

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR



TRABAJO FIN DE MÁSTER

TRANSCRIPTIONAL AND STRUCTURAL OUTCOMES OF GENOME-WIDE CTCF DEPLETION IN B CELLS

Máster Universitario en Bioinformática y
Biología Computacional

Autora: Ana Rodríguez Ronchel
Tutora: Almudena Rodríguez Ramiro
Ponente: Luis del Peso Ovalle

Febrero de 2021

TRANSCRIPTIONAL AND STRUCTURAL OUTCOMES OF GENOME-WIDE CTCF DEPLETION IN B CELLS

Autora: Ana Rodríguez Ronchel
Tutora: Almudena Rodríguez Ramiro
Ponente: Luis del Peso Ovalle

Centro Nacional de Investigaciones Cardiovasculares (CNIC)
Febrero de 2021

Agradecimientos

Quisiera agradecer este trabajo de fin de máster a todas las personas que lo han hecho posible.

En primer lugar, a Almudena, por su dirección científica y su confianza en mi trabajo. Gracias a todo el grupo por acogerme desde el primer momento y, en particular, a Ester, por guiarme en el proceso y por su cercanía.

Por último, gracias a mis compañeros de máster, por su apoyo cuando lo he necesitado, por los buenos momentos en la cafetería de la facultad y por su amistad.

Resumen

Resumen

La proteína estructural CTCF está implicada en el establecimiento de interacciones entre regiones distales del genoma que, además de definir la arquitectura de la cromatina, regulan el programa transcripcional de la célula. En células B, CTCF participa en la recombinación VDJ y en el cambio de isotipo, dos procesos esenciales en la respuesta inmune. Sin embargo, todavía no se conoce en detalle la relación mecanística entre la unión de CTCF al ADN y la regulación transcripcional que se deriva de ella. Para abordar esta cuestión, en este trabajo utilizamos un modelo de ratón en el que CTCF se elimina de forma específica en células B maduras. A través de experimentos de ChIP-Seq, hemos identificado un grupo de sitios de unión de CTCF resistentes a la eliminación de la proteína. Estos sitios resistentes tienen motivos de unión de CTCF más similares a la secuencia consenso y están enriquecidos en las regiones límite de dominios asociados topológicamente (TADs). Por otra parte, nuestros estudios de ARN-Seq indican que los cambios transcripcionales causados por la pérdida de CTCF son moderados. Para relacionar los cambios de expresión con los sitios de unión diferencial de CTCF hemos realizado un estudio de la unión de CTCF en regiones promotoras y la formación de bucles. Para ello, hemos desarrollado un algoritmo capaz de identificar regiones del genoma con alta probabilidad de ser reguladas transcripcionalmente por bucles de CTCF. Creemos que esta aproximación supone un avance en el estudio de la regulación transcripcional mediada por bucles de CTCF, que se reforzará con la verificación experimental de las regiones identificadas en este trabajo.

Palabras Clave

CTCF, célula B, bucles de ADN, TAD, transcriptoma, ARN-Seq, ChIP-Seq

Abstract

CTCF is involved in establishing long-range interactions that define chromatin architecture and regulate transcriptional programs. In B cells, CTCF participates during VDJ recombination and class switch recombination, both critical processes for the immune response. However, to date, the relationship between CTCF-mediated contacts and their transcriptional implications in mature B cells is not completely understood. Here we used a conditional mouse model where CTCF is eliminated specifically in mature B cells and found a subset of CTCF-binding sites that are resistant to protein depletion. These "retained" CTCF sites have a higher proportion of consensus like CTCF motifs and are preferentially localized at topologically associating domains (TADs) boundaries. In addition, we found that CTCF deletion causes few transcriptional changes in mature B cells. To link CTCF differential binding with changes in gene expression we studied CTCF binding to promoter regions and the formation of CTCF-mediated loops. With that aim, we developed an algorithm that identifies regions that can be transcriptionally regulated by CTCF-dependent loops. We consider that this approach represents a step forward in the understanding of transcriptional regulation mediated by CTCF loops, which will be further strengthen with the experimental validation of the regions identified in our study.

Key words

CTCF, B cell, DNA loops, TAD, transcriptome, RNA-Seq, ChIP-Seq

Contents

List of figures	ix
Abbreviations	1
1 Introduction	3
1.1 Genome organization	3
1.2 CTCF and gene expression regulation	4
1.3 CTCF in B cells	5
2 Objectives	7
3 Methods	9
3.1 Mouse model	9
3.2 B cell selection	9
3.3 ChIP-Seq	9
3.4 ChIP-Seq analysis	10
3.5 CTCF motif analysis	10
3.6 TADs boundaries enrichment analysis	10
3.7 RNA-Seq	11
3.8 RNA-Seq analysis	11
3.9 Analysis of CTCF binding in promoter regions	11
3.10 CTCF-mediated loop prediction algorithm	11
3.11 Statistics	12
3.12 Code availability	12
4 Results	13
4.1 Characterization of CTCF deficient mouse model	13
4.2 CTCF binding sites study	13
4.2.1 Higher amount of CTCF consensus motif in retained peaks	16
4.2.2 Retained peaks are enriched in TAD boundaries	16
4.3 Transcriptional effects of CTCF deletion	16

List of figures

1.1	Structural organization of chromatin	4
1.2	Subclasses of CTCF-mediated chromatin contacts involved in transcription	5
1.3	Chromatin extrusion model	6
4.1	CTCF deletion mouse model	14
4.2	CTCF lost and retained binding sites have distinct features	15
4.3	Retained peaks are enriched in CTCF consensus motif	16
4.4	TAD boundaries are more enriched in CTCF retained than lost peaks	17
4.5	CTCF deletion alters gene expression in mature B cells	18
4.6	Steps followed to determine the relationship between proximal peaks and differential gene expression	19
4.7	Steps followed to determine the relationship between distal peaks and differential gene expression	20
4.8	Loop and no-loop regions have distinct features	22
4.9	Predicted loop regions are enriched in genes coordinately expressed	23

Abbreviations

- **CES:** Coordinated Expression Score
- **ChIA-PET:** Chromatin Interaction Analysis by Paired-End Tag
- **ChIP:** Chromatin Immunoprecipitation
- **CSR:** Class Switch Recombination
- **CTCF:** CCCTC-binding factor
- **DB:** Data Base
- **DEG:** Differentially Expressed Gene
- **DER:** Differentially Expressed Region
- **ESC:** Embryonic Stem Cell
- **FISH:** Fluorescence *In Situ* Hybridization
- **GC:** Germinal Center
- **GFP:** Green Fluorescent Protein
- **GO:** Gene Ontology
- **IgH:** Immunoglobulin Heavy chain
- **logFC:** logarithm of the Fold Change
- **PCA:** Principal Component Analysis
- **SHM:** Somatic Hypermutation
- **TAD:** Topological Associating Domain
- **TSS:** Transcription Start Site
- **YY1:** Yin Yang 1

1

Introduction

1.1 Genome organization

Genomes do not only encode for genetic information in their linear sequence, but their three-dimensional architecture is also critical for the cell biology. Genome architectural environment regulates different nuclear processes, such as transcription, DNA replication or cell division, which in turn impact multiple biological functions. These events depend on complex spatial arrangement of the chromatin, which involves folding the DNA into hierarchical and dynamic structures involving different layers of regulatory information [1].

Fluorescence labeling techniques like fluorescence *in situ* hybridization (FISH), have shown that each chromosome occupies a specific territory within the nucleus [2]. Chromosomes are organized into two compartments, called A and B based on chromatin state. Transcriptional active, euchromatin and highly accessible regions are located at the inner nuclear space, conforming the A compartment, whereas inactive, heterochromatin and little accessible regions reside near the nuclear lamina, in the B compartment [3]. At a smaller scale, genomic regions are divided into topological associating domains (TADs). TADs are \sim 1 Mb sized, contiguous chromosomal regions. They are characterized by high interaction frequency between loci located within their boundaries, as compared to interactions with other regions of the genome (Fig. 1.1). TADs also represent a regulatory unit with specific-gene expression profiles [4]. TAD boundaries are suggested to have barrier activity, which would stop heterochromatin spreading from neighboring domains [5]. These boundaries are enriched in proteins such as cohesin and CCCTC-binding factor (CTCF) and are very conserved across different cell types [6].

Zooming in into the genome organization, chromatin forms intra-TAD loops. The length of these loops can vary from a few kilobases to megabases and, in contrast to TADs, their conformation changes among cell types or during development [7]. As is the case for TADs boundaries, loop anchors are predominantly bound by CTCF [8]. Chromatin loops bring together sequences that are located far apart in the genome, which may have a variety of functional implications. They allow, for example, the interaction of enhancer or silencer elements with promoter regions. Loops can also be formed around a group of genes, establishing gene clusters where transcription is coordinated in a more efficient way [9].

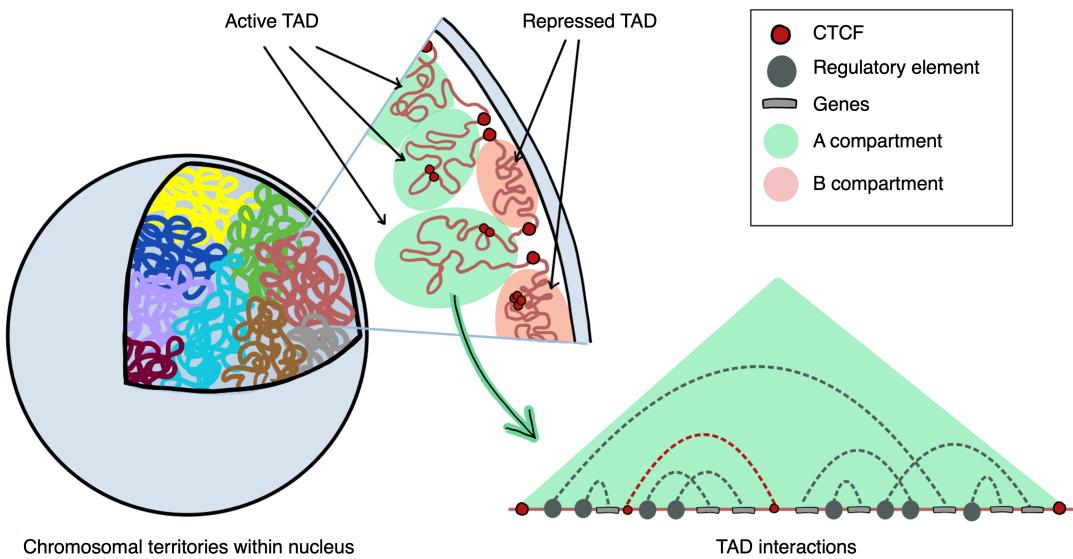


Figure 1.1: Structural organization of chromatin. Chromosomes are found to occupy specific nuclear spaces, called chromosomal territories. Each chromosome is subdivided into TADs that can be transcriptionally active or repressed, depending on whether they are located in the A (green) or B (red) compartment. TADs present preferential intradomain interactions (dotted lines) compared to interdomain interactions and have insulating boundaries enriched in CTCF among other structural proteins. Inside TADs, there are several chromatin loops that allow the interaction between distal regulatory elements and genes (grey dotted lines) helping to modulate gene expression. This intra-TAD loops can be established by CTCF (red dotted lines).

