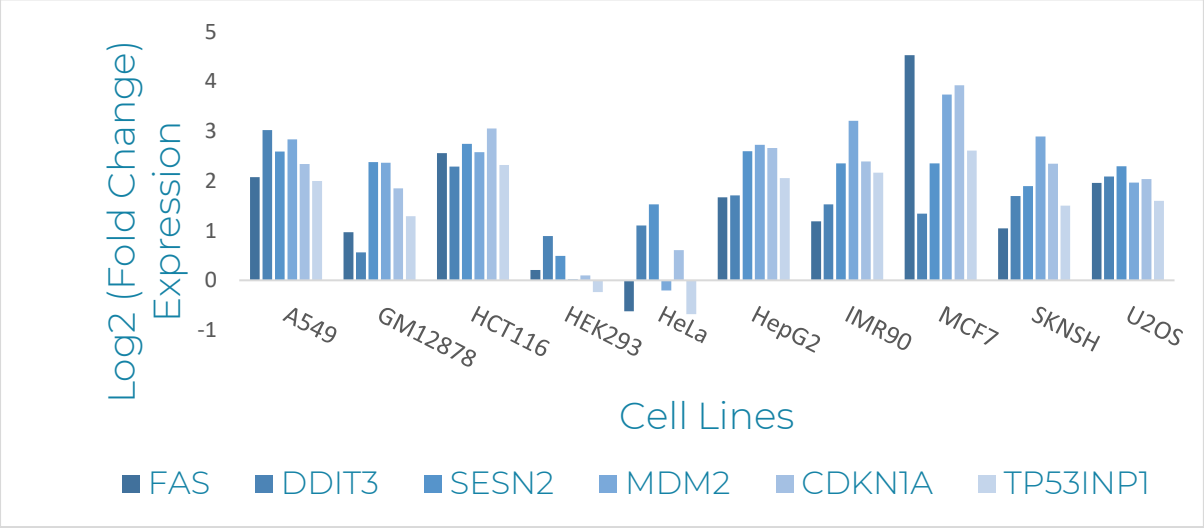


Figure 23. | Expression levels of well-known target genes of p53. The plot shows the fold change in Nutlin-treated cells vs controls. All six genes are induced in most or all cell lines. HEK293 and HeLa cell lines are less responsive as they do not carry a functional p53.

We next sought changes in chromatin looping in the vicinity of these canonical p53 targets that occur upon p53 activation. We therefore examined the Micro-C data in these loci using resolution of 5kb. In line with our previous results, no change in chromatin interactions was detected in these regions in response to Nutlin treatment (Figure 24). This is despite the marked induction of their expression levels, as well as the induced binding of p53 in these loci (p53 ChIP-seq peaks). Applying similar analysis to genes that showed cell type-specific expression, detected very strong cell type-specific chromatin 3D organization (Figure 25).

Next, we characterized the promoter loops associated with these six canonical p53 target genes. We found that most of these loops (14 out of 21) promoter-promoter loops (P-P loops) (Figure 26). We therefore focused on the other gene that was linked to the canonical p53 target on the other anchor of such P-P loop (see the examples of MED18 for SENS2 and CCNE2 for TP53INP1, Figure 26). We tested if these 'paired genes' were induced too upon Nutlin treatment. In contrast to our expectation, we did not find any significant induction for these genes whose promoters physically interact with the promoter of the induced genes. Furthermore, for six out of the six canonical genes, we found that the promoter anchor of its P-loop also contains a p53 ChIP-seq peak. That is, the p53 binding site is located less than 5kb (the bin resolution we worked with) from the gene's TSS.

Given these results, we decided to further increase the resolution analysis of the Micro-C, and produced data at bin resolution of 1000 bases (increasing the resolution decreases the number of reads assigned to each bin and therefore lowers sensitivity of detecting chromatin interactions). Since such resolution has a lot of background noise, we used a sliding window mean and median of bins intensity for five consecutive bins (moving the window by one bin at a time). Contrary to our expectation, this analysis too did not detect any increase the intensity

of chromatin interactions that involve the promoters of these p53 canonical target genes upon Nutlin treatment, despite the robust activation of these promoters by this treatment (Figure 27, S3-S6).

In contrast, applying this analysis to genes that show a very strong differential expression between different cell lines, we detected very strong cell type-specific E-P loops. Figure 28 shows the results for APOC3, which is specifically expressed in HepG2. Supplementary Figures S7 and S8 show the results for CCR7 and CD80, which are specifically expressed in GM12878.

Taken together, these results demonstrate that transcriptional changes induced upon activation of p53 are not accompanied by massive remodeling of chromatin interactions. This is in stark contrast to differential expression between different cell types, where cell type-specific transcriptional programs are accompanied by major changes in 3D organization of the genome. This difference can be explained by the marked difference in the magnitude of differential expression in these two cases: DE between different cell types is two orders of magnitude larger than DE in response to p53 activation. That is, cell type-specific genes show 100-1000 fold-change in expression between cell types, while genes induced by p53 activation typically show 2-4 fold change (Figure 25).

