Aix-Marseille University
The Faculty of Sciences
MSc : Computational and Mathematical Biology
2020-2021

Study of the 3D chromatin organization in the context of OAS3 Epromoter regulation inside the OAS cluster upon IFNα stress stimulation

Author : Ayachi Youssef
Internship Supervisor : Salvatore Spicuglia
Co-Superviser : Charbel Souaid
June 7, 2021

Laboratoire TAGC/Inserm U1090. Theories and Approaches of Genomic Complexity Parc scientifique et technologique de Luminy 163, avenue de Luminy - Case 928 13288 Marseille cedex 09

ABSTRACT

Inside the nucleus of the mammalian cells, the chromatin organization in the 3D space is controlled by intra-chromatin interactions and this special arrangement of the genome plays a crucial role in gene regulation. By interacting together towards transcription factors, promoters and enhancers participate in the 3D organization of the chromatin. In addition, the existence of promoters with both gene promoter and enhancer activity, called Epromoters, increase the complexity of this chromatin organization. To study those interactions and to understand more the 3D organization of the chromatin inside the nucleus, several techniques have been developed, especially the Chromosome Conformation Capture (3C) and 3C-derived methods. Here, we focus on the study of the chromatin interactions happening inside and nearby the OAS cluster of the human chromosome 12 which comprises the *OAS1*, *OAS2* and *OAS3* genes. These three genes are known to be activated after a Type-I Interferon-alpha (IFNα) stress response and this cluster is known for its *OAS3* Epromoter which, once activated, has an enhancer activity on the *OAS1* and the *OAS2* gene promoters. Here, we use a 3C based technique, Hybrid Capture Hi-C, to comprehensively detect the *OAS3* Epromoter interactions with its target *OAS1* and *OAS2* gene promoters before and after IFNα stress stimulation and to quantify the intra-chromatin interactions happening inside and nearby the OAS gene cluster upon the IFNα stress stimulation.

INTRODUCTION

During the development of multicellular organisms, the fertilized egg cell undertakes successive divisions that lead to the formation of the embryo then to the complete organism. Thereby, despite their different functions in specialized tissues, almost all cells share the identical genomic content. This is due to a distinctive gene regulation mechanism that acts essentially on the transcriptional level. At this level, multiple Cis-Regulatory Elements (CREs) abundantly found in the intergenic regions (previously called "Junk DNA"), were shown to play an important role in gene regulation. The most outstanding CREs are promoters and enhancers.

Promoter elements are DNA sequences localized upstream the Transcription Starting Site (TSS) of the gene that are mandatory for transcription. Their activity is orientation dependent and confers a basal transcription level of the gene. Even so, many promoters have shown to have a bidirectional activity[1]. The length of the promoter sequence can extend up to 1 kilobase (kb) upstream the TSS, however the core promoter usually corresponds to the region of 50 base pair (bp) surrounding the TSS. Upon gene activation, the core promoter recruits the preinitiation complex (PIC) composed of general transcription factors (GTFs) and the RNA polymerase II. Different elements were described to compose the core promoter. The most known is the TATA box, which is a T/A rich sequence (TATAA) located 30 bp upstream the TSS and that is essential for the positioning of the pre-initiation complex via the recruitment of the TATA-binding protein (TBP). Core promoters are often rich in CpGs, and roughly 50% of human promoters overlap CpG islands. Other elements like the initiator element (INR) sequence that overlap

TSS and downstream promoter elements (DPE) are also found to play a role in the recruitment of the PIC[2].

Enhancer elements are DNA sequences localized distantly -up to 1 Mb- from their target genes[3]. They are important for the enhancement of the basal promoter transcription activity and they can act independently from their orientations. As promoters, enhancer sequences' length is around hundreds of bp and they are mainly associated with the regulation of tissue-specific genes[4]. Super-enhancers are larger regions that can extend to several kb where multiple enhancers are clustered in close proximity. They are involved in the regulation of cell identity[5]. Similarly to promoters, enhancers recruit transcription factors and RNA polymerase II[6]. Thus, they are subjected to an active transcription of their sequences into enhancer RNAs (eRNAs)[7]. However, the transcription of enhancers is generally not productive, leading to short unstable bidirectional transcripts[8]. Another marked difference between mammalian gene promoters and enhancers is the overall CG content whereas almost no enhancers do overlap CpG islands[9,10].

Multiple epigenetic modifications are associated with the activity of promoters and enhancers. When promoters and enhancers become active, nucleosomes surrounding their regions are evicted creating nucleosome depleted regions (NDR) allowing the binding of the transcription machinery. Histone tails (N-terminal) of certain nucleosomes are subjected to specific post-translational modifications associated with the activity of CRE. Thus, nucleosomes in the vicinity of active promoters and enhancers were found to have a high level of lysine-27 acetylation at the N-terminal of their Histone 3 component (H3K27ac)[11]. Nucleosomes on active promoters are preferentially highly enriched in Histone-3 kinase-4 trimethylation (H3K4me3) at their N-terminal, whereas in active enhancers they are more enriched in Histone-3 kinase-4 monomethylation (H3K4me1)[11]. These modifications are often used as epigenetic marks to identify active promoters and enhancers[11]. Active promoter and enhancer regions can be predicted using ChIP-seq technique. In this technique, genomic regions of active CRE are co-immunoprecipitated using an antibody targeting a specific histone modification and are then identified by high-throughput sequencing[12].

Enhancers activate transcription on promoters by coming into the proximal vicinity of the promoter and the formation of a DNA loop. The formation of the DNA loop model was validated in 2002 via the development of the 3C techniques[13]. The 3C experiments are based on three steps (Fig.2). DNA interactions between different regions are established via proteins. Thus, the first step comprises a crosslinking of DNA and proteins in the cell population. This allows the creation of covalent bonds between DNA fragments and proteins and subsequently the fixation of the 3D organisation of the chromatin inside cells' nuclei. Secondly, the chromatin is digested using a restriction enzyme. This digestion allows the creation of overhang DNA fragments around DNA crosslinked interacting fragments. A third step of proximity ligation is done by adding a ligase enzyme, under a diluted condition that will specifically promote ligation events between the sticky ends of fragments engaged in the same DNA loop. This leads to the formation of a chimeric DNA molecule between fragments that were forming a DNA

loop. The more the interaction is present in the cell population, the more proximal ligation events will be favored and the more chimeric DNA molecules are obtained. The assessment of two DNA fragments interaction is then computed as a proportion of the number of copies of the chimeric molecules found in the final 3C library which represents the interaction frequency between those two fragments. In the conventional 3C technique, the interaction frequency is measured by a real-time quantitative polymerase chain reaction (qPCR) using primers designed to amplify specific chimeric molecules. New 3C based techniques were derived by the development of high throughput sequencing technologies such as 4C (Circular Chromosome Conformation Capture), Hi-C (High throughput sequencing 3C) and Hybrid Capture Hi-C. In the 4C, an additional ligation step is added to the 3C protocol in order to obtain a circular library. Then, an inverse PCR is performed from one single locus in the genome. This allows the amplification of all the genomic contacts of this single locus. Those contacts are then identified by high throughput sequencing technology. In Hi-C, all genomic contacts of the library are sequenced. This led to the identification in an unbiased manner of the complete genomic interactions. In order to study interactions in an exclusive specific region (usually around 3 megabase Mb) or in a subset of CREs, RNA probes are used to capture specific interactions from the 3C-based library followed by sequencing[14].

The advance of both 3C-based and microscopy techniques has provided new discoveries about genome organisation in the three dimensional space. By microscopy, mammalian chromosomes occupy different territories inside the nucleus[15]. These are known as chromosome territories. On the other hand, Hi-C experiments in mammals confirmed the low abundance of interchromosomal interactions[16]. In eukaryotes, the genome is compartmentalized inside the nucleus into domains at sub-megabase level where higher intra domain interactions are promoted. Those domains are called Topologically Associated Domains (TADs) and are marked by boundaries. In mammals, TAD boundaries are often enriched in CTCF protein and cohesin complexes, or are transcriptionally active regions. Inside those domains, promoter-enhancer DNA looping is favored. In some cases, the deletion of a TAD boundary region has led to a non-specific activation of genes in the neighboring TAD, leading to polydactyly and brachydactyly malformations[17]. The complete abolition of TADs structure upon CTCF deletion did not lead to an abnormal global gene expression in macrophages, however the specific proinflammatory response genes were severely affected[18].