This multi-level genome architecture can be regulated by several components such as architectural proteins, transcription factors and non-coding RNAs in order to coordinate gene expression and cell fate [1]. In this project we are going to focus on CTCF as a regulator of gene expression by analyzing transcriptome changes in a conditional model for CTCF deficiency.

1.2 CTCF and gene expression regulation

CTCF is an essential, highly conserved and ubiquitously expressed protein in higher eukaryotes [10]. CTCF binds DNA through 11 Zn-finger motifs that form a DNA-binding central domain. CTCF can form long-range chromatin interactions and has been considered a multivalent protein responsible for bridging the gap between nuclear organization and gene expression [11]. Although CTCF is enriched at TAD boundaries, these regions are only a small proportion of all CTCF binding sites genome-wide. Indeed, CTCF can modulate gene expression independently of forming TAD boundaries, for instance by promoter binding and the recruitment of cofactors (Fig. 1.2.A) [12]. However, out of more than 50,000 CTCF binding sites identified by genome-wide experiments, only 12% lie near promoters, while 53% lie within intergenic regions and 35% in intragenic regions [13]. This indicates that in most of its binding sites, CTCF does not act as a classical transcription factor. In fact, it is known that CTCF can also act as insulator when positioned between an enhancer and gene promoter, by blocking their communication and preventing transcriptional activation (Fig. 1.2.B) [14]. Furthermore, CTCF-dependent chromatin loops regulate gene expression by various mechanisms that are determined by the nature of the sequences that are brought together. For example, CTCF loops facilitate the interaction between enhancer or silencer regions with the promoter, thus acting as transcriptional activator or repressor, respectively (Fig. 1.2.C and 1.2.D). In addition, there is evidence that CTCF may be

involved in sequestering clusters of co-regulated genes into independently regulated chromatin domains, thanks to its ability to establish boundaries between active and repressive chromatin domains (Fig. 1.2.E) [15, 16]. Therefore, CTCF exact function at a given genomic site is currently difficult to predict [17].

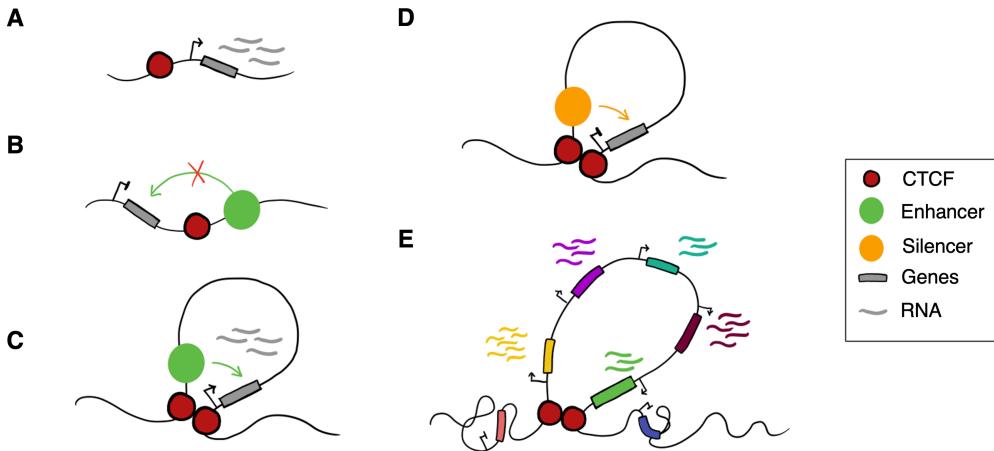


Figure 1.2: Subclasses of CTCF-mediated chromatin contacts involved in transcription. A) CTCF binding to a promoter region, acting as a transcription factor. B) CTCF role as an insulator, preventing the interaction between enhancer and promoter regions. C) CTCF loop that approximates enhancer and promoter regions, acting as a transcriptional activator. D) CTCF loop that approximates silencer and promoter regions, acting as a transcriptional repressor. E) CTCF-mediated interactions forming an independently regulated chromatin domain containing a co-regulated gene-dense cluster.

Not surprisingly, CTCF is involved in multiple cellular processes. CTCF deficiency is embryonic lethal, underlying a critical role in development [18]. CTCF is also essential in the development of several cell types, ranging from stem cells to neural or cardiac cells. Recent studies show CTCF as a developmentally regulated protein, suggesting that it plays a role in cell type-specific genome organization and expression via chromatin looping [19].

CTCF-mediated contacts that result in the formation of TADs and chromatin loops tend to be in convergent orientations [8]. Based on this feature, the chromatin extrusion model has been proposed to explain the formation of CTCF chromatin loops (Fig. 1.3). The extrusion complex, formed by two DNA-binding subunits (cohesin rings), is loaded onto the DNA and starts forming the loop by allowing the DNA thread to slide through them. Both cohesin subunits move in opposite directions (forward and reverse) causing the extrusion and the looping of DNA until the complex encounters CTCF-bound sites arranged in a convergent orientation, thus giving rise to TADs and chromatin loops [19, 20].

1.3 CTCF in B cells

B cells are the core of the adaptive humoral immune system by the production of antibodies, molecules that can specifically bind and inactivate pathogens. Each B cell expresses one type of antibody molecule with a unique specificity, such that all B cells combined can produce a huge repertoire of antibodies which can recognize virtually any antigen. This diversity is achieved at two points during the development of B cells that involve important chromatin conformation and transcriptional changes. First, during bone marrow differentiation, B cells rearrange their immunoglobulin genes in a process called V(D)J recombination, which assembles V, D and J gene segments through a site-specific recombination reaction; such combinatorial diversity al-

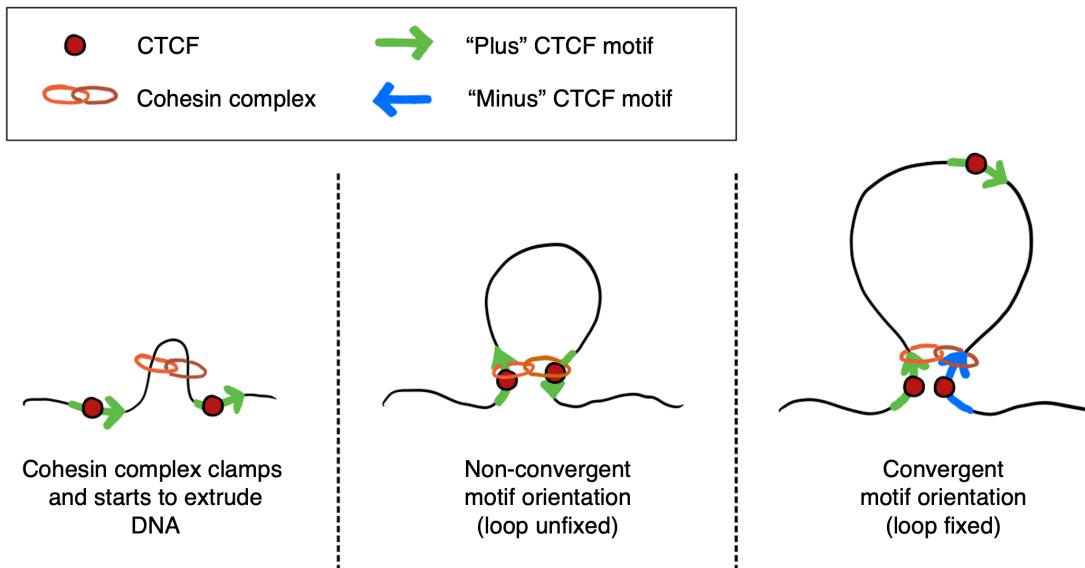


Figure 1.3: Chromatin extrusion model. Schematic representation of the loop extrusion dynamics. The extrusion complex, conformed by two cohesin rings, binds the DNA and starts forming the loop. It only stops when encounters with CTCF molecules in opposite directions, meaning that the 5' motif must be on the forward (plus) strand and the 3' motif on the reverse (minus) strand.

lows the generation of a broad repertoire of immunoglobulin genes [21]. V(D)J recombination requires the immunoglobulin heavy chain gene (IgH) to undergo conformational changes such as contraction and looping in order to ensure a proper region assembly. CTCF-mediated looping is considered a main regulator of IgH locus contraction and V(D)J recombination [22]. After V(D)J recombination, B cells exit the bone marrow and become mature B cells in the spleen. Upon antigen encounter, mature B cells are activated, proliferate, and engage in the germinal center (GC) reaction, where they can further diversify their antibody repertoire by somatic hypermutation (SHM) and class switch recombination (CSR). GCs allow the generation of memory B cells and high affinity plasma cells, which are critical for the immune response and underlie the mechanism of action of vaccines [23].

The host lab has previously shown that CTCF is a key regulator for the GC response. Thus, CTCF is involved in CSR by regulating long-range DNA loops in the IgH locus and limiting chromatin accessibility prior to CSR [24]. In addition, CTCF transcriptionally regulates the proliferation rate of GC B cells and represses the expression of Blimp-1, thus preventing premature terminal differentiation to plasma cell [25]. However, how CTCF regulates gene expression in mature B cells has not been addressed so far.

2

Objectives

CTCF is a DNA-binding protein that can regulate gene expression by forming long-range chromatin loops. The establishment of interactions between distal areas of the genome is essential in a myriad of cellular and developmental events. In B cells, CTCF loops are essential both in immature B cells during VDJ recombination and in activated mature B cells during class switch recombination. However, to date, the relationship between CTCF-mediated contacts and their transcriptional implications in mature B cells has not been studied in detail. Thus, in this TFM we aimed at exploring the role of CTCF in mature B cells with the following specific objectives:

- To analyze the features of CTCF binding sites in mature B cells.
- To relate changes in gene expression with CTCF binding patterns, both in proximal and distal regions.

3

Methods

3.1 Mouse model

Conditional CTCF-deficient mice ($\text{CTCF}^{\text{fl}/\text{fl}}\text{CD19-Cre}^{\text{ki}/+}$) were obtained by breeding $\text{CTCF}^{\text{fl}/\text{fl}}$ mice [26] with $\text{CD19-Cre}^{\text{ki}/+}$ mice [27]. Mice were housed in pathogen-free conditions, under a 12 h dark/light cycle with food *ad libitum*. All animal procedures were conducted in accordance with EU Directive 2010/63/UE, enforced in Spanish law under Real Decreto 53/2013. The procedures have been reviewed by the Institutional Animal Care and Use Committee (IACUC) of Centro Nacional de Investigaciones Cardiovasculares and approved by Consejería de Medio Ambiente, Administración Local y Ordenación del Territorio of Comunidad de Madrid (Ref: PROEX 341/14) [28].