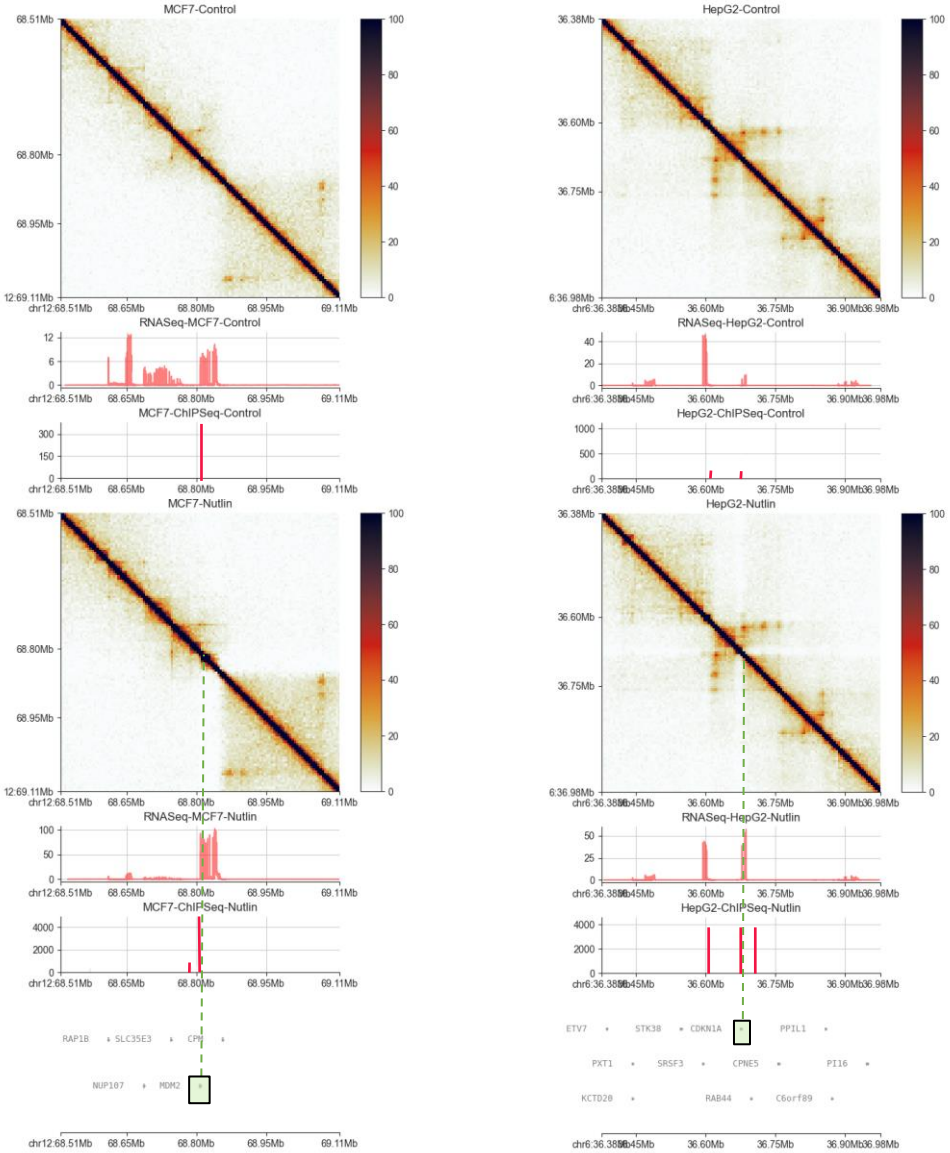


Figure 24. | Hi-C Plot of chromatin loops, at bin resolution of 5kb: Left: MDM2 gene in MCF7 cell line. Right: CDKN1A gene in HepG2 cell line. The tracks below each contact map shows the gene expression and p53 binding. The contact maps show no difference in 3D loops between basal and treated samples, although there is a strong induction of gene expression.

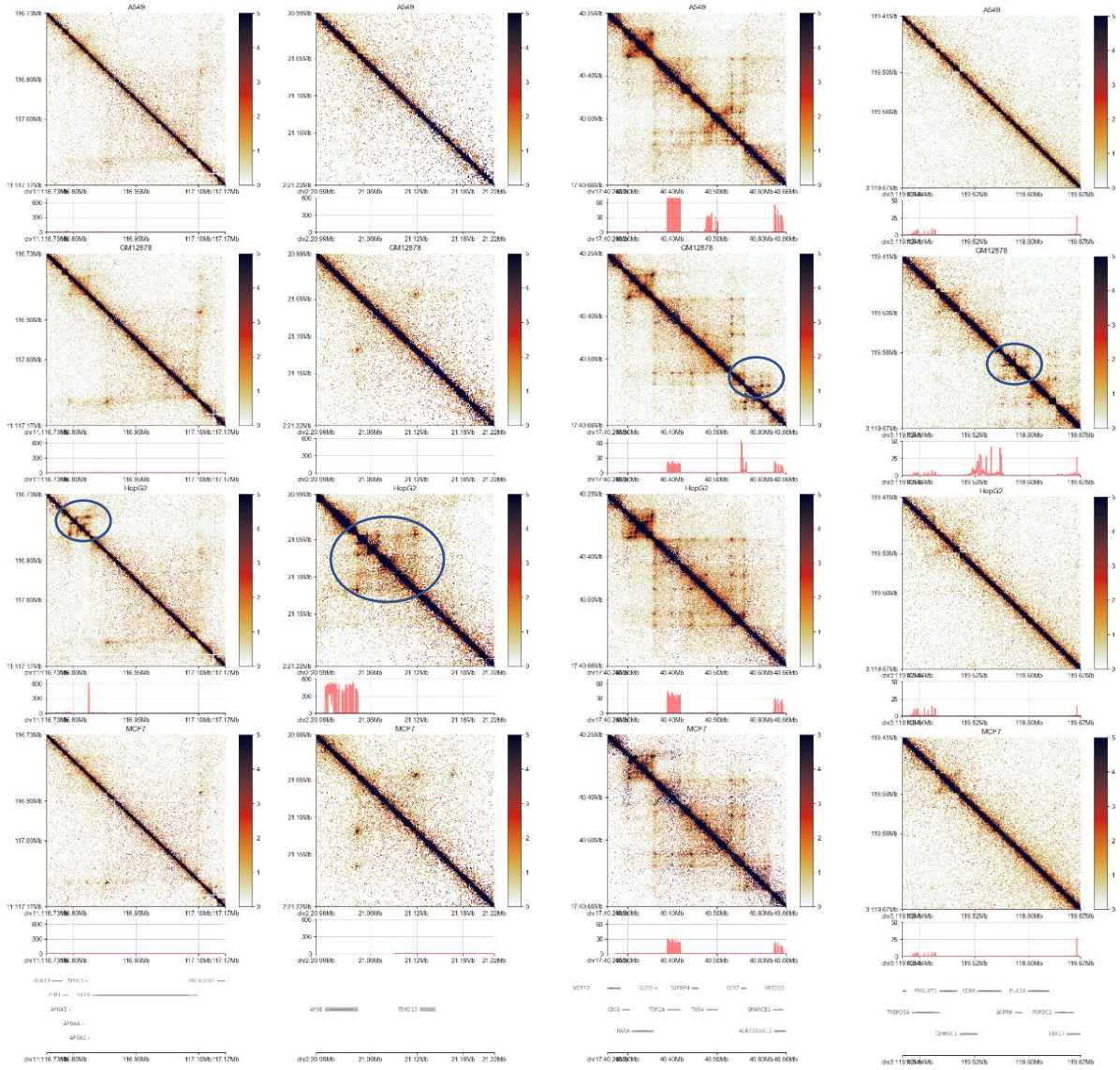


Figure 25. | HiC contact maps and gene expression across cell lines. Each column shows the interaction maps in four cell lines in the same genomic region, and the track below each map shows the mRNA expression, in CPM. Cell type-specific expression as manifest in cell type-specific peak in the track is coupled with the formation of cell type-specific chromatin loops observed in the interaction map. Blue circles indicate cell type-specific loops.


Figure 26. | UCSC Genome Browser display of regions that include loops anchored at the TSS of p53 canonical targets. Top: SESN2 gene. Bottom: TP53INP1 gene. Each display includes, from top to bottom, loop anchors, p53 ChIP-seq peaks, TADs, the TSS of the p53 canonical gene, and all gene models in the region. Data presented includes loops and peaks from all cell types, both basal and treated. It can be seen that TSS and the p53 ChIP-seq peaks are in the same loop anchor (light blue). Moreover, most of the loop anchors reside near TSSs of different genes (blue). The dynamic loops were not detected at 5000bp bin resolution and higher resolution is needed.

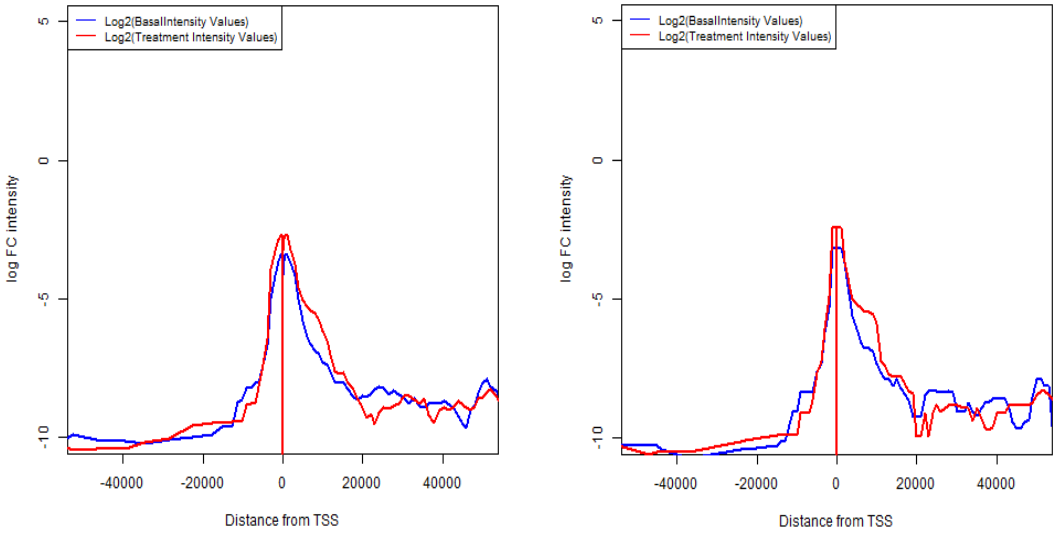


Figure 27. | Loop Intensities around the promoter of the gene TP53INP1: Loops intensity values in sliding windows of chr8 in A549 cell line are shown. Intensity values were calculated as the mean or the median of 5 bins in 1000bp resolution. Left: mean intensity values. Right: median intensity values. Blue: basal; red: treated.

