Characterization of the interplay between the genome architecture and gene co-expression

Audrey Baguette

Department of Human Genetics

McGill University

Montreal, Quebec

April 2020

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science

A. Abstract

The relation between the structure of the genome and gene regulation is critical to normal and disease development, but the molecular details of how they are interconnected are still unknown. Chromatin conformation capture (Hi-C) studies discovered several layers of chromatin organization. However, the way those structures impact or are impacted by regulation is unclear. We thus wanted to clarify the links between chromatin architecture and transcription regulation. In this study, we use two types of domains, one having a structural definition and the other a functional definition and compare them to find their differences and similitudes. Topologically Associating Domains (TADs) have been selected to represent the genomic architecture. They have a more static nature and their boundaries have been suggested to limit the spread of regulatory signals. Co-expression Domains (CODs) were chosen to represent the aspects of gene regulation. CODs are defined as domains within which genes have correlated expression. By definition, CODs are thus very dynamic and more likely to change from cell to cell.

In this study, we analyze the effect of TAD boundaries on nearby genes. Here we show that TADs and CODs have distinct functions and are delimited by different boundaries. We confirm that TAD boundaries disrupt co-expression. We also characterize COD boundaries and find that they seem to be marked by a switch of strand on which genes are located and they are independent of structural proteins. We use expression quantitative trait loci (eQTL) data to confirm the observations and find that genes affected by the same eQTL are preferentially located on the same strand and are less likely to be separated by barriers such as TAD boundaries. We thus propose a model for human cells in which the gene conformation impacts gene co-regulation. We suggest that strand position of genes affects their co-expression probability and the introduction of barrier elements further disrupts it. That model would serve as a simple principle to which more complex mechanisms may rely.

B. Résumé

La relation entre la structure du génome et la régulation des gènes est critique au développement normal et à celui des maladies, mais on sait encore bien peu à propos des détails qui les relient au niveau moléculaire. Des études de capture de la conformation de la chromatine (Hi-C) ont découvert plusieurs couches d'organisation de la chromatine. Toutefois, il n'est pas clair de quelle manière ces structures impactent et sont impactées par la régulation. Nous voulions donc clarifier les liens entre l'architecture de la chromatine et la régulation de la transcription. Dans cette étude, nous utilisons deux types de domaines, l'un défini de manière structurellement et l'autre défini de manière fonctionnelle et les comparons pour trouver leurs différences et similitudes. Les "Topologically Associating Domains" (Domaines d'Association Topologique, TADs) ont été sélectionnés pour représenter l'architecture du génome. Ils ont une nature plus statique et il a été suggéré qu'ils limitent la propagation des signaux de régulation. Les "Co-expression Domains" (Domaines, de Co-expression, CODs) ont été choisis pour représenter les aspects de la régulation des gènes. Les CODs sont définis comme étant des domaines au sein desquels les gènes ont une expression corrélée. Par définition, les CODs sont donc très dynamiques et plus enclins à changer d'une cellule à l'autre.

Dans cette étude, nous analysons les effets des frontières de TAD sur les gènes adjacents. Ici nous montrons que les TADs et les CODs ont des fonctions distinctes et qu'ils sont délimités par des frontières différentes. Nous confirmons que les frontières de TAD perturbent la co-expression. Nous caractérisons aussi les frontières de COD et trouvons qu'elles semblent marquées par le changement de brin sur lequel les gènes se situent et qu'elles sont indépendantes des protéines structurelles. Nous utilisons les données d' "expression quantitative trait loci" (eQTL) pour confirmer les observations et trouvons que les gènes affectés par le même eQTL se situent préférentiellement sur le même brin et sont moins enclins à être séparés par des barrières telles que les frontières de TAD. Nous proposons donc un modèle pour les cellules humaines dans lequel la conformation des gènes impacte leur co-régulation. Nous suggérons que la position des gènes sur les différents brins affecte la probabilité qu'ils soient co-exprimés et que l'introduction d'éléments-barrière diminue davantage celle-ci. Ce modèle servirait de principe de base sur lequel des mécanismes plus complexes pourraient reposer.

C. Table of Contents

D. List of Abbreviations

A549	Cell line of adenocarcinomic human alveolar basal epithelial cells
BCL3	B-cell lymphoma 3-encoded protein
Br-UTP	Brome-Uracil TriPhosphate
CEBPB	CCAAT/Enhancer-Binding Protein Beta
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-seq	Chromatin Immuno-Precipitation followed by sequencing
COD	Co-expression domain
CRD	Cis-regulatory domain
CTCF	CCCTC-Binding Factor
Ctrl	Control
DEG	Differentially Expressed Genes
E2	Estradiol
ENCODE	The Encyclopedia of DNA Elements
EP300	E1A-associated Protein p300
eQTL	Expression Quantitative Trait Loci
FISH	Fluorescence In Situ Hybridization
FOSL2	Fos-related Antigen 2
GO	Gene Ontology
GR	Glucocorticoid receptor
GTEx	Genotype-Tissue Expression
H3K4me1	Mono-methylation at the 4th lysine residue of the histone H3 protein
H3K4me2	Di-methylation at the 4th lysine residue of the histone H3 protein
H3K4me3	Tri-methylation at the 4th lysine residue of the histone H3 protein
H3K9me3	Tri-methylation at the 9th lysine residue of the histone H3 protein
H3K27ac	Acetylation at the 27th lysine residue of the histone H3 protein
HES2	Hes family bHLH transcription factor 2
Hi-C	Hi-throughput sequencing for chromatin conformation capture
JUN	Transcription factor jun
JUNB	Transcription factor jun-B
kb	kilobase (1000 base pairs)
Mb	Megabase (1 000 000 base pairs)
MCF-7	Michigan Cancer Foundation-7 cell line
mRNA	Messenger RiboNucleic Acid
NR3C1	Glucocorticoid receptor
PCA	Principal Component Analysis
RAD21	Double-strand-break repair protein rad21 homolog
RNAPol2	RNA Polymerase II
SMC3	Structural Maintenance of Chromosomes protein 3
snRNP	small nuclear RiboNucleoProteins
t-SNE	t-distributed Stochastic Neighbor Embedding
TAD	Topologically Associating Domain
TF	Transcription Factor

E. List of Figures

F. List of Tables

G. Acknowledgments

I would like to thank Dr. Guillaume Bourque and Dr. Steve Bilodeau for their support through my Master degree. Their insightful comments made me to understand how to do better research: make sure you understand the data and techniques before trying to produce results. Under their supervision, I learned to develop my independence as a student, but I also learned that I should not be afraid to ask for help when I need it.

I am grateful to Dr. Yasser Riazalhosseini and Dr. Hamed S. Najafabadi, my Supervisory Committee members, for the pertinent questions they brought up during our annual meetings.

I would like to acknowledge the Fonds de Recherche du Québec en Santé (FRQS) funding agency, as well as McGill University's Department of Human Genetics, which gave me the financial resources needed to support my research. McGill University provided me the required environment to grow and become an accomplished student.

A special mention is needed to the ENCODE Consortium and the laboratory of Dr. Tim Reddy in Duke, the GTEx Projet, Compute Canada and Calcul Québec, without which I would not have had data to analyze nor the tools to do it efficiently. The Hi-C, RNA-seq and ChIP-seq data was the downloaded from the ENCODE Portal. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this thesis were obtained from Single-Tissue cis-eQTL Data on the GTEx Portal, dbGaP accession number phs000424.v8.p2 on 03/02/2020.

Special thanks to Dr. Rola Dali to whom I can attribute most of my gained knowledge of Hi-C data and techniques in my first months, but who was also open to share professional experiences and helped me set my future academic goals.

On a personal note, I would like to thank my parents for their unconditional love and support all these years. I would not be the woman I am today without them.

I would also like to thank Dr. Guillaume Bourque's students and the Canadian Centre for Computational Genomics (C3G) team, Montreal node, for making me feel included in the lab, and more specifically Mr. François Lefebvre, Mr. Edouard Henrion. Mr. José Héctor Gálvez López,

Dr. Alain Pacis and Mr. Pubudu Nawarathna for their help understanding and using specific bio-informatics tools.

H. Format of the Thesis

This thesis is manuscript-based.

I. Contribution of Authors

Chapters 1, 3 and 4:

I wrote these chapters in their entirety, and they were revised and edited by Dr. Guillaume Bourque and Dr. Steve Bilodeau.

Chapter 2:

The pre-processed data was retrieved from ENCODE[1,2], and I conducted all further processing and analyses presented (normalization of RNA-seq read counts, PCA and t-SNE, identification of differentially expressed genes, analysis of distance between pairs of genes, analysis of prevalence of TAD boundaries or TF between pairs of genes).

For eQTL analysis, eQTL-gene pairs and the list of tested genes were retrieved from GTEx[3] and I conducted further analyses.

I wrote the manuscript and it was revised and edited by Dr. Guillaume Bourque, Dr. Steve Bilodeau and Dr. Rola Dali.

Chapter 1: General introduction

1.1 Elements of the Nucleus Architecture

The nucleus is only a few microns in diameter, while the total length of DNA is around 2 meters[4]. Chromosomes thus need to be greatly compacted and that compaction is achieved through several layers of architecture. The first level consists in sections of 146 base pairs of DNA rolled around an histone to form the chromatin. During interphase, when chromosomes are the most condensed and in their well-known "X" shape, the chromatin forms nucleosomes, then coils and supercoils. However, in that conformation, genes might not as easily be accessed by the transcription machinery. Active genes must thus have another architecture, highly structured to allow for controlled regulation, but flexible enough to be able to switch quickly from repressed state to active state and vice-versa. The several layers of organization include compartments, topologically associating domains (TADs) and chromatin loops, but also several less-characterized sub-TAD domains.

Compartments are the largest structure. Their size ranges from a few to a dozen Mb long[4]. Compartments are divided in two types: A and B. They are identified by their interaction patterns: A compartments interact more often with other A compartments, while B tend to interact with B compartments[4,5], A compartments have found to be enriched in genes, more specifically active genes, and harbor more histone marks of open chromatin[4,6–8]. In contrast, B compartments show an enrichment in closed histone marks. A study suggested they could be further divided compartments into sub-compartments, each having specific histone marks signatures[5]. Compartments and sub-compartments themselves contain TADs. Topologically associating domains are regions, usually less than 1Mb long[9–12], defined by a high concentration of interactions: sections within TADs have a high frequency of interactions with other sections within the same TAD but not with sections outside of it[9,11–13]. Interactions are mediated through chromatin loops. Their role is simply to bring distal elements located on the same chromosome, often enhancers and promoters, in contact so they can interact[7,14–17].

1.1.1 Topologically Associating Domains

TADs are structures smaller than compartments that have been vastly explored and are well characterized. TAD boundaries are dynamic but have found to be in great part conserved between single cells[10], across species[13,18–20] and cell types[7,12,13,18,20] and would even be resistant to heat

shock[21]. Some of those boundaries may be shared by compartments, while they are distinct structures[6,13,19]. On the contrary to TADs, the level of conservation of compartments is debatable, as their boundaries seem similar across cell lines[22], but show variation around key genes to activate them or repress them[19,23–25]. Because of their low level of conservation, relative to TADs, compartments have been suggested to be a statistical entity reflecting preferential contacts between TADs, rather than being physical entities[26].

Intra-TAD interactions are less well understood than TADs. They have been reported to vary from cell to cell[10,16,20,27], yet a heat shock has been shown not to affect contacts between promoters and enhancers[21] and some loop might even be formed before different stimulus are applied on cells[28,29]. Some loops are thought to be very dynamic, forming and breaking depending on the cell needs[7,30]. Others, more static loops, sometimes referred to as "CTCF loops", as they are stabilized by structural proteins CTCF and Cohesin[18,28], the two main structural proteins in human cells.

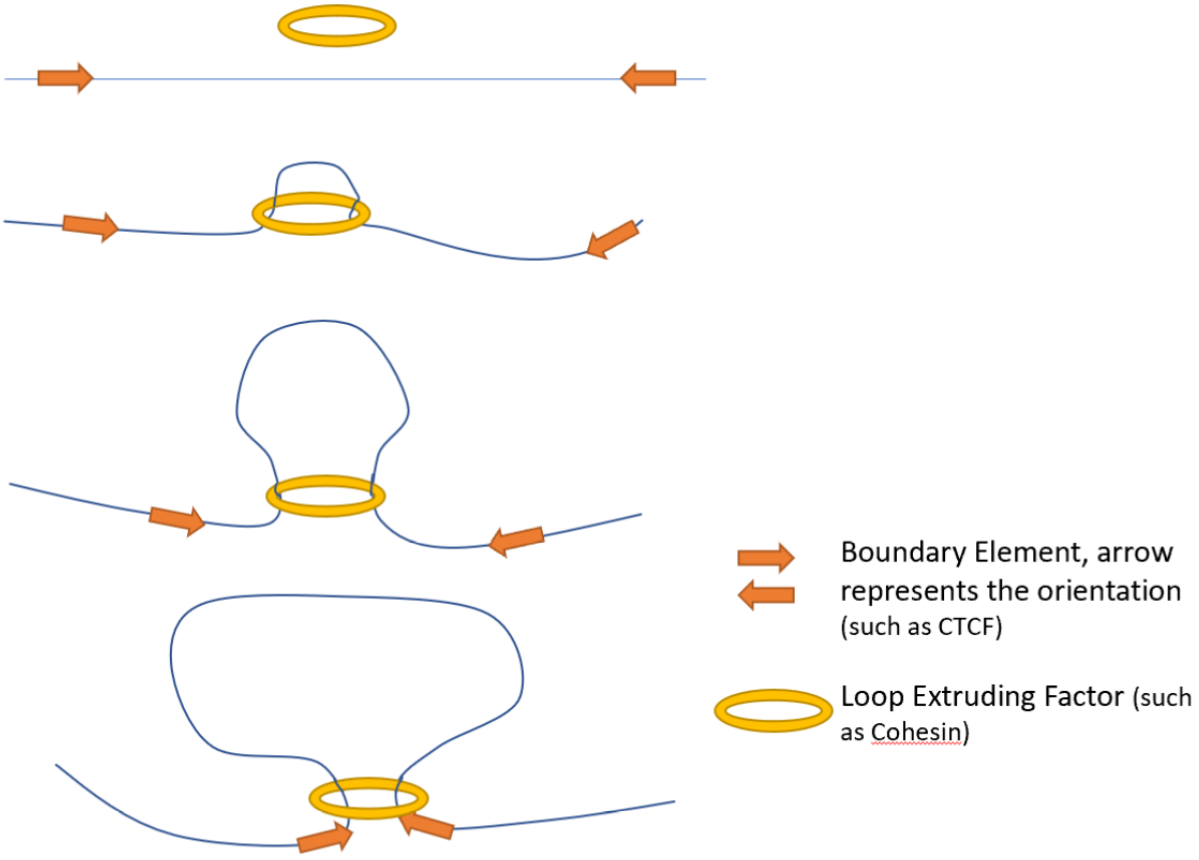
Because of their well-defined, structural description, TADs can serve as a basis to describe the 3-dimentional organization of the genome. Their positions have been reported in several cell types and there seems to be a consensus about their nature. They can be easily identified, seem to reflect physical entities[26], and while they can fluctuate through time and conditions, there is evidence they are present even before differentiation and on inactive chromosomes[9].

1.1.2 Sub-TAD Domains

As technologies improved, studies tried to explore the internal structure of TADs and found several types of sub-TAD domains. First, TAD-like structures with similar properties have been identified. They are often simply called sub-TADs[5,16,31] but can also be referred to as insulation neighborhoods[31]. They are defined as chromatin loops formed by a CTCF homodimer and by Cohesin containing at least one gene. Their boundaries have insulating properties and their perturbation lead to gene expression dysregulation. The CTCF binding sites forming insulated neighborhoods also have been showed to be conserved in human germline and primates. Other studies moved on from the architectural definition and tried to find domains with genes having correlated expression. Thus, came co-expression domains (CODs)[32] and cis-regulation domains (CRDs)[33]. Both of them are defined using correlation, but the former correlates gene expression and the latter chromatin peaks.

1.1.3 Structural proteins

Many nuclear proteins help to stabilize the various chromatin structures such as TADs, insulation neighborhoods, chromatin loops. Indeed, TAD formation is thought to be driven by a loop-extrusion model: chromatin is pulled through a loop extruding factor, a ring-shaped protein complex, until boundary elements are reached[16,34–36] (General figure 1). The main loop extruding factor involved in TAD formation is Cohesin, and two CTCF proteins positioned in convergent manner usually serve as boundary elements. TADs have thus also been defined as corner-dot domains that are formed by that model and delimited by architectural proteins[16].The strength of TAD boundaries has also been correlated with the number of structural proteins found within those boundaries[18].



General figure 1: Illustration of the loop extrusion model. *DNA is extracted through a loop extruding factor to form a loop until boundary elements are reached. Inspired by: Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Reports 15, 2038–2049 (2016).*

1.1.3.1 CTCF

CTCF can act as a transcription factor, but is also one of the main structural proteins in eukaryotic cells[36]. It helps forming TADs, insulated neighborhoods and chromatin loops. As such, CTCF has been found to be enriched at TAD boundaries[9,13,27,37], but CTCF binding sites can also be found within TADs[9,14,18,38,39]. It could thus have multiple roles, depending on its location[36]. A recent study found that some CTCF binding sites are resistant to depletion, and that those resistant sites are more often found at TAD boundaries[37]. Moreover, a study using Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) with CTCF antibodies found clusters that seem to correspond to TADs[40]. CTCF is thus definitely involved in TAD formation and stabilization, but it cannot explain everything by itself.