Classically, the assessment of the activity of cis-regulatory elements is done by using a single gene-reporter assay for each element[19]. Briefly, after cloning the CRE candidate in a gene-reporter vector, the assessment of element activity can be done according to the activity of the emitted signal of the reporter gene itself. Combining gene-reporter assays with high-throughput sequencing have enabled quantitative measurements of enhancer activity of thousands of regulatory elements simultaneously. One of the most commonly used approaches is the self-transcribing active regulatory region sequencing (STARR-seq) and its derivative Capture STARR-seq (CapSTARR-seq) technique developed in our lab by the Spicuglia Team. The principle of these techniques is to clone candidates inside the open reading frame (ORF) of a reporter gene that is under the control of a promoter with a basal activity. Candidates can be

fragments from the whole genome (as in STARR-seq) or selected candidates using a microarray capture approach (as in CapStarr-seq). If the candidate acts as an enhancer element, the basal promoter activity will be enhanced and consequently the candidate region will be transcribed itself. Due to the high correlation between enhancer activity and the transcription of the element itself, a final transcriptomic analysis targeting transcripts from the reporter gene led to the quantification of the enhancer activity.

By quantifying the potential enhancer activity after a capture selection of all human promoters, our team has shown that at least 3% of human promoters displayed an enhancer activity in a given cell line[20]. This suggests that certain human promoters act as enhancers and coregulate nearby promoters which adds further complexity of the gene regulation process. This type of promoters was classified as a new class of cis-regulatory elements called Epromoters and was identified in different organisms such as mice and Drosophila in subsequent studies[21]. Epromoters are gene promoters that share both promoters and enhancers characteristics. They exhibit a promoter activity on the proximal downstream gene and an enhancer activity on a distal promoter by forming a DNA loop[22]. The deletion of these elements by CRISPR/Cas9 leads to the inactivation of both proximal and distal regulated genes[23]. However, the inversion of the Epromoter leads only to the abolition of the proximal-promoter activity but not the enhancer distal activity. Epromoters were shown to play a role in the regulation of ubiquitously expressed genes, but interestingly also in the regulation of the stress response and in particular the Type-I interferon-response genes[24].

Depending on the stimulus and the cell-type, the response to stress requires a highly specific and dynamic gene regulation. During the inflammatory response, cells communicate via cytokine signaling proteins. Cytokines bind to their specific cell surface receptors which leads to the activation of a particular transcription regulatory process . The most well studied inflammatory regulation process is the regulation of interferon (IFN) response genes. There are mainly two types of interferons that regulate the activation of immune response: the type I (IFNα/IFNβ) and the type II (IFNγ). Type-I IFN cytokines activate the interferon-response genes in almost all cells during viral infection. Its aim is to inhibit the replication of the virus inside the host infected cell and thus to increase the antiviral defense. However, type-II IFN is restricted to the immune cells and is activated by cytokines such as IL-12. Both types require specific transcription factors regulators to act on the targeted genes such as STAT1/2 and IRFs factors for type-I IFN response.

Recently in the laboratory, the team have investigated the role of the newly discovered Epromoter elements in the interferon stress-response. For that, in recent unpublished data, the team have firstly identified Epromoters in Human chronic myeloid K562 cells upon 6 hours of stimulation with type I IFNα. During this stimulation, the team validated using CapSTARR-seq, RNA-seq and ChIP-seq techniques that Epromoters involved in the IFN response acts in clusters of induced genes. They were found to be surrounded in less than 100kb distance by one or more induced genes that do not exhibit

Epromoter activity. Inside those clusters, Epromoter preferentially recruits the specific interferon-response transcription factors IRFs and STAT1/2. Those clusters were classified as clusters of Epromoters. In this model, after the stimulation of IFNα, Epromoters recruit the specific regulator factors and activate nearby genes in order to achieve a coordinated rapid induction. Several interferon-response genes clusters were identified such as the 2'-5'-oligoadenylate synthetase (OAS) gene cluster. The OAS gene cluster comprises 3 genes: *OAS1*, *OAS2* and *OAS3*. Those genes are activated after IFN response and exhibit an antiviral activity by the activation of RNaseL that degrades RNAs in the cells including viral RNA[25]. In this cluster, *OAS3* promoter exhibits a highly induced Epromoter activity in the Cap-STARR-seq assay, and preferentially binds the stress-specific STAT1/2 and IRF factors upon stimulation (unpublished data). The deletion using CRISPR/Cas9 technique of *OAS3* Epromoter, but not of the *OAS1* and *OAS2* promoters, leads to the simultaneous inactivation of the three OAS genes in the cluster.

3D chromatin organization is known to play a role in the regulation of transcription[26]. Promoter-enhancer pairs are found inside specific TADs domains, and they interact together to form a hub of DNA loops[27]. Intra and inter-chromosomal DNA long distance interactions were shown to be formed between promoters that bind the same transcription factors and to form a hub of transcription factories using microscopy and 3C based techniques. However, it is still unclear how the chromatin is organized upon the stimulation by the Epromoter. In this context, I will try at first to identify if the intra-chromatin interactions known as DNA loops are induced by IFNα stress stimulation and then to figure out the role of the *OAS3* Epromoter in the DNA loop formation. We hypothesize that upon stimulation, the intra-chromatin interactions inside the OAS cluster will increase and that the Epromoter activity is responsible for the formation of DNA loops inside the cluster.

Thus, to understand the dynamic of 3D chromatin interactions involved by Epromoters inside the induced gene clusters, the team has generated Hybrid Capture Hi-C experiments targeting a 3 Mb region around the OAS cluster. The experiments were performed before and after the stimulation in the Wild-Type (WT) and in the Knock-Out (KO) deletion of the *OAS3* Epromoter in K562 cells. In order to answer our hypothesis, the aim during the internship was to analyze the Hybrid Capture Hi-C experiments at a different level. First, a statistical test that aims to identify highly enriched DNA loops was applied. Second, an analysis of the intradomain interactions within the cluster of *OAS3* was applied. Through this work, I have determined that even so DNA loops were not statically confirmed, in the WT stimulated K562 *OAS3* promoter induce a local increase of interactions within the cluster.

METHODS

To analyze the generated Hybrid Capture Hi-C interaction matrices, I chose to use different statistical and bioinformatic tools. As is the case for most of the data analysis pipelines, I started by using R software. R is a programming language developed specifically for organising, manipulating and visualising data. It also allows us to compare different datasets using statistical tests. The choice of the statistical test depends on the compared datasets. Knowing that in addition our datasets have a large number of values (N>30), the optimal statistical test to choose in this case have to be one of the two most used one : the Two-sample T-test (also called the independent samples t-test) or the Wilcoxon test (also known as Wilcoxon–Mann–Whitney test). The two-sample T-test requires that the compared datasets should be normally distributed. Since our datasets do not follow a normal distribution (Annexe, Fig.1.) and knowing that the Wilcoxon test does not require the mentioned condition, my choice has turned to this latter. As an alternative to the two-sample T-test, the Wilcoxon test is a non parametric test, i.e. can be used when the distributions of the studied datasets are not specified. Performing a Wilcoxon test aims to verify the null hypothesis which announces that there is no significant difference between the two compared datasets. Here, I will compare two by two both independent datasets, i.e. datasets that are randomly selected from two different cell populations[28], and dependent datasets. Thus, I will use respectively the unpaired and the paired Wilcoxon test to compare independent and dependent datasets.

Among its other numerous functions, R software is also used for its graphical representations of large datasets. Infact, visualisation is the center of data analysis and R provides several packages that are useful to this aim. To visualise the generated Hybrid Capture Hi-C interaction matrices, I used the ggplot2 package in addition to the R basic graphic commands. Using these tools, I was able to build histograms, boxplots and heatmaps that will be interpreted and discussed across the following sections. The R script that has led to the following results is available on http://bit.ly/YA2k21HiC.

To be able to interpret our Hybrid Capture Hi-C datasets in terms of biological meaning and more specifically in terms of intra-chromatin interactions, the use of a genome browser is required. Here, I chose the WashU Epigenome Browser (see https://epigenomegateway.wustl.edu/). Basically, a genome browser is a graphical interface that enables researchers to navigate across the genome of a given species and to visualise annotated data generated after a biological experiment such as our Hybrid Capture Hi-C experiment. Here, I used this powerful tool to visualise the intra-chromatin interactions that happen inside the studied genome region, especially the interaction of the *OAS*3 Epromoter with the nearby genome fragments.

RESULTS

Hybrid Capture Hi-C data matrices

Previously, our team have validated that the *OAS3* Epromoter recruit the stress specific transcription factors after the IFNα stimulation in order to activate the transcription of the *OAS3* and also *OAS1* and *OAS2* genes located in a cluster less than 100 kb distant from *OAS3* Epromoter. The enhancer activity of *OAS3* Epromoter was validated in both CapSTARR-seq and gene reporter assay. In order to determine the changing of the 3D organization of the chromatin in this cluster upon the stress response, the laboratory have generated Hybrid Capture Hi-C data before and after the stimulation in the Wild-Type (WT) and in the Knock-Out (KO) deletion of the *OAS3* Epromoter in K562 cells.