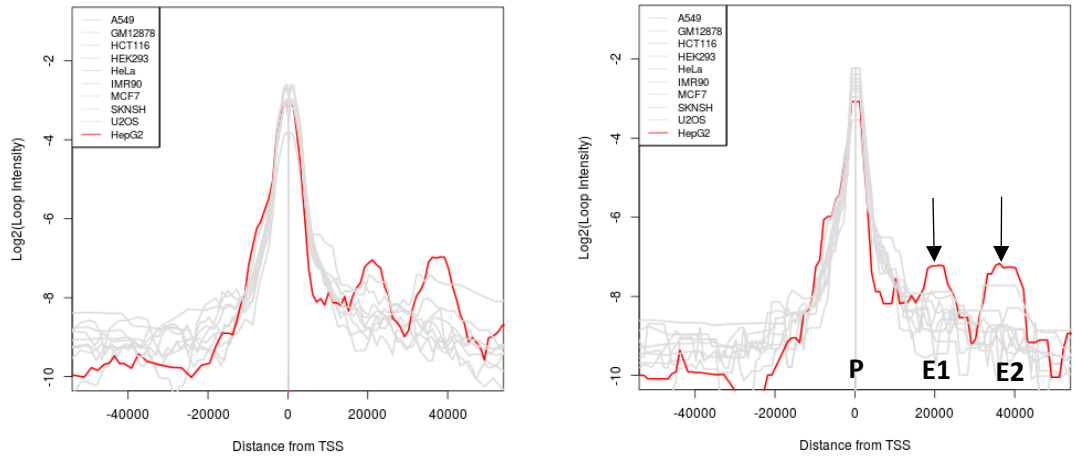


Figure 28. | Loop Intensities around the promoter of the gene APOC3: Loops intensity values in sliding windows of chr11 in all cell lines are shown. Basal intensity values were calculated as the mean and the median of 5 bins in 1000bp resolution. Left: mean intensity values. Right: median intensity values. HepG2 is highlighted in red color to show its different intensity basal value from the rest of cell lines. The arrows point to the two enhancers in the vicinity of the promoter.

Finally, aggregated peak analysis (APA), which examines the strength of a set of loops compared to their genomic surrounding, revealed no increase in interactions between induced p53 peaks and their putative target promoters (closets to the p53 peak) (Figure 29A). In contrast, as expected, cell type-specific loops showed strong aggregated peaks by this analysis (Figure 29B).

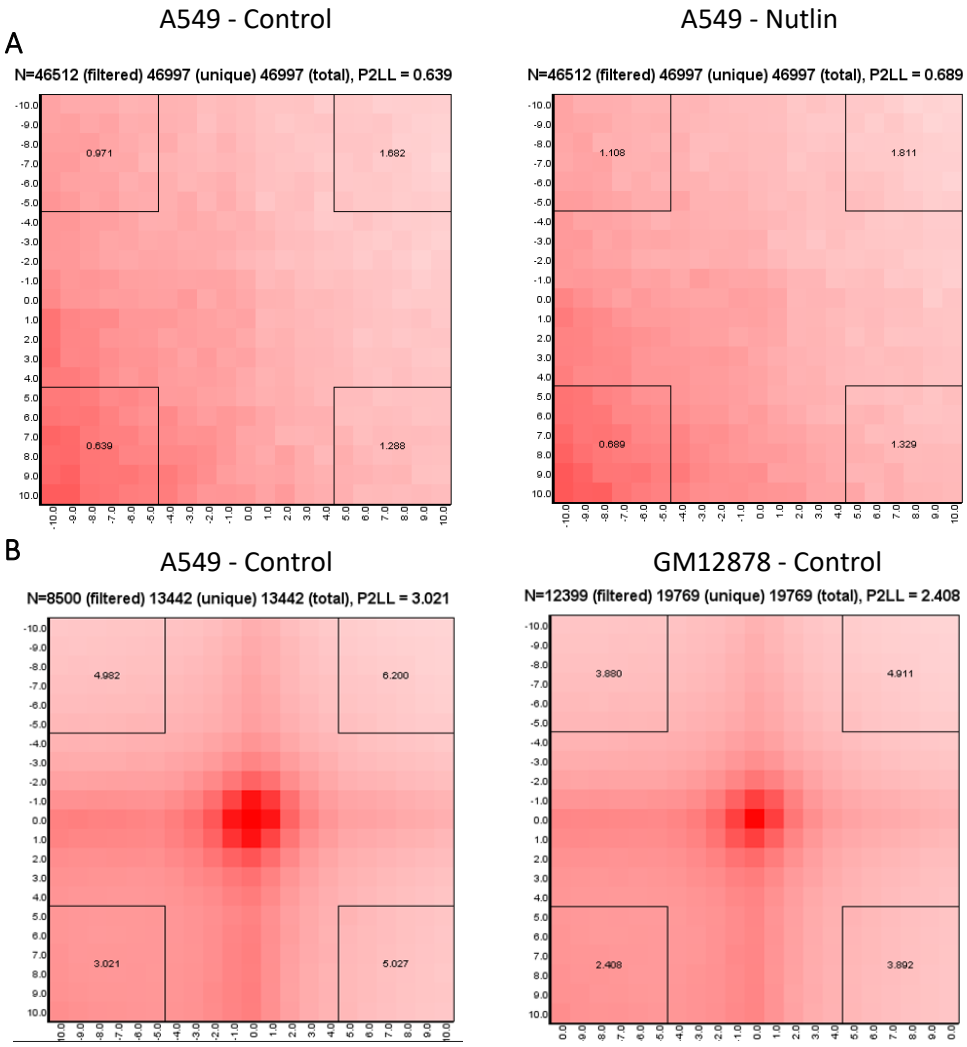


Figure 29. | (A) APA plots of interaction intensity between p53-ChIP-seq peaks and closest gene promoters. Left: control sample. Right: Nutlin treated sample of A549 cell line. No signal is detected. (B) Aggregated peaks in a contact map of loops detected in A549 control sample (left) and loops detected in GM12878 control sample (right). In this positive control analysis, a very strong interaction signal is evident.

## 8. Discussion

In this thesis, we investigated how the spatial structure of the genome correlates with the response of cells to stress, in the context of p53 activation.

First, we collected three different layers of data: RNA-seq, ChIP-seq, and Micro-C from ten different cell lines, before and after p53 activation using Nutlin, a potent activator of p53. The results of the RNA-seq analysis were in line with our hypothesis: As befits an activator, there were more induced genes than repressed genes after the treatment. Also, two cell lines in our panel have a defective p53 profile, and in those cell lines little response was seen compared to the other cell lines (Figure S1).