3.2 B cell selection

Naive B cells were isolated from spleen. Spleens were meshed through 70 μm pore nylon cell strainers (BD Falcon) in complete RPMI medium (supplemented with 10% FBS and penicillin (50 U/ml) and streptomycin (50 $\mu\text{g}/\text{ml}$)). Erythrocytes were lysed using erythrocyte lysis buffer (ACK Lysing Buffer, BioWhittaker) for 4 minutes at room temperature. After washing with cold complete RPMI, B cells were isolated by immunomagnetic depletion using anti-CD43 beads (Miltenyi Biotec).

3.3 ChIP-Seq

ChIP was performed according to the Diagenode protocol (iDeal ChIP-seq Kit for Transcription Factors C01010055). In brief, 5 million cells were crosslinked in 1% formaldehyde (Sigma) for 10 min at 37°C and quenched with 0.125 M cold glycine. Cell pellets were lysed in 1 mL RIPA buffer (10 mM Tris-HCl, 1 mM EDTA, 0.1% sodium deoxycholate, 0.1% SDS, 1% Triton X-100, pH 8.0) at 4°C during 20 min and centrifuged at 2,300 x g for 5 min at 4°C. Nuclei were suspended in 500 μL of 0.5% SDS lysis buffer (0.5% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.0) and sonicated using Covaris system (shearing time 15 min, 10% duty cycle,

200 cycles per burst and 175 W PIP). 1% of the sheared chromatin was set apart (input) and the rest of the sonicated chromatin was incubated overnight at 4°C with anti-CTCF antibody (Diagenode). Immunoprecipitated chromatin was eluted and decrosslinked for 8 hours. DNA was purified and quantified using Invitrogen Qubit Fluorometer. Finally, 3-4 ng of DNA were used to prepare libraries using NEBNext Ultra II DNA Library Prep Kit for Illumina. ChIP-seq and input control libraries from three biological replicates per genotype were sequenced on a HiSeq2500 (Illumina).

3.4 ChIP-Seq analysis

ChIP-seq analysis was performed using a custom pipeline based on a course developed by the Harvard Chan Bioinformatics Core (HBC) [29]. First, quality of the sequencing was checked using **FastQC** (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Next, reads were mapped to reference genome (mm10, GRCm38, December 2011) using **Bowtie2** with local alignment mode to perform soft-clipping for the removal of poor quality bases or adapters [30]. The resulting SAM files were converted to BAM format with **samtools v.1.8** [31]. Then, reads were sorted and filtered to remove duplicates, multimappers and unmapped reads with **sambamba v.0.6.8** [32]. Peaks were identified using **MASCS2** [33] and their quality was checked with **ChIPQC** [34]. Next, **DiffBind** [35] was used to identify differential binding sites and a consensus peakset was generated including peaks found in a minimum of 2 of the replicates (minOverlap=0.66). Bed files with lost and retained peaks were generated in this step. Peak size distributions were analyzed with custom R code [36]. For density track visualization, Bam files were indexed with **samtools** [31] and then, BigWig files were generated with **bamCompare** from the suite of python tools **deeptools** [37]. BigWig files were used to obtain a global evaluation of enrichment around the lost and retained peaks regions for each replicate using **computeMatrix** and **plotProfile** functions from **deeptools** [37]. Both bigwig (signal) and bed (peak calls) files were visualised using the Integrative Genomics Viewer (IGV)[38]. Additionally, peaks were annotated with **ChIPseeker** [39] using the genome annotation version vM23 from GENCODE, allowing to relate peaks and genome features. More details can be found in the code used to perform this analysis (3.12).

3.5 CTCF motif analysis

HOMER [40] motif discovery software was used to assess motif presence at each binding site. We used the **findMotifs.pl** function to find out what percentage of binding sites contained CTCF motifs. In addition, we used the **-find** option with the HOMER CTCF-motif matrix to extract the matching score for the best motif instance at each binding site.

3.6 TADs boundaries enrichment analysis

Overlap between peaks and TADs boundaries was calculated and plotted with custom bash and R code [36]. mm10 TAD coordinates from ESCs were obtained from The 3D Genome Browser [41, 42]. The genomic region surrounding the TAD boundaries in a window of ± 500 kb was scanned in 10 bp segments. We calculated the total genome area occupied by peaks of each group and then, which percentage of that total area overlapped at each of the segments. Thus, we generated a plot showing the peak enrichment at each position relative to the TAD boundary.

3.7 RNA-Seq

RNA was extracted using the Qiagen RNeasy kit and was treated with DNase. Five-hundred nanograms of total RNA were used to generate libraries using the TruSeq RNA sample preparation kit v2 (Illumina). Briefly, poly-A RNA was purified using poly-T oligo-attached magnetic beads using two rounds of purification followed by fragmentation and first and second cDNA strand synthesis. Then, cDNA 3' ends were adenylated and the adapters were ligated followed by PCR library amplification. Finally, the size of the libraries was checked using the Agilent 2100 Bioanalyzer DNA 1000 chip and their concentration was determined using the Qubi® fluorometer (Life Technologies). Libraries were sequenced on a HiSeq2500 (Illumina) to generate 60 bases single reads [28].

3.8 RNA-Seq analysis

Processing of RNA-Seq data was performed as described in a Galaxy training course [43, 44]. Briefly, **FastQC** (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to assess read quality, and **Cutadapt** [45] to eliminate Illumina adaptor remains, trimming the reads and discarding reads shorter than 20 bp or with poor quality (Phred score < 20). Mapping was performed using **STAR** [46] and the Ensembl mouse genome annotation (GRCm38 v.M22). **MultiQC** [47] was used to assess that more than 80% of the reads were mapped. Next, mapped reads per gene were quantified using the tool **featureCounts** [48] (multi-mapping reads and reads with a mapping quality < 10 were excluded). For comparison between conditions, the R package **DESeq2** [49] with default parameters was used to normalize total read count per sample and to identify the differentially expressed genes (DEGs) (adjusted p-value < 0.05). The DEGs were annotated with **Annotate DESeq2 output tables** Galaxy tool and their Z-scores were computed and plotted with **heatmap2** tool. Functional enrichment analysis of the DEGs was done with **goseq** [50], that allows performing Gene Ontology (GO) analysis, and with Ingenuity Pathways Analysis (IPA) software [51] to study the affected pathways. All tools used in this analysis except IPA were executed from galaxy platform [52]

3.9 Analysis of CTCF binding in promoter regions

The list of DEGs with a lost peak in their promoter region was obtained using a custom bash code. First, the annotated lost peak file obtained from the ChIP-Seq analysis was filtered to select only those peaks that lie in the promoter region of a gene. Then, these genes were compared with DEG list and the common genes were selected and quantified. More details can be found in the code used to perform this analysis (3.12).

3.10 CTCF-mediated loop prediction algorithm

Lost and retained peaks were annotated with CTCF-motif orientation using the **HOMER** program **findMotifsGenome** [40]. When different oriented motifs were identified in the same peak, both orientations were annotated. Then, using a custom python code (<https://www.python.org/>), pairs of peaks were selected to form the loop group according to the following conditions: 1) one of the peaks should be in the lost group, 2) CTCF motif has to be in forward orientation for the peak located at 5' and in reverse orientation for the peak located at 3' and 3) distance between both peaks must be lower than 1 kb. The generation of the no-loop group was done

using the `bedtools` [53] `shuffle` function. A custom python script was generated to associate each of these regions with the genes that lie within them and to calculate the Coordinated Expression Score (CES) using the annotated DESeq2 output file. Then, differentially expressed regions (DERs) were selected and genes within them were compared with DEG list to quantify the number of common genes using custom code. Boxplot and histogram representations were done with ggplot2 [54].

3.11 Statistics

Wilcoxon test was applied to asses whether the mean of two data populations differed. Two-sample Kolmogorov-Smirnov test was performed to detect significant differences between two distributions. P-values were corrected for multiple hypothesis testing by Benjamini-Hochberg method where appropriate [55]. All statistical analyses were performed using R [36].

3.12 Code availability

The custom code used to analyse and represent the data is available in GitHub: <https://github.com/AnaRonchel/TFM>

4

Results

4.1 Characterization of CTCF deficient mouse model

To address the role of CTCF in mature B cells, we used a conditional mouse model where CTCF was deleted during B cell maturation: $\text{CTCF}^{\text{fl}/\text{fl}}\text{CD19-Cre}^{\text{ki}/+}$ (CTCF-deficient mice, hereafter CTCF $^{\text{fl}/\text{fl}}$ group) (Fig. 4.1.A). CD19 expression starts in the bone marrow and progressively increases during B cell differentiation, thus allowing complete Cre-mediated deletion of CTCF floxed alleles in mature naive B cells (Fig. 4.1.B) [56]. $\text{CTCF}^{\text{fl}/+}\text{CD19-Cre}^{\text{ki}/+}$ were used as control (control mice, hereafter CTCF $^{\text{fl}/+}$), since it has been previously shown that CTCF is haplosufficient in mature B cells [25]. RT-qPCR analysis of CTCF mRNA levels showed that the CTCF $^{\text{fl}/\text{fl}}$ naive B cells did not completely lack CTCF, but showed reduced expression compared to the control group (Fig. 4.1.C).

Since our mouse model shows a progressive CTCF deletion starting soon in B cell development, experiments were carried out to assess whether CTCF deletion affected B cell differentiation. Flow cytometry analysis showed that there were no differences in neither the proportions of bone marrow B cell subsets nor the number of total B cells between CTCF $^{\text{fl}/\text{fl}}$ and CTCF $^{\text{fl}/+}$ groups (data not shown) indicating that CTCF deletion using a CD19-Cre strain does not cause major defects in B cell development.