During Hybrid Capture Hi-C, as all the 3C-based techniques, the generated library contains all the possible interactions in the genome. This is due to the ligation of close proximal DNA fragments cut with a specific restriction enzyme. Restriction enzymes are enzymes that cut the DNA after the recognition of a specific restriction site usually consisting of several bp. In a Hybrid Capture Hi-C library, the most commonly used enzyme is DpnII (MboI) which is a 4bp-cutter recognizing 'GATC' palindromic sites. Approximately, the DpnII enzyme cuts each 200bp of the human genome, generating around 16 millions fragments. The hypothetical total number of possible interactions is then up to 2.5e+16 interacted chimeric molecules. Each sequenced chimeric molecule represents an interaction, and the fraction of sequenced reads of each interaction represents the frequency (or score) of this interaction inside the cell population. In order to achieve a complete coverage at the restriction enzyme resolution, a high sequencing depth covering several times all the possible interactions is needed. However, high throughput sequencing is highly cost-effective. Thus, Hybrid Capture Hi-C experiments are not sequenced at their maximum resolution. In order to overcome this problem, the genome is subdivided into specific windows or bins where the scores inside those bins are summed. In the litterature, it is commonly accepted that at a given resolution, at least 1000 interactions per bin have to be summed to obtain a good signal[29]. The highest human Hybrid Capture Hi-C resolution achieved until today is at a 5kb window after sequencing of around 1 billion reads[30].

In order to obtain a high resolution Hybrid Capture Hi-C on a selected region without the need of a high depth sequencing, Hybrid Capture Hi-C experiments were developed[30]. Briefly, in these experiments, RNA probes are designed to target the region of interest, and permit the capture via hybridization on the Hybrid Capture Hi-C library of specific interactions that occurs in the region. As mentioned above, these experiments were done before and after the stimulation in the Wild-Type (WT) K562 cells and in the Knock-Out (KO) deletion of the *OAS3* Epromoter in K562 cells. After the enrichment of the Hybrid Capture Hi-C library with interactions that fall in the 3 Mb region surrounding the OAS gene cluster (chr12:113340000-113450000, *hg19*), a high throughput sequencing of around 60 million reads per sample is sufficient to achieve a 5kb Hybrid Capture Hi-C resolution exclusively in the region. In the

absence of capture, up to 1 billion total reads were needed to achieve the same resolution on this region in the published Hi-C data[30]. The Hybrid Capture Hi-C data were previously mapped in the lab using Juicer bioinformatic tools. Hybrid Capture Hi-C biases such as the variability of the number of restriction fragments inside bins were previously normalized using the classical proportional iterative fitting (IPF) that aims to equally normalize the signal between bins and thus, the sum of the total interaction scores of every bin is equal to 1.

Generation of triangular Hybrid Capture Hi-C heatmaps

The final output of the Hybrid Capture Hi-C experiment is a square symmetric matrix that is also called contact map. For each chromosome, a Hybrid Capture Hi-C matrix contains identical rows and columns that represent bins of chromosome coordinates. The matrix is then filled with the interactions score between each bins. The diagonal of the matrix contains high signals due to the short distance between fragments. Far away from the diagonal the scores are generally low, and are high specifically at the DNA interacted bins. Due to its symmetry, the Hi-C matrix is often visualised as a triangular matrix heatmap. (Fig.3).

To visualise these 4 Hi-C interaction matrices heatmaps, I used the ggplot2 R package. Basically, this heatmap is a graphic representation of the contact map where each interaction score is assigned to a proportional red colour intensity of the cell where a red colour is assigned to each cell and whose colour intensity is proportional to the corresponding interaction score. The R script uses the classical square matrix as an input, but since the matrix is symmetric the output is a triangular heatmap in order to facilitate the interpretation of the figure. The base of this triangle corresponds to intervals of 5kb bins of the selected genomic region (chr12:111875000-114915000, *hg19*) and the interactions of one single bin are represented in the diagonal starting from the base of the triangle heatmap (Fig.3). The interaction score between two specific bins is represented by the intersection of their diagonals starting from the base of the triangle. The heatmap corresponding to the 4 analysed samples are presented in Fig.4. Additionally, using the WashU epigenome browser, I added to each heatmap the Refseq gene annotation that corresponds to the genomic coordinates of the studied genome region. In line with the literature, at the 3 Mb region, the Hi-C contact maps in the 4 matrices show a similar pattern: Local domains, i.e. domains where intra-domain interactions are favored, are observed as small triangles all along the studied region. These small triangles correspond to TADs which are often delimited by CTCF proteins in the human genome. Thus, I added a published ChIP-seq on CTCF protein in K562 cells to the heatmap representation. In general, CTCF binds specifically at the boundary of the observed domains. To this aim, the representation of the CTCF proteins on Hi-C triangular heatmaps is important as a proof of the good quality of the generated Hi-C data.

Loop calling on Hybrid Capture Hi-C matrices

The graphical representation of the Hybrid Capture Hi-C matrices as triangular heatmaps are essential in order to have a visual validation of the quality of our data. Here, my goal is firstly to understand how the *OAS3* Epromoter regulate *OAS1* and *OAS2* genes at the 3D organization level of the chromatin, secondly to identify if the *OAS3* Epromoter interact with *OAS1* and *OAS2* promoters upon the IFNα stimulation, and thirdly to figure out if this interaction is interrupted after the knock-out of the *OAS3* Epromoter.

For that, I started by analyzing the Hybrid Capture Hi-C contact matrix using a statistical test to detect highly significant DNA loops from the matrix. In fact, the Hybrid Capture Hi-C experiment generates a contact matrix containing all the interaction scores between regions. The more two DNA regions are found in close 3D space proximity in the cell population, the more chimeric molecules are obtained and the higher the score is. Two regions can be in close 3D space when they are in a functional DNA loop, like the classical enhancer-promoter loops. However, regions of the neighboring DNA sequences can also be found randomly in close 3D space proximity. The latter leads to the generation at a high level of random ligation events. This explains the high score of interactions observed at few kilobases around the diagonal in the heatmaps. Thus, a high interaction score between two neighboring regions is very difficult to analyse. In contrast, a low interaction score with a region located afar can be more informative. To overcome this problem, statistical-based bioinformatic tools have been developed in order to distinguish highly significant interactions within Hi-C contact matrices taking into account Hi-C biases. This analysis is thus called ‘’loop calling’’ and can be performed by the FitHiC tool.

FitHiC is an R package that computes statistical confidence estimates for Hi-C matrices allowing the identification of statistically significant intra-chromosomal Hi-C contacts. FitHiC relies on the comparison of the ‘’observed’’ interaction score from the Hi-C matrix to the ‘’expected’’ interaction score according to the random polymer looping model. In the random polymer looping model, the Hi-C matrix scores are highly enriched in the diagonal and they become homogeneously low far away from the diagonal (Annexe, Fig.2). Thus, using a specific algorithm, FitHiC recomputes Hi-C matrix interaction scores by taking into account the genomic distance between the two considered bins.

In this context, I have run the FitHiC tool on the 4 Hi-C contact matrices. I have obtained respectively 3640, 320, 3565 and 585 significant interactions of *OAS3* Epromoter (q-value > 0.05) in the WTNS, WTS, KONS and KOS Hi-C matrices. None of the detected loops were established by the *OAS3* Epromoter, neither was the case of *OAS2* or *OAS1* promoters. These results suggested that there is no specific DNA loop engaging *OAS3* Epromoter towards the two regulated genes *OAS2* and *OAS1*. However, one of the Hybrid Capture Hi-C limitations, and thereby the FitHiC tool, is the detection of functional DNA loops that are happening between neighboring regions. This is due to the high signal of the background around the diagonal. The *OAS3* Epromoter is indeed located about 30 kb away from the *OAS1* gene, and 40 kb away from the *OAS2* gene. Taking into account that our Hi-C matrix is 5kb-bin

resolution, this means that we are looking at interactions at 6 and 8 bins away from the *OAS3* Epromoter, which is too close to the diagonal of the matrices. This might explain why the FitHiC could not detect any interaction between those close promoters.