Second, we conducted p53 ChIP-seq analysis. There, too, we saw many differential p53 ChIP-seq peaks that were highly enriched for the p53 binding motif and for potential motifs of cofactors. We considered the closest genes to p53 ChIP-seq peaks as putative targets of p53 and showed that their overlap with the set of DEGs was highly significant. This test was performed twice, where putative targets were selected based on two possible maximal distances from a promoter, and in both tests we saw a significant overlap before and after treatment.

Third, we investigated the three-dimensional structure of the genome in the cell nucleus using Micro-C data. First, we quantitatively characterized the number of TADs and loops, the numbers were similar before and after p53 activation. After that, we sought loops whose levels of intensity changed significantly following the treatment with Nutlin. Unexpectedly, using a statistical tool for this task (HiCDC+) we did not find differential loops. On the other hand, when we conducted the same tests to compare two different basal cell types, dozens and sometimes hundreds of differential loops were detected.

In light of the above results, we decided to increase the resolution and study in the vicinity of the promoter of several known p53 genes that were robustly induced in our dataset. In each cell type, we calculated the average intensity level as well as the median intensity level over a sliding window of 5 windows of 1000 bases at a time. We found no significant change in the interaction strength of these loops before and after the activation of p53. On the other hand, looking at the basal level of the promoter-interaction intensities of a certain promoter of a gene that was induced in response to Nutlin only in one cell line and not in the other cells, we saw a change in the spatial structure between the cell in which where the induction took place and all the other types of cells tested, even before exposure to Nutlin (Figure 28).

In conclusion, our work shows a clear difference between the impact of the 3D organization of the genome on cell-identity and stress-induced transcriptional programs. While cell-identity programs are highly correlated with cell type-specific genome organization, we did not detect any similar correlations with the transcriptional response to p53. The most probable explanation for this difference is that changes in transcriptional activity between different cell types are orders of magnitude larger than changes induced within a cell type in response to stress. Therefore, we conclude that chromatin loops that are associated with cell identity are markedly stronger and more stable than chromatin loops induced by p53 activation. The current sensitivity of HiC (Micro-C) is sufficient for detecting the former, but misses most of the latter. Much higher sequencing depth, or revised protocols, are needed for this technique to detect also the chromatin structures that are associated with transcriptional programs modulated by stress responses.

## 9. References


[1] W. A. Bickmore, "The spatial organization of the human genome," *Annual Review of Genomics and Human Genetics*, vol. 14. pp. 67–84, Aug. 2013. doi: 10.1146/annurev-genom-091212-153515.

[2] H. bin Sun, J. Shen, and H. Yokota, "Size-dependent positioning of human chromosomes in interphase nuclei," *Biophys J*, vol. 79, no. 1, pp. 184–190, 2000, doi: 10.1016/S0006-3495(00)76282-5.

[3] M. Jackson, L. Marks, G. H. W. May, and J. B. Wilson, "The genetic basis of disease," *Essays in Biochemistry*, vol. 62, no. 5. Portland Press Ltd, pp. 643–723, Dec. 03, 2018. doi: 10.1042/EBC20170053.

[4] S. Nurk *et al.*, "The complete sequence of a human genome." [Online]. Available: https://www.science.org

[5] K. Pruitt, "Molecular and Cellular Changes During Cancer Progression Resulting From Genetic and Epigenetic Alterations," *Prog Mol Biol Transl Sci*, vol. 144, pp. 3–47, 2016, doi: 10.1016/BS.PMBTS.2016.09.001.

[6] Nature Education, "Transcription factors definition".

[7] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA Sequencing with Chain-Terminating Inhibitors," 1977.

[8] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1. pp. 57–63, Jan. 2009. doi: 10.1038/nrg2484.

[9] Y. Zhang *et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol*, vol. 9, no. 9, Sep. 2008, doi: 10.1186/gb-2008-9-9-r137.

[10] P. J. Park, "ChIP-seq: Advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10. pp. 669–680, Oct. 2009. doi: 10.1038/nrg2641.

[11] E. Lieberman-Aiden *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science (1979)*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, doi: 10.1126/science.1181369.

[12] T. H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, "Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C," *Cell*, vol. 162, no. 1, pp. 108–119, Jul. 2015, doi: 10.1016/j.cell.2015.05.048.

[13] M. Imakaev *et al.*, "Iterative correction of Hi-C data reveals hallmarks of chromosome organization," *Nat Methods*, vol. 9, no. 10, pp. 999–1003, Oct. 2012, doi: 10.1038/nmeth.2148.

[14] P. A. Knight and D. Ruiz, "A fast algorithm for matrix balancing," *IMA Journal of Numerical Analysis*, vol. 33, no. 3, pp. 1029–1047, 2013, doi: 10.1093/imanum/drs019.

[15] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, May 2012, doi: 10.1038/nature11082.

[16] L. Costantino, T.-H. S. Hsieh, R. Lamothe, X. Darzacq, and D. Koshland, “Cohesin residency determines chromatin loop patterns,” 2018, doi: 10.1101/2020.06.11.146902.

[17] N. Parikh *et al.*, “Effects of TP53 mutational status on gene expression patterns across 10 human cancer types,” *Journal of Pathology*, vol. 232, no. 5, pp. 522–533, 2014, doi: 10.1002/path.4321.

[18] L. A. Donehower *et al.*, “Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas,” *Cell Rep*, vol. 28, no. 5, pp. 1370-1384.e5, Jul. 2019, doi: 10.1016/j.celrep.2019.07.001.

[19] C. L. Brooks and W. Gu, “New insights into p53 activation,” *Cell Research*, vol. 20, no. 6. pp. 614–621, Jun. 2010. doi: 10.1038/cr.2010.53.

[20] N. C. Durand *et al.*, “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments,” *Cell Syst*, vol. 3, no. 1, pp. 95–98, Jul. 2016, doi: 10.1016/J.CELS.2016.07.002.

[21] M. Ganji *et al.*, “Real-time imaging of DNA loop extrusion by condensin Downloaded from,” 2018. [Online]. Available: http://science.sciencemag.org/

[22] M. Sahin, W. Wong, Y. Zhan, K. van Deynze, R. Koche, and C. S. Leslie, “HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP,” *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-23749-x.

[23] K. Kruse, C. B. Hug, and J. M. Vaquerizas, “FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data,” *Genome Biol*, vol. 21, no. 1, Dec. 2020, doi: 10.1186/s13059-020-02215-9.

[24] M. Ringnér, “What is principal component analysis?,” 2008. [Online]. Available: http://www.nature.com/naturebiotechnology

[25] M. Fischer, “Census and evaluation of p53 target genes,” *Oncogene*, vol. 36, no. 28. Nature Publishing Group, pp. 3943–3956, Jul. 13, 2017. doi: 10.1038/onc.2016.502.

[26] I. Ben-Sahra *et al.*, “Sestrin2 integrates Akt and mTOR signaling to protect cells against energetic stress-induced death,” *Cell Death Differ*, vol. 20, no. 4, pp. 611–619, Apr. 2013, doi: 10.1038/cdd.2012.157.

[27] 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. 576–589, May 2010, doi: 10.1016/j.molcel.2010.05.004.

[28] T. L. Bailey, “DREME: Motif discovery in transcription factor ChIP-seq data,” *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, Jun. 2011, doi: 10.1093/bioinformatics/btr261.

[29] N. C. Grieder, D. Nellen, R. Burke, K. Basler, and M. Affolter, “schnurri Is Required for Drosophila Dpp Signaling and Encodes a Zinc Finger Protein Similar to the Mammalian Transcription Factor PRDII-BF1,” 1995.

[30] A. Roayaei Ardakany, H. T. Gezer, S. Lonardi, and F. Ay, “Mustache: Multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation,” *Genome Biol*, vol. 21, no. 1, Sep. 2020, doi: 10.1186/s13059-020-02167-0.