4.2 CTCF binding sites study

CTCF binds thousands of sites throughout the whole genome. However, how CTCF depletion in B cells affects the occupancy of its binding sites still remains unknown. To explore the function of CTCF in mature B cells, we carried out ChIP-Seq experiments in spleen mature B cells from CTCF $^{\text{fl}/\text{fl}}$ and CTCF $^{\text{fl}/+}$ mice (Fig. 4.1.D). We processed the ChIP-Seq samples to identify peaks and evaluated their quality with the CHIPQC program. We performed a Principal Component Analysis (PCA) to study sample similarity and found that replicates mostly lie close to each other (Fig. 4.2.A). It is observed that principal component 1 (PC1), which explains 95% of the variance, separates the replicates based on the group to which they belong. This indicates that the genotype of the mouse from which the cells come is responsible for most of the variance observed between samples. We did differential binding site analysis and obtained one set of peaks

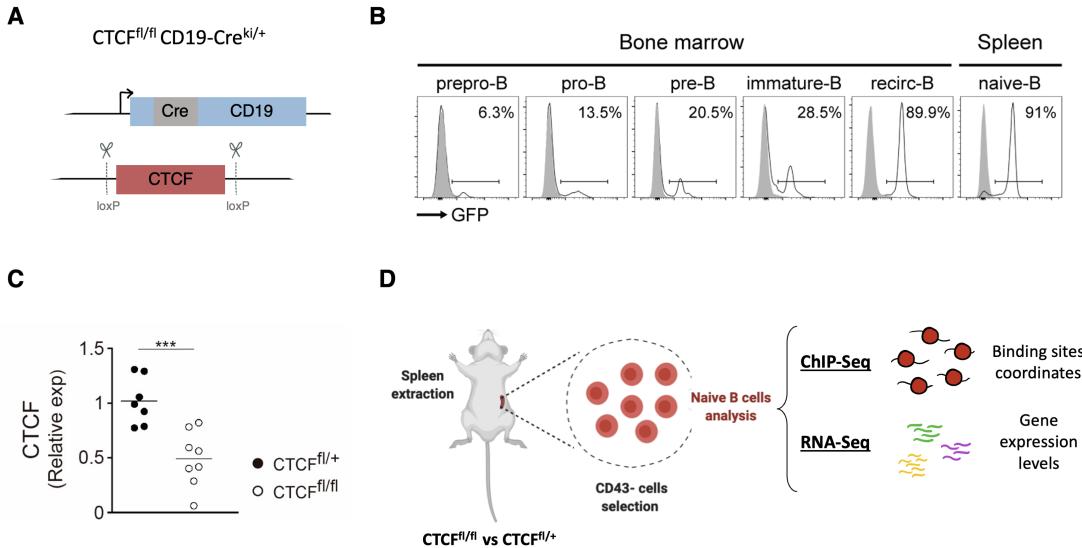


Figure 4.1: CTCF deletion mouse model. A) Representation of the constructs used for conditional depletion of CTCF in mature B cells. CD19^{+/Cre} mice were crossed to mice carrying the CTCF allele flanked by LoxP sites. B) Progressive CD19 expression during B cell development using a R26-GFP reporter mice where the Cre excision is showed by GFP expression. Representative FACS analysis of GFP in bone marrow and spleen B cells from R26^{+/GFP} CD19^{+/Cre} (black empty line) and R26^{+/+} CD19^{+/Cre} (grey shade) (extracted from [56]). C) RT-qPCR analysis of CTCF expression in naive B cells from CTCF^{fl/fl} and CTCF^{fl/+} mice. *** symbol indicates a p-value = 0.0009. Statistical analysis was done with the two-tailed unpaired Student's t-test. Each dot represents an individual mouse (modified from [28]). D) Experiment design scheme highlighting the steps followed prior to performing the ChIP-Seq and RNA-Seq experiments.

for CTCF^{fl/fl} cells and one set of peaks for CTCF^{fl/+} cells (Fig. 4.2.B). We found that most of the peaks present in the control group were not present in the CTCF^{fl/fl} group (hereafter, lost peaks), but a set of peaks was conserved between both conditions (hereafter, retained peaks). There was also a very small group of peaks present only in the CTCF^{fl/fl} cells, but further analysis suggested that it was noise signal (not shown).

We next explored the distinctive features between lost and retained CTCF-binding sites that could account for those binding sites resistant to CTCF depletion in CTCF^{fl/fl} cells. We observed that retained peaks had on average 2-fold higher signal than lost peaks (Fig. 4.2.C). This indicates that a higher proportion of cells had CTCF molecules bound in the regions of retained peaks, suggesting that CTCF has more affinity for these regions. It should be noted that the peak signal is also higher in the control group than in the CTCF^{fl/fl} cells even inside retained peaks. This is consistent with the lower amount of CTCF in the CTCF^{fl/fl} cells (Fig. 4.1.D).

We studied the size of the regions of lost and retained peaks and found that retained peaks were significantly wider than lost peaks (Fig. 4.2.D). This finding probably reflects that retained peaks are often regions in which several tandem peaks overlap. To test this hypothesis, we examined the profile of retained and lost peaks. As expected, we observed that a big proportion of the retained regions actually contained several peaks in tandem (Fig. 4.2.E).

To analyze peak distribution across the genome, we studied the genomic annotation associated with lost and retained peaks. We found that retained peaks were more enriched in distal intergenic regions (34.1% vs 29.8%) and less in promoter regions (20.1% vs 24.9%) compared to lost peaks (Fig. 4.2.F). In addition, we studied the distribution of the peaks relative to their distance to the transcriptional start sites (TSS) of their nearest gene (Fig. 4.2.G). In agreement

with the previous result, we observed that retained peaks were located predominantly in distal gene regions as compared to lost peaks. This finding suggests that retained peaks are less involved in gene regulation by promoter interaction and more involved in other types of distal regulation, such as loops.

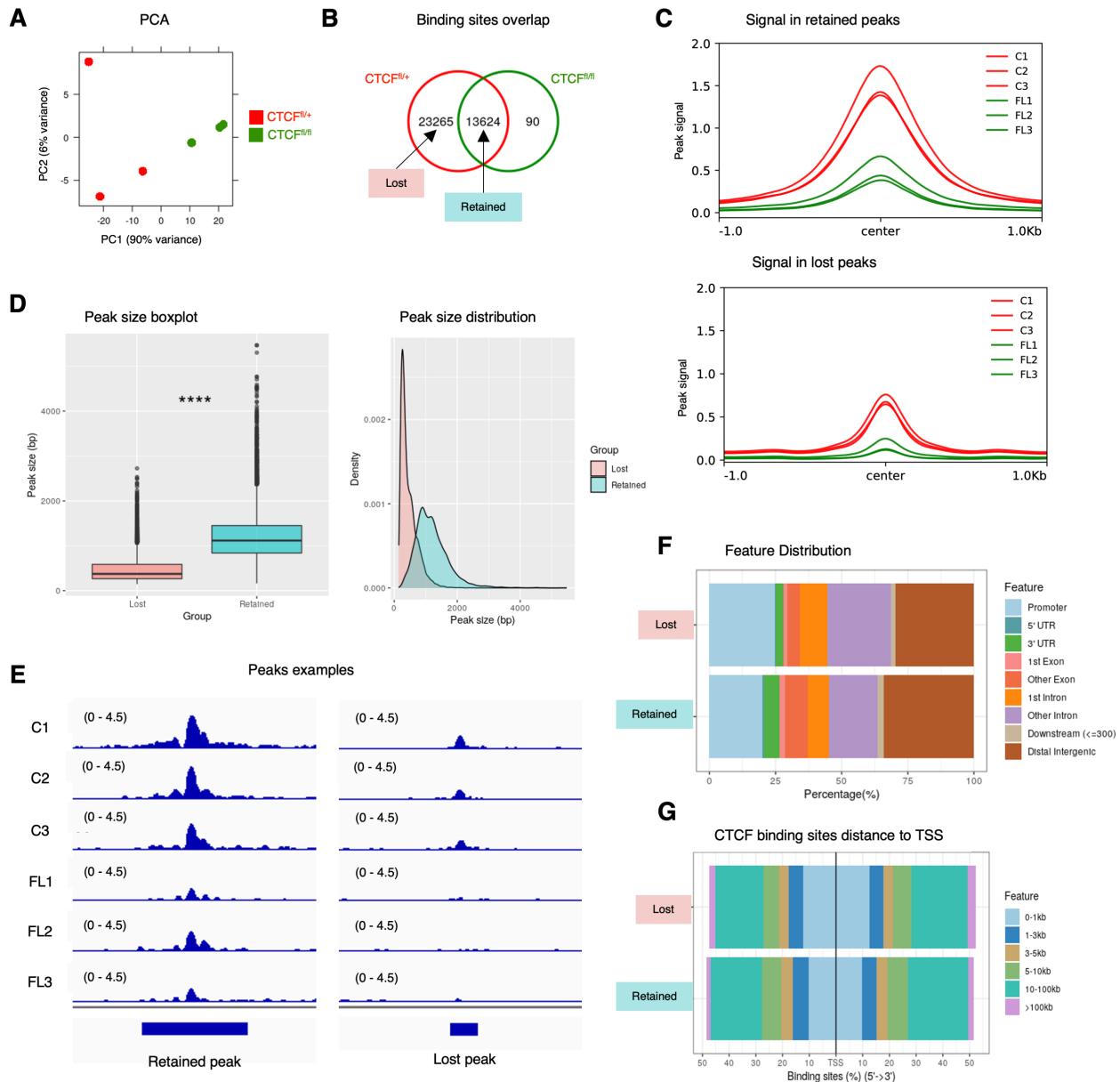


Figure 4.2: CTCF lost and retained binding sites have distinct features. A) PCA of CTCF^{fl/+} and CTCF^{fl/fl} replicates generated with ChIPQC program. B) Venn diagram showing the number of retained, lost and gained binding sites between CTCF^{fl/+} and CTCF^{fl/fl}. C) Plot showing the average peak signal at CTCF retained (top) and lost (bottom) peaks from each control (C, CTCF^{fl/+}) and CTCF^{fl/fl} (FL) replicates. D) Peak size distribution of CTCF retained and lost groups shown as a boxplot (left) or a density plot (right). Lost peak size mean = 461 bp, Retained peak size mean = 1,194 bp. **** symbol indicates a p-value < 2.2e-16. Statistical analysis was done with the Wilcoxon test. E) CTCF ChIP-Seq tracks of a representative retained (left) or lost (right) peak. CTCF signal (normalized read counts) for each control (C, CTCF^{fl/+}) and CTCF^{fl/fl} (FL) replicate and peak coordinates (blue bottom boxes) were visualised using IGV software [38]. Retained peak: chr4 154,796,953-154,799,255 (2,302 bp length), Lost peak: chr4 156,025,873-156,026,457 (584 bp length). F) Distribution of CTCF lost (top) and retained (bottom) peaks across genomic regions. G) Distribution of CTCF lost (top) and retained (bottom) peaks relative to gene TSS.

4.2.1 Higher amount of CTCF consensus motif in retained peaks

CTCF binding site analysis showed that retained peaks have a higher peak signal, suggesting that CTCF has more affinity for those regions. Genome-wide studies have revealed a degenerate CTCF consensus DNA binding motif composed of a ~20-bp core (Fig. 4.3.A) [57]. To approach whether variations in the nucleotide sequence affect the binding affinity of CTCF, we studied the CTCF binding motifs present in lost and retained peaks. We observed that the percentage of peaks containing a CTCF motif was higher in the retained than in the lost group (Fig. 4.3.B). Furthermore, the average score of these motifs was also higher in the retained peaks (Fig. 4.3.C). This suggests that retained peaks are more enriched in the CTCF consensus motif, which would indicates that the affinity of CTCF for those sequences is higher.