1.1.3.2 RAD21 and SMC3: subunits of Cohesin

Cohesin is a protein complex containing multiple sub-units that form a ring shape, through which DNA slides to form loops. Its core sub-units are SMC1A, SMC3, RAD21 and STAG1 or STAG2[41]. It participates in cohesion and segregation of sister chromatids before mitosis and DNA repair but is also important for genome organization. Its main roles for gene regulation are to bring regions in close proximity, favorizing contacts between elements for co-regulation purposes and help CTCF isolating elements from each other. Cohesin is one of the proteins that helps connecting promoters and enhancers[19,42] and its knockdown causes changes in gene expression[42]. RAD21 and SMC3 can be used used as proxies in ChIP-seq experiments to find the binding sites of Cohesin along the chromatin.

1.2 Interplay between genomic structure and gene transcription

1.2.1 Transcription Factories

Transcription has long been thought to be happening in a stochastic manner. There is however mounting evidence that this process is highly controlled, but the mechanisms are still poorly understood. In 1993, a study conducted by Jackson and collaborators used imaging techniques to visualize the production of mRNA[43]. They treated cells with Br-UTP to detect nascent RNA. When ongoing transcription, a portion of the new RNA molecules incorporated Br-UTP instead of normal UTP. A combination of two special antibodies, one targeting RNA with incorporated Brome, and the second linked to Texas red and binding to the first one, were used to

visualize the location of newly created RNA molecules with fluorescence microscopy. Rather than seeing points of red, witnessing the location of nascent RNA, spread randomly across the nucleus, they found distinct foci. This suggest that mRNA is not created everywhere is the nucleus, but that genes that must undergo transcription regroup themselves at specific locations. To confirm that the observations are indeed due to a transcription hub, they double-labelled cells, adding Sm antigens labelling to the previous protocol. Sm antigens targets small nuclear ribonucleoproteins (snRNP), which form spliceosomes. The majority of snRNPs co-localize with the nascent RNA foci, confirming that mRNA is produced at defined loci. Thus, came the idea of transcription factories, clusters of DNA and transcription machinery whose purpose is to produce mRNA.

Since their discovery, various studies tried to uncover the secrets of transcription factories. Surprisingly, they have been found to be resistant to transcription inhibition[44,45]. Some suggested that transcription factories are not only clusters of DNA, RNAPol2 and spliceosomes, but also that they could regroup TFs such that genes having similar needs in would co-cluster into specialized factories[46–48]. Others suggested that genes are not bound to one transcription factory but might come and go freely[45,49], or even that transcription factories are not "hovering" into the nucleic space, but that RNAPol2 might be attached to some kind of structure[47,50]. A study used Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) to find clusters of genes interacting through RNAPol2 and discovered that they tend to be enriched for the same Gene Ontology categories or share similar pathways, suggesting transcription factories could co-transcribe genes needed for a typical cell response[51,52]. However, the exact definition of "transcription factory" remains blurry and changes slightly from paper to paper. It is also unclear what constitutes a transcription factory: is it only DNA, RNA and RNAPol2 that make up its core, or should the whole transcription machinery, including TFs, be considered when identifying them? As the interaction of genes in transcription induces a clustering, and changes the DNA architecture, the answers to uncovering gene expression regulation seem to be hidden in how the nucleus is organized in the 3D space.

1.2.2 TADs as Regulatory Units

The most controversy around TADs concerns their functional role. Do they serve directly a regulatory purpose or not? TAD boundaries show an enrichment in promoter-associated histone marks, gene transcription start sites, (TSS) and more specifically housekeeping genes[13–15,17]. TAD

boundaries have been found to have insulator properties: they seem to limit the effect of enhancer to genes within the same TAD[13,27,34,37]. Indeed, disruption of TAD boundaries leads to changes in gene regulation[14,16,39]. TAD boundaries thus seem essential for transcription regulation. A study centered around the HoxD gene and its expression through limb development showed that the gene is situated at the boundary of two TADs and interacts preferentially with one or the other depending on the cell requirements[53]. CTCF has been suggested to act as an insulator protein, limiting the spread of regulatory elements outside of the designated zone at TAD boundaries[9,13,27,37]. However, knock-down of CTCF does not entirely disrupts TADs[15,37] and effects on gene expression take time to be seen[28].

Besides the suggested function of their boundaries, TAD have been suggested to act as regulation units[29]. Indeed, some TADs showed coordinated expression changes of the genes they contained following progesterone and estrogen induction in MCF7 cells. However, all TADs do not follow that pattern. Moreover, the correlation of expression between genes contained in the same TAD is usually similar to the correlation of expression between genes inside randomly created regions of similar sizes as TADs[32]. Thus, while they are structurally well defined, the function of TADs, from a regulatory point of view, needs further exploration.

1.2.3 Co-expression Domains

To find the links between chromatin architecture and expression regulation, we selected a structure to best represent the co-regulation patterns. The TADs have been selected to best represent the structure, but an expression-oriented structure is needed to serve as a comparison point. The exact differences between insulated neighborhoods, CODs and CRDs are not well defined, it should thus be risky to consider them all, as they might overlap. We chose to focus on CODs. Indeed, we want to identify groups of genes with similar levels of expression, and CODs are large domains within which gene have coherent gene expression. They are retrieved by comparing pairwise correlation of gene expression and merging all genes with similar expression within the same domain[32]. By definition, CODs are thus variable between cell types and best capture the expression changes.

1.2.4 Distal Regulation and Gene Orientation

The easiest way to imagine gene regulation is to assume regulatory elements influencing transcription are located in close proximity to gene promoters and transcription start sites.

However, it is rarely the case, and most enhancers seem to be distal[51,54]. That observation strengthens the idea that structural elements are needed to guide the interaction between elements that are far from each other, in linear distance along the genome. Close proximity may have an influence, but it would be wrong to assume that distal elements never interact and cannot influence each other. Moreover, even at a closer range, structural elements seem to play an important part in transcription regulation as the relative orientation of genes seems to influence co-expression. Indeed, some gene that are positioned in a divergent conformation ("back-to-back") in yeast share a promoter, resulting to co-expression[55]. Occurrences of those bidirectional promoters have been found in human[56,57], suggesting that gene orientation plays a major part in transcription regulation and gene co-expression.

1.3 Methods for Exploring the Nucleus

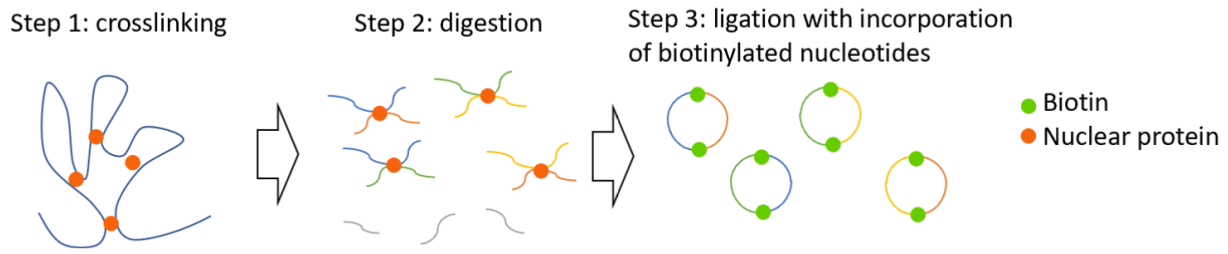
The relation between the structure of the genome and transcription requires the integration of various data types, and several tools to analyze them. For example, information about contacts between chromatin regions (TADs), the variations in gene expression and the position of key nuclear proteins can be retrieved through using Hi-C, RNA-seq and Chromatin Immuno-Precipitation followed by sequencing (ChIP-seq), respectively.

1.3.1 Hi-C: a Chromatin Conformation Capture Method

Hi-C is a genome wide conformation capture derived from 3C[58] (General figure 2). First, the cells are crosslinked, such that contacts between DNA regions, with or without the help of a nuclear protein, are stabilized[4,58,59]. DNA is then digested using a restriction enzyme and the fragments are ligated[4,58]. After ligation, the crosslinking is reversed. The goal is to obtain circular DNA fragments containing the two regions that were in contact in the nucleus. In the Hi-C protocol, biotinylated nucleotides are added to the junction of the fragments, such that DNA participating in contacts can be extracted from all fragments. The extracted fragments are then directly sequenced. During the following alignment, reads should align to two regions of the genome, which means those two regions were in close contact in the initial nucleus.

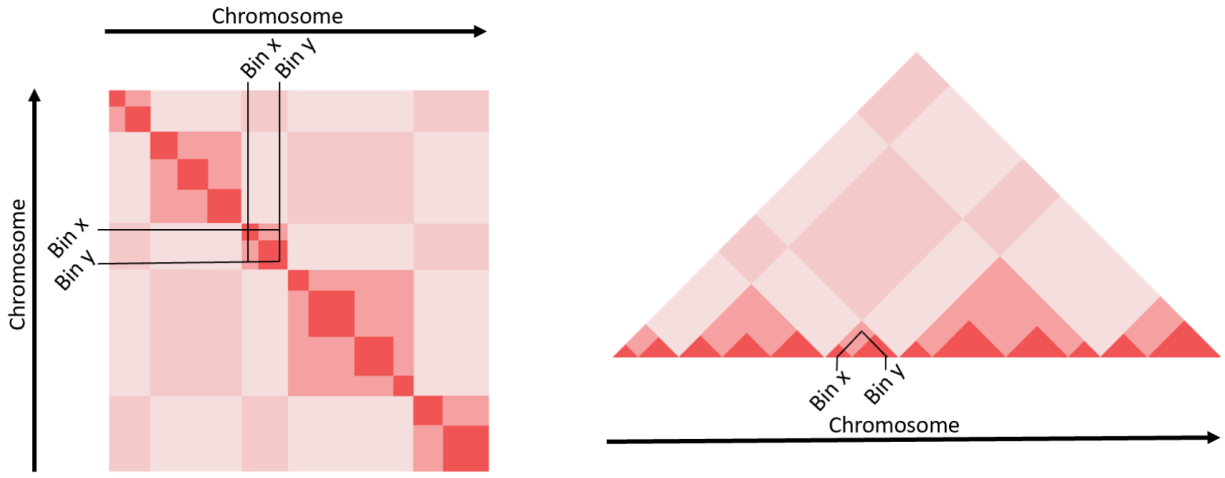
However, as the number of chromatin contacts in the genome is hard to determine, and a minimum number of observations must be made to distinguish true contacts from noise, there

needs to be a compromise made between scaling and depth. All interactions within bins of fixed length are summed, and the length of the bins corresponds to the resolution of the Hi-C experiment, such that a high number corresponds to a lower resolution[4]. Therefore, high-resolution experiments (10kb-long bins or less) need a very high sequencing depth, but permit to find structures at a fine scale, while low-resolution experiments are easier to produce but only retrieve large structures.



General figure 2: General workflow of a Hi-C experiment. *DNA and nuclear proteins are crosslinked, then DNA gets digested using a restriction enzyme. After digestion, the fragments are ligated and biotin is introduced so that fragments resulting from a ligation event can be sub-selected from all DNA fragments.*

Hi-C experiments are usually visualized trough red-tinted heatmaps (General figure 3). The x and y axes of the heatmap correspond to the bins defined by the resolution. At their intersection, the intensity of the color corresponds to the strength of the interaction: blank if there is no recorded interaction between the regions, red if there is a strong interaction. The number of interactions that need to be observed to be qualified as "strong" may vary from one experiment to another and according to the resolution. Hi-C heatmaps are symmetrical, as the interactions are not directed (the number of interactions from region 1 to region 2 is the same as the number of interactions from region 2 to region 1). Therefore, Hi-C heatmaps are often cut along the diagonal and represented as triangles, sitting on their hypotenuse. The bins are labeled along the hypothenuse and to find the regions corresponding to a red dot, one must trace a line from the dot to the hypothenuse, parallel to the first edge of the triangle (first interacting region) and a second line parallel to the second edge (second interacting region). Due to their interaction patterns, compartments create a checkerboard-like pattern on Hi-C heatmaps[6], TADs can be seen as triangles and chromatin loops are represented by a single red dot (General figure 4).

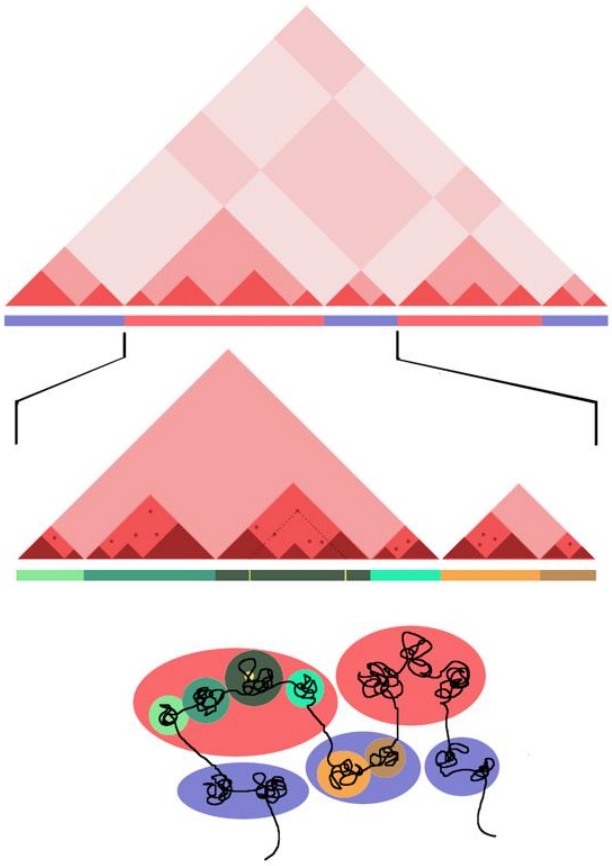


General figure 3: Schema of a Hi-C heatmap. *X and Y axis represent bins along the genome and the strength of the red color represents the frequency of contacts between regions.*

1.3.2 Capturing Differential Gene Expression using RNA-seq

Transcription is one of the most important process as it is the first step leading to the production of proteins. Gene transcription must be tightly regulated, such that key proteins are produced in sufficient amounts when they are needed. In addition to producing necessary proteins, the production of non-pertinent proteins must be repressed to avoid using resources present in very limited amounts in the cell. To explore how the architecture of the genome interplays with transcription, it is necessary to be able to capture precisely changes in gene expression.

Changes in transcription can be captured using RNA-seq. The two main tools to find and quantify gene expression changes are edgeR and DESeq2[60]. These methods use different models to normalize read counts per gene and compare them across conditions. In order to be the most conservative possible and label differentially expressed genes with a low rate of false positives, the consensus of both methods was considered.



General figure 4: Structures found by the Hi-C heatmap and their biological correspondence. *Top panel: The largest structures correspond to compartments (blue and red). They show a higher frequency of contacts between compartments of the same type. Middle panel: At a smaller scale, we can identify TADs (green and orange). They show a high frequency of intra-TAD interactions, but less inter-TAD interactions. Each red dot corresponds to an individual chromatin loop (one of them is highlighted in yellow). Bottom panel: The structures identified with Hi-C suggest that DNA forms loops that aggregate into TADs, and TADs interact to form compartments.*

1.3.3 Detecting Protein Binding with ChIP-seq

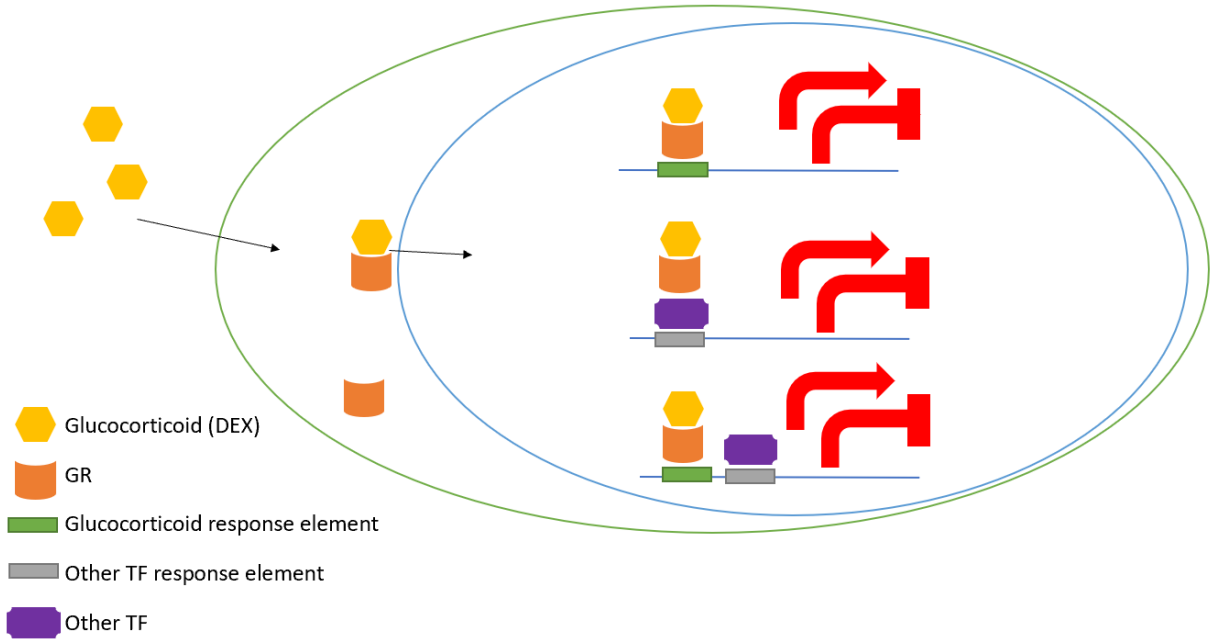
The regions to which nuclear protein bind can carry a lot of information about nuclear processes. The best way to find the protein binding sites along the genome is Chromatin Immuno-Precipitation followed by sequencing (ChIP-seq). As for Hi-C the first steps consist in crosslinking followed by digestion or sonication. The DNA strands linked to nuclear proteins are then immunoprecipitated, using an antibody targeting the protein of interest. This permits to retrieve all regions interacting with a specific protein. The extracted part is then cleaned to remove the protein

and the resulting DNA fragments are sequenced. After sequencing, the reads are mapped to a genome of reference, which creates peaks: regions with many mapped reads that represent the binding site of the protein of interest.