Changes in interaction within the OAS cluster

Since the loop calling was not informative concerning the presence of interactions between the *OAS3* Epromoter and the *OAS1* and *OAS2* gene promoters in the clusters due to the small distance, I then thought that if the *OAS3* Epromoter is interacting with the two other promoters, this will lead to an increase of the global interactions within the cluster. In fact, if the Epromoter was in a close proximity to both *OAS1* and *OAS2* promoters, all the regions within the cluster that contain the 3 genes would be in close proximity, resulting in the formation of a small domain of interactions. To test this hypothesis, I have calculated the fold change (FC) of the interaction score obtained by dividing scores of the matrices after the IFNα stimulation over scores before the stimulation in the WT condition and the KO of the *OAS3* Epromoter condition. The log2 of the fold change (Log2FC) values were plotted as a triangular heatmap (Fig.5). Surprisingly, I had realized that upon stimulation, in the WT condition, a global increase of interactions was observed specifically in terms of short range interactions (Fig.5.A). Very closeby regions interactions increased in all the 3 Mb interactions, and this is shown by the increase of the red colour in very small triangles around the diagonal, i.e. the base of the triangle. However, no such increase was observed in the case of the KO of the *OAS3* Epromoter (Fig.5.B). This argues for a possible compaction of the chromatin upon the stress condition leading to a better definition of the TAD domains.

After the generation of the Log2 FC heatmaps, I wanted to quantify if there is a specific increase of interactions between the OAS cluster upon the stimulation of IFNα. For that, I developed an R script in order to select the interaction scores that fall only within the OAS cluster (chr12:113340000-113450000, *hg19*). I have applied this script to the Log2FC matrices and I have represented those values within a violin plot in the WT and KO of the *OAS3* Epromoter conditions (Fig.6). In order to know if the increase of interaction is specific to the OAS cluster, I have divided the 3 Mb region in 100kb cluster windows excluding the OAS cluster, and have applied the script to these clusters and represented the values within a violin plot (Fig.6). I have realised that in the OAS cluster, there is a small but non-significant increase of interactions of the WT condition comparing the *OAS3* Epromoters KO condition (Fig.6.A., paired-samples Wilcoxon test, p=9.2e-2). These results are the same in the 100kb cluster where a small but non-significant increase of the interactions in the WT condition is shown compared with the *OAS3* Epromoter KO condition (Fig.6.B, paired-samples Wilcoxon test, p=8.4e+3). Thus, by comparing these results with those shown by the Log2FC interaction scores heatmaps, one can deduce that the global increase of interactions inside the OAS cluster observed in the Log2FC heatmap is thus general in the WT condition and thus is not specific to the OAS cluster.

Next, I decided to compare if the increase of the interactions inside the OAS cluster, even if it is global, is higher than the one observed in the 100 kb cluster. By comparing the log2FC interaction scores between the OAS cluster and the 100 kb cluster (Fig.6.C,D), I have found a significant difference in both WT and Epromoter KO conditions (both unpaired-samples Wilcoxon test, respectively p=0.59 and p=0.58).

CONCLUSION

To analyse the 3D organisation of the chr12 studied region (chr12:111875000-114915000, *hg19*) and to study the interaction of the *OAS3* Epromoter with its target *OAS1* and *OAS2* promoters, I started by visualising the generated Hybrid Capture Hi-C interaction matrices of the 4 tested conditions using a triangular heatmap. This representation has shown the existence of TADs, delimited by CTCF proteins, inside the studied regions and so suggesting the existence of interactions between different fragments including those inside the OAS gene cluster. However, these heatmaps have provided little information about the interactions of a specific sequence, especially the *OAS3* Epromoter interactions with the *OAS1* and *OAS2* gene promoters upon the IFNα stimulation.

Thus, I continued my analysis by performing a loop calling in order to identify an eventual interaction between the studied Epromoter and its two target promoters upon the IFNα stimulation. This step did not identify the suspected DNA loops. Indeed, no specific DNA loop was observed between *OAS3* Epromoter and the *OAS1* and *OAS2* in the 4 tested conditions.

As both of the previous steps did not provide interesting information about the existence of an interaction between the *OAS3* Epromoter and the *OAS1* and *OAS2* gene promoters inside the OAS cluster, I decided next to study this interaction at the OAS cluster level. In this context, I wondered if the IFNα stimulation of the *OAS3* Epromoter increases the interactions inside the OAS gene cluster, which enclose both *OAS1* and *OAS2* genes. To this aim, I observed the two fold change heatmaps representing respectively the WT and the KO conditions. Unexpectedly, I noticed that in the WT condition, the IFNα stimulation had led to a global increase of short range interactions inside the 3Mb studied region while in the KO condition, no such increase was observed. These results suggest that the IFNα stress stimulation increases the compaction of the chromatin which explains the well distinguishable small TADs in the corresponding Log2 FC WT heatmap.

To verify this suggestion, I continued then by searching for an eventual increase of the chromatin interactions that are happening inside the OAS cluster before and after IFNα stress stimulation. In parallel, I divided the 3Mb studied region to 100kb cluster windows excluding the OAS cluster.
In order to compare the Log2FC interaction scores of both clusters -the OAS cluster and the 100kb cluster-, I built a violin plot for each cluster comparing the Log2FC interaction scores in the WT and the KO condition (Fig.6.A.B). The resulting plot has shown a small but non-significant increase of the Log2

FC interaction scores between the WT and the KO condition in the OAS cluster and the 100kb cluster (paired-samples Wilcoxon test, respectively p=9.2e-2 and p=8.4e-3).
In order to compare the Log2 FC interaction scores between the OAS cluster and the 100kb cluster, I represented the corresponding scores in two violin plots, one for each condition: WT and Epromoter KO (Fig.6.C,D). The resulting plot has shown a significant difference of the Log2 FC interaction scores between the OAS cluster and the 100kb cluster in both WT and Epromoter KO conditions (unpaired-samples Wilcoxon test, respectively p=0.59 and p=0.58).

DISCUSSION

While I was studying the *OAS3* Epromoter and its interactions with *OAS1* and *OAS2* gene promoters, I realized that what makes the expected results so difficult to get is the very short distance separating the studied Epromoter with its target promoters. Infact, as mentioned above, the more the considered bins are close to each other, the more difficult it is to correctly quantify their interaction score. Knowing that the distance separating the *OAS3* Epromoter and the *OAS1* and the *OAS2* gene promoters is respectively 30 and 40 kb and knowing that our Hybrid Capture Hi-C matrix is 5kb-bin resolution, it is predictable to face serious challenges to identify the studied interactions. This means that the *OAS3* Epromoter is distant from the *OAS1* and the *OAS2* Epromoters respectively by only 6 and 8 bins. This may explain the little information provided by the 4 heatmaps (Fig.4) , by the loop calling experiment and also by other analysation tools that I used during the internship as the non-discussed Virtual Hi-C plot (Annexe, Fig.3.)

In the context of these generated Hybrid Capture Hi-C interaction heatmaps (Fig.4.) and by comparing the heatmaps of the WTNS and the WTS interaction matrices, one can immediately notice the difference between the TADs where they are more distinguishable in the WTS than in its counterpart (Fig.4.A,B). It is also the case when comparing respectively the KONS and the KOS interaction matrices (Fig.4.C,D.). Thus, one can consequently suggest that these differences represent an increase of the interactions in the IFNα stimulated conditions (WTS and KONS) compared to the non-stimulated conditions (respectively WTNS and KONS). Thus, it is important to precise that these differences in resolution are due to the differences in the biological experiment. Therefore, these differences should not be taken into account in the analysis of the obtained heatmaps.

FIGURES

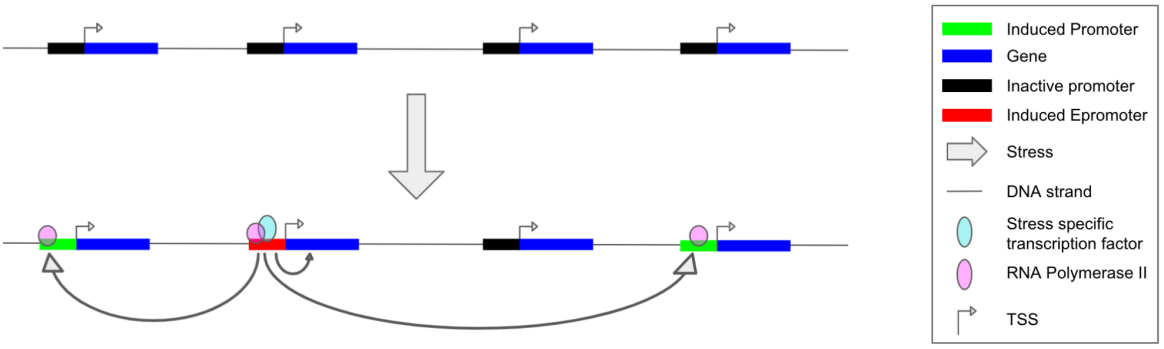


Fig.1. Model of activation of Epromoter cluster gene upon IFNα-stress response: Upon stimulation with IFNα, Epromoter (red) recruits exclusively the stress specific transcription factors (mainly IRF and STAT, in cyan) within a gene cluster. They act by activating the transcription of downstream gene (promoter activity) and neighboring clustered distal genes located in less than 100kb distance (enhancer activity).