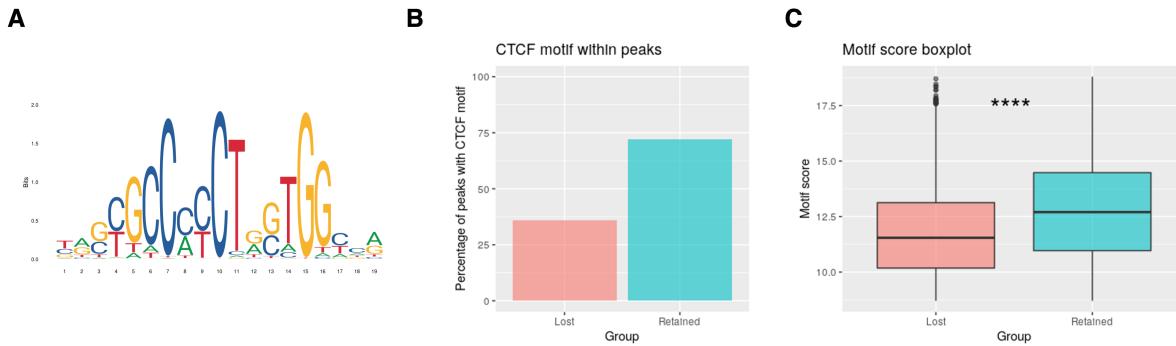


Figure 4.3: Retained peaks are enriched in CTCF consensus motif. A) CTCF consensus motif logo (from JASPAR database, matrix profile MA0139.1). B) Percentage of peaks that have in their sequence a CTCF motif with a score higher than 8.7 C) Boxplot showing motif scores within lost and retained peaks that have a score higher than 8.7. **** symbol indicates a p-value < 2.2e-16. Statistical analysis was done with the Wilcoxon test.

4.2.2 Retained peaks are enriched in TAD boundaries

Since CTCF is involved in the maintenance of the global chromatin architecture by binding to TAD boundaries, persistence of some peaks could be due to their structural involvement in DNA organization. To explore this hypothesis, we analyzed the overlap between TAD boundaries and CTCF retained and lost peaks. Given that TADs are highly conserved across different cell types, we used TAD boundaries data from a previous study in embryonic stem cells (ESCs) [42]. We analyzed the peak overlap within a window of ± 500 kb (Fig. 4.4). We found that both lost and retained peaks showed an enrichment in TAD boundaries; however we observed a significantly higher proportion of retained peaks at TAD boundaries. These results suggest that the retained CTCF sites may be more involved in higher order chromatin architecture. Therefore, TAD formation would be less sensitive to CTCF depletion than other CTCF-dependent mechanisms, which would be mediated to a greater extent by binding sites in the lost group.

4.3 Transcriptional effects of CTCF deletion

To explore the role of CTCF in gene regulation in mature B cells, we performed a transcriptomic analysis of spleen mature B cells isolated from CTCF^{fl/fl} and CTCF^{fl/+} mice (Fig. 4.1.D). Differential expression analysis showed significant changes in 138 genes (adjusted p-value <0.05), identified as differentially expressed genes (DEGs). Fifty-two of those genes were downregulated and 86 upregulated in CTCF^{fl/fl} cells (Fig. 4.5.A). This result suggests that CTCF acts both as negative and positive regulator of gene expression in B cells.

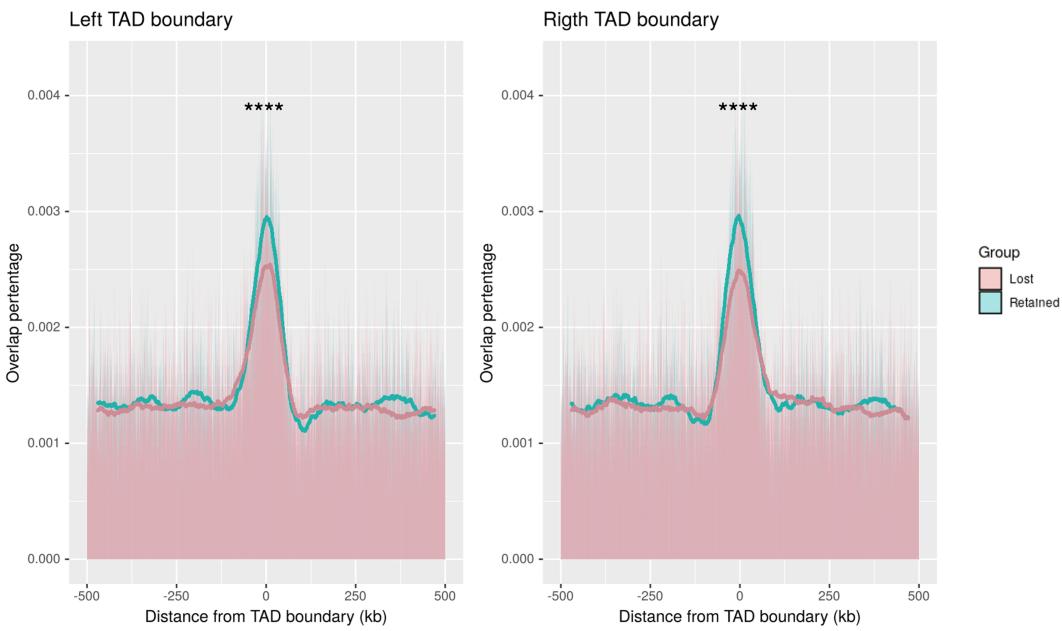


Figure 4.4: TAD boundaries are more enriched in CTCF retained than lost peaks. Plots showing CTCF binding sites distribution around TAD boundaries. The plot shows the percentage of peak overlap in the ± 500 kb region around the TADs boundaries. **** symbol indicates a p-value $< 2.2\text{e-}16$. Statistical analysis was done with two-sample Kolmogorov-Smirnov test.

To determine the functional consequences of the gene expression changes after CTCF deletion we made a Gene Ontology (GO) analysis using the GOseq package (Fig. 4.5.B) [50]. Several gene categories related with the extracellular matrix status and cell mobility were significantly enriched in CTCF deficient B cells. Ingenuity Pathway Analysis (IPA) [51] identified pathways related with secretion and cell mobility functions, as well as inflammatory response and cellular proliferation process. Thus, we conclude that CTCF deficiency promotes transcriptional changes that affect various biological processes of mature B cells.

4.4 Integrative analysis of gene expression and CTCF binding site data

A priori, prediction of the functional role of a particular CTCF binding site or the genes whose expression is regulated by CTCF occupancy is virtually impossible. The strategy most commonly used to extrapolate which genes are controlled by a transcription factor has been to analyze whether they bind near the promoter region of a gene and then test whether the expression of those genes is altered when those binding sites are lost [58]. However, CTCF not only binds to promoter regions but can mediate long-range DNA interactions that either active or repress transcription. In this project, we have sought to address how the changes in CTCF occupancy relate to transcriptional deregulation considering both CTCF capabilities: promoter binding and loop formation.

4.4.1 Proximal peak-mediated regulation: CTCF as a transcription factor

To analyze the DEGs regulated by CTCF proximal binding sites, we assessed CTCF peaks in DEG promoter regions, defined as a 1-kb window around the gene TSS. We focused on lost

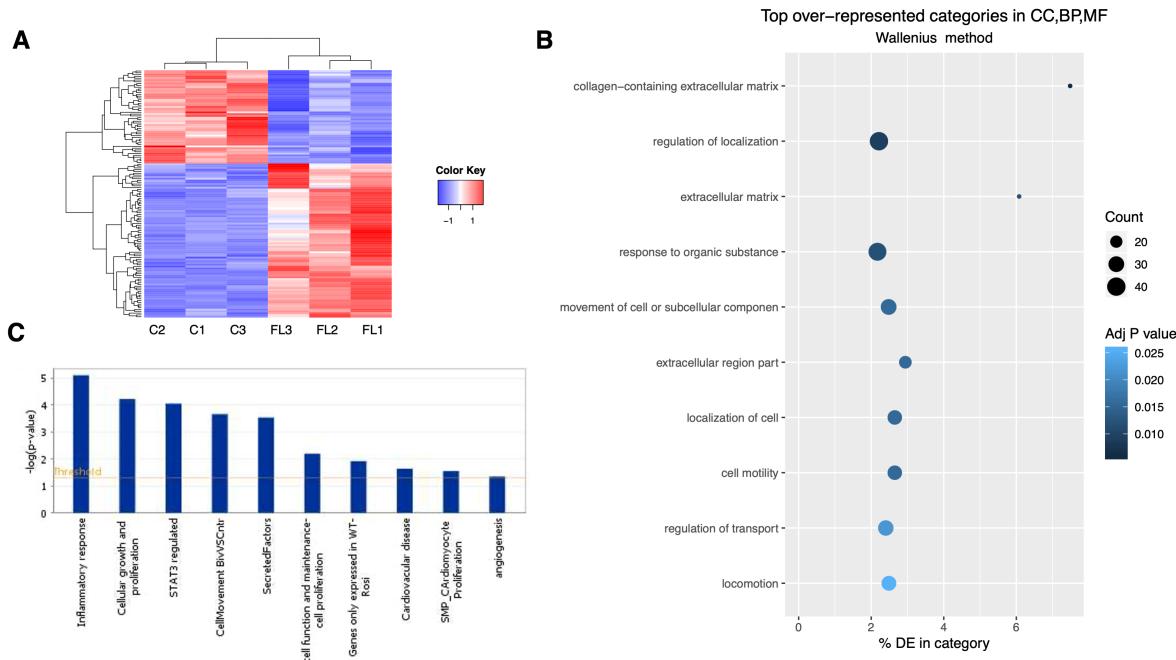


Figure 4.5: CTCF deletion alters gene expression in mature B cells. A) Heatmap of the z-scores of DEGs in each control (C, CTCF^{fl/fl}) and CTCF^{fl/fl} (FL) replicate . B) Graph with the top 10 over-represented GO terms including Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Different categories are indicated in y-axis. The x-axis shows the percentage of genes in those categories that belong to the group of DEGs. The size of the dots represent the number of DEGs associated with the category and the color represent the adjusted p-value (p-value for over-representation of the term in DEGs, adjusted for multiple testing with the Benjamini-Hochberg procedure). C) Graph with pathways significantly altered in CTCF-deficient cells compared to control cells generated with IPA software [51].

peaks, as the lack of binding may be the cause of gene expression changes between control and CTCF-deficient cells. We found 4,750 genes harboring lost peaks in their promoter regions; of these genes, 53 were DEGs (Fig. 4.6). This indicates that 53 out of 138 DEGs may be regulated by proximal CTCF sites.