# Supplementary

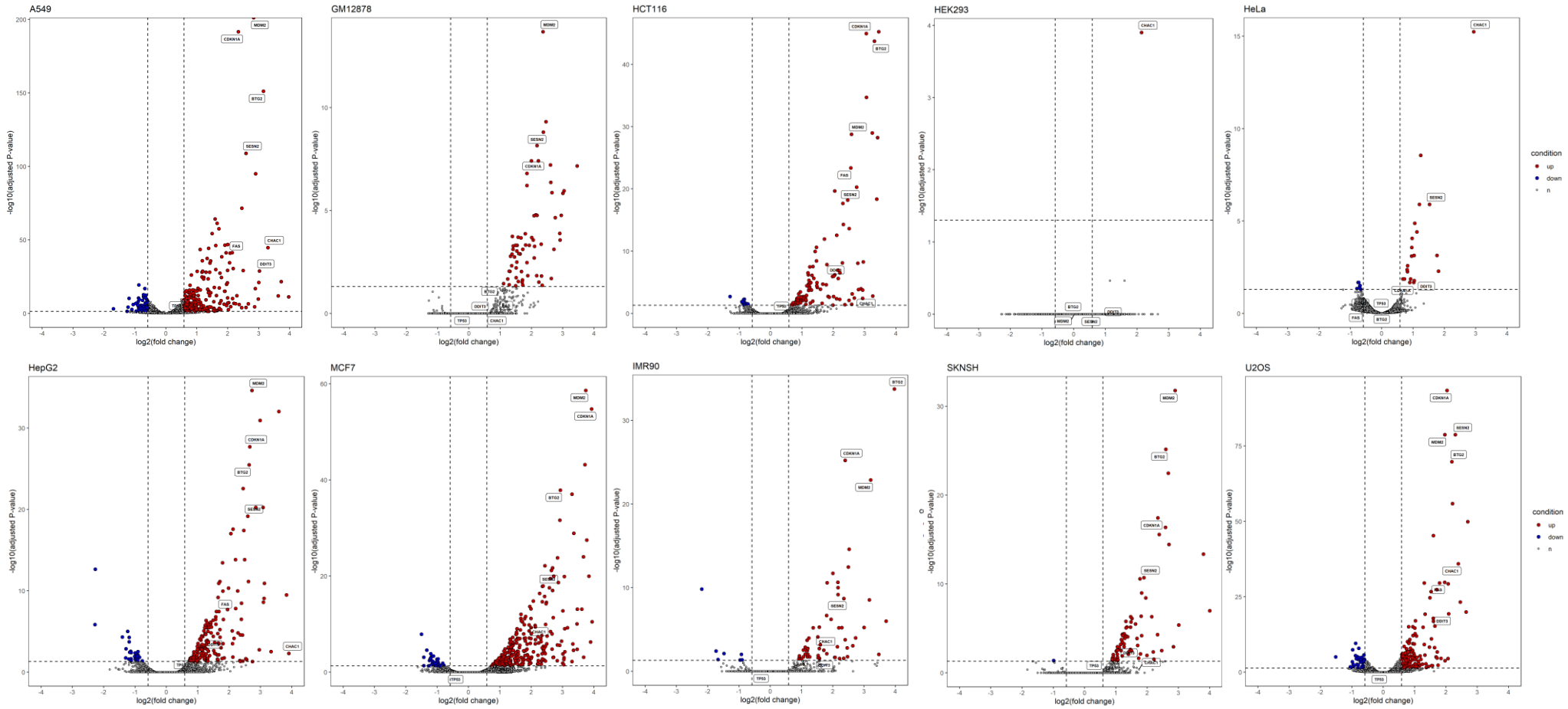


Figure S1. | Volcano plots of all 10 cell lines. Each dot represents a differentially expressed gene. Red dots represent up-regulated genes, blue dots represent down-regulated genes, and genes in grey are not significantly differentially expressed. Labeled genes are well known for P53.

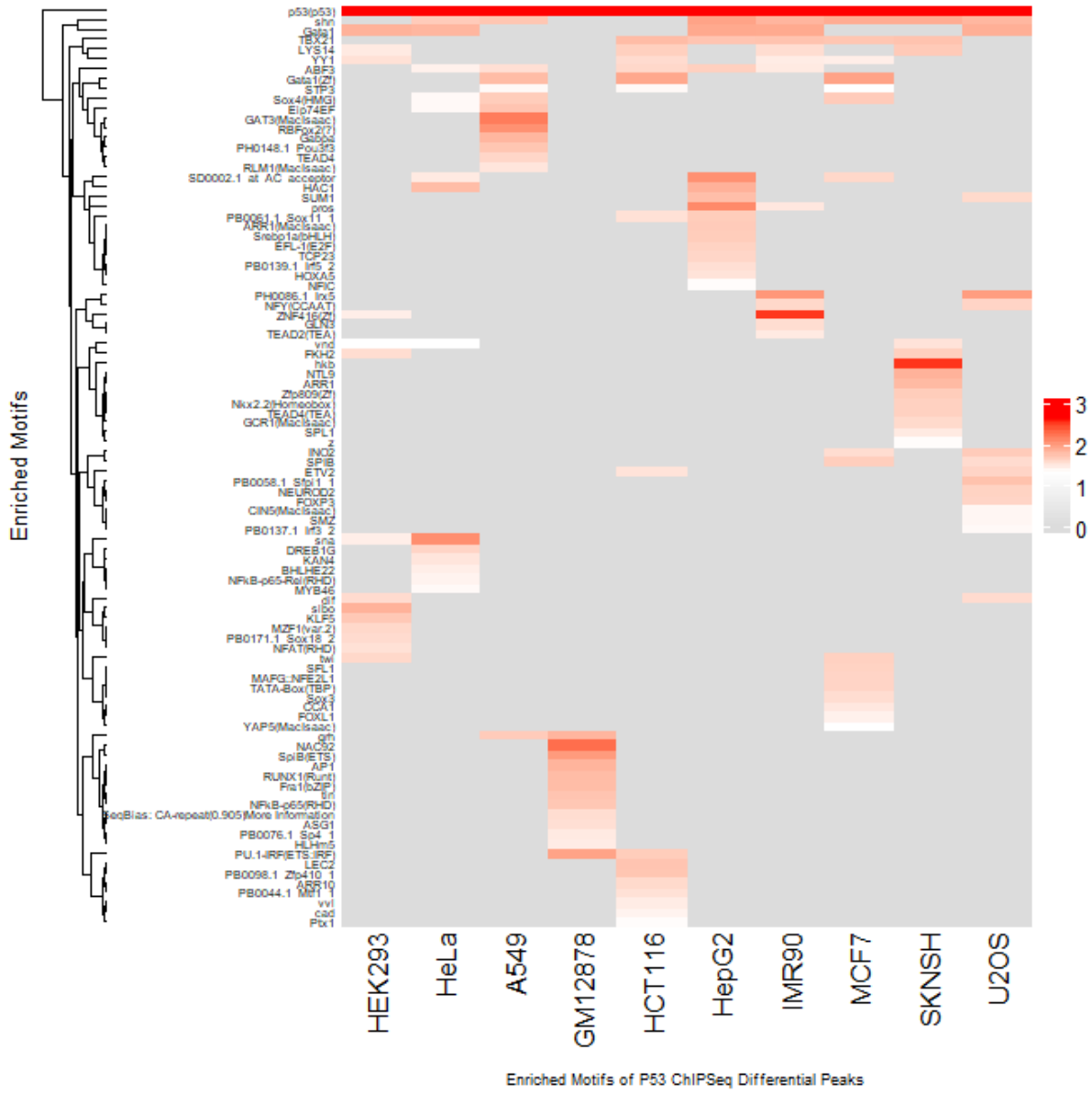


Figure S2. | P-values of de novo motifs similar to known ones. An enriched motif is defined as one with p-value<=1e-20. Scale is between 0 to 3 such that zero in non-enriched and 3 is highly enriched motif. The scaling of the p-values to the range [0,3] was done as follows: $p-value = 10^{-n} \rightarrow log10(10^{-n}) = -n \rightarrow (-n)\cdot(-1) = n \rightarrow log10(n) = [0,3]$. P53 is highly enriched in all cell lines.