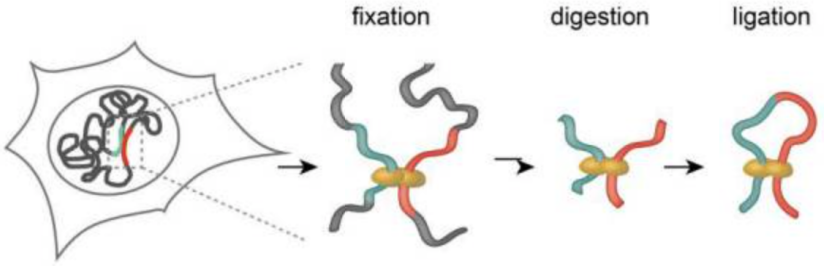


Fig.2. Chromatin Conformation Capture (3C) experiment. The 3C experiments are three-based steps. Firstly DNA interactions mediated by proteins are fixed by crosslinking. Secondly, the chromatin is digested using a restriction enzyme leading to the generation of overhang fragments. Thirdly, close proximal regions are ligated with a ligase enzyme. This leads to the formation of a chimeric DNA molecule between fragments that were forming a DNA loop. (Figure reproduced with modifications, from Dekker et al., 2002.)

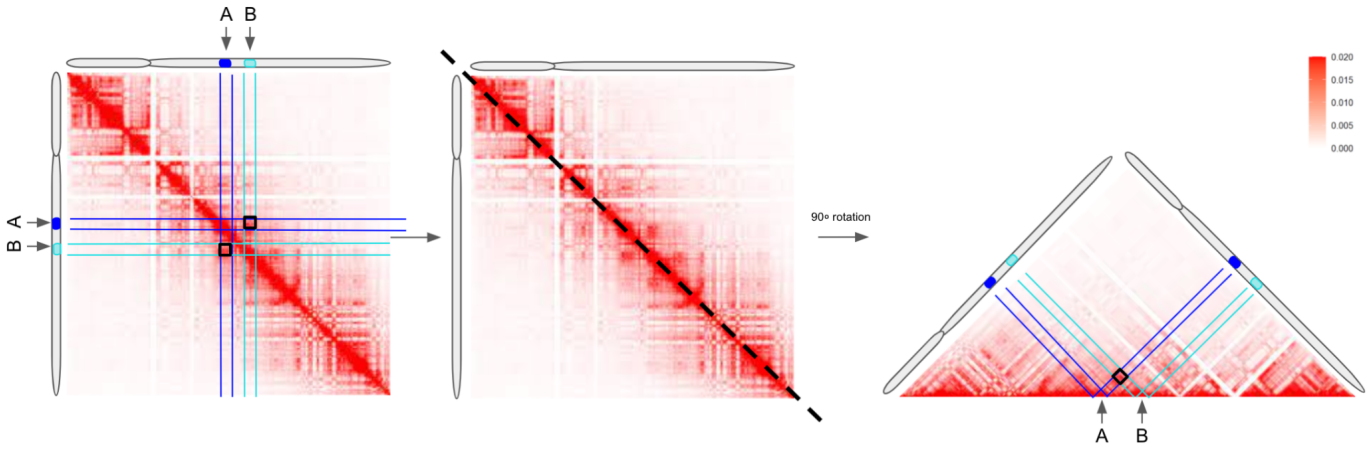
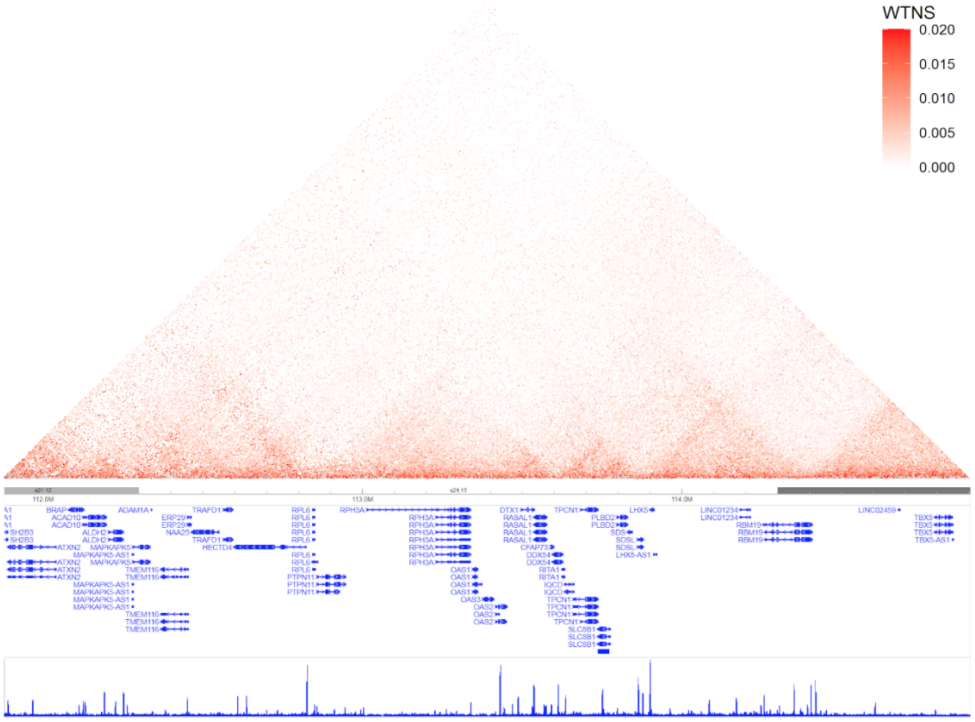
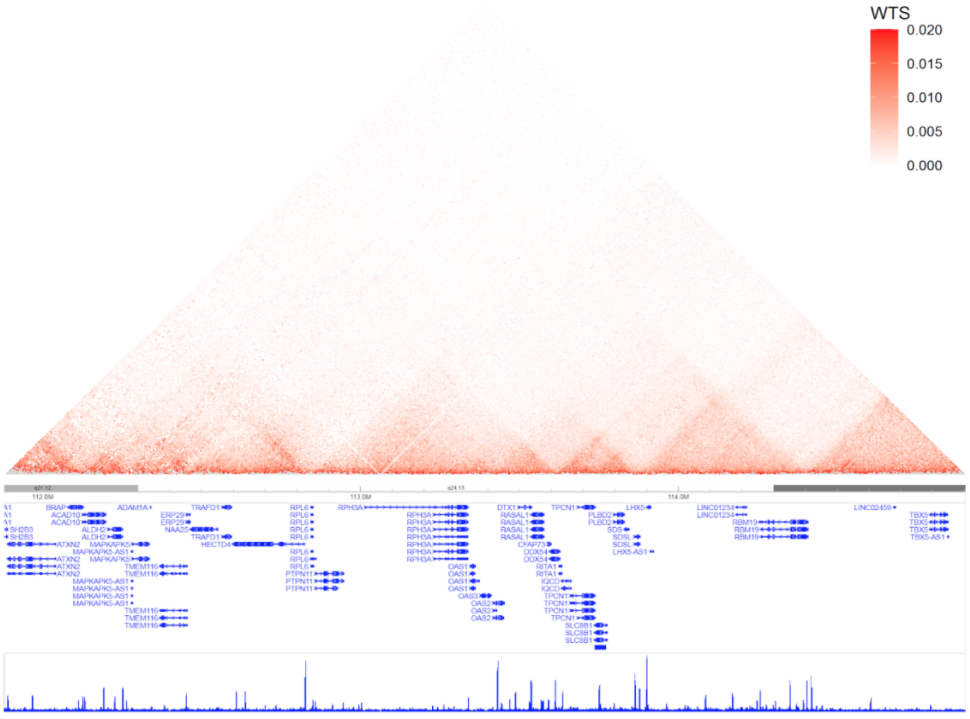