ChIP-seq is known to be noisy and thus trying to find differential binding sites between conditions can be challenging[61]. For example, DiffBind is a package that creates pseudo RNA-seq data, with read counts associated to regions resulting from the peak calling. It itself can use edgeR or DESeq2 to perform the differential binding analysis, but the results vary depending on the method use, even with the exact same data[61]. However, as DiffBind needs to pre-define regions derived from the called peaks to create its pseudo-counts to compare various conditions, it is useful to obtain a consensus list of called peaks across conditions. In this study, DiffBind has thus been used as a tool to create a consensus peak set and not to analyze the differential binding across timepoints.

1.4 Hormonal Induction

Stimulating cells with a hormone is a good way to obtain quick changes in gene transcription, while theoretically not affecting chromatin structure in short-term. In this study, we explore the effects on architecture and transcription of cells after hormonal induction. We use data coming from A549 cells, a lung cancer cell line, induced with dexamethasone (DEX) as an example. Dexamethasone is a synthetic hormone belonging to the glucocorticoids family[62]. Glucocorticoids are cholesterol-derived, which permits them to diffuse directly through the cellular membrane. Once in the cell, the glucocorticoid receptor (GR), a transcription factor, usually mediates the effects of glucocorticoids[62–64]. In the nucleus, GR can affect gene transcription using various mechanisms and activate or repress genes through direct binding to glucocorticoid-response elements or by interfering with other TFs (General figure 5).



General figure 5: Illustration of the mechanisms by which glucocorticoids influence transcription. *Glucocorticoids such as dexamethasone enter the cytoplasm, where they bind glucocorticoid receptors (GR) and activate them. The complex can then enter the nucleus and activate or repress transcription of genes through direct binding, tethered sites or interaction with other transcription factors. Inspired by: Lin, K.-T. & Wang, L.-H. New dimension of glucocorticoids in cancer treatment. Steroids 111, 84–88 (2016).*

1.4.1 Effects on Transcription

After entering the nucleus, GR binds to several thousands of regions, and affects a few thousands of genes[54,62,63], with a similar proportion of downregulation and upregulation[54]. GR tends to mainly bind distal enhancers[54], but not all DEX-responsive loci are direct binding sites[63]. Some GR-mediated differential expression is due to secondary GR binding and require long-range interactions with a direct binding site. Those two mechanisms of action, direct or tethered binding, could make the difference between rapid response and slower response in differential expression[54,62]. Tethered sites have been seen to cluster around direct binding sites, and their interaction could be mediated through chromatin looping[33,63]. In addition, genes within CTCF-bound regions have coordinated dynamics[63], which reinforces the idea that secondary targets of GR are defined by chromatin architecture. Glucocorticoids response varies from cell type to cell

type[54] and tethered sites, due to their dependence on interactions, have been suggested to be the main cause in differential expression seen due to DEX in different cell types[63].

1.4.2 Effects on Architecture

Chromatin architecture has been shown to be important in mediating transcriptional response after DEX induction as tethered sites interact with direct binding sites through long-range chromatin contacts. DEX-induced genes were reported to have increased 3D contacts after stimulus, More specifically, activated genes interact more with active compartments and repressed genes with the inactive ones[65]. However, DEX does not seem to create *de-novo* chromatin interactions, but only to change the contact frequencies of pre-existing interactions[63,65]. TAD boundaries are not affected by DEX[65]. Taken together, those observations suggest that while individual loops may form more frequently when they involve DEX-responsive genes, DEX induction does not affect the overall chromatin architecture in short term, meaning the *a priori* structure may influence changes in expression. Contradictorily, compartments switches have been reported in MCF7 cells induced with estrogen (E2)[66]. E2 and DEX are both steroid-derived hormones and are thus expected to affect cells in a similar manner. It might thus be that the absence of clear changes in TADs observed in A549 cells is due to technical limitations and methods more precise than Hi-C would reveal an effect of DEX induction on TAD boundaries location.

1.5 Objectives and hypotheses

During the past years, great advances have been made regarding the nuclear architecture, gene regulation, and the interplay between the two. New technologies such as Hi-C permitted a more precise characterization of the 3D contacts between chromatin regions while the RNA-seq tools continue to gain accuracy. There remain, nevertheless, many unanswered questions such as:

1) What is the functional role of chromatin organization, more specifically of TADs or sub-TAD structures such as CODs[11,16]?
2) How does the “transcriptional ecosystem” and the proteins it contains influence response to stimuli[28,67]?
3) Are there different types of boundaries, demarcating different domains with different roles at the sub-TAD level[16]?

The purpose of our work is to structurally and functionally characterize architectural boundaries, between TADs and inside them, and find how their composition affect their function. The hypothesis is that there exist sub-TAD boundaries, different in structure and in function from TAD boundaries. TAD boundaries' role is to limit the spread of regulatory elements and to insulate regions. This has been suggested before, but we argue that the insulation at TAD boundary consists in their main, indispensable function, and not only a property. The second type of boundaries, found inside TADs, mark the delimitation of CODs but are less strict than TAD boundaries to account for different needs in gene expression.

To achieve this goal, we collected and analyzed data coming from A549 cells induced with glucocorticoids. Ideally, to reduce variation between the data types, they should all come from the same set of experiments. For that reason, we used the data coming from the A549 cell line, a lung cancer cell line, induced with dexamethasone (DEX) on ENCODE[1,2,54,65]. ENCODE harbors a large set of data produced on that cell line, and most of them have been produced consecutively by the same laboratory (Dr. Tim Reddy, Duke). The data was mainly processed with R, but Hi-C required some steps using Juicer[68]. The most computing-heavy tasks were performed on the servers of Compute Canada.

ENCODE contains Hi-C data performed on A549 cells that follow a time-course made of a control (0h) and 4 timepoints after DEX induction (1h, 4h, 8h, 12h), each with 4 biological replicates (ENCSR842RTB, ENCSR435JUA). The pre-processed files were used to perform a quality control on the data, after which the already-analyzed files containing the position of TADs were directly used to find the position of TAD boundaries.

It also contains a complete time-course of A549 cells following DEX induction (control, 30m, 1h, 2h, 3h, 4h, 5h, 6h, 7h, 8h, 10h, 12h) (ENCSR897XFT). Each timepoint has 3 or 4 biological replicates and can thus capture changes in expression at short and long term. The files containing the raw read counts were normalized using edgeR[69,70] to control the quality of the RNA-seq data and verify that there are indeed changes in gene expression that follow the time-course with a Principal Component Analysis (PCA) and a t-distributed Stochastic Neighbor Embedding (t-SNE) analysis.

Finally, ChIP-seq was available for 11 different nuclear proteins and 5 histone modifications: BCL3 (ENCSR022IHB), CEBPB (ENCSR625DZB), CTCF (ENCSR738NGQ),

EP300 (ENCSR738NGQ), FOSL2 (ENCSR447VJR) H3K4me1 (ENCSR180FFI), H3K4me2 (ENCSR868FCL), H3K4me3 (ENCSR342NKR), H3K9me3 (ENCSR476OXC), H3K27ac (ENCSR375BQN), HES2 (ENCSR790OOG), JUN (ENCSR588JLN), JUNB (ENCSR483SDK), NR3C1 (ENCSR210PYP), RAD21 (ENCSR501UJL) and SMC3 (ENCSR376GQA). They follow the same time-course that RNA-seq with 3 or 4 biological replicates. Files with the pre-aligned reads and the called peaks were used

The main objective of our study is thus to structurally and functionally characterize boundaries at the sub-TAD level. It was achieved by first characterize TAD boundaries to serve as a reference point, then characterizing new boundaries. We found that those were marked by the changing of strand on which genes are positioned and seem to be independent of structural proteins. The observations were confirmed by eQTL data.

Chapter 2: Manuscript

Genes on different strands mark boundaries associated with co-expression domains

Authors and affiliations:

Audrey Baguette
Department of Human Genetics, Faculty of Medicine, McGill University

Dr. Steve Bilodeau
Centre de recherche du CHU de Québec – Université Laval, Axe Oncologie, Québec, Québec, Canada, G1V 4G2.
Centre de Recherche sur le Cancer de l'Université Laval, Québec, Québec, Canada, G1R 3S3.
Centre de recherche en données massives de l'Université Laval, Québec, Québec, Canada, 1V 0A6
Département de biologie moléculaire, biochimie médicale et pathologie, Faculté de Médecine, Université Laval, Québec, Québec, Canada, G1V 0A6.

Dr. Guillaume Bourque
Department of Human Genetics, Faculty of Medicine, McGill University
Canadian Center for Computational Genomics, McGill University

Currently in the final stages of preparation for submission

Abstract

Gene regulation is influenced by chromatin folding, but the precise mechanisms guiding their interconnection remain unclear. Topologically associating domains (TADs) have been suggested to act as regulatiory units, yet they also have been proven to be distinct from co-expression domains (CODs), structures in which genes have correlated expression. What exactly are the roles of CODs and how do their boundaries and function differ from TADs? We use a combination of available RNA-seq, ChIP-seq and Hi-C data from A549 cells stimulated with the glucocorticoid dexamethasone to answer that question. We find that while TAD boundaries act as insulators and are significantly enriched between up and down-regulated genes (odds ratio of 1.85), they are not the only boundaries limiting co-expression. Indeed, we find that divergent and convergent pairs of genes create boundaries at the sub-TAD level. Moreover, when such gene pairs are not separated by a TAD boundary, we find that they are depleted for structural proteins, with odds ratios between 0.53 and 0.75. This suggests that COD boundaries could be demarcated by a switch of strand independently of structural proteins. Aligned with this idea we show, using eQTL data from lung cells, that genes affected by the same strong variant tend to be found on the same strand and to lack any barrier (TAD boundary or CTCF/Cohesin) between them (33% fall in that category, while 16% were expected). This enrichment was even stronger when a subset of pairs comprising those most affected by the eQTLs are considered. Based on these results, we propose a model in which same-strand genes form small sub-TAD domains that are the building blocks of CODs. Further exploration of this model could help better understand and anticipate changes in transcription in different cell types and conditions.

1. Introduction

The nucleus of each cell is highly compacted, full of molecules and an overcrowded environment. To fit the 2 meters-long human genome in such a small space, DNA requires several layers of organized structures. Genome-wide chromosome conformation capture (Hi-C) is a technique that elucidates chromatin contacts in the 3-dimensional nuclear space[1–3]. At the lowest level of resolution, compartments can be identified from Hi-C data. Compartments are divided in two types, A and B, defined by their interaction frequencies; compartments tend to interact more often with compartments of the same type[1,4]. Compartment A has been shown to contain more

genes and to be enriched for active genes and open chromatin marks, while compartment B is enriched for closed chromatin marks[1,5–7]. Compartments are not to be mixed with topologically associating domains (TADs), another structure found with high-resolution Hi-C. TADs are defined by their high concentration of contacts within their domain, relative to their low level of interaction across different TADs[8–11]. Intra-TAD contacts vary from cell to cell[3,12–14], but most TAD boundaries are conserved across single cells[3], cell types[7,9,11,15] and species[9,15,16]. In addition to TADs, that range from a few kilo bases to around 1 million bases long[8–11], smaller TAD-like structures, called sub-TADs have been discovered within TADs[4,12–14]. Finally, at the highest resolution, chromatin loops can be observed. They represent contacts between DNA regions, including promoters and enhancers[7,13,17–19], usually stabilized by CTCF or other structural proteins[15,16,20].

Among all those structures, parts of the chromatin must remain flexible to allow for transcription. Several studies have confirmed that TADs construct an environment favoring gene co-regulation. Genes within the same TAD tend to be co-expressed[10,16,17,21]. Indeed, some TADs act as regulatory units after hormonal induction[22] and hormone responsive genes are found within the same interaction networks[23] or in-between TAD boundaries[24]. Moreover, paralogs are usually co-regulated and found within the same TADs[25]. However, there is also evidence that some genes resist the intra-TAD co-regulation[26]. Co-regulation might thus be driven by sub-TAD structures such as cis-regulatory domains[27], insulation neighborhoods[28] or co-expression domains (CODs)[26]. Those seemingly contradictory observations show the shortcomings of our understanding of transcription with respect to chromatin architecture[8,13]. CODs are especially interesting to us as they are purely defined from a transcriptional point of view. Indeed, they are defined as regions of consecutive genes with correlated expression. Their definition does not include any structural aspect, besides the linear aspect of consecutive genes. It would thus be interesting to compare CODs to the well-known TADs and to better understand their similitudes and differences, in structure and in function.

Here we propose a detailed characterization of regulation boundaries at the TAD and sub-TAD level using a combination of available RNA-seq, ChIP-seq and Hi-C data from A549 cells induced with dexamethasone. We hypothesize that TADs and CODs have different types of boundaries. We looked at the properties of genes around these boundaries and found that the strand

position of genes had a significant and differential impact. We further tested these observations using eQTL in lung cells. Our results lead us to propose a model for human cells where genes on the same stand have a high probability of co-expression. COD boundaries, marked by the change of strand, reduce the co-expression probability independently of structural proteins.

2. Results

2.1 Gene expression and chromosome architecture changes associated with glucocorticoid stimulation

To understand how chromatin structure and gene expression are related, we used RNA-seq, ChIP-seq and Hi-C datasets coming from A549 cells induced with 100nM dexamethasone (DEX) were retrieved from ENCODE[29,30] (Methods). The literature shows that stimulating cells with a hormone should mainly affect expression levels while keeping the global nuclear architecture rather stable[23]. A t-SNE analysis of the expression data confirmed that the RNA-seq replicates have genes with differential expression after DEX induction and that the observed changes between timepoints follow the time-course (Figure 1A). This observation was further confirmed by a principal component analysis (Supplementary figure 1). Differentially expressed genes that were reported by both edgeR[31,32] and DESeq2[33] were identified at each timepoint, relative to the “0h” timepoint (Methods). A consensus set was then retrieved in order to find genes that had an upregulated or downregulated behavior across the time-course. Doing this, 1716 genes were labeled as “Up”, 1810 as “Down” and 10751 as “Stable” (Supplementary table 1).

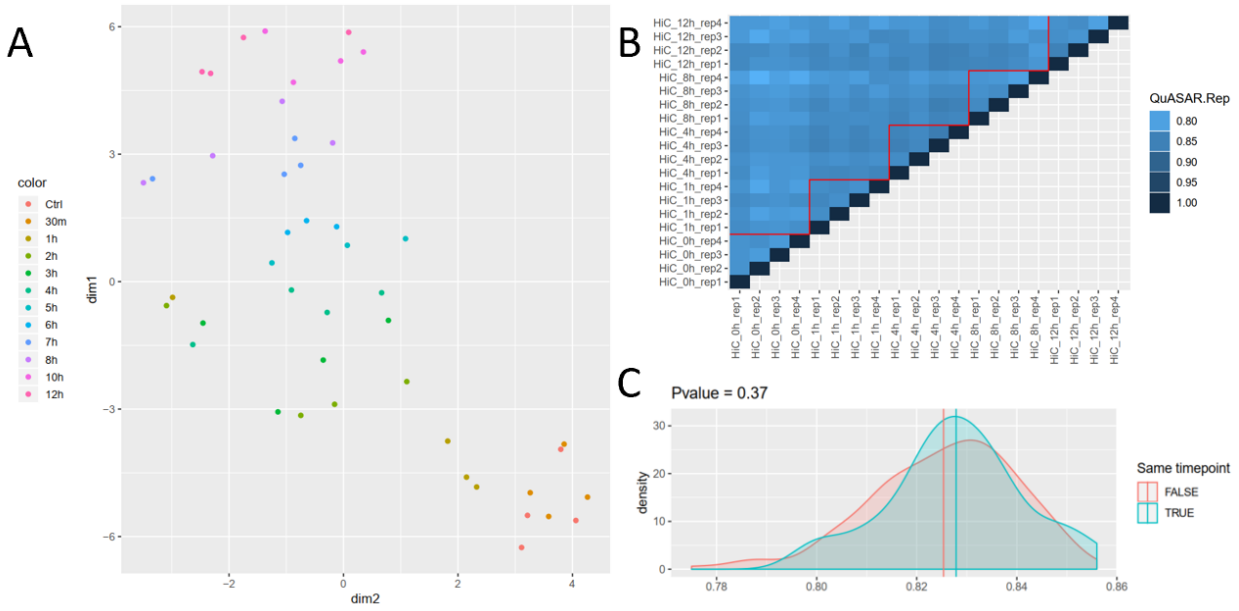


Figure 1: Gene expression changes in the RNA-seq samples and Hi-C reproducibility scores*. (A) t-SNE of the normalized RNA-seq data. (B) Heatmap of the replication scores given by QuASAR-Rep. Comparaisons inside the same timepoint are under the red lines and comparisons across timepoints are over it. (C) Distributions of the scores between replicates of the same timepoint (blue) or different timepoints (red). The vertical lines represent the means of the distributions. The p-value comes from a t-test.*

The quality of Hi-C heatmaps obtained from the same dataset[23] was assessed using QuASAR-QC[34]. QuASAR uses transformed matrices, based on read count matrices and enrichment matrices, corrected for distance, to produce quality scores for all chromosomes and replicates. The goal of the matrix transformation is to find regions showing deviation from the surrounding regions, as a high deviation is probably due to random ligation and is thus certainly noise. The quality scores for all chromosomes was found to vary between 0.015 and 0.02 (Supplementary figure 1B), which is characteristic of somewhat noisy data at this resolution of 10kb, but not uncommon[34,35]. That resolution was selected as it was the middle resolution the three resolutions used to call TADs (5kb, 10kb and 25kb)[23]. The reproducibility between pairs of replicates, from the same timepoints and across timepoints, was then compared using three methods: QuASAR-Rep[34], GenomeDISCO[36] and HiC-Spector[37]. The reproducibility scores were used to quantify the similarity of the Hi-C maps through the time-course. Indeed, Hi-C maps and the TADs derived from them were available for five timepoints and we wanted to see if the maps

are similar enough to only use a single timepoint as reference to locate TAD boundaries. The reproducibility scores varied from method to method, but all methods confirm that there is no more variability between replicates from different timepoints than between replicates from the same timepoint (Figure 1B, Supplementary figure 1C and E). This can be visually assessed with the heatmaps of the scores and is confirmed by a two-sample t-test comparing the distributions of the pairwise scores (Figure 1C, Supplementary figure 1D and F). None of the p-values for QuASAR-Rep (p-value = 0.37), GenomeDISCO (p-value = 0.36) and HiC-Spector (p-value = 0.22), respectively, was found to be significant. We cannot draw conclusions as to whether the TADs are completely stable after DEX induction using only those scores. However, as the scores report no major change in chromatin architecture, they justify the use of a single timepoint as reference for TAD boundaries to facilitate the analyses below.