Median - chr8 TP53INP1 gene:

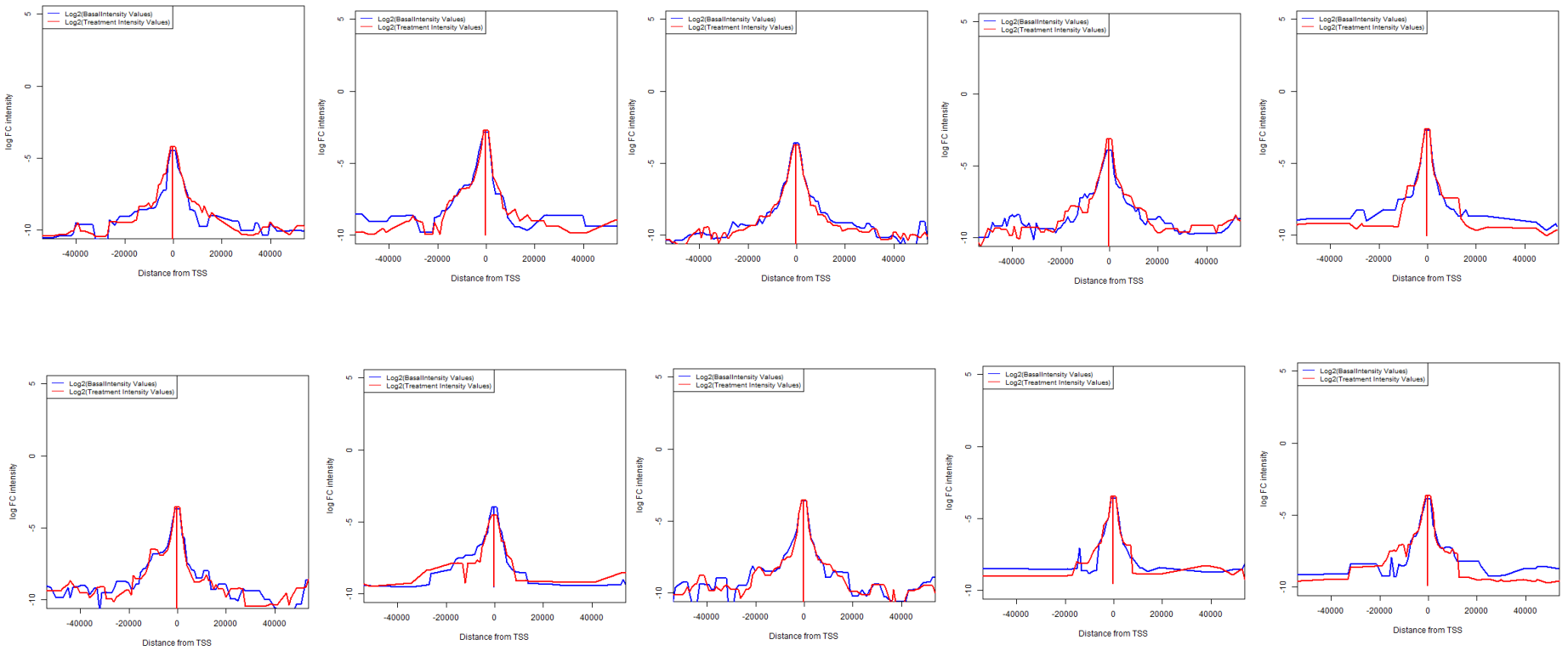


Figure S3. | Loop Intensities around Promoter: Sliding window of loops intensity values of chr8 in all cell lines are shown. Intensity values were calculated as the median of 5 bins in 1000bp resolution. 0 is the promoter of TP53INP1 gene. From left to right: A549, GM12878, HCT116, HEK293, HeLa, HepG2, IMR90, MCF7, SKNSH, U2OS.

Mean – chr8 TP53INP1 gene:

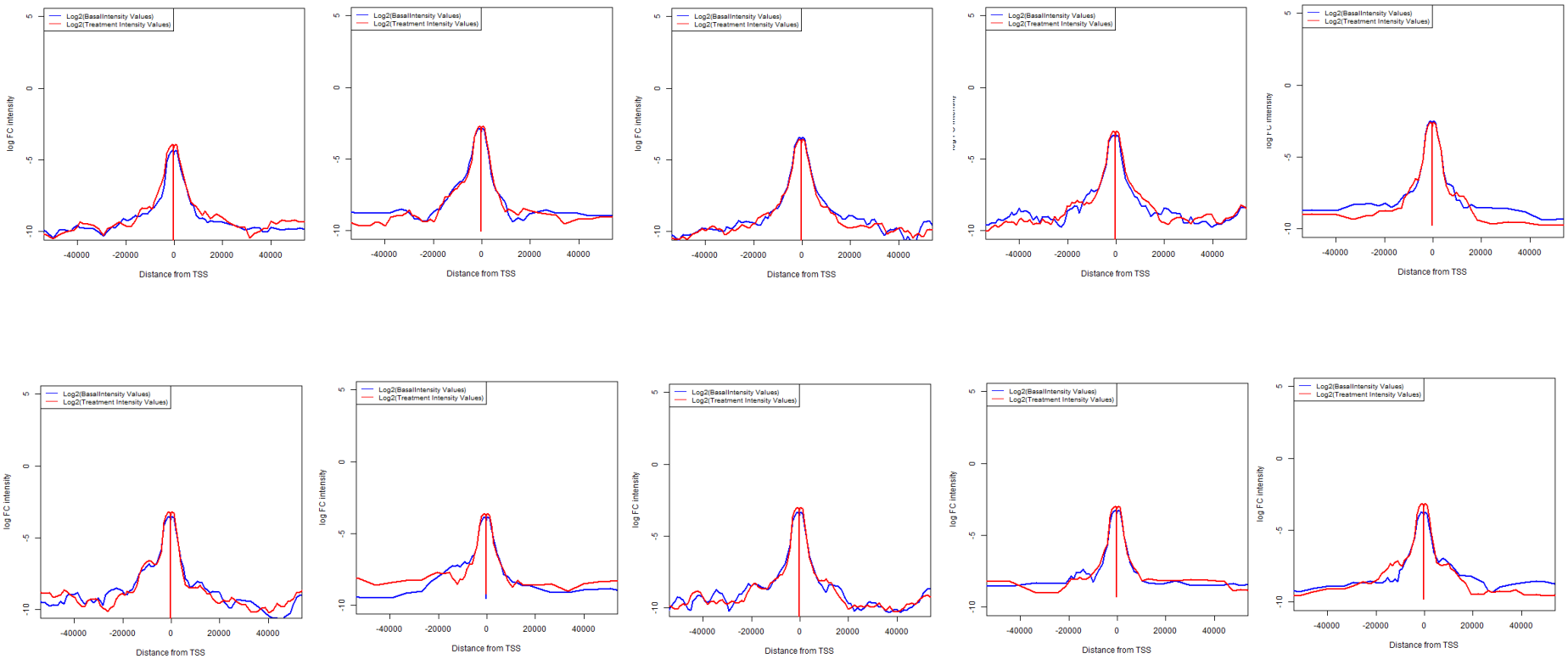


Figure S4. | Loop Intensities around Promoter: Sliding window of loops intensity values of chr8 in all cell lines are shown. Intensity values were calculated as the mean of 5 bins in 1000bp resolution. 0 is the promoter of TP53INP1 gene. From left to right: A549, GM12878, HCT116, HEK293, HeLa, HepG2, IMR90, MCF7, SKNSH, U2OS.