Fig.3. Rectangular and triangular heatmap representation of Hybrid Capture Hi-C interaction matrix. The output of Hi-C experiments is an interaction matrix summarizing all different interacted DNA fragments. Each interaction is assigned with a score that represents the frequency of this interaction within the cell population. A heatmap graphic representation can be done as a square symmetric matrix where rows, as columns, represent the similar range of bins of chromosome coordinates. Cells of the heatmap are assigned to a red intensity color that is proportional to the score of interaction. The generated heatmap matrix is symmetric (the interaction between A and B region is the same as the interaction between B and A, black rectangle in the square heatmap). Thus, it is more convenient to limit observations to one of the half parts of the matrix. The resulting triangular heatmap is then rotated by 90° leading to a conventional representation of Hi-C results where the chromosome coordinates regions occupies the base of the triangle heatmap. An interaction score between A and B region can be then visualized by following the intersection of the diagonal starting from each region (black rectangle in the triangular heatmap)

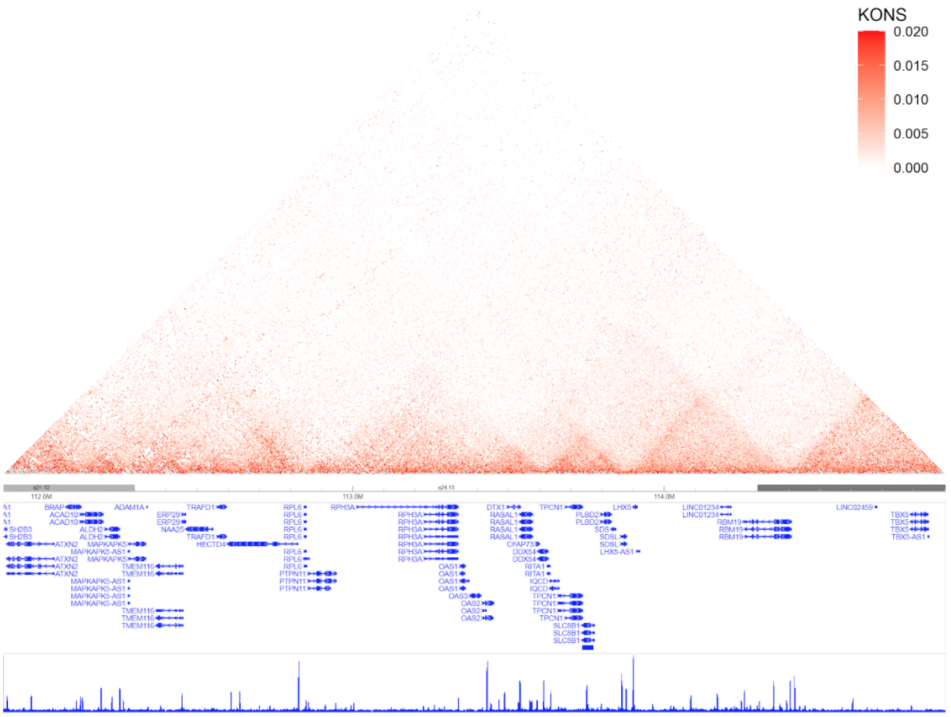
A



B



C



D

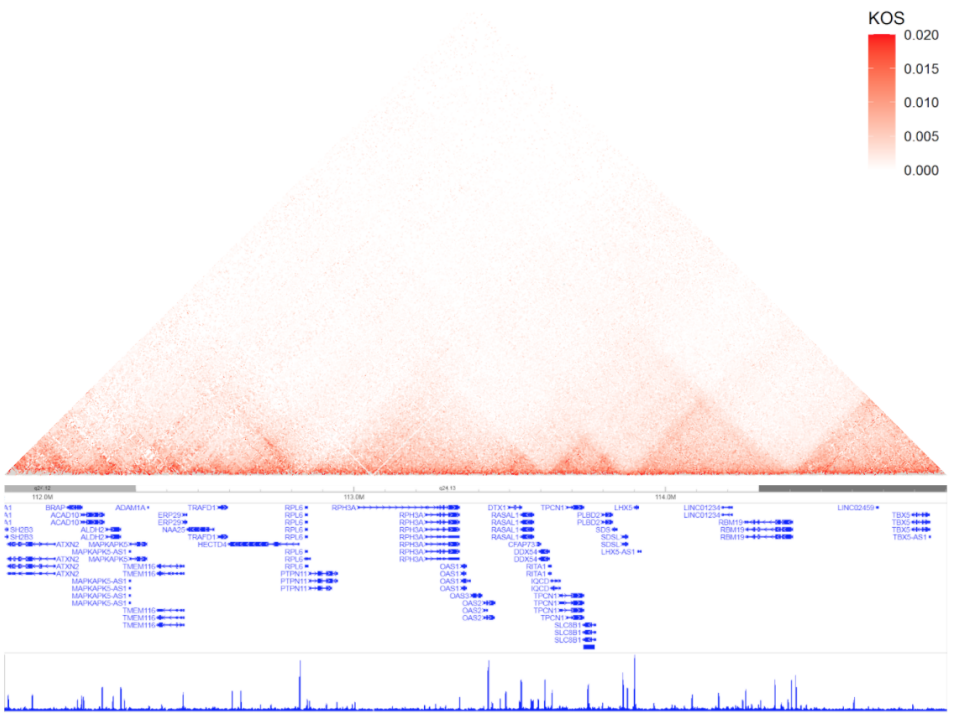
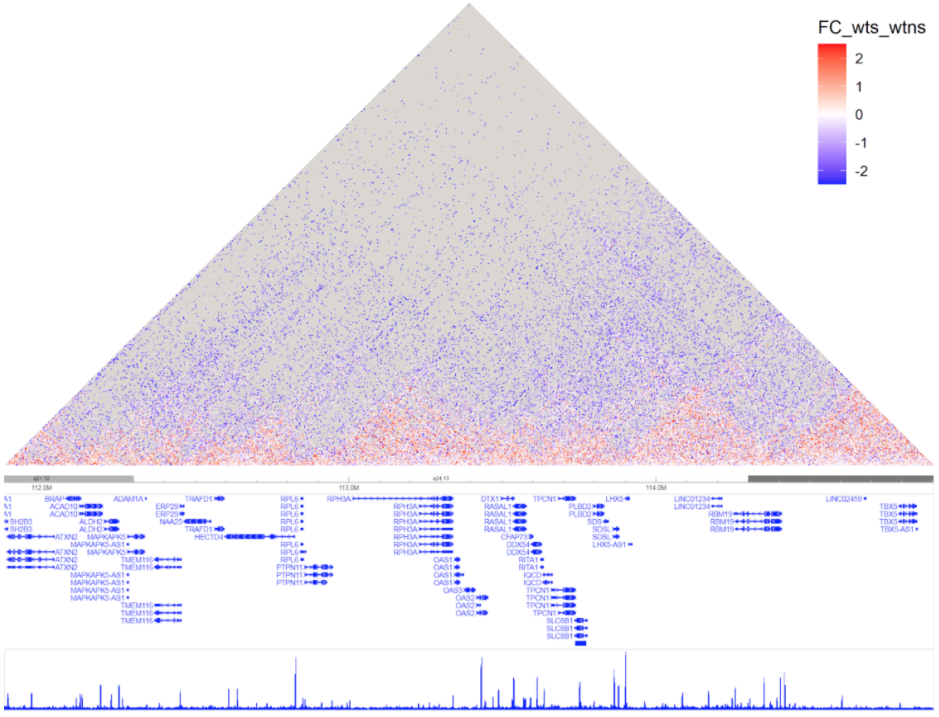


Fig.4. Heatmaps of the Hybrid Capture Hi-C interactions matrices obtained before or after IFNα stress stimulation in the 4 tested conditions. From the top to the bottom on each graph is represented the triangular heatmap, the genomic coordinates, the Refseq gene annotation and the CTCF localisation -in frequencies- obtained from already published data. **A** Heatmap of the Hi-C interaction matrix of the Wild Type Non-Stimulated (WTNS) condition. **B** Heatmap of the Hi-C interaction matrix of the Wild Type Stimulated (WTS) condition. **C** Heatmap of the Hi-C interaction matrix of the Knock Out Non-Stimulated (KONS) condition. **D** Heatmap of the Hi-C interaction matrix of the Knock Out Stimulated (KONS) condition.