2.2 Closer genes tend to be co-expressed but TAD boundaries act as barriers

TAD boundaries are known to have insulator properties, limiting the effects of enhancers to nearby genes on the other side of the boundary[9,12,14,38]. If true, the prevalence of TAD boundaries between genes having opposite behaviors following DEX induction should be higher. We retrieved the list of TADs, called with Juicer's Arrowhead[39], at the earliest time-point (0h)[23,29,30]. We then transformed the TADs list in a list of 11454 TAD boundaries (Methods, Supplementary figure 2). To test the insulation effect of TAD boundaries in regard to other structural elements, the location of several nuclear proteins was compared to the location of TAD boundaries. Indeed, the barrier properties of some TAD boundaries are attributed to CTCF and Cohesin. By comparing the binding sites of CTCF, RAD21 and SMC3 (two subunits of Cohesin) to the location of TAD boundaries, we want to verify that TAD boundaries have a unique property that is not attributable to a specific structural protein only. The other nuclear proteins were used as controls. For example, H3K4me1 marks open chromatin and should be found around activated genes. On the other hand, NR3C1 should be found near differentially expressed genes. The available raw reads and pre-called peaks of the ChIP-seq data corresponding to the dataset[24] were further processed through DiffBind[40,41] in order to find the binding sites of the 16 proteins along the time-course (Methods). This resulted in, for example, 52729 CTCF peaks, 14699 NR3C1 peaks, 72433 RAD21 peaks and 71220 SMC3 peaks (Supplementary figure 2).

The previously defined gene lists ("Up", "Down" and "Stable") were paired with each other, which resulted in 147248 pairs of genes separated by less than 1 Mb, from transcription start site (TSS) to TSS (Methods). Since the further two genes are from each other, the higher the chances are that they contain a TAD boundary or a protein binding site between them, we explored the effect of linear distance on the relative comportment of gene pairs (based on the previously defined categories of "Up", "Down" and "Stable"). This showed that genes with the same behavior tend to be closer than genes behaving differently after DEX induction (Figure 2A). Indeed, pairs separated by less than 100kb make up 13.72% of all pairs, but as much as 19.08% of pairs showed concurrent upregulation compared to only 6.8% of pairs where one gene is activated and the other is repressed. The threshold of 100 kb was chosen as it is a distance long enough to account for long-range interaction, but we assume that over that distance, genes are too far away from each other to effectively be co-expressed because they share transcription machinery[24]. This is seen at a finer scale too, as even when only the pairs separated by less than 100 kb are considered, the distributions of distances changes depending on the relative behaviors of the genes within the pair (Figure 2B). The mode of the distribution density for activated pairs is at 14651 bp, while it is at 73207 bp for pairs were genes go in opposite directions. Finally, the enrichment of finding a TAD boundary between two genes was computed for each type of pair, using the pairs where both genes are "Stable" as reference.

To account for the identified distance bias, the reference pairs were sub-sampled 1000 times, such that the distribution of distances of the sub-sampling matches the distribution of distances of the pairs of interest and the query and reference contain the same number of pairs (Supplementary figure 3). Relative to stable gene pairs, TAD boundaries were found to be significantly enriched between pairs of genes having a different behavior, especially pairs where one gene is activated and the other is repressed (odds ratio of 1.85, empirical p-value < 0.001) (Figure 2C, Supplementary figure 4). There is also a small depletion of TAD boundaries between upregulated genes (odds ratio of 0.89). The enrichment between genes with opposite behaviors is weaker for CTCF (odds ratio of 1.51, p-value < 0.05). In addition, pairs of upregulated genes and pairs of downregulated genes do not show any enrichment for TAD boundaries but only small, significant enrichments for CTCF (odds ratios of 1.21 in both cases), confirming CTCF alone is not enough to create the insulation property of TAD boundaries. SMC3 and RAD21 are enriched around activated genes (odds ratio of 1.69 and 1.54, respectively, empirical p-value < 0.001),

which is coherent with their function; those are two subunits of Cohesin, a ring-shaped protein complex that helps bringing promoters and enhancers in close proximity[15,16,20]. NR3C1 is directly activated by DEX and is thus expected to be located around differentially expressed genes, especially activated genes, and this is confirmed by its enrichment pattern across the gene pairs. NR3C1 is also very significantly enriched between pairs of downregulated genes, confirming GR acts as a repressor and not only as an activator. As EP300 binds to enhancers, it is found around expressed genes. Indeed, we see an enrichment of EP300 between genes that were highly expressed at the start of the time-course (downregulated genes) or at the end (upregulated genes).

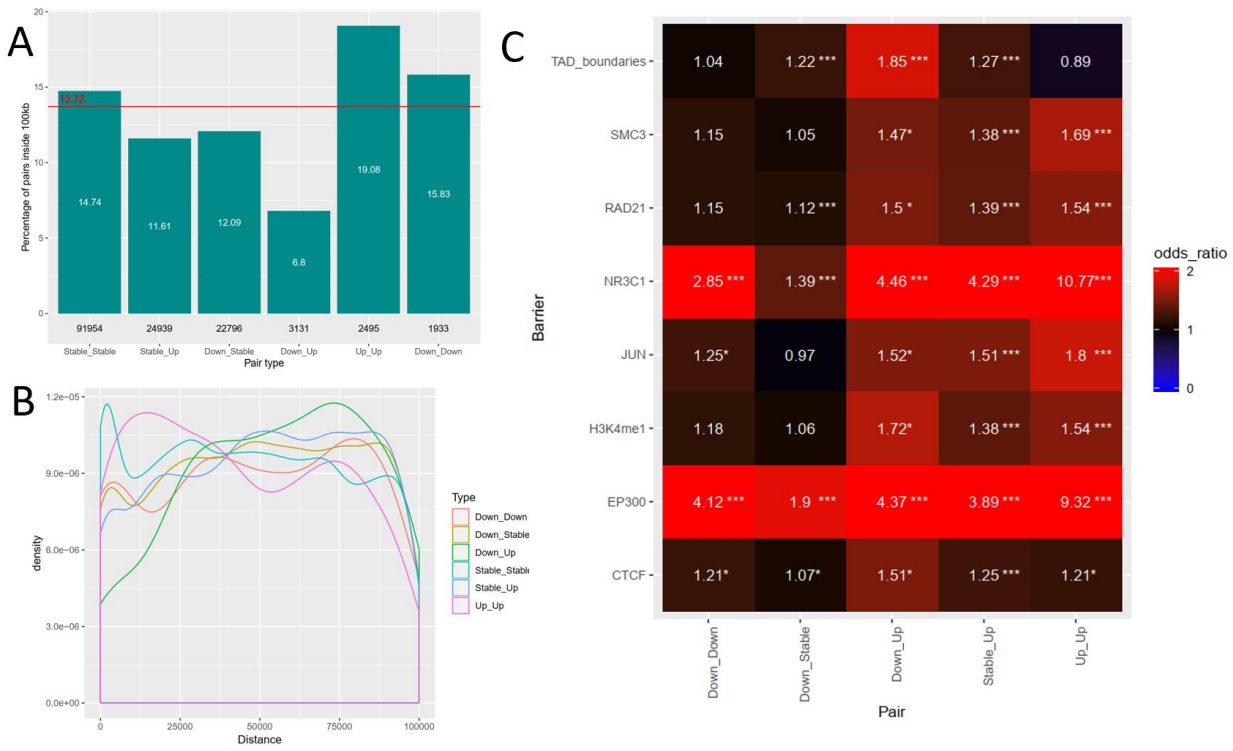


Figure 2: Genes with the same behavior tend to be closer from each other and pairs of genes going in opposite directions are more often separated by TAD boundaries than pairs of stable genes. *(A) Proportion of pairs in which genes are separated by less than 100kb among all pairs in which genes are separated by less than 1Mb. The red line represents the proportion when all pairs are considered. (B) Distribution of distance (in bp) between the genes of the pairs, for pairs separated by less than 100kb. (C) Heatmap of odds ratios for the presence of a physical barrier between the genes of the pairs. The "Stable_Stable" pairs are used as reference. *P-value < 0.05; ***P-value < 0.001. The p-values are empirical, computed with 1000 resampling events*

2.3 Gene conformation marks small sub-TAD boundaries

Another structural element that can be explored in regard to co-expression is gene "strandedness". Indeed, genes can be found on the same strand, but also on opposite strands and facing each other (convergent) or "back-to-back" (divergent). How are same-strand, convergent and divergent gene pairs represented around boundaries? Does conformation have an impact on boundary strength? To answer these questions, four new categories of gene pairs were created, depending on whether they were (1) around a TAD boundary, (2) inside the same TAD but had different behaviors (potentially associated with a COD boundary), (3) inside the same TAD but with the same behavior (thus probably within the same COD), or (4) outside TADs (Methods). To characterize the relation of genes at the boundaries, only the pairs of genes that were the closest to the boundaries were considered. Analyzing the distribution of pairs in different conformations according to those categories showed that at TAD boundaries, gene pairs are usually separated by both Cohesin (SMC3 and RAD21) and CTCF, regardless of the relative position of genes (Figure 3A). However, while inside TADs, divergent and convergent pairs of genes are usually less often separated by CTCF and Cohesin than same-strand pairs. Indeed, among pairs of genes found within the same TAD that show a different behavior ("Down-Stable", "Down-Up" and "Stable-Up"), 64.6% of same-strand pairs have CTCF and Cohesin between them, while it is only true for 50.5% of the convergent and 42.2% of divergent pairs. If we consider pairs of genes found in the same TAD that also have the same behavior ("Down-Down", "Stable-Stable" and "Up-Up"), we find that 50.8% of the same-strand pairs have both structural proteins, against as few as 42.1% for convergent pairs and 31.1% for divergent pairs. Using the "same strand" pairs as reference, while accounting for any possible distance bias (Methods), divergent and convergent genes were found to be significantly depleted of CTCF and Cohesin everywhere (p-value < 0.001) but at TAD boundaries as compared to same strand pairs with a same distribution of distances between the genes (Figure 3B). The odds ratios for finding CTCF and Cohesin between divergent genes are of 0.52 when genes are in the same TAD but have a different behavior, of 0.64 when genes are in the same TAD and show the same behavior and of 0.58 when genes are outside TADs (p-values < 0.001). For convergent genes, the odds ratios are of 0.61, 0.75 and 0.56 (p-values < 0.001), in the same order. Assuming pairs of genes with a different behavior mark COD boundaries, the depletion of CTCF and Cohesin between divergent and convergent pairs suggests that those pairs could mark sub-TAD domains without the need of Cohesin or CTCF.

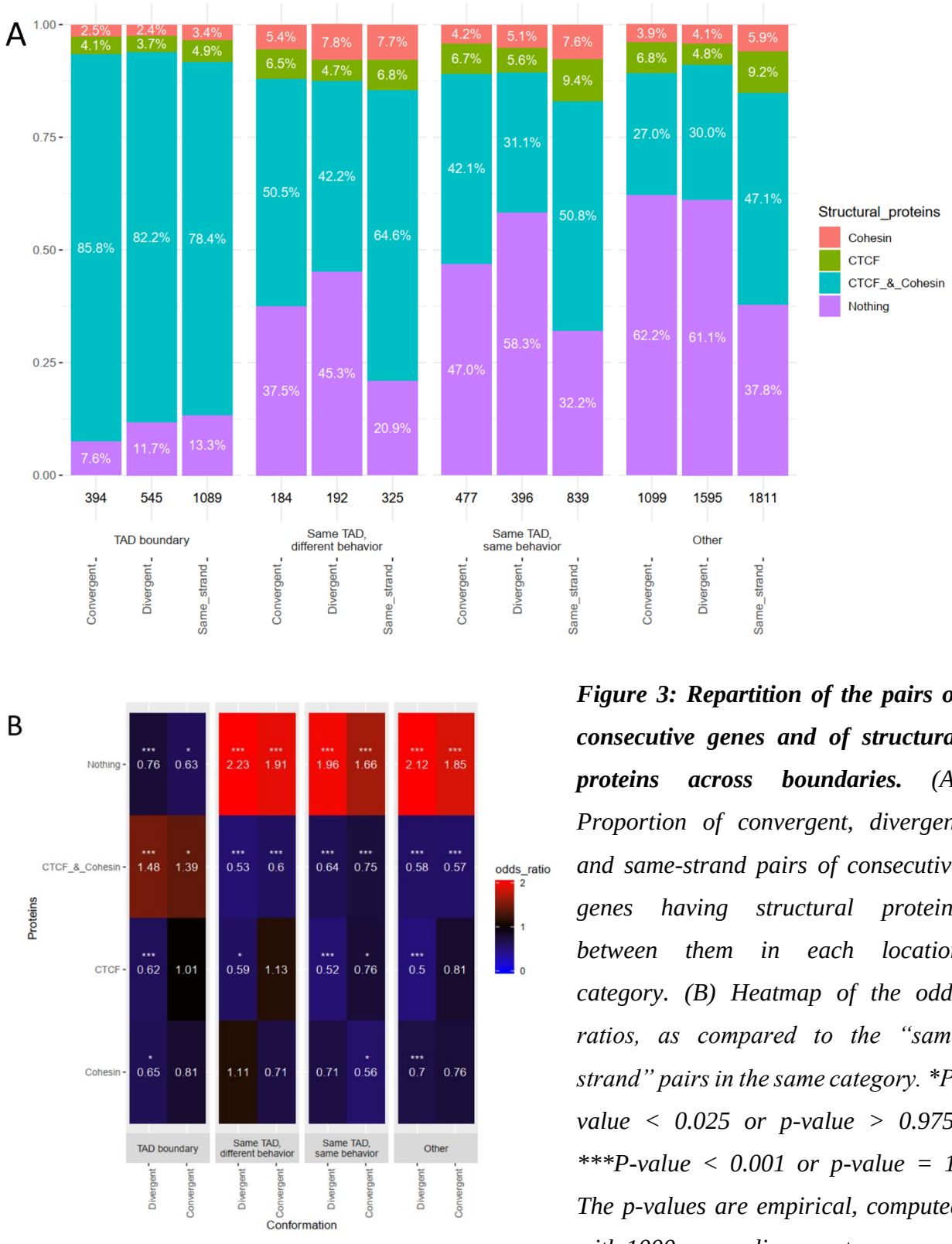


Figure 3: Repartition of the pairs of consecutive genes and of structural proteins across boundaries. *(A) Proportion of convergent, divergent and same-strand pairs of consecutive genes having structural proteins between them in each location category. (B) Heatmap of the odds ratios, as compared to the "same strand" pairs in the same category. *P-value < 0.025 or p-value > 0.975; ***P-value < 0.001 or p-value = 1. The p-values are empirical, computed with 1000 resampling events.*

2.4 eQTL gene targets are preferentially on the same strand

To look for supportive evidence that the change of strand influences co-expression probability, expression quantitative traits loci (eQTL) found in lung cells and their target genes were retrieved from GTEx Portal[42]. The A549 cell line being a lung cancer cell line, we chose to use data coming from lung cells as it is the closest cell type available on GTEx. We hypothesized that genes affected by the same variant would be more likely to be found on the same strand and less likely to be separated by a barrier than random pairs of genes, in accordance with our model. We thus compared 9088 pairs of genes affected by the same eQTL to 24035 control pairs using the same resampling technique to limit the distance bias as before (Methods). We found that pairs of genes affected by the same eQTL show a preference for being on the same strand, without barrier (Figure 4A). Indeed, 19% of pairs fall under that category, while the expected proportion is of around 14% for pairs of genes not affected by the same eQTL (odds ratio of 1.44, p-value < 0.001). The difference is even more marked when pairs of genes affected by the strongest variants are considered (Methods). The proportion of pairs affected by the same genes that are found on the same strand without barrier between them goes as high as 33%, resulting in an odds ratio of 2.68(p-value < 0.001, Figure 4B). Moreover, the pairs show an aversion for being on opposite strands and separated by a barrier than control pairs. 58% of all pairs of interest fall under that category, while the expected value is around 66% (odds ratio of 0.73, p-value < 0.001). Once again, considering the most affected genes makes the contrast even stronger, as the proportions drops to 39% and the odds ratio to 0.4 (p-value < 0.0001). Specific examples of those observations include the eQTLs chr1_109678559_T_A_b38 and chr3_195614883_T_A_b38 (Figure 4C-D). When we consider all genes having their TSS within 100kb of those variants, they seem to affect preferentially genes located on the positive strand and their action seems blocked by barriers, especially by TAD boundaries. These analyses support the observation that the change of strand can serve as a boundary limiting co-expression.