To test if there is an enrichment of DEGs in the group of genes with lost peaks in their promoters, we calculated the number of DEGs that would be expected to be found in a set of 4,750 genes randomly chosen. We estimated this number based on the number of genes contained in the peak annotation database (DB) and the total number of DEGs, as shown in the equation 4.1. We found that it would be expected to obtain around 12 DEGs by chance which, compared to the 53 DEGs we detected, indicates that there is an enrichment of DEGs among genes with lost peaks in their promoters. This suggests that there is a relationship between loss of binding in promoter regions and differential expression changes in the corresponding genes.

$$\text{Nº genes randomly match} = \frac{\text{Nº DEGs}}{\text{Nº genes in the annotation DB}} \cdot \text{Nº genes analyzed}$$

$$\text{Nº genes randomly match} = \frac{138}{55,401} \cdot 4,750 = 11.83 \text{ genes} \quad (4.1)$$

4.4.2 Distal peak-mediated regulation: CTCF loops

To analyse the role of distal binding sites on promoting changes in gene expression, we developed an algorithm to identify, among all the possible pairs of binding sites, those most likely to form loops that regulate the expression of the genes inside them. Our algorithm selects potential loop

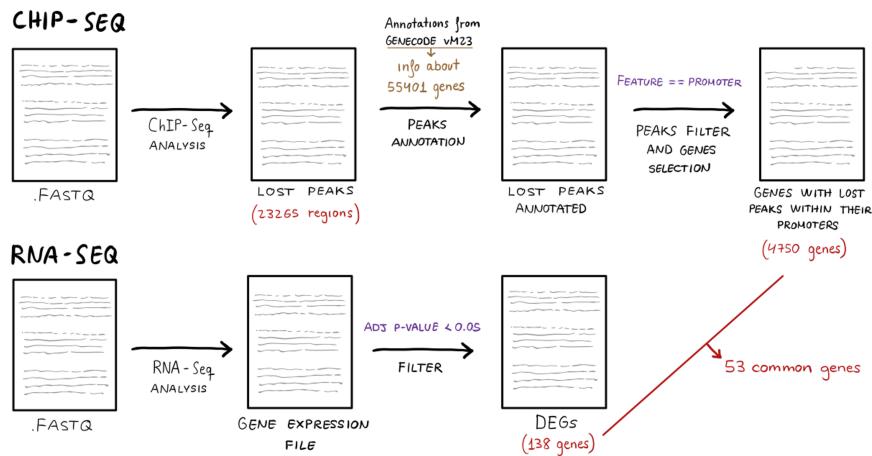


Figure 4.6: Steps followed to determine the relationship between proximal peaks and differential expression. Lost peaks bed file from the ChIP-Seq analysis was annotated using the ChIPseeker package from Bioconductor with GENECODE vM23 database, containing information about the genomic location of the regions (ex: promoter, 3'UTR, distal intergenic region, etc). After annotation, a filtration step enabled selecting only those peaks located in a promoter region and make a list with the names of the corresponding genes. This list was compared with the DEG list, obtaining the number of common genes between them.

regions based on the binding data and then identifies those with coordinated expression changes based on the RNA-Seq data.

As mentioned in the introduction, the vast majority of CTCF-mediated loops are formed between two convergent binding motifs. Therefore, the first step of our algorithm was to select only those peaks with a binding motif close enough to the consensus (score > 8.7) and annotate their forward (+) or reverse (-) orientation (Fig. 4.7 Step 1). Then, we selected pairs of peaks with convergent motif orientation (+ orientation in 5' end and - orientation in 3' end) in which at least one of them belonged to the lost peaks group (Fig. 4.7 Step 2). This restriction was imposed under the assumption that expression changes are due to the loss of loops and that this requires losing CTCF binding in at least one loop anchor. We restricted the loop size to up to 100 kb as we want to study only those intra-TAD loops that are involved in the coordinated expression of the genes between their boundaries. Thus, we obtained a group of 3,895 predicted loops, regions potentially capable of forming loops in control B cells that would be lost in the CTCF-deficient B cells (hereafter, loop group).

To evaluate the coordinated expression changes that took place in each of the predicted loops we established a metric that integrates the expression changes of set of genes (Fig. 4.7 Step 4). We used the average of the logarithm of the fold change (logFC) of the genes included in the region weighted with their p-value, thus reducing the contribution of genes with high variance between replicates. We called this metric Coordinated Expression Score (CES):

$$\text{CES} = \frac{1}{n} \sum_{i=1}^n [\log_2 \text{FC}_i \cdot (-\log_{10} p_i)] \quad (4.2)$$

where p_i is the p-value of the i gene and n is the number of genes included in the region analyzed. Note that, the lower p_i is, the more weight is given to the logFC of the gene in the CES calculation.

To determine, which of our predicted loops could be considered differentially expressed regions (DERs) -based on their CES-, we generated a control group that contains regions with the same size distribution but random locations in the genome and thus, not expected to form

chromatin loops (hereafter, no-loop group). This group was composed by 100 times more regions than the loop group in order to generate a reliable distribution of CES values of regions not identified as loops (Fig. 4.7 Step 3).

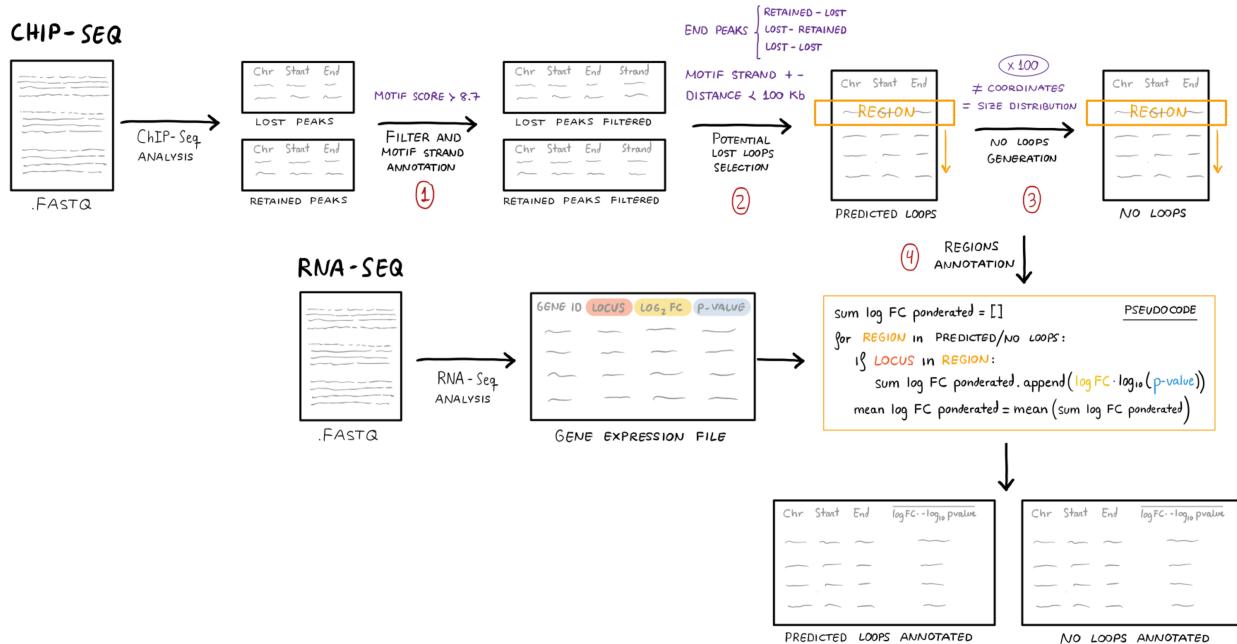


Figure 4.7: Steps followed to determine the relationship between distal peaks and differential gene expression. **Step 1:** Lost and retained peaks bed files were filtered to only contain those regions with a motif score > 8.7 . A new column with the motif orientation was added. **Step 2:** The peaks that passed the previous step were sorted by coordinates and scanned to establish pairs of peaks satisfying the following constraints: 1) one of the peaks should be in the lost group, 2) CTCF motif has to be in forward orientation for the peak located at 5' and in reverse orientation for the peak located at 3' and 3) distance between both peaks must be lower than 1 kb. In this way potential loop regions (loop group) were established and their coordinates were written to a new file. **Step 3:** The predicted loop regions were randomly permuted to generate the no-loop file, which contains 100 times more regions. **Step 4:** For each region in both loop and no-loop files the CES was calculated using the logFC and p-value information contained in the differential expression data file obtained in the RNA-Seq analysis (Eq. 4.1) and the number of genes included in each region was also annotated.

We assessed that the size distribution of the regions contained in both groups were identical (Fig. 4.8.A). However, we found that the no-loop regions had a significantly higher percentage of regions without genes compared to the loop group (Fig. 4.8.B and 4.8.C). This suggests that the loops predicted with our algorithm are enriched in gene-containing regions, as opposed to randomly selected regions. Then, we compared the CES distribution in predicted loops versus the no-loop regions and observed that loop regions had more widespread CES values, while the no-loop group had CES values closer to 0 (Fig. 4.8.D). This demonstrates that the loop group contains a higher proportion of regions with coordinated expression changes. To identify those regions with a significant coordinated expression change -based on CES-, we established a null hypothesis (H_0) distribution using the no-loop data so that it could be used to calculate p-values. In our case, H_0 is that the region analyzed does not exhibit coordinated expression changes. We established a CES threshold that was only exceeded by 5% of the null distribution data. Therefore, analyzed regions with a CES value over that threshold had a probability lower than 5% of belonging to the H_0 distribution ($p\text{-value} < 0.05$) (Fig. 4.8.E). We then calculated which percentage of the loop and no-loop regions exceeded the threshold on each side of the distribution. We observed that almost 20% of the loop regions passed the threshold, compared with 5% of the no-loop regions, thus confirming that loop group was enriched in regions with significant CES. In addition, we observed that in the loop group, a higher proportion of regions exceed the CES

threshold on the right side (overexpressed regions) than on the left side (underexpressed regions) (Fig. 4.8.F). Thus, the loss of certain loops in CTCF deficient cells results in more groups of genes coordinately upregulated than downregulated, suggesting that loop-mediated regulation of gene expression by CTCF may be more predominantly repressive.

To select differentially expressed regions from all the predicted loop regions analyzed, we corrected by Benjamini-Hochberg multiple testing [55] and set a new threshold that would allow us to select those loop regions with a significant CES (Fig. 4.9.A). We found that 208 out of the 3,895 predicted loop regions exceeded the new established threshold and, therefore, were considered DERs. From the 276 genes located within the DERs, 17 corresponded to DEGs. To test for the enrichment of DEGs, we also calculated the number of DEGs in the DERs selected from no-loop group. We obtained an average of 4.25 DEGs per group of 3,895 no-loop regions analyzed, 4 times less than those obtained with the loop group. We conclude that our algorithm is able to select regions that, in addition to being potential loops, are enriched in DEGs.