A



B

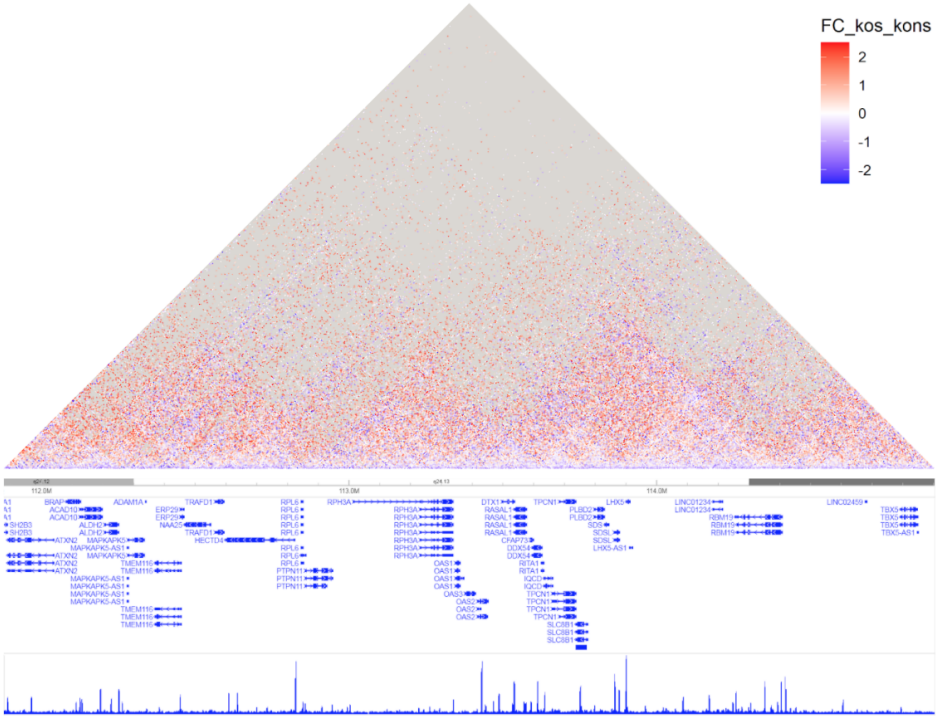


Fig.5. Heatmaps of the Log2 FC of the Hybrid Capture Hi-C interaction scores upon the IFNα stress stimulation in the Wild Type and the *OAS3* Epromoter Knock Out condition. A Heatmap of the Wild Type Log2 FC of the interaction scores of the stimulated condition (WTS) over the non stimulated condition (WTNS), i.e. log2 of the ratio WTS/WTNS interaction scores. **B** Heatmap of the *OAS3* Epromoter Knock Out Log2 FC of the interaction scores of the stimulated condition (KOS) over the non stimulated condition (KONS), i.e. log2 of the ratio KOS/KONS interaction scores.

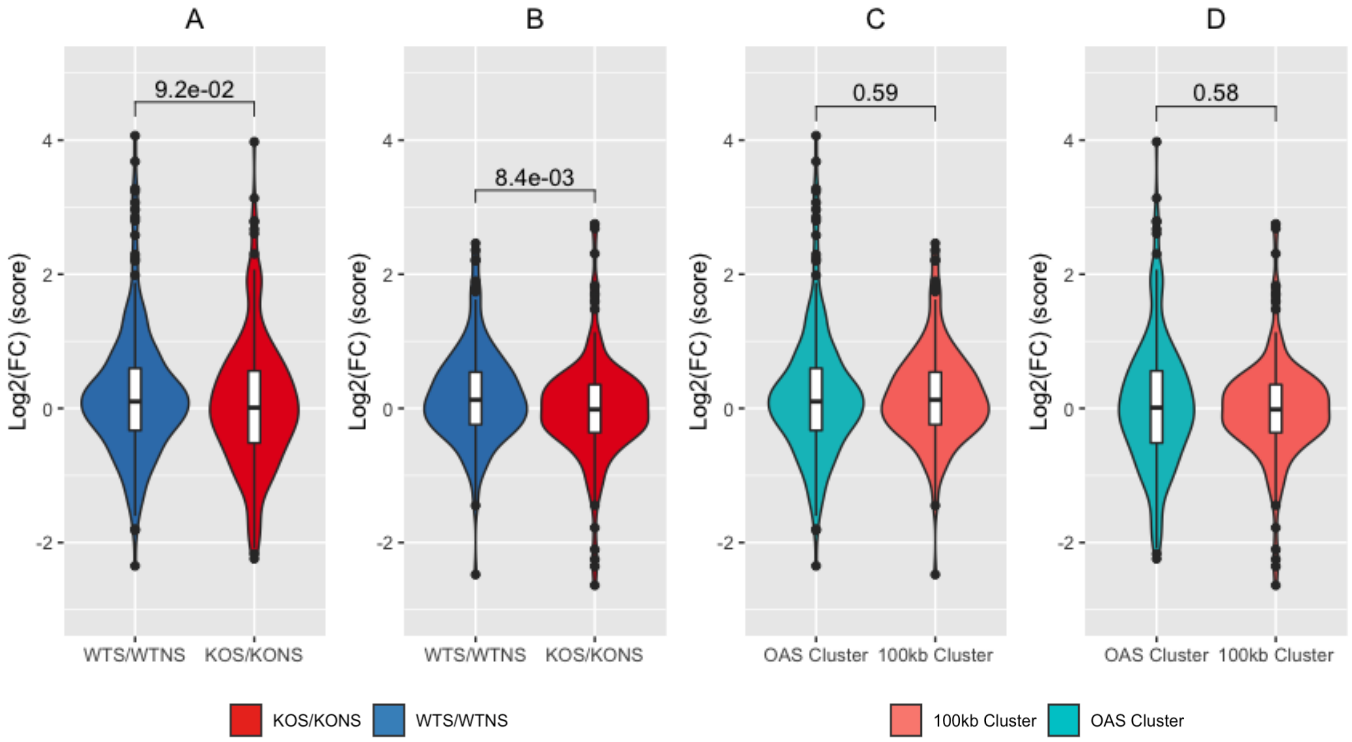


Fig.6. Violin plots of the Log2FC of the Hybrid Capture Hi-C interaction scores before and after IFNα stress stimulation inside the OAS gene cluster and the 100kb cluster. The plotted scores are the log2 of the fold change (Log2FC) of the Hi-C interaction scores before the stimulation (NS) over those after the stimulation (S) in the Wild Type (WT) or the Epromoter Knock Out (KO) condition. **A** Violin plot of the Log2FC of the Hi-C interaction scores of the WT (blue) and the KO (red) conditions in the OAS cluster, showing a non-significant difference between the compared data (paired-samples Wilcoxon test, p=9.2e-2). **B** Violin plot of the Log2FC of the Hi-C interaction scores of the WT (blue) and the KO (red) conditions in the 100kb cluster, showing a non-significant difference between the compared data (paired-samples Wilcoxon test, p=8.4e-3) **C** Violin plot of the Log2FC of the Hi-C interaction scores of the WT condition of the OAS cluster (cyan) and the 100kb cluster (pink)m showing a significant difference between the two compared data (unpaired-samples Wilcoxon test, p=0.59). **D** Violin plot of the Log2FC of the Hi-C interaction scores of the Epromoter KO condition of the OAS cluster (cyan) and the 100kb cluster (pink) showing a significant difference between the two compared data (unpaired-samples Wilcoxon test, p=0.58).

ANNEXE

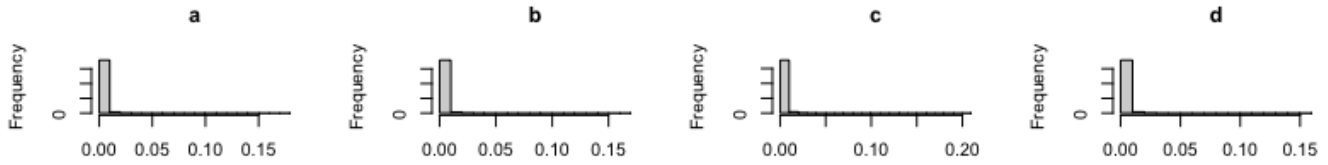


Fig.1. Histograms of Hybrid Capture Hi-C interaction scores after the log2 transformation in 4 different conditions. These histograms show that our Hi-C datasets are not normally distributed and thus comparing them to each other statistically cannot be done by using the T test. Thus, the test used for this aim is the Wilcoxon test. **a** Wild Type Non-Stimulated condition. **b** Wild Type Stimulated Condition. **c** Knock Out Non-Simulated condition. **d** Knock Out Stimulated condition.

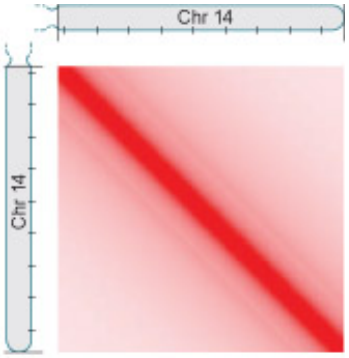


Fig.2. Heatmap of the expected interaction matrix according to the random polymer looping model. In the random polymer looping model, the expected interaction matrix corresponds to what would be observed in case of no long-range intra-chromatin interactions. (Figure reproduced with modifications, from van Berkum, Nynke L et al., 2010.)