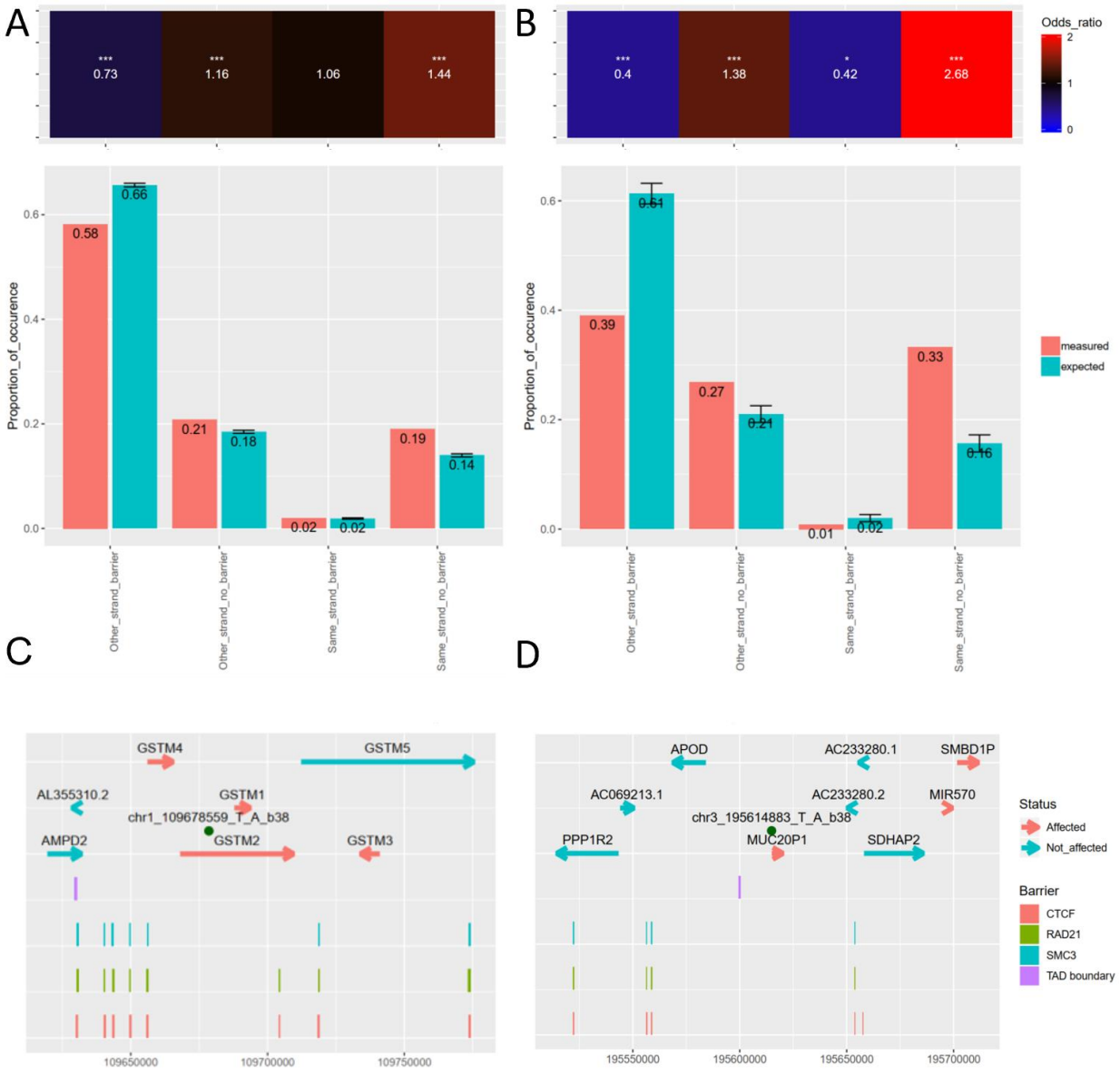


Figure 4: Analyzing gene conformation with eQTLs. *Pairs of genes where both genes are affected by the same eQTL are enriched for being on the same strand without barrier (CTCF and Cohesin or a TAD boundary) and depleted for being on different strands and separated by a barrier, as compared to control pairs. This is true when considering (A) all pairs of genes separated by 100kb affected by the same eQTL or (B) a subset with only the pairs with the strongest association (regression slope > 0.7 or < -0.7) to their eQTL. *P-value < 0.025 or p-value > 0.975; ***P-value < 0.001 or p-value = 1. The p-values are empirical, computed with 1000 resampling events. (C-D) Specific examples of variants (green dots) and the genes they affect depending on their orientation and the position of barriers.*

2.5 A new, probabilistic model for gene co-expression

The previous observations suggested a model for human cells where consecutive same-strand genes make up small sub-TAD domains within which genes are likely to be co-expressed (Figure 5A). The boundaries of those domains are marked by the change of strand, as genes on different strands are less likely to be co-expressed. The cell can introduce CTCF and Cohesin that tend to create insulation boundaries such as TAD boundaries that disrupt the possibility of co-expression (Figure 5B). Those small sub-TAD domains are the building blocks of co-regulation domains. CODs are, by definition, statistical entities and not physical entities, as they are found by aggregating all genes that have correlated expression. We propose that sub-TAD boundaries which are protein-independent mark the preferential split point of CODs. In a specific condition, genes on either side of the boundary can have similar expression level and make up a single COD. After a stimulus, the expression levels of the two regions might not be the same anymore and they would split into different CODs according to the position of the boundary (Figure 5C). The proposed model is not mutually exclusive with previous ones, such as transcription factories, and could help to understand how genes that do not seemingly receive the same signal could have similar expression levels.

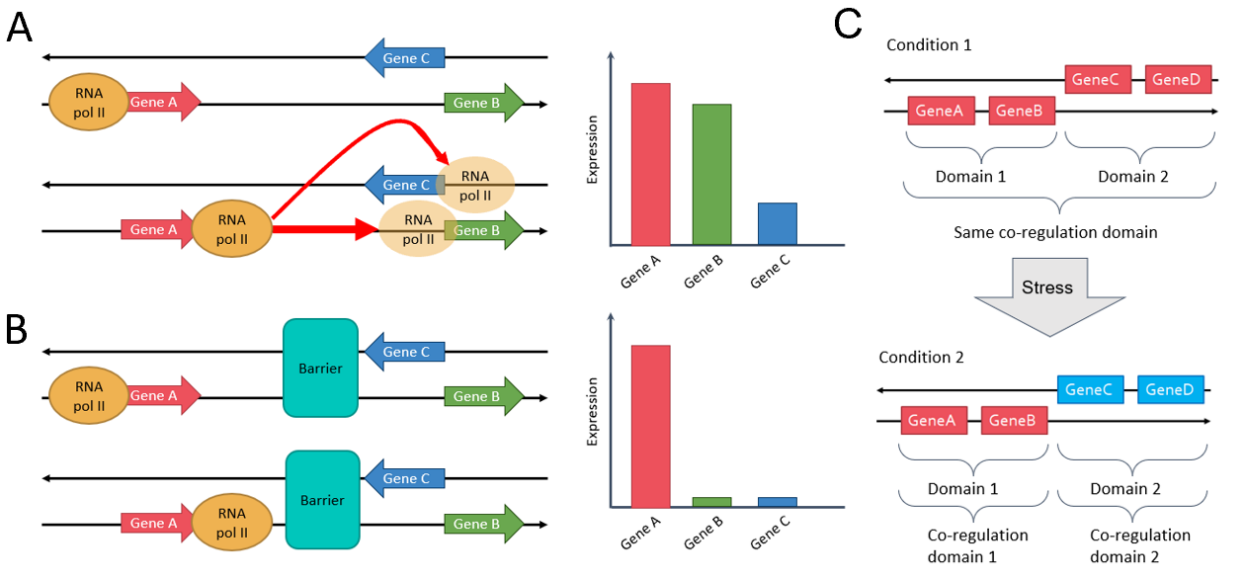


Figure 5: Model of co-expression likeliness. *(A) Same-strand genes are very likely to be co-expressed, as the RNA pol II just needs to continue its path along the strand. Divergent and convergent genes are less likely to be co-expressed, as the RNA pol II would needs to detach and reattach itself to go from one gene to the other. (B) When genes are separated by a barrier (CTCF and Cohesin or TAD boundary), there is complete disruption of co-expression. (C) Illustration of how domains - physical entities - can make up a single co-regulation domain - a statistical entity - or split up in two, depending on the condition.*

3. Discussion

The relation between TADs and transcription is a subject that many studies tried to understand[10,22,26,27]. The boundaries of co-expression domains were not fully characterized and TADs were lacking a functional definition[13]. Using publicly available data of A549 cells induced with dexamethasone[23,24,29,30], we found that the strand position of genes influences the probability of their co-expression and that TAD boundaries disrupt co-expression.

In this study, we compared the insulation property of TAD and COD boundaries. Genes having an opposite behavior after DEX induction are enriched to have a TAD boundary between them, with a significant odds ratio of 1.85 as compared to pairs of stable genes. The odds ratio is less significant and of 1.51 for having CTCF between them. We also show that the relative position

of genes seems to create structural protein-independent boundaries inside TADs. At TAD boundaries, same-strand, convergent and divergent gene pairs are usually separated by CTCF and Cohesin. However, when pairs are inside the same TAD, more than half of the same-strand pairs still contain CTCF and Cohesin, but the convergent and divergent pairs are not separated by those proteins as often. Finally, we further supported the observations using eQTL data and found that genes affected by the same eQTL are enriched for being found on the same strand, without barrier, than genes not affected by the same variant (odds ratio of 1.44, p-value < 0.001). Moreover, genes affected by the same eQTL are depleted for being on different strands and separated by a barrier (odds ratios 0.73, p-value < 0.001). When the genes most affected by eQTLs are considered, the tendencies are even more marked. Taken together, all those results suggest that convergent and divergent pairs mark the boundaries between co-expression domains, at the sub-TAD level, without the need of CTCF and Cohesin. This led us to propose a model for co-expression probabilities based on the relative conformation of genes and the presence of barriers. Same-strand genes are likely to be co-expressed by chance, divergent and convergent genes less so, and barriers such as CTCF and Cohesin or TAD boundaries greatly reduce the chances of co-expression.

The proposed model has only been observed in A549 cells and is thus limited to that cell line for now. It would be interesting to test it further using other conditions, in other cell lines or even different organisms, to see if it can be applied to all eukaryotic cells. In addition, eQTL data comes from normal lung cells of over 850 individuals, while A549 cells come from a lung cancer cell line. Cancer cells usually contain structural variants which could impact the genomic architecture. The position of TAD boundaries, CTCF, RAD21 and SMC3 used during validation with eQTL comes from A549 cells. Those elements have been reported to be greatly conserved, but there still could remain differences that could influence our validation step. Ideally, for future validation of the model, Hi-C, ChIP-seq and eQTL data should come from the same cells. Using different cell types would also help COD boundaries detection. We used consecutive genes with a different behavior as proxy to COD boundaries. However, two genes could be found to be upregulated after DEX induction, but it does not mean they have completely correlated expression. They might thus be found to be into different CODs while our method would not identify it as boundary. We may thus potentially be missing COD boundaries. CODs also change depending on the cell needs and therefore, to have a complete mapping of CODs and their relative splitting or merging from one condition to another would be useful. Still, the fact that we can support our

model with significant values knowing our definition of COD boundary is more stringent reinforces our confidence in the model.

An important control for the analyses presented in this study was to check for properties that could be different between the tested pairs of genes and the control sets that would introduce a bias. Looking at the distribution of distances between pairs of genes depending on their relative behavior showed that there is a strong distance bias. This was accounted for but still consists the main limitation of the study, as it is reduced but cannot be completely removed. A second technical limitation is that the effect of TAD boundaries on co-expression has been assessed using TAD boundaries derived from the TADs called before DEX induction (time 0). We used the reproducibility scores of the Hi-C maps to justify that selection. However, all possible changes in structure are then contained within a single number. Since the regions covered by TAD boundaries add up to a small portion of the genome, it is possible that small changes in TAD boundaries location would be lost in the general score given for the whole genome. It might thus be better to account for those small changes in future explorations of the model. Lastly, we assumed all genes have their own promoter. However, there are a few reported case of divergent genes sharing a promoter, that thus have coordinated expression[43–45]. Those remain exceptional cases, but it would be interesting to treat genes sharing a promoter separately from all others when further exploring the model, as they constitute another mechanism for controlling co-regulation.

Despite the limitations of our method, the model for human cells we propose is supported by different types of analyses, using different data types, which leads us to believe in its robustness. Exploring it further with improved bias correction and supplementary data is more likely to improve the results than disprove them. Our model is also compatible with other suggested models. One rising paradigm change in how active transcription is seen suggests that RNA polymerase II is not the moving part during transcription, but rather that it is fixed in the nuclear space[46,47]. Multiple RNA polymerase II would cluster into transcription factories[46,48–50] and DNA would cluster to factories, or detached from them, depending on the cellular needs[51]. Our model has been explained such that RNA polymerase II is the mobile part, for ease of understanding, but it is not mutually exclusive with the idea of fixed polymerase. If the transcription machinery is indeed fixed, the model stays the same and can be explained as such: when two genes are located on the same strand, they are very likely to be co-expressed as the DNA could not detach itself from the

transcription factory. When genes are divergent or convergent, they are slightly less likely to be co-expressed as the DNA would have to detach then re-attach itself to the factory, but slightly shifted to attain the TSS on the other strand. CTCF and TAD boundaries greatly decrease the probability of co-expression as they serve as barrier, preventing the DNA to slide freely. The model we propose is thus compatible with the other models of transcription factories and might even help understanding how transcription is regulated in those hubs of active transcription. As such, it would serve as a basis for more complex gene regulation mechanisms and could well be the key to understand unresolved biological questions: What is the regulatory function of the genomic architecture[8]? What is the functional definition of TADs and sub-TAD domains[13]? How would cells lacking structural proteins behave following a stimulus demanding a change in the expression program[52]? We suppose that the regulatory function of the chromatin organization differs at depending on the level considered; gene "strandedness" affects the probability of co-regulation while TAD function is to disrupt the spread of expression signals through its boundaries. The functional definition is TADs would thus be centered on the insulation properties of the boundaries, while the sub-TAD domains definition would rather relate to gene position. Cells that do not have the necessary structural proteins would have expression patterns reflecting the relative position of the genes, with close, same-strand genes behaving similarly. The model thus serves as a stable ground on which complex hypotheses can be constructed and tested in the near future.

4. Conclusion

In this paper, we describe the existence of intra-TAD boundaries, delimited by the changing of strand on which genes are placed, that change the probability of co-expression. The regions bordered by those new boundaries act as the building block of co-expression domains (COD). Indeed, CODs are statistical entities, created by clustering all consecutive genes having a correlated expression. However, correlation is different from causality and genes could have correlated expression by chance. If two nearby genes need to be expressed in similar amounts in condition A, they would be part of the same COD. If the cell enters condition B, the two genes might not be transcribed in similar amounts anymore and they would be split up into different CODs. There are thus regions bordered by physical boundaries (the change of strand), independent of structural proteins, within which genes have a probability of co-regulation, and that can be

"assembled" to form CODs. To completely disrupt the possibility of having co-expression of nearby genes, TAD boundaries or the co-localization of CTCF and Cohesin are introduced. Such model was validated using eQTL data, but further work would be needed to exactly determine if the model can be extended to other human cells or other organisms.

5. Methods

5.1 Data origins and pre-processing

The data sets used in this paper come from previous studies on A549 cells and are available in public databases. The cells were all treated with 100nM of dexamethasone (DEX) or only a vehicle (for controls). RNA-seq (ENCSR897XFT), ChIP-seq (ENCSR571KWZ, ENCSR375BQN, ENCSR588JLN, ENCSR210PYP, ENCSR022IHB, ENCSR625DZB, ENCSR738NGQ, ENCSR447VJR, ENCSR180FFI, ENCSR868FCL, ENCSR342NKR, ENCSR476OXC, ENCSR790OOG, ENCSR483SDK, ENCSR501UJL, ENCSR376GQA) and Hi-C samples (ENCSR842RTB, ENCSR435JUA) were produced by the same laboratory (Dr. Tim Reddy, Duke) and the pre-processed files, annotated with the GRCh38 reference assembly, were retrieved from ENCODE[23,24,29,30] (Supplementary tables 2-5).

For polyA$^+$ RNA-seq, the trimmed, aligned and quantified files containing the raw read counts for each gene were downloaded, for an hourly time-course (Control, 30m, 1h, 2h, 3h, 4h, 5h, 6h, 7h, 8h, 10 and 12h), each timepoint having three or four replicates. Genes are labeled with their ENSEMBL name and read counts for the multiple isoforms were summed. The most upstream base of all isoforms was selected as the start of the gene and the most downstream as the end. Quality control was made by normalizing the read counts with edgeR[31,32] and removing batch effects with svaseq[53] following the method used by McDowell et al. [24]. A principal components analysis (PCA) analysis and a t-distributed stochastic neighbor embedding (t-SNE) analysis were performed on normalized read counts. The replicates clustered correctly by timepoint.