Mean – chr6 CDKN1A gene:

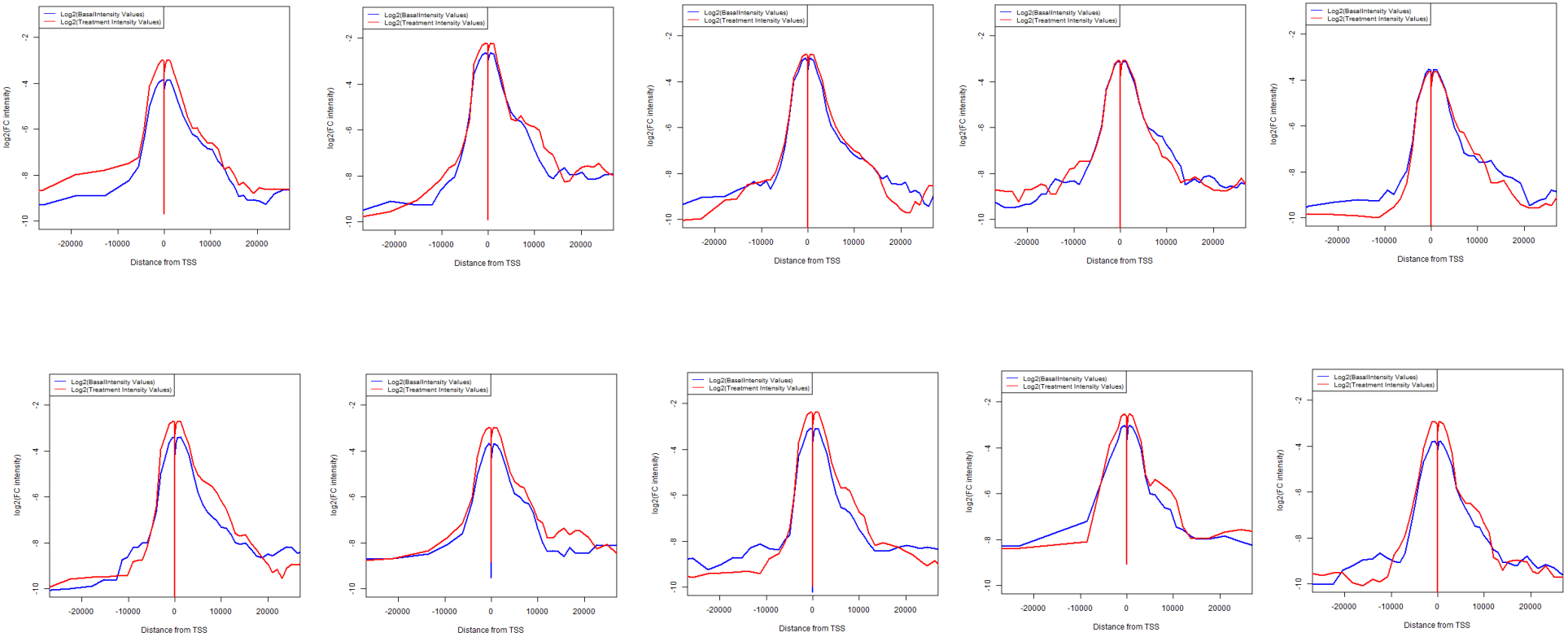


Figure S5. | Loop Intensities around Promoter: Sliding window of loops intensity values of chr6 in all cell lines are shown. Intensity values were calculated as the mean of 5 bins in 1000bp resolution. 0 is the promoter of CDKN1A gene. From left to right: A549, GM12878, HCT116, HEK293, HepG2, HeLa, HepG2, IMR90, MCF7, SKNSH, U2OS.

Median – chr6 CDKN1A gene:

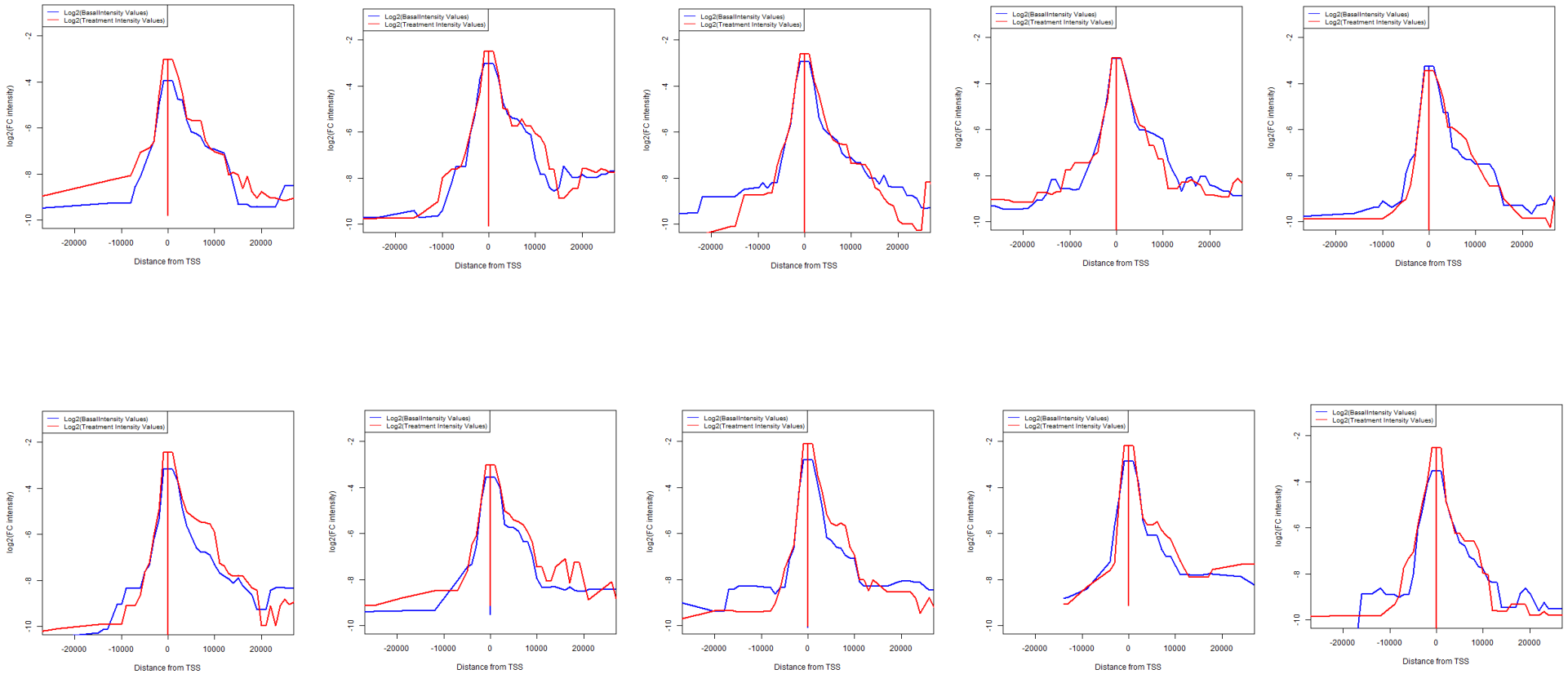


Figure S6. | Loop Intensities around Promoter: Sliding window of loops intensity values of chr6 in all cell lines are shown. Intensity values were calculated as the median of 5 bins in 1000bp resolution. 0 is the promoter of CDKN1A gene. From left to right: A549, GM12878, HCT116, HEK293, HepG2, HeLa, HepG2, IMR90, MCF7, SKNSH, U2OS.