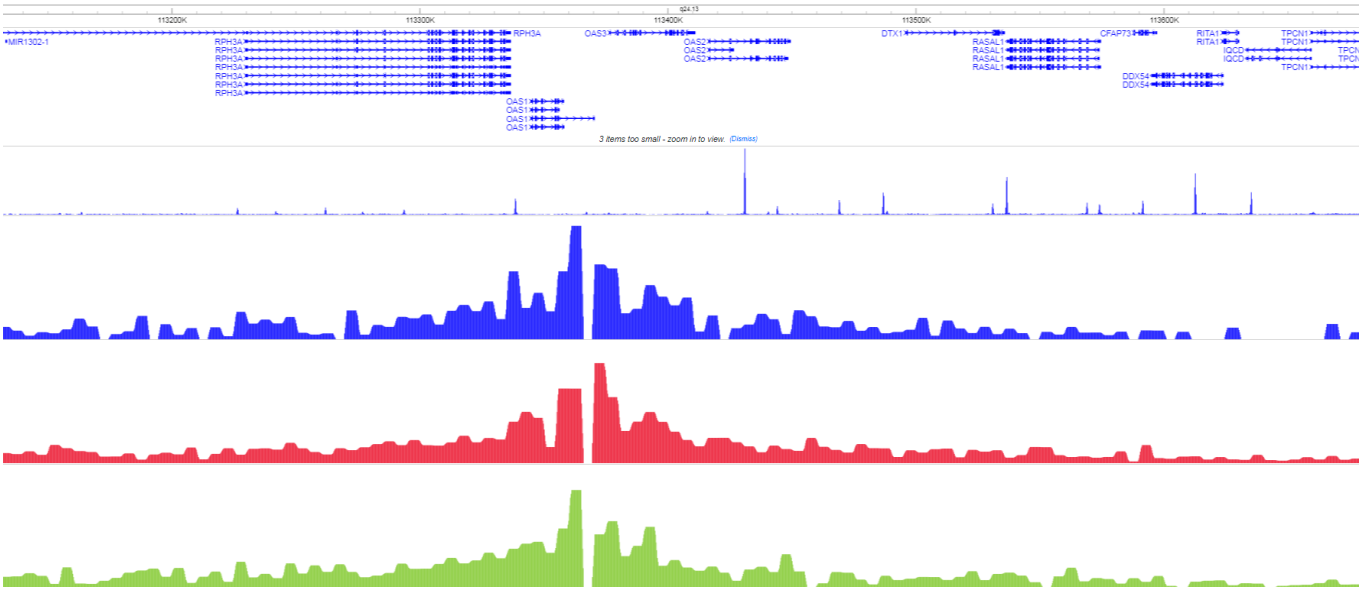


Fig.3. Virtual Hybrid Capture Hi-C showing the interactions of one chosen bin of the *OAS3* Epromoter (chr12:113365000-113370000) with the rest of the bins of the studied 3Mb region. Aligned with the Refseq gene annotation, the 3 presented Virtual Hi-C plots colored in blue, red and green represent respectively the Wild Type Non Stimulated, the Wild Type Stimulated and the Knock Out Stimulated interaction scores of the selected *OAS3* Epromoter fragment with the rest of the 3Mb region fragments.The three Virtual Hi-C plots do not show significant difference between each other.

REFRENCES

1. Liu B, Chen J, Shen B. Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs. *BMC Syst Biol*. 2011;5 Suppl 1(Suppl 1):S2. Published 2011 May 4. doi:10.1186/1752-0509-5-S1-S2	11. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011;473:43–49. doi: 10.1038/nature09906	21. Hoskins RA, Landolin JM, Brown JB, et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res*. 2011;21(2):182-192. doi:10.1101/gr.112466.110
2. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. 2018;19(10):621-637. doi:10.1038/s41580-018-0028-8	12. Zehnder, T., Benner, P. & Vingron, M. Predicting enhancers in mammalian genomes using supervised hidden Markov models. *BMC Bioinformatics* 20, 157 (2019). https://doi.org/10.1186/s12859-019-2708-6	22. Stadhouders R, van den Heuvel A, Kolovos P, et al. Transcription regulation by distal enhancers: who's in the loop?. *Transcription*. 2012;3(4):181-186. doi:10.4161/trns.20720
3. Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev*. 2012;22(2):79-85. doi:10.1016/j.gde.2011.11.001	13. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. 2002;295(5558):1306-1311. doi:10.1126/science.1067799.	23. Tam KT, Chan PK, Zhang W, et al. Identification of a novel distal regulatory element of the human Neuroglobin gene by the chromosome conformation capture approach. *Nucleic Acids Res*. 2017;45(1):115-126. doi:10.1093/nar/gkw820
4. Visel A, Blow MJ, Li Z, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009;457(7231):854-858. doi:10.1038/nature07730	14. Han J, Zhang Z, Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Mol Cytogenet*. 2018;11:21. Published 2018 Mar 9. doi:10.1186/s13039-018-0368-2	24. Sprooten J, Garg AD. Type I interferons and endoplasmic reticulum stress in health and disease. *Int Rev Cell Mol Biol*. 2020;350:63-118. doi:10.1016/bs.ircmb.2019.10.004
5. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155(4):934-947. doi:10.1016/j.cell.2013.09.053	15. Meaburn, K., Misteli, T. Chromosome territories. *Nature* 445, 379–381 (2007). doi.org/10.1038/445379a	25. Lee WB, Choi WY, Lee DH, Shim H, Kim-Ha J, Kim YJ. *OAS1* and *OAS3* negatively regulate the expression of chemokines and interferon-responsive genes in human macrophages. *BMB Rep*. 2019;52(2):133-138. doi:10.5483/BMBRep.2019.52.2.129
6. Jared S. Stees, Fred Varn, Suming Huang, John Strouboulis, Jörg Bungert Biology (Basel) 2012 Dec; 1(3): 778–793. Published online 2012 Dec 5. doi: 10.3390/biology1030778	16. Zhang C, Xu Z, Yang S, et al. tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. *Cell Rep*. 2020;32(13):108206. doi:10.1016/j.celrep.2020.108206	26. van Steensel B, Furlong EEM. The role of transcription in shaping the spatial organization of the genome. *Nat Rev Mol Cell Biol*. 2019;20(6):327-337. doi:10.1038/s41580-019-0114-6
7. Sartorelli, V., Lauberth, S.M. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol 27, 521–528 (2020). doi.org/10.1038/s41594-020-0446-0	17. Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161(5):1012-1025. doi:10.1016/j.cell.2015.04.004	27. Kooren J, Palstra RJ, Klous P, et al. Beta-globin active chromatin Hub formation in differentiating erythroid cells and in p45 NF-E2 knock-out mice. *J Biol Chem*. 2007;282(22):16544-16552. doi:10.1074/jbc.M701159200
8. Miguel-Escalada I, Pasquali L, Ferrer J. Transcriptional enhancers: functional insights and role in human disease. *Curr Opin Genet Dev*. 2015;33:71-76. doi:10.1016/j.gde.2015.08.009	18. Ouboussad L, Kreuz S, Lefevre PF. CTCF depletion alters chromatin structure and transcription of myeloid-specific factors. J Mol Cell Biol. 2013 Oct;5(5):308-22. doi: 10.1093/jmcb/mjt023.	28. McGee M. Case for omitting tied observations in the two-sample t-test and the Wilcoxon-Mann-Whitney Test. *PLoS One*. 2018;13(7):e0200837. Published 2018 Jul 24. doi:10.1371/journal.pone.0200837
9. Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol Cell. 2015 Jun 18;58(6):1101-12. doi: 10.1016/j.molcel.2015.04.006.	19. Chabot A, Shrit RA, Blekhman R, Gilad Y. Using reporter gene assays to identify cis regulatory differences between humans and chimpanzees. *Genetics*. 2007;176(4):2069-2076. doi:10.1534/genetics.107.073429	29. Yan, F., Powell, D.R., Curtis, D.J. *et al.* From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 21, 22 (2020). doi.org/10.1186/s13059-020-1929-3
10. Song L, Zhang Z, Grasfeder LL, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res*. 2011;21(10):1757-1767. doi:10.1101/gr.121541.111	20. Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet*. 2017;49(7):1073-1081. doi:10.1038/ng.3884	30. Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping [published correction appears in Cell. 2015 Jul 30;162(3):687-8]. *Cell*. 2014;159(7):1665-1680. doi:10.1016/j.cell.2014.11.021