For ChIP-seq data sets, trimmed and aligned *bam* files and *bed* files resulting from peak calling were downloaded for all available replicates, conditions and targets. The ChIP-seq datasets follow the same hourly time-course as the RNA-seq data set. A consensus peakset was produced using DiffBind[40,41]. DiffBind uses the peaks called at each timepoint and the sequenced reads to

create RNA-seq-like read counts, with the number of reads found in each replicate, at each timepoint, for all called peaks. We were not interested in performing a differential binding analysis, so the read counts were not considered. We only used the consensus peakset produced by DiffBind to identify binding sites across the time-course.

Concerning Hi-C data, in addition to the *hic* files available for all replicates, the already identified topologically associated domains contained in *bedpe* files were also downloaded. The Hi-C time-course contains only 5 points: 0h, 1h, 4h, 8h and 12h. The changes in chromatin were quantified using 3DChromatin_ReplicateQC[35], which itself implements four different quality control methods for Hi-C data. First, the pre-processed *hic* files were downloaded from encode and dumped using the *dump observed* method of Juicer[39] using bins of 10kb. 3DChromatin_ReplicateQC first computes quality scores using QuASAR-QC[34], then it uses QuASAR-Rep[34], GenomeDISCO[36] and HiC-Spector[37] to produce reproducibility scores.

TAD boundaries were created by directly taking the list of found TADs at the time 0 available on ENCODE and creating regions 500 bp upstream and 500 downstream of all beginning and ends of TADs. Overlapping boundary regions were merged, such that the 5935 TADs resulted in 11454 TAD boundaries.

5.2 Categorizing and pairing genes

First, differentially expressed genes were identified individually for each timepoints using edgeR and DESeq2[33]. A gene is identified as upregulated if it is differentially expressed according to both methods (FDR < 0.05 for edgeR and padj < 0.05 for DESeq2) and that have a higher expression level at the considered timepoint than the reference. Downregulated genes are those which are differentially expressed and have a higher expression level in the controls. A consensus was created across the time-course such that “Up” genes are genes that are found to be upregulated for at least two timepoints but are never downregulated. The same principle is applied to identify “Down” genes. Genes that did not fall in either of those categories were said to be “Stable”.

Genes were then paired, and their distance was calculated, from TSS to TSS. Only pairs of genes separated by less than 1Mb, from TSS to TSS were kept. This resulted in six categories of pairs, depending on their comportment: “Up-Up”, “Down-Down”, “Stable-Up”, “Down-Stable” and “Down-Up”. With the 14493 genes having detectable transcription, 147248 pairs separated by less than 1Mb could be formed. In addition, pairs were labeled according to the relative position

of the genes: “Same strand” if both genes are located on the same DNA strand, “Divergent” if the genes are on different strands and back-to-back and “Convergent” if they are on different strands and facing each other.

5.3 Odds ratios and distribution matching

As one of the main objectives is to find whether genes with opposite behaviors are separated by a physical barrier, the enrichment for finding a barrier between such pairs had to be computed. Enrichments were expressed in odds ratios, where the “Stable-Stable” category was used as reference. However, the distribution of distances is not the same between pair types. To avoid a bias where pairs separated by a larger distance are found to have a barrier between them by chance, the reference pairs were sub-sampled such that the distribution of distances of the resampling matches the distribution of distances of the pairs of interest, and the query and reference contain the same number of pairs. The distribution matching algorithm consists in dividing the distributions of distance from the interest pairs and the control pairs into bins of 5kb, then counting the number of interest pairs falling in each bin and sampling as many control pairs in the corresponding bin. The resampling was done 1000 times, to allow the calculation of empiric p-values. Odds ratios are defined as follows:

$$OR = \frac{x(1-\bar{y})}{(1-x)\bar{y}}$$

In the above equation, *x* corresponds to the proportion of pairs of interest containing a barrier between the two genes and *y* represents the proportion of pairs of interest, after resampling, with a barrier. The region “between the two genes” is considered as the regions between gene bodies, thus overlapping genes cannot contain a barrier between them, even if their TSS are not at the same position. For this analysis and the following, only the pairs with genes separated by less than 100kb were kept, which resulted in 20197 pairs.

5.4 Characterizing boundaries

To characterize the boundaries, only the pairs of consecutives genes were kept, which totalize 8946 pairs of genes. The pairs of genes were subdivided in categories depending on whether they were found in the same TAD, around a TAD boundary or outside TADs. Pairs within the same TAD were further subdivided between pairs where both genes have the same behavior

("Up-Up", "Down-Down" and "Stable-Stable") or opposite behavior ("Down-Up", "Down-Stable" and "Stable-Up"). The relative abundance of same-strand, convergent and divergent pairs was compared for those four categories: (1) pairs at TAD boundary, (2) pairs outside TADs, (3) pairs inside the same TAD with the same behavior and (4) pairs inside the same TAD with opposite behaviors. That last category serves as a proxy to COD boundaries. The pairs were then analyzed separately depending on the "strandedness" to find the repartition of structural proteins such as CTCF and Cohesin. As Cohesin does not have direct ChIP-seq data, the co-localization of RAD21 and SMC3, two of its subunits for which there are available ChIP-seq replicates, was used instead.

The odds ratio heatmap has been by comparing convergent and divergent pairs to "same-strand" pairs after consecutive resampling to match distance distributions as before. Unlike in the first heatmap, the empirical p-values were computed considering both tails, rather than just the upper tail, to be able to have p-values associated with depletion too and not only with enrichment.

5.5 Predicting eQTL targets

The significant eQTL and their gene targets found in healthy lung cells were retrieved from GTEx[42]. The data used for the analyses described in this manuscript were obtained from Single-Tissue cis-eQTL Data on the GTEx Portal, dbGaP accession number phs000424.v8.p2 on 03/02/2020. As we wanted to analyze the relation between genes affected by the same eQTL, all eQTLs with a single target were discarded. This resulted in a total of 477420 selected eQTLs and 9434 gene targets. All genes having their TSS within 100 kb of each other were paired, but pairs involving the same genes were only considered once. Indeed, two genes can be both affected by multiple eQTLs and we wanted to consider each pair of genes uniquely, even if multiple variants are involved. We thus obtained 9088 interest pairs, where both genes are affected by the same eQTL, and 24035 control pairs, where genes are affected by different eQTLs or one gene is affected by an eQTL and the other is not. The prevalence of finding those genes on the same strand or not, or to be separated by a barrier or not was assessed following the distribution matching technique and the odds ratio formula described previously. Two genes were said to be separated by a barrier if there was either a TAD boundary or if there was evidence of CTCF, RAD21 and SMC3 between them. The pairs of genes affected by the same eQTL were compared to pairs in which one gene is affected by an eQTL and the other is not significatively impacted by it and is within 100kb of the first TSS. To account for the variations in which eQTL affect genes, we tested

all pairs in addition to the following subsets: (1) only pairs of consecutive genes, (2) all pairs affected by strong eQTLs (regression slope of the gene-eQTL association > 0.7 or < -0.7) and (3) consecutive pairs affected by strong eQTLs. The pairs were then further categorized into (1) pairs on the same strand without barrier (both CTCF and Cohesin or a TAD boundary) between them, (2) pairs on opposite strands without barrier between them, (3) pairs on the same strand with a barrier between them and (4) pairs on opposite strands with a barrier between them. P-values associated with both depletion and enrichment were computed empirically as before.

6. References

1. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods* **72**, 65–75 (2015).

2. Grob, S. & Cavalli, G. Technical Review: A Hitchhiker's Guide to Chromosome Conformation Capture. in *Plant Chromatin Dynamics: Methods and Protocols* (eds. Bemer, M. & Baroux, C.) 233–246 (Springer New York, 2018). doi:10.1007/978-1-4939-7318-7_14.

3. Nagano, T. *et al.* Single cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, (2013).

4. Rao, S. S. P. *et al.* A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

5. Rowley, M. J. *et al.* Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**, 837-852.e7 (2017).

6. Nurick, I., Shamir, R. & Elkon, R. Genomic meta-analysis of the interplay between 3D chromatin organization and gene expression programs under basal and stress conditions. *Epigenetics Chromatin* **11**, (2018).

7. Wendt, K. S. & Grosveld, F. G. Transcription in the context of the 3D nucleus. *Current Opinion in Genetics & Development* **25**, 62–67 (2014).

8. Adriaens, C. *et al.* Blank spots on the map: some current questions on nuclear organization and genome architecture. *Histochem Cell Biol* (2018) doi:10.1007/s00418-018-1726-1.

9. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376–380 (2012).

10. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

11. Paulsen, J. *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nature Genetics* (2019) doi:10.1038/s41588-019-0392-0.

12. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: the Unit of Chromosome Organization. *Mol Cell* **62**, 668–680 (2016).

13. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat Genet* **52**, 8–16 (2020).

14. Chang, L.-H., Ghosh, S. & Noordermeer, D. TADs and Their Borders: Free Movement or Building a Wall? *Journal of Molecular Biology* **432**, 643–652 (2020).

15. Gómez-Díaz, E. & Corces, V. G. Architectural proteins: Regulators of 3D genome organization in cell fate. *Trends Cell Biol* **24**, 703–711 (2014).

16. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* **14**, 762–775 (2014).

17. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32**, 225–237 (2016).

18. Sexton, T. & Cavalli, G. The Role of Chromosome Domains in Shaping the Functional Genome. *Cell* **160**, 1049–1059 (2015).

19. Tang, B. *et al.* Advances in Genomic Profiling and Analysis of 3D Chromatin Structure and Interaction. *Genes (Basel)* **8**, (2017).

20. Kagey, M. H. *et al.* Mediator and Cohesin Connect Gene Expression and Chromatin Architecture. *Nature* **467**, 430–435 (2010).

21. Smith, E. M., Lajoie, B. R., Jain, G. & Dekker, J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *Am J Hum Genet* **98**, 185–201 (2016).

22. Le Dily, F. *et al.* Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* **28**, 2151–2162 (2014).

23. D'Ippolito, A. M. *et al.* Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids. *Cell Systems* **7**, 146-160.e7 (2018).

24. McDowell, I. C. *et al.* Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res* **28**, 1272–1284 (2018).

25. Ibn-Salem, J., Muro, E. M. & Andrade-Navarro, M. A. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Res* **45**, 81–91 (2017).

26. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput Biol* **13**, (2017).

27. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).

28. Hnisz, D., Day, D. S. & Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* **167**, 1188–1200 (2016).

29. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74 (2012).

30. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* **46**, D794–D801 (2018).

31. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

32. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288–4297 (2012).

33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, (2014).

34. Sauria, M. E. & Taylor, J. QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data. *bioRxiv* 204438 (2017) doi:10.1101/204438.

35. Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol* **20**, 57 (2019).

36. Ursu, O. *et al.* GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).

37. Yan, K.-K., Yardımcı, G. G., Yan, C., Noble, W. S. & Gerstein, M. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199–2201 (2017).

38. Khoury, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat Commun* **11**, 1–13 (2020).

39. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).

40. Stark, R. & Brown, G. DiffBind: Differential binding analysis of ChIP-Seq peak data. 33 (2011).

41. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).

42. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

43. Yan, C., Wu, S., Pocetti, C. & Bai, L. Regulation of cell-to-cell variability in divergent gene expression. *Nat Commun* **7**, (2016).

44. Yang, W. *et al.* Promoter-sharing by different genes in human genome – CPNE1 and RBM12 gene pair as an example. *BMC Genomics* **9**, 1–16 (2008).

45. Zhang, L.-F., Ding, J.-H., Yang, B.-Z., He, G.-C. & Roe, C. Characterization of the bidirectional promoter region between the human genes encoding VLCAD and PSD-95. *Genomics* **82**, 660–668 (2003).

46. Rieder, D., Trajanoski, Z. & McNally, J. G. Transcription factories. *Front Genet* **3**, (2012).

47. Sutherland, H. & Bickmore, W. A. Transcription factories: gene expression in unions? *Nature Reviews Genetics* **10**, 457–466 (2009).

48. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO J* **12**, 1059–1065 (1993).

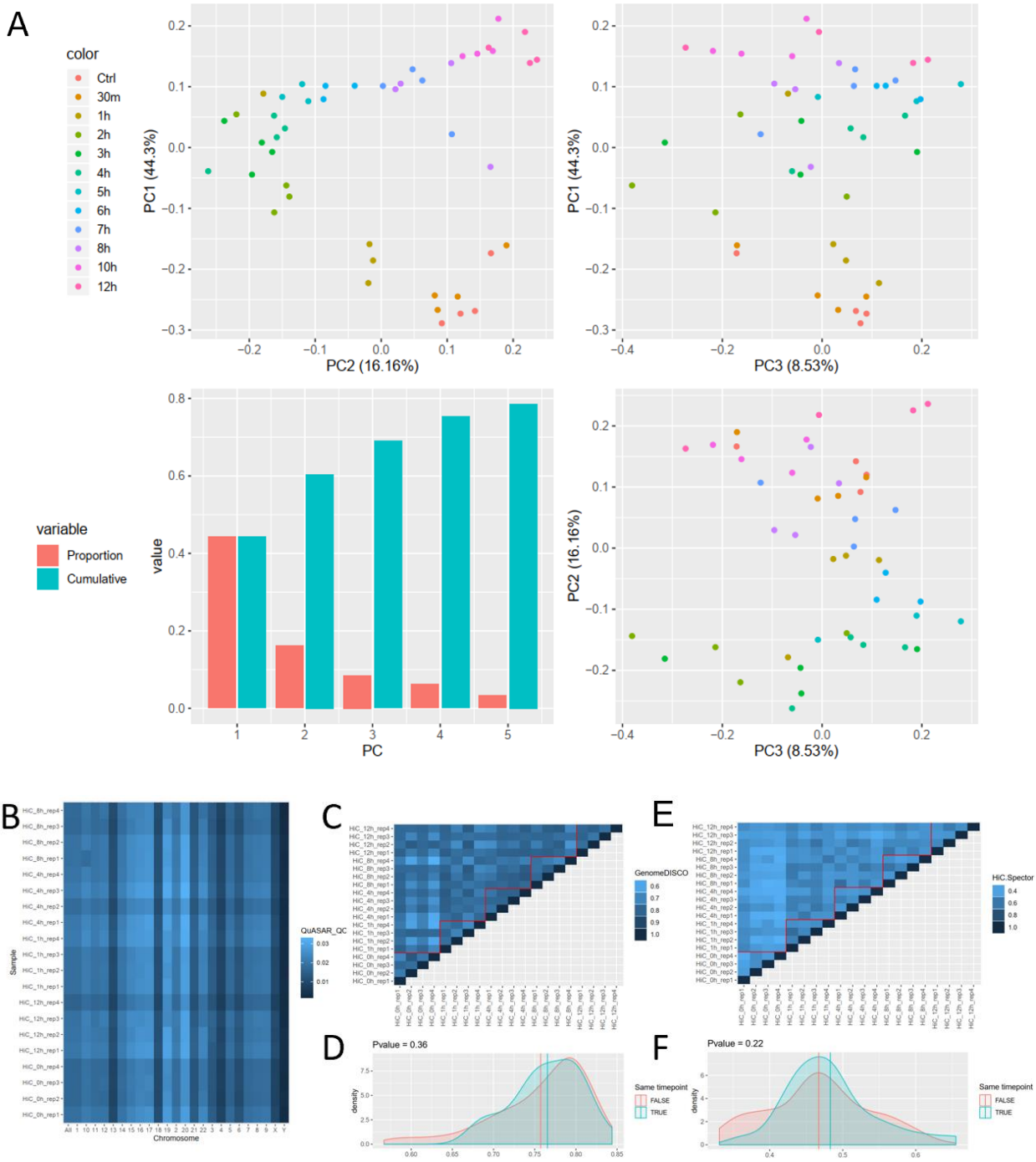
49. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53–61 (2010).

50. Ma, X., Ezer, D., Adryan, B. & Stevens, T. J. Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome Biology* **19**, 174 (2018).

51. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics* **36**, 1065–1071 (2004).

52. Rada-Iglesias, A., Grosveld, F. G. & Papantonis, A. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol Syst Biol* **14**, (2018).