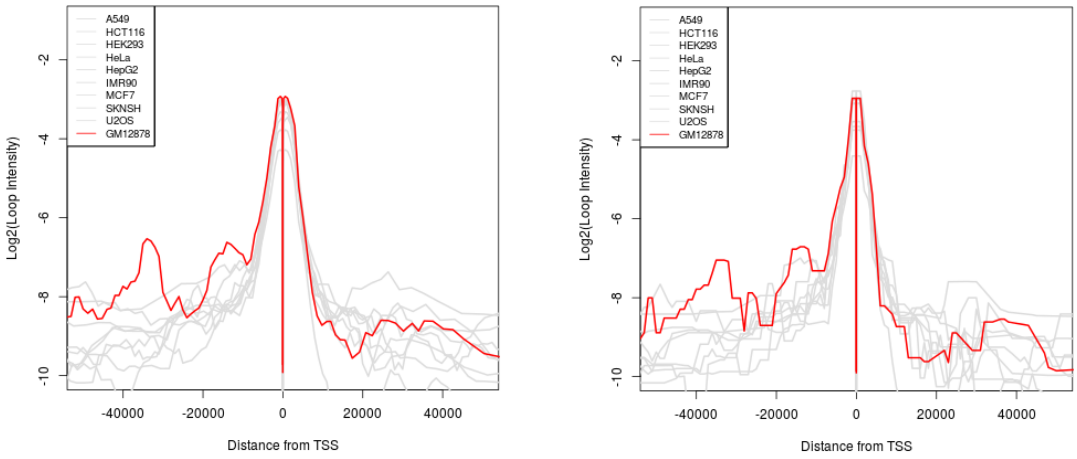


Figure S7. | Loop Intensities around Promoter of the gene CCR7: Loops intensity values in sliding windows of chr17 in all cell lines are shown. Basal intensity values were calculated as the mean or the median of 5 bins in 1000bp resolution. Left: mean intensity values. Right: median intensity values. GM12878 is highlighted in red color to show its difference intensity basal value from the rest of cell lines.

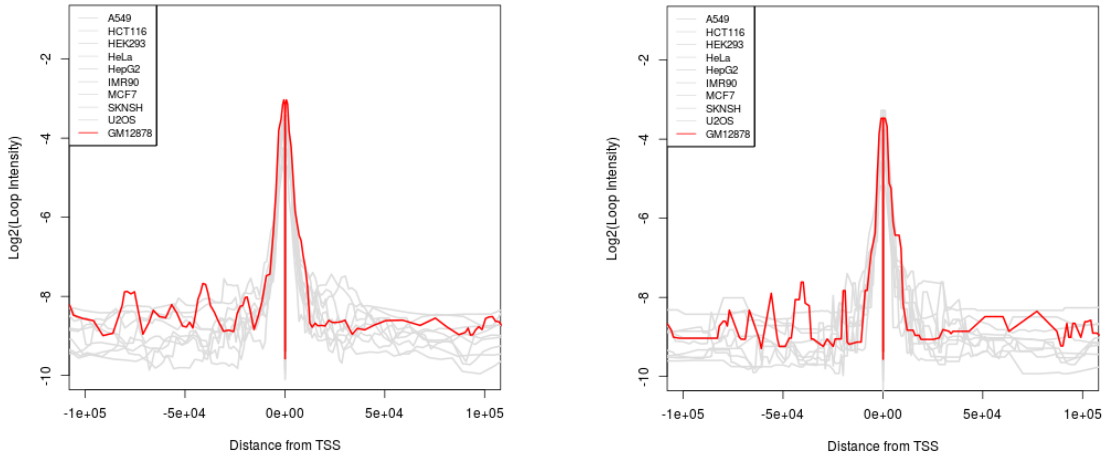


Figure S8. | Loop Intensities around Promoter of the gene CD80: Loops intensity values in sliding windows of chr3 in all cell lines are shown. Basal intensity values were calculated as the mean or the median of 5 bins in 1000bp resolution. Left: mean intensity values. Right: median intensity values. GM12878 is highlighted in red color to show its difference intensity basal value from the rest of cell lines.

# 10. תקציר

מחקרים רבים הראו כי תגובות שעתוק למצבי עקה הינן ספציפיות מאוד לסוג התא. עם זאת, המנגנונים העומדים בבסיס ספציפיות רקמה זו נותרו לא ידועים במידה רבה. במחקר שלנו, אנו מתמקדים, כמערכת מודל, ברשת השעתוק המופעלת על ידי p53.

p53, המכונה "שומר הגנום", הוא הגן העיקרי המדכא גידולים בגנום שלנו, והוא משמש כמנגנון הגנה מרכזי מפני טרנספורמציה של סרטן. הפעלת p53 בסוגי תאים ורקמות שונים מביאה להשראה של רשתות שעתוק שונות מאוד, לצד הפעלה של תגובת ליבה אוניברסלית של p53. המטרה העיקרית של המחקר שלנו היא למצוא גורמים שקובעים תגובות ספציפיות לסוג תא. המחקר שלנו השתמש בשלוש שכבות של נתוני omics: RNA-seq, ChIP-seq ו-Micro-C (גרסה משופרת של Hi-C, עם רזולוציה גבוהה), שנאספו על עשר שורות תאים שונות. עבור כל שורת תאים בוצעו מדידות הן בבקרה והן בתאים שטופלו על ידי Nutlin, מפעיל חזק של p53.

בניתוח של נתונים נרחבים אלה, זיהינו (1) עשרות אירועי קישור מסוג p53 ספציפיים לסוג תא; (2) קו-פקטורים מסוג p53 ספציפיים לתא; (3) אירועי קישור מסוג p53 ספציפיים לסוג תא בקורלציה עם ביטוי גנים ספציפי המושרה על ידי p53, ו-(4) אינטראקציות פיזיות של מגביר-תחל ספציפיות לסוג תא. בדקנו באיזו מידה תגובות הנגרמות על ידי p53 והספציפיות לסוג התא נמצאות בקורלציה עם תכונות הארגון המרחבי של הגנום הספציפיות לסוג התא. לא מצאנו שינויים בארגון המרחבי בשום שורת תאים בתגובה ל-Nutlin, לעומת שינויים מרחביים רבים שנצפו בין שורות תאים שונות שלא טופלו. השערתנו היא שהשינויים בתגובה לעקה קטנים בסדרי גודל לעומת השינויים בין שורות תאים שונות ואינם בטווח הזיהוי של השיטות הנוכחיות ל Hi-C.

אוניברסיטת תל אביב

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"י בלווטניק

# ניתוח הארגון המרחבי של הגנום והשפעתו על תגובות שעתוק של p53 הספציפיות לסוג תא

חיבור זה הוגש כעבודת גמר לתואר "מוסמך האוניברסיטה"

בבית הספר למדעי המחשב

על ידי

## הדר עמירה-חכם

בהנחייתם של

## פרופ' רון שמיר

## פרופ' רן אלקון

שבט תשפ"ג