Finally, we examined the selected regions to verify that they were indeed DERs and that they could constitute a loop in the control cells that would be lost in the CTCF-deficient ones. Here we show a representative example Fig. 4.9.B). Note that the region contains 3 genes that exhibit coordinated expression changes (all have a positive logFC), 2 of them being DEGs (adjusted p-value < 0.05). It is a <100 kb length region flanked by a lost peak at one end and the CTCF motifs of the peaks flanking the region are in a convergent orientation. Thus, this region fulfills the requirements to constitute a loop that regulates the expression of genes inside it, although it must be experimentally verified.

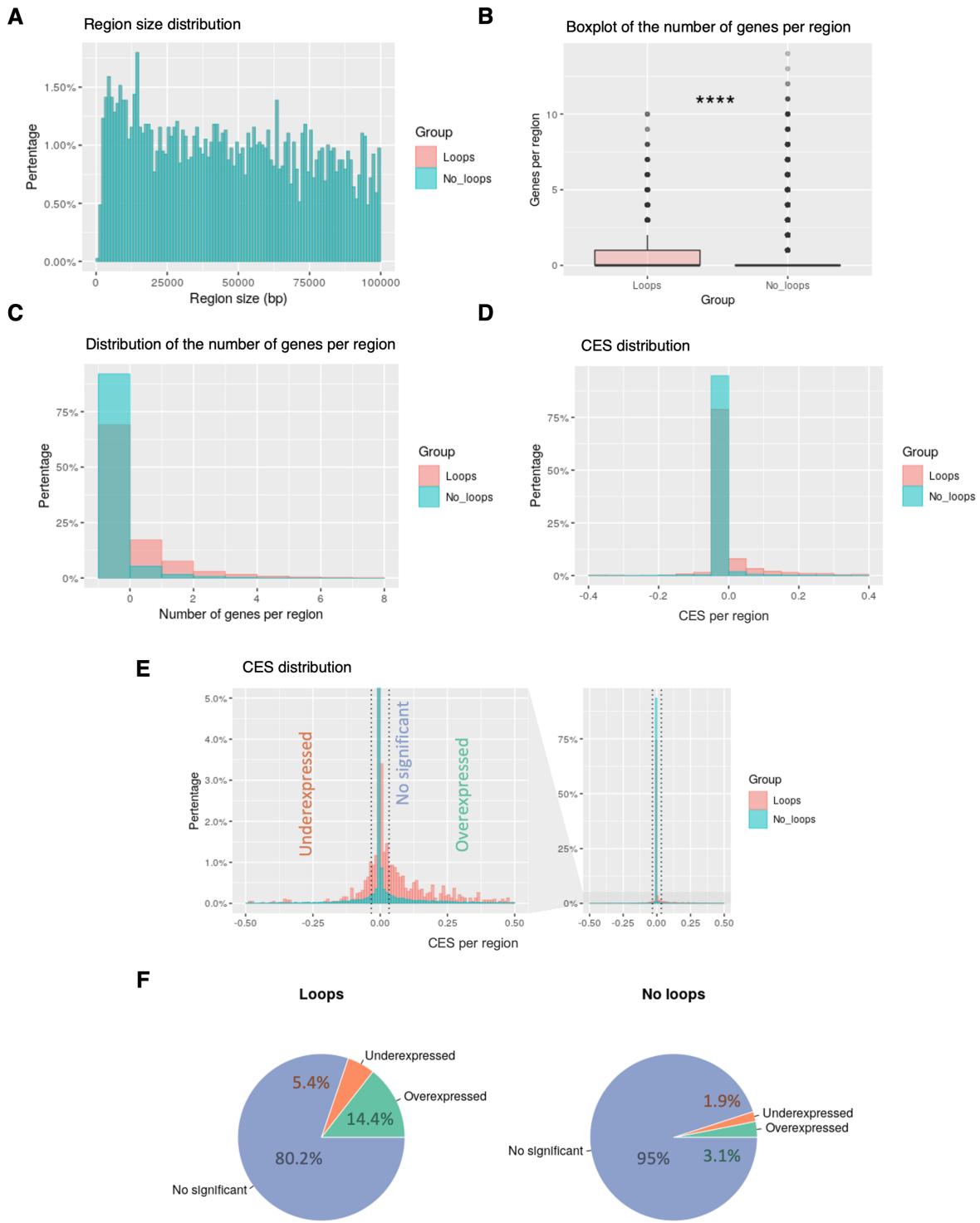


Figure 4.8: Loop and no-loop regions have distinct features. A) Size distribution of loop and no-loop regions. B) Boxplot of the number of genes included in loop and no-loop regions. **** symbol indicates a p-value $< 2e-16$. Statistical analysis was done with the Wilcoxon test. C) Histogram showing the distribution of the number of genes per region. D) CES values distribution for both groups of regions (bin size = 0.05). The x-axis has been cropped for easier interpretation, leaving less than 1% of the data out of the representation. E) CES values distribution for both groups of regions (bin size = 0.01) with zoom into the 0-0.5 region of the y-axis. Dotted lines highlight a threshold of ± 0.033 , established to leave out 5% of the regions from the no-loop group distribution. Regions with a CES value higher or lower than 0.033 are considered over- or underexpressed, respectively. F) Pie charts representing the percentage of regions belonging to previously defined categories within the loop (left) and no-loop (right) groups.

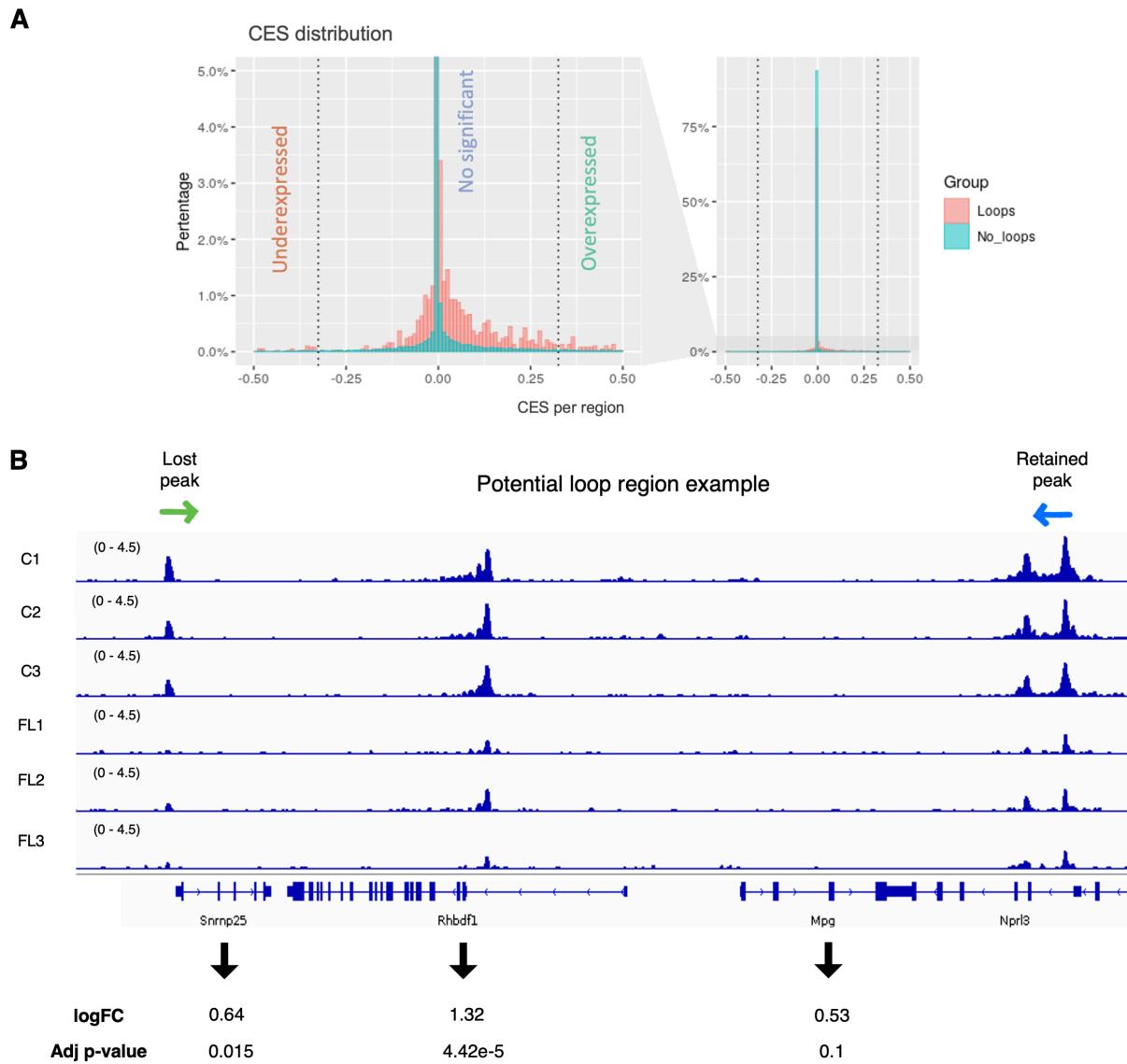


Figure 4.9: Predicted loop regions are enriched in genes coordinately expressed. A) CES values distribution for both groups of regions with zoom into the 0-0.5 region of the y-axis. The x-axis has been cropped for easier interpretation, leaving less than 1% of the data out of the representation. Dotted lines highlight the new threshold at ± 0.32 value (after Benjamini-Hochberg correction, FDR = 0.25). Regions with a CES value higher or lower than 0.32 are considered significantly over- or underexpressed, respectively. B) Representative CTCF ChIP-Seq profile of an overexpressed potential loop region (CES = 4.4, chr11:32,200,317-32,239,186). CTCF signal (normalized read counts) for each control (C, CTCF^{fl/+}) and CTCF^{fl/fl} (FL) replicate is shown. Refseq genes in the region are also indicated. The orientation of the CTCF-motif included in end peaks of the regions are indicated with green (forward) and blue (reverse) arrows. logFC and adjusted p-value of genes included in the region are listed below them.

5

Discussion

In this project, we have analyzed how CTCF depletion impacts on the occupancy of CTCF binding sites genome-wide and the subsequent transcriptional changes it causes in mature B cells. We found, by ChIPseq analysis in CTCF deficient B cells, that some CTCF peaks were lost while others were resistant to CTCF depletion, which suggests that there are CTCF binding sites with different affinities. The regions containing the retained CTCF binding sites are significantly wider, reflecting that retained peaks harbor tandem CTCF binding sites. This finding suggests that regions with clustered CTCF-binding sites are essential for maintaining chromatin architecture and cell functions. We also observed that the retained regions are more enriched in CTCF motifs sharing a higher similarity with the consensus CTCF binding sequence, indicating that variations in the binding sequence can regulate CTCF recruitment. Together, the arrangement and the sequence of retained CTCF binding sites can explain the differences in CTCF binding (retained versus lost peaks), although other mechanisms of CTCF regulation, such as DNA methylation, histone marks or protein partners, are also compatible with our findings.