53. Leek, J. T. *et al.* sva: Surrogate Variable Analysis. R package version 3.30.1. (2019).

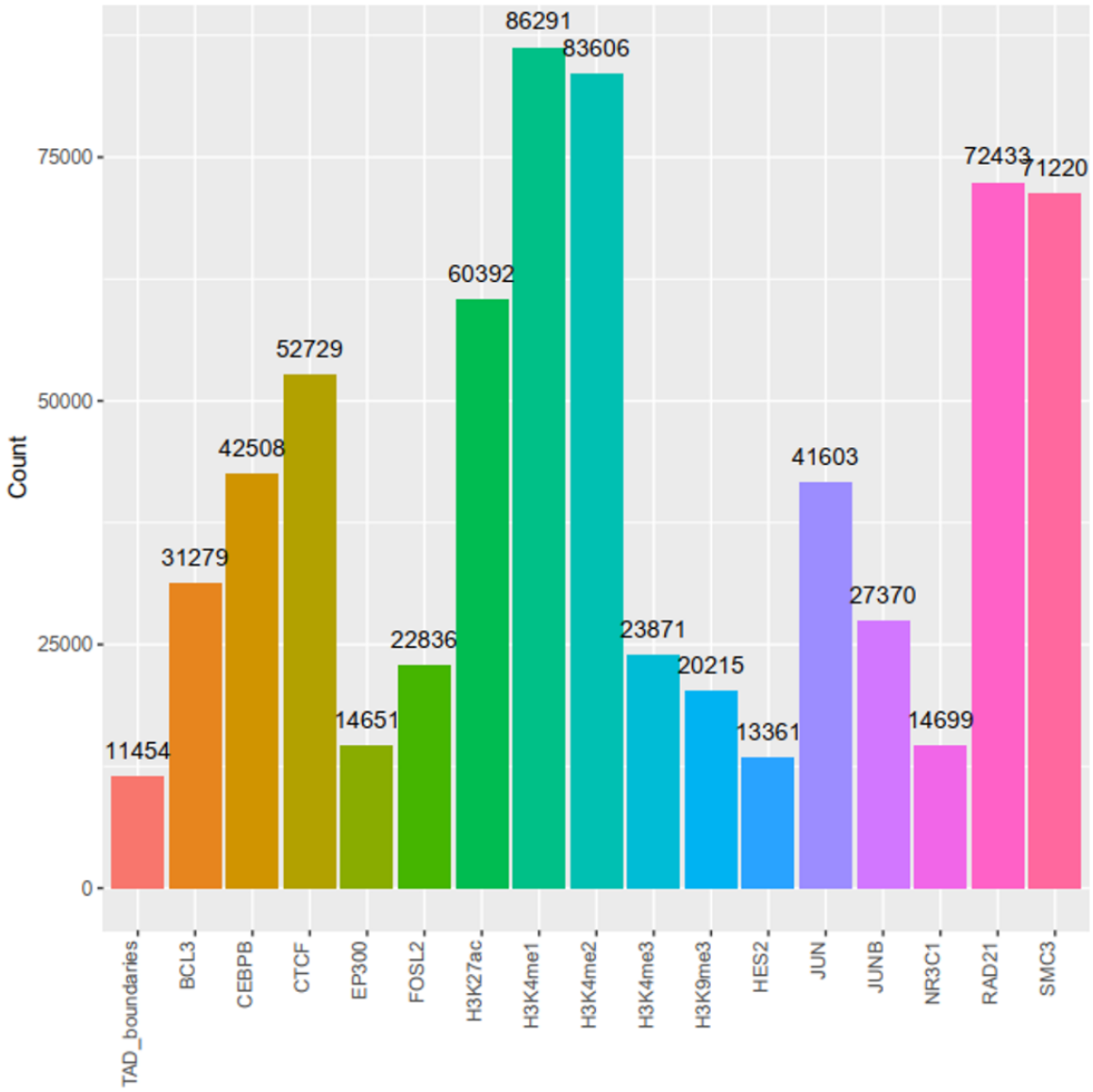


Supplementary figure 1: Gene expression changes in the RNA-seq samples and Hi-C reproducibility scores. *(A) PCA of the normalized RNA-seq data. The top-left, top-right and bottom right panels show the repartition of replicates in the space produced by the first three principal components (PC1 and PC2, PC1 and PC2, then PC2 and PC3, respectively). The bottom-left panel shows the proportion and cumulative proportion of the variance explained by*

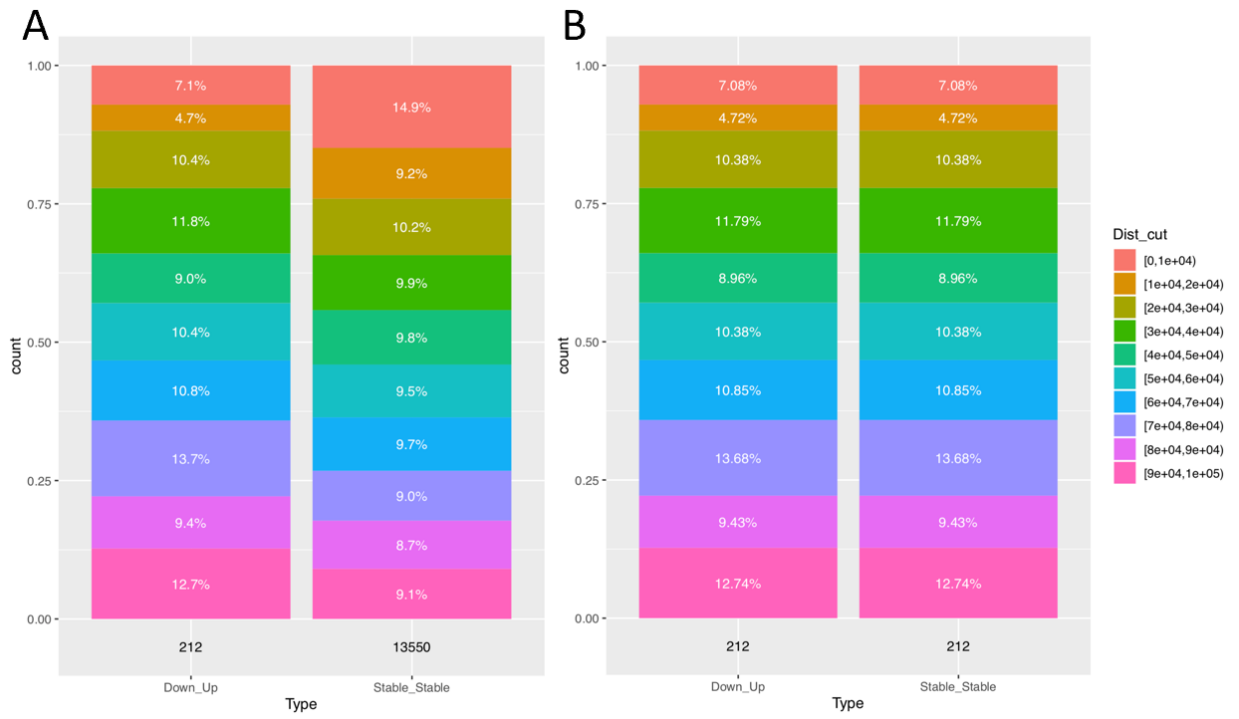
the first five PCs. (B) Heatmap of the quality control scores given by QuASAR-QC. (C) Heatmap of the replication scores given by GenomeDISCO. Comparaisons inside the same timepoint are under the red lines and comparisons across timepoints are over it. (D) Distributions of the scores given by GenomeDISCO between replicates of the same timepoint (blue) or different timepoints (red). (E) Heatmap of the replication scores given by HiC-Spector. (F) Distributions of the scores given byHiC-Spector.

	Up	Stable	Down
30m	0	13768	0
1h	144	13512	84
2h	667	12623	468
3h	913	12267	653
4h	1026	12034	746
5h	1186	11604	957
6h	1308	11338	1139
7h	1193	11548	1076
8h	887	12272	687
10h	1468	10878	1508
12h	1528	10670	1677
Consensus	1716	10751	1810

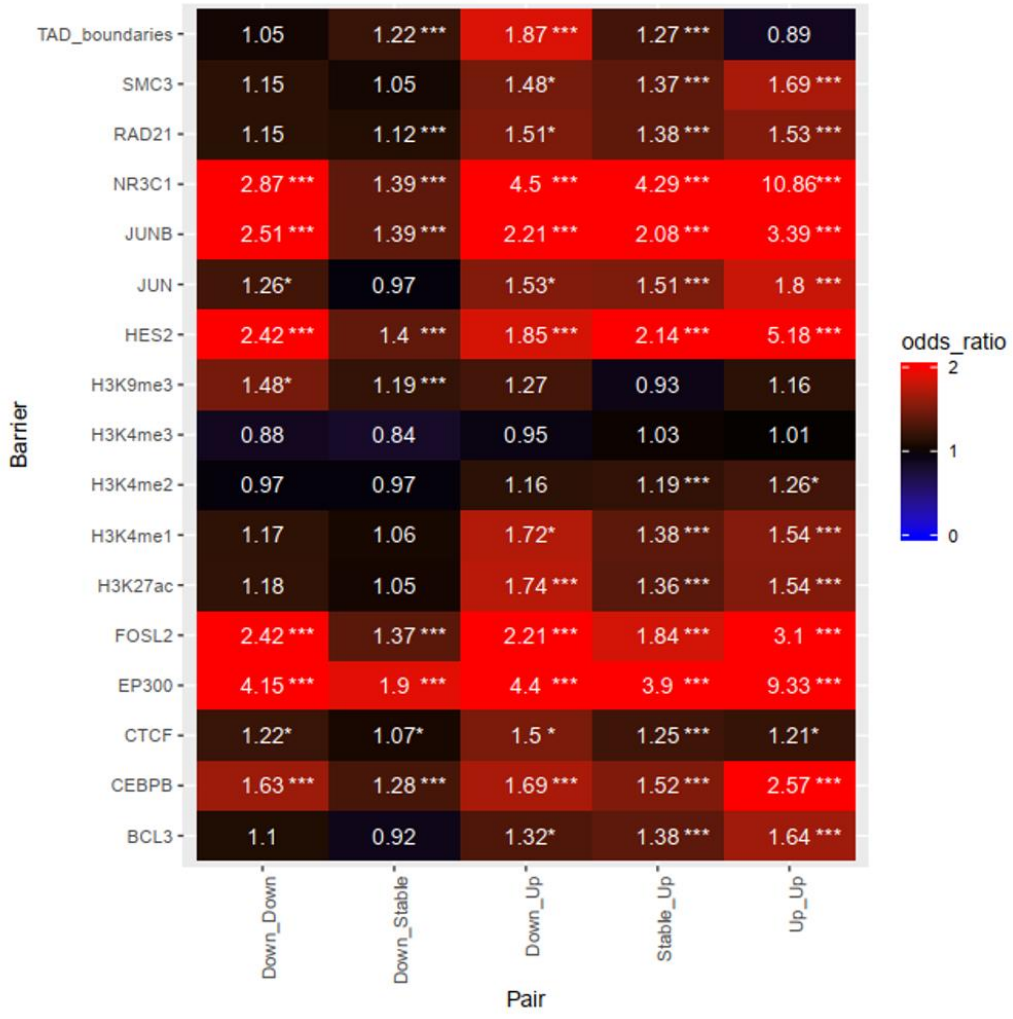
Supplementary table 1: Number of genes labeled as "Up", "Stable" and "Down" at each timepoint and consensus.



Supplementary figure 2: Count of the nuclear proteins and TAD boundaries. *Number of peaks (for nuclear proteins) and of TAD boundaries found in the data.*



Supplementary figure 3: The resampling step limits distance bias. *(A) Distribution of distances of the "Down-Up" pairs and reference ("Stable-Stable") pairs before sub-sampling. (B) Distribution of distances of the "Down-Up" pairs and reference ("Stable-Stable") pairs after the sub-sampling.*



Supplementary figure 4: Complete heatmap of odds ratios for the presence of a physical barrier between the genes of the pairs, for all available TFs. *The "Stable_Stable" pairs are used as reference. *P-value < 0.05; ***P-value < 0.001. The p-values are empirical, computed with 1000 resampling events*

Chapter 3: General Discussion

3.1 Analysis of the Boundaries in the Genome

3.1.1 Strand Position Affects Co-Expression Probability

In Chapter 2, we have shown that TAD boundaries serve as barriers to co-expression, as they are more prevalent between genes that have opposite behavior after DEX induction. We also found a second type of boundary determined by a structural element. The switch of strand on which genes are located mark intra-TAD boundaries, independently of CTCF and Cohesin. Co-expression of genes across those boundaries is reduced. The regions bordered by the boundaries seem to serve as building blocks for CODs. eQTL data seem to corroborate the previous findings, as genes affected by the same eQTL tend to be located on the same strand and to be less separated by a barrier such as a TAD boundary.

We thus proposed a model where genes on the same strand tend to by highly co-expressed by chance. The change of strand reduces that probability and the introduction of barriers disrupts co-expression even more. TADs seem to work as a frame in which CODs, more variable, change in function of the conditions.

3.1.2 Limitations of the method

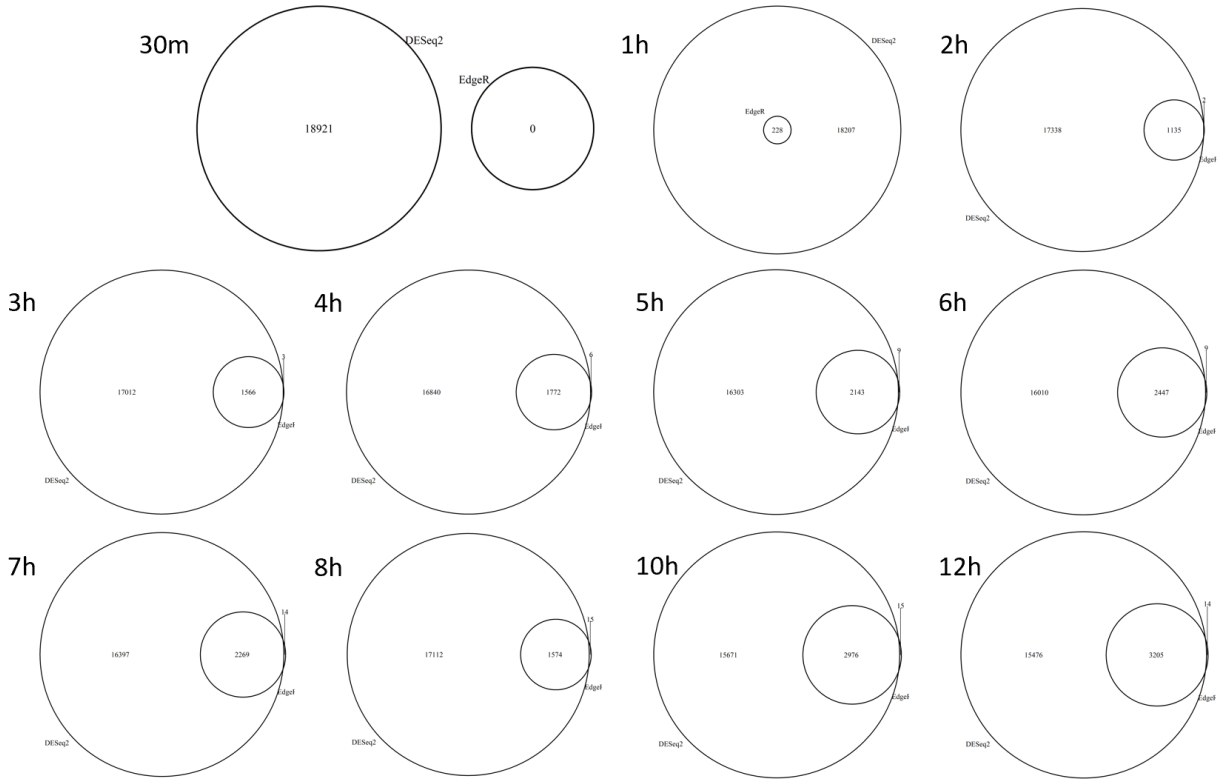
3.1.2.1 Distance Distribution Correction

The results produced during this project can be divided into three main parts: 1) the effect of nuclear proteins and TAD boundaries on co-expression, 2) the effect of gene orientation on co-expression probability and 3) validation of previous observations using eQTL data. One important element all these results share is the distance-bias correction. Indeed, we have seen that distance between genes greatly influences co-expression. For example, genes that are activated tend to be found closer from each other than genes having opposite behaviors. The distance difference between various types of pairs is especially important for the first analysis, as two genes have a higher probability of having a nuclear protein between them, by chance, as the distance between them increases. There was thus the need of correcting that distance bias before doing any analysis. The distance bias has been accounted for by resampling a control set of paired genes to match the distribution of distances of the interest set of paired genes. Yet, the distribution matching is not perfect and allows for a maximum of 5kb difference. In other words, the distribution matching

permits to obtain a control set that is closer to the test set, regarding the distances between the pairs, but there is not a perfect match and the distance bias is reduced but not completely removed. However, a perfect distribution matching, meaning there would be a pair in the control set having the exact same distance between the genes for each pair of the tested set, would certainly lead to an important loss of data, as the probability of finding matching pairs with the exact same distance is low. The distribution-matching technique used in the project is thus a needed compromise between lowering the distance bias while not losing to much data. That compromise is to keep in mind while commenting the results of the following analyzes.

3.1.2.2 Conservative Gene Labeling

The first analysis used six categories of paired genes, labeled according to how genes behave after DEX induction. Two different R packages for identifying differentially expressed genes (DEG), edgeR and DESeq2, were used. DESeq2 usually identified more genes than edgeR did, and almost all edgeR-identified DEG were also found by DESeq2 (General figure 6). Only genes that were DEG according to both methods were kept as true DEG. Their "direction" was then attributed for each time-point before a consensus was made. The identification of "Up", "Down" and "Stable" genes is thus rather stringent, as they have to be found by two methods and there is no fine categorization of behavior; genes that are rapidly upregulated and those that are upregulated after a few hours only all under the same "Up" label. This means that the "Down-Up", "Down-Down" and "Up-Up" categories contain genes that can be trusted to be differentially expressed, but that two "Up" genes do not necessarily have the exact same expression pattern across the time-course. In addition, some genes falling into the "Stable-Stable" category might actually have expression changes that were too small to be detected with certitude by both methods. While it could be seen as a limitation, the strict criteria the genes must respect to be labeled as "Up" or "Down" actually strengthens the results. Indeed, the analyses show significant enrichment despite the possibility of not taking all DEG into account. It would however be interesting to create more categories of pairs, with more fine behavior descriptions, to see if different upregulation patterns also tend to be more often separated by a barrier than upregulated genes with the same pattern.



General figure 6: Venn diagrams of the number of differentially expressed genes*. Most genes identified by edgeR are also identified by DESeq2, regardless of the timepoint.*

3.1.2.3 Position of TAD Boundaries

The barrier property of TAD boundaries was assessed using the TAD called at time 0 as reference. This was justified by using three methods to compute reproducibility and concluding the scores are similar enough. However, all chromatin changes cannot fully be represented by a single number. TAD boundaries only cover a small portion of the complete genome. Indeed, the boundaries used in this project cover less than 11.5 million of base pairs, while the whole human genome contains approximately 3 billion of them. TAD boundaries used in this study cover thus a little less than 0.38% of the whole genome and the changes that may happen in those 0.38% is likely to be lost in all other possible changes captured by a single reproducibility score. It is thus possible that changes happening in a subset of the boundaries are not represented by the returned reproducibility score of any of the methods.