Interestingly, we found a higher percentage of CTCF retained peaks that overlap with TAD boundaries, suggesting that the retained subset of binding sites sites is more likely to be establishing higher order chromatin architectures. A recent study of CTCF genome-wide depletion with siRNA has shown very similar results in a human prostate cancer cell line, although the differences in the percentage of overlap were less pronounced in our study [59]. This can be explained because we have used the boundary coordinates from ESCs instead of B cells, which can lead to imprecisions in TAD definition. Thus, it would be interesting to perform this analysis making use of existing Hi-C B cell data [60], or ideally, to perform our own HiC experiment in CTCF proficient and deficient B cells, which would also allow to detect TAD disruption triggered by CTCF deficiency.

Transcriptomic analysis shows 138 DEGs after CTCF depletion in mature B cells, most of which are upregulated in CTCF deficient B cells. This could indicate that, at the level of the regulation of gene expression, CTCF could have a predominant repressor role in mature B cells. Intriguingly, the loss of more than 20,000 CTCF binding sites results in a relatively mild gene expression phenotype, which has been previously discussed [19]. One of the hypothesis to explain these observations is that additional structural proteins, such as YY1, continue establishing long-range interactions between enhancers and promoters in the absence of CTCF. Alternatively, the interactions between regulatory elements mediated by CTCF may not be critical in a stable state (i.e. in resting, mature B cells) but rather for the dynamic control of gene transcription

at transition processes such as cell activation, differentiation, etc. Some studies have shown that CTCF-associated chromatin loops tend to comprise regions of enhancer-regulated stimulus responsive genes, thus insulating them from neighboring regions of housekeeping genes [61]. In this scenario, only those genes involved in the response to stimuli would display an expression change upon CTCF deletion. Thus, the quiescent G_0 state of naive B cells, with relatively little transcriptional activity, would be compatible with a low impact of CTCF deletion on gene expression [62]. In contrast, when naive B cells become activated, their transcriptional program changes dramatically to face the GC reaction which involves cell proliferation, cell death, migration, DNA recombination and repair and cell differentiation. In addition, during activation, B cells undergo a massive architectural change in their chromatin [60]. In agreement with this, previous data from the lab showed that CTCF deletion in activated B cells triggers a dramatic transcriptome change [28]. Thus, we are currently performing this ChIPseq/RNAseq integrative analysis on activated B cells.

Regarding the relationship between expression changes and CTCF binding sites, we approached the analysis in two different, complementary ways: a) genes that could be regulated by CTCF binding to their promoter region and b) groups of genes that could be coordinately regulated by CTCF-mediated loops. For this second approach, we have developed an algorithm to link CTCF-mediated chromatin organization to transcriptional regulation so that we could predict which regions are involved in forming loops and whether they are coordinately regulated by CTCF. Some studies have attempted to predict intra-TAD loops mediated by CTCF based on binding information [18, 61]. However, we expect that combining binding site information with gene expression data, we will be able to make a more accurate prediction, since potential loop regions that show coordinate differential expression are more likely to be loops.

We think that our algorithm could be very useful as an alternative to promoter region analysis to integrate chromatin organization and gene expression. However, our results need to be followed up by experimentally testing these coordinated regulation by, for example, disrupting a loop with mutants and studying the consequent expression changes. Additionally, several features of the algorithm could be optimized as detailed in the next paragraphs.

- We have established 100 kb as a size limit for the regions in order to focus on intra-TAD loops and avoid detecting patterns of coordinated regulation caused by higher order genomic structures. One-hundred kb is the resolution limit of most techniques to study chromatin interactions, such as DNA-FISH and Hi-C [5]. Therefore, our algorithm could complement the results obtained from standard resolution Hi-C data to increase the confidence in the interaction of certain small regions. However, ChIA-PET studies have shown several >100 kb intra-TAD CTCF-anchored loops that regulate gene expression [63]. Therefore, increasing the maximum region size allowed by our algorithm could allow the detection of additional loops.
- Since the smaller the number of genes in a region, the more likely it is to find them coordinately regulated by chance, we could optimize CES calculation by integrating a term that increases proportionally to the number of genes included in the region analyzed.
- We cannot rule out that our no-loop group contains regions that are involved in loop formation due to random location, thus distorting our results. We are currently working on establishing a more appropriate control group formed by regions for which there is evidence that they are not part of any CTCF-mediated intra-TAD loop based on Hi-C data.

Our analysis does not distinguish between direct gene expression changes (caused by CTCF binding) and indirect changes (those arising from action of proteins encoded by CTCF-regulated genes). To solve this problem, we would ideally have used an inducible model in which we were able to analyze the changes soon after CTCF deletion, so that mostly direct expression changes were detected.

Despite the limitations mentioned above, we consider that the algorithm developed in this project provides a first approach to understand the relationship between CTCF-mediated loops and gene expression changes. Moreover, this way of integrating ChIP-Seq and RNA-Seq data allows us to establish *a priori* hypotheses about the function of specific CTCF binding sites in the absence of DNA structural data, which could then be validated experimentally.

Bibliography

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* **17**, 661–678 (2016).
2. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics* **2**, 292–301 (2001).
3. Van Schoonhoven, A., Huylebroeck, D., Hendriks, R. W. & Stadhouders, R. 3D genome organization during lymphocyte development and activation. *Briefings in Functional Genomics* **19**, 71–82 (2020).
4. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* **62**, 668–680 (2016).
5. Matharu, N. & Ahituv, N. Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS Genetics* **11**, e1005640 (2015).
6. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
7. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
8. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
9. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genetics* **42**, 53–61 (2010).
10. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the β -globin locus. *Genes and Development* **20**, 2349–2354 (2006).
11. Ong, C. T. & Corces, V. G. CTCF: An architectural protein bridging genome topology and function. *Nature Reviews Genetics* **15**, 234–246 (2014).
12. Lutz, M. *et al.* Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Research* **28**, 1707–1713 (2000).
13. Chen, H., Tian, Y., Shu, W., Bo, X. & Wang, S. Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome. *PLoS ONE* **7**, e41374 (2012).
14. Liu, F., Wu, D. & Wang, X. Roles of CTCF in conformation and functions of chromosome. *Seminars in Cell and Developmental Biology* **90**, 168–173 (2019).
15. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. *Cell* **137**, 1194–1211 (2009).
16. Barski, A. *et al.* High-Resolution Profiling of Histone Methylation in the Human Genome. *Cell* **129**, 823–837 (2007).
17. Bastiaan Holwerda, S. J. & de Laat, W. CTCF: The protein, the binding partners, the binding sites and their chromatin loops. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368** (2013).

18. Matthews, B. J. & Waxman, D. J. Computational prediction of CTCF/ cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *eLife* **7**, e34077 (2018).
19. Arzate-Mejía, R. G., Recillas-Targa, F. & Corces, V. G. Developing in 3D: the role of CTCF in cell differentiation. *Development* **145** (2018).
20. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E6456–E6465 (2015).
21. Schatz, D. G. & Swanson, P. C. V(D)J recombination: Mechanisms of initiation. *Annual Review of Genetics* **45**, 167–202 (2011).
22. Ebert, A., Hill, L. & Busslinger, M. Spatial Regulation of V-(D)J Recombination at Antigen Receptor Loci. *Advances in Immunology* **128**, 93–121 (2015).
23. Methot, S. P. & Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Advances in Immunology* **133**, 37–87 (2017).
24. Marina-Zárate, E., Pérez-García, A. & Ramiro, A. R. CCCTC-Binding Factor Locks Premature IgH Germline Transcription and Restains Class Switch Recombination. *Frontiers in Immunology* **8**, 1076 (2017).
25. Pérez-García, A. *et al.* CTCF orchestrates the germinal centre transcriptional program and prevents premature plasma cell differentiation. *Nature Communications* **8**, 1–12 (2017).
26. Heath, H. *et al.* CTCF regulates cell cycle progression of $\alpha\beta$ T cells in the thymus. *EMBO Journal* **27**, 2839–2850 (2008).
27. Rickert, R. C., Roes, J. & Rajewsky, K. B lymphocyte-specific, Cre-mediated mutagenesis in mice. *Nucleic Acids Research* **25**, 1317–1318 (1997).
28. Marina-Zárate, E. *CTCF regulates transcriptional programs and antibody diversification in mature B cells*. PhD thesis (Universidad Autónoma de Madrid, 2020).
29. Teaching team at the Harvard Chan Bioinformatics Core (HBC). *Introduction to ChIP-Seq using high-performance computing* <https://github.com/hbctraining/Intro-to-ChIPseq> (09/07/2020).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
33. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
34. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in Genetics* **5** (2014).
35. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
36. R Core Team. *R: A Language and Environment for Statistical Computing* <https://www.r-project.org/>.
37. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160–W165 (2016).
38. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).

39. Yu, G., Wang, L. G. & He, Q. Y. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
40. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
41. Wang, Y. *et al.* The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology* **19**, 151 (2018).
42. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557–572.e24 (2017).
43. Batut, B. *et al.* Reference-based RNA-Seq data analysis (Galaxy Training Materials) <https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html> (12/05/2020).
44. Batut, B. *et al.* Community-Driven Data Analysis Training for Biology. *Cell Systems* **6**, 752–758.e1 (2018).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
46. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
50. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**, R14 (2010).
51. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
52. Blankenberg, D. *et al.* Dissemination of scientific software with Galaxy ToolShed. *Genome Biology* **15**, 403 (2014).
53. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
54. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
55. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
56. Delgado, P. *et al.* Interplay between UNG and AID governs intratumoral heterogeneity in mature B cell lymphoma. *PLoS Genetics* **16**, e1008960 (2020).
57. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
58. Kelley, D. Z. *et al.* Integrated analysis of whole-genome ChIP-Seq and RNA-Seq data of primary head and neck tumor samples associates HPV integration sites with open chromatin marks. *Cancer Research* **77**, 6538–6550 (2017).

59. Khouri, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nature Communications* **11** (2020).
60. Kieffer-Kwon, K. R. *et al.* Myc Regulates Chromatin Decompaction and Nuclear Architecture during B Cell Activation. *Molecular Cell* **67**, 566–578.e10 (2017).
61. Oti, M., Falck, J., Huynen, M. A. & Zhou, H. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics* **17**, 252 (2016).
62. Myers, D. R., Zikherman, J. & Roose, J. P. Tonic Signals: Why Do Lymphocytes Bother? *Trends in Immunology* **38**, 844–857 (2017).
63. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics* **43**, 630–638 (2011).