3.1.2.4 Position of COD Boundaries

Once the TAD boundaries were confirmed to act as a barrier, we searched for intra-TAD boundaries. The pairs of consecutive genes that were inside the same TAD but having different behaviors were used as proxy for COD boundaries. CODs are usually identified using correlation of gene expression, and genes with uncorrelated expression mark the boundaries[32]. It is safe to assume that genes with different behaviors have uncorrelated expression. Using genes inside the same TAD with different behavior is thus a good proxy for finding COD boundaries without risking having false positives. However, as the genes are not more finely labeled, there is a risk of missing COD boundaries. Two genes having the "Up" label does not mean they are correlated expression. One of them could be activated early on during the time-course, and the other later, resulting in a lesser correlation of their expression. This might explain why there is only a slight difference between the pairs of genes in the same TAD with different behavior, and those in the same TAD with the same behavior: that last category is likely to contain COD boundaries that were not identified by our strict labeling.

3.1.2.5 eQTL Provenance

The last step combines the position of barriers and eQTL data to validate the model. TAD boundaries are highly conserved between single cells[10] and cell types[7,12,13,19]. Nonetheless, they are not perfectly static, and a few changes may occur[18]. The TAD boundaries and ChIP-seq data comes from A549 cells, while the eQTL data comes from lung cells of patients. The small differences in cellular organization between the lung cells and the A549 lung cancer cell line may affect the results.

3.2 Results Put in Context

The insulation property of TADs we show had been reported before[20,27,34,37]. However, the regulatory properties of TADs were unclear and TADs were missing a functional definition[11,16]. With this study, we answer partially that question by distinguishing TAD boundaries and COD boundaries. Different levels of architecture have different functions and different strengths. TADs are a stricter, less mobile structures that include CODs, more variable regulatory units that change according to the cell needs. We thus suggest that the genome expression regulation is linked to genomic architecture, through boundaries that affect the co-expression probability. This would be compatible with previous observations. Indeed, some smaller TADs may, in a specific condition,

harbour only one COD and thus the whole TAD would act as a regulation unit as seem by Le Dily et al[29]. On the other hand, in other conditions or in larger TADs, multiple CODs may exist, explaining why randomly created TADs seem to have as much co-expression as real TADs[32].

The proposed model is also compatible with previous models. Indeed, transcription factories have been suggested to be an important element of the transcription process, but their composition is still unresolved. They seem to be structured through both RNA polymerase II loops and CTCF loops[40]. The model has been explained as if the DNA ere static and the transcription machinery mobile, for ease of comprehension. But it can be understood as being part of transcription factories. Assuming the transcription machinery forms the core of factories and are static, the proposed model can be explained in the following manner: TADs serve as the structural basis of factories, multiple TADs aggregate to a single factory, thus resulting to what was interpreted as compartments, and CODs would correspond to a single or a subset of RNA polymerases in the factory. In other words, when a gene enters a factory and binds itself to a polymerase to start transcription, nearby genes on the same strand are likely to get transcribed as well by chance. When there is no promoter sharing, the switch of strands introduces a small disruption in co-expression. Transcription of genes on opposite strands would probably require the DNA to detach the strand being transcribed and completely change its orientation in order to have the opposite site transcribed. It seems less likely than transcription of genes on the same strand, that requires the DNA to "forget" to detach itself after the transcription of one gene and thus getting the downstream gene transcribed or shifting a little to much while going back to the promoter and getting the upstream gene transcribed. Genes that are co-expressed by the same factory would thus be interpreted as CODs and consecutive CODs would form TADs. The aggregation of multiple TADs to the same factory might lead to the interaction patterns seen in Hi-C maps that were interpreted as compartments.

3.3 Perspectives

3.3.1 Promoter Sharing and Tethered Sites

For this study, the position of genes, and more specifically their TSS, was considered, but no other element was. It was assumed that each gene has its own promoter. Studies showed a few cases of promoter sharing between divergent genes in in yeast[55] and humans[56,57]. Promoter sharing

leads to co-regulation and in future analyzes, it might be better to consider only one gene by promoter, or at least acknowledge genes sharing the same promoter.

In this study, we argue that the activation or repression of one gene affects nearby genes, depending on their relative position and what barriers are present. Previously, it has been shown that some transcription factors such as GR have primary targets, directly bound by GR, and secondary targets, bound to GR *via* tethered sites and required co-factors. It would be interesting to explore those two mechanisms together, as they could create an entire hierarchy of responses for each direct TF binding: one primary target, a few by-products, secondary targets through tethering and finally by-products of the secondary targets. Adding information on what regulatory elements affect what genes, and which of those regulatory elements are bound, directly or not, would certainly greatly improve the model and our general knowledge on transcription. Promoters and enhancers of some genes can be found in annotation packages, can be deduced using 3D chromatin contacts or retrieved using special methods[71]. Direct and tethered TF binding sites are obtained using a mix of ChIP-seq and Hi-C data[65]. It is thus a variety of possibilities that exist to improve the model, but they require multiple types of data sets and should ideally all come from the same cells.

3.3.2 Expansion of the Data

In addition to obtaining more data types in one cell type after one stimulus, the model would also benefit from the use of different cell types and different stimuli. This would help to confirm the validity of the model and would also have the advantage of confirming COD boundaries. Indeed, CODs are, by definition, variable and change from condition to condition, and from cell to cell. By only using one stress as in this study, we certainly only retrieved a subset of all possible COD boundaries positions. The lack of difference in the repartition of the structural proteins and the strand conformations between the “same TAD, same behavior” and the “ same TAD, different behavior” can be attributed to the lack of fine labelling, as explained before, but it could also be due to the variable nature of CODs and thus their boundaries. As the cells needs change from one condition to another and the CODs get re-organized to satisfy those needs, it may be that some COD boundaries might be seen only after a specific stimulus and not another. We argue that gene orientation and strand change pre-determine the location of COD boundaries. It would be interesting to confirm that convergent and divergent gene pairs that were not observed

to be COD boundaries when inducing A549 cells with DEX become COD boundaries in different cells or in A549 cells induced with a different hormone or a heat shock.

3.3.3 Exploring the Nucleus Environment

Hi-C data is very useful to explore the nuclear architecture. However, it remains noisy and can thus be imprecise at very high resolutions. Micro-C is a new technique, based on Hi-C, that retrieves the same elements (compartments, TADs and chromatin loops) with a better signal-to-noise ratio[72,73]. As such, it would be interesting to produce data using Micro-C rather than Hi-C to have a more precise mapping of chromatin interaction. Moreover, Micro-C detect accurately interaction between enhancers and promoters and between promoters of different genes[72]. Micro-C would thus be useful to better characterize how the proximity of genes affects their co-expression in a 3D context. In this study, we only used the linear distance, due to technical limitations, but Micro-C might permit to use special distance instead.

The direct environment of the genes seems important, as it determines how genes are regulated, depending on the expression of nearby genes, and which transcription factors are present. Now that as basis to link architecture and transcription has been proposed, it would be interesting to explore it further, keeping transcription factories in mind. Indeed, if the proposed model and the model of transcription factories can be confirmed as different points of view of the same mechanism, changes in transcription could be better anticipated. It would also serve as a starting point to better explore if some transcription factories are indeed specialized or not[46–48]. It would also permit to explore how genes compete for resources in the closed environment that is the nucleus[67].

Chapter 4: Conclusions and Future Directions

The model presented in this study could help better anticipate changes in cell behavior due to a stimulus or a change in cell environment. But, before predictions can be made, further work is needed to test the model under different conditions and different cell types. First, A549 cells induced with DEX were used. A549 cells come from a cell line, more specifically from a cancer cell line, and could thus have dissimilarities from patient-extracted cells. It is thus necessary to verify the validity of the model in other cell types. Second, the model has been discovered using a hormonal stimulus. There is the need to confirm that the model is robust regardless of the hormone tested. The model should also be tested using a different kind of stimulus. For example, a heat shock has been seen to induce changes in gene expression, but no to change the overall chromatin architecture, the contacts between enhancers and promoters, nor the TAD boundaries[21]. For those reasons, a heat shock would be another type of stress to put cells under to account for the robustness of the model. Third, bulk RNA-seq was used. As single-cell RNA-seq techniques have become more powerful, it would be interesting to see how individual changes in singular cells support the model. It could further confirm the statistical nature of CODs.

Once the model will be more accurately defined, and its strength improved by multiple analyses in different cell lines under various conditions, it could be used to predict secondary targets simply by knowing the 3D nuclear architecture and the primary targets. The unwanted by-product effects could be prevented by knowing before-hand which genes will be affected by a certain stimulus more precisely than before.

By augmenting the data types (more complete RNA-seq, ChIP-seq and Hi-C) and the data sources (different stimuli, different cell types), we hope to eventually find the key to understanding how genomic architecture dictates transcription. Indeed, while there are many evidences of the existence and function of transcription factories, not much is known concerning how they are regulated, what signals permit the recruiting or release of genes onto them or how they are related to TADs. By exploring our model more in depth, we hope to answer those questions.

Chapter 5: References: Master reference list

1. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

2. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*. 2018;46(D1):D794-D801. doi:10.1093/nar/gkx1081

3. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

4. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods*. 2015;72:65-75. doi:10.1016/j.ymeth.2014.10.031

5. Rao SSP, Huntley MH, Durand NC, et al. A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665-1680. doi:10.1016/j.cell.2014.11.021

6. Rowley MJ, Nichols MH, Lyu X, et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell*. 2017;67(5):837-852.e7. doi:10.1016/j.molcel.2017.07.022

7. Wendt KS, Grosveld FG. Transcription in the context of the 3D nucleus. *Current Opinion in Genetics & Development*. 2014;25:62-67. doi:10.1016/j.gde.2013.11.020

8. Nurick I, Shamir R, Elkon R. Genomic meta-analysis of the interplay between 3D chromatin organization and gene expression programs under basal and stress conditions. *Epigenetics Chromatin*. 2018;11. doi:10.1186/s13072-018-0220-2

9. Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381-385. doi:10.1038/nature11049

10. Nagano T, Lubling Y, Stevens TJ, et al. Single cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 2013;502(7469). doi:10.1038/nature12593

11. Adriaens C, Serebryannyy LA, Feric M, et al. Blank spots on the map: some current questions on nuclear organization and genome architecture. *Histochem Cell Biol*. Published online September 20, 2018. doi:10.1007/s00418-018-1726-1

12. Paulsen J, Liyakat Ali TM, Nekrasov M, et al. Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nature Genetics*. Published online April 22, 2019. doi:10.1038/s41588-019-0392-0

13. Dixon JR, Selvaraj S, Yue F, et al. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature*. 2012;485(7398):376-380. doi:10.1038/nature11082

14. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*. 2016;32(4):225-237. doi:10.1016/j.tig.2016.01.003

15. Sexton T, Cavalli G. The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*. 2015;160(6):1049-1059. doi:10.1016/j.cell.2015.02.040

16. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. *Nat Genet*. 2020;52(1):8-16. doi:10.1038/s41588-019-0561-1

17. Tang B, Cheng X, Xi Y, Chen Z, Zhou Y, Jin VX. Advances in Genomic Profiling and Analysis of 3D Chromatin Structure and Interaction. *Genes (Basel)*. 2017;8(9). doi:10.3390/genes8090223

18. Gómez-Díaz E, Corces VG. Architectural proteins: Regulators of 3D genome organization in cell fate. *Trends Cell Biol*. 2014;24(11):703-711. doi:10.1016/j.tcb.2014.08.003

19. Gorkin DU, Leung D, Ren B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell*. 2014;14(6):762-775. doi:10.1016/j.stem.2014.05.017

20. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: the Unit of Chromosome Organization. *Mol Cell*. 2016;62(5):668-680. doi:10.1016/j.molcel.2016.05.018

21. Ray J, Munn PR, Vihervaara A, Ozer A, Danko CG, Lis JT. *Chromatin Conformation Remains Stable upon Extensive Transcriptional Changes Driven by Heat Shock*. Molecular Biology; 2019. doi:10.1101/527838

22. Schmitt AD, Hu M, Jung I, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*. 2016;17(8):2042-2059. doi:10.1016/j.celrep.2016.10.061

23. Barutcu AR, Lajoie BR, McCord RP, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol*. 2015;16. doi:10.1186/s13059-015-0768-0

24. Liu S, Chen H, Ronquist S, et al. Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming. *iScience*. 2018;6:232-246. doi:10.1016/j.isci.2018.08.002

25. Zhou Y, Gerrard DL, Wang J, et al. Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. *Nature Communications*. 2019;10(1). doi:10.1038/s41467-019-09320-9

26. Szabo Q, Jost D, Chang J-M, et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci Adv*. 2018;4(2). doi:10.1126/sciadv.aar8082

27. Chang L-H, Ghosh S, Noordermeer D. TADs and Their Borders: Free Movement or Building a Wall? *Journal of Molecular Biology*. 2020;432(3):643-652. doi:10.1016/j.jmb.2019.11.025

28. Rada-Iglesias A, Grosveld FG, Papantonis A. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol Syst Biol*. 2018;14(6). doi:10.15252/msb.20188214

29. Le Dily F, Vidal E, Cuartero Y, et al. Hormone-control regions mediate steroid receptor–dependent genome organization. *Genome Res*. 2019;29(1):29-39. doi:10.1101/gr.243824.118

30. Tan L, Xing D, Chang C-H, Li H, Xie XS. Three-dimensional genome structures of single diploid human cells. *Science*. 2018;361(6405):924-928. doi:10.1126/science.aat5641

31. Hnisz D, Day DS, Young RA. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell*. 2016;167(5):1188-1200. doi:10.1016/j.cell.2016.10.024

32. Soler-Oliva ME, Guerrero-Martínez JA, Bachetti V, Reyes JC. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput Biol*. 2017;13(9). doi:10.1371/journal.pcbi.1005708

33. Delaneau O, Zazhytska M, Borel C, et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*. 2019;364(6439):eaat8266. doi:10.1126/science.aat8266

34. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. *Cell*. 2016;164(6):1110-1121. doi:10.1016/j.cell.2016.02.007

35. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*. 2016;15(9):2038-2049. doi:10.1016/j.celrep.2016.04.085

36. Nishana M, Ha C, Rodriguez-Hernaez J, et al. Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. *Genome Biol*. 2020;21(1):1-34. doi:10.1186/s13059-020-02024-0

37. Khoury A, Achinger-Kawecka J, Bert SA, et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat Commun*. 2020;11(1):1-13. doi:10.1038/s41467-019-13753-7

38. Poterlowicz K, Yarker JL, Malashchuk I, et al. 5C analysis of the Epidermal Differentiation Complex locus reveals distinct chromatin interaction networks between gene-rich and gene-poor TADs in skin epithelial cells. Ezhkova E, ed. *PLoS Genet*. 2017;13(9):e1006966. doi:10.1371/journal.pgen.1006966

39. Valton A-L, Dekker J. TAD disruption as oncogenic driver. *Current Opinion in Genetics & Development*. 2016;36:34-40. doi:10.1016/j.gde.2016.03.008

40. Tang Z, Luo OJ, Li X, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*. 2015;163(7):1611-1627. doi:10.1016/j.cell.2015.11.024

41. Hill VK, Kim J-S, Waldman T. Cohesin Mutations in Human Cancer. *Biochim Biophys Acta*. 2016;1866(1):1-11. doi:10.1016/j.bbcan.2016.05.002

42. Kagey MH, Newman JJ, Bilodeau S, et al. Mediator and Cohesin Connect Gene Expression and Chromatin Architecture. *Nature*. 2010;467(7314):430-435. doi:10.1038/nature09380

43. Jackson DA, Hassan AB, Errington RJ, Cook PR. Visualization of focal sites of transcription within human nuclei. *EMBO J*. 1993;12(3):1059-1065.

44. Mitchell JA, Fraser P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev*. 2008;22(1):20-25. doi:10.1101/gad.454008

45. Ghamari A, van de Corput MPC, Thongjuea S, et al. In vivo live imaging of RNA polymerase II transcription factories in primary cells. *Genes Dev*. 2013;27(7):767-777. doi:10.1101/gad.216200.113

46. Schoenfelder S, Sexton T, Chakalova L, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*. 2010;42(1):53-61. doi:10.1038/ng.496

47. Rieder D, Trajanoski Z, McNally JG. Transcription factories. *Front Genet*. 2012;3. doi:10.3389/fgene.2012.00221

48. Ma X, Ezer D, Adryan B, Stevens TJ. Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome Biology*. 2018;19(1):174. doi:10.1186/s13059-018-1558-2

49. Osborne CS, Chakalova L, Brown KE, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*. 2004;36(10):1065-1071. doi:10.1038/ng1423

50. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? *Nature Reviews Genetics*. 2009;10(7):457-466. doi:10.1038/nrg2592

51. Li G, Ruan X, Auerbach RK, et al. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*. 2012;148(1):84-98. doi:10.1016/j.cell.2011.12.014

52. Sandhu KS, Li G, Poh HM, et al. Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. *Cell Reports*. 2012;2(5):1207-1219. doi:10.1016/j.celrep.2012.09.022

53. Rodríguez-Carballo E, Lopez-Delisle L, Zhan Y, et al. The *HoxD* cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev*. 2017;31(22):2264-2281. doi:10.1101/gad.307769.117

54. McDowell IC, Barrera A, D'Ippolito AM, et al. Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res*. 2018;28(9):1272-1284. doi:10.1101/gr.233346.117

55. Yan C, Wu S, Pocetti C, Bai L. Regulation of cell-to-cell variability in divergent gene expression. *Nat Commun*. 2016;7. doi:10.1038/ncomms11099

56. Zhang L-F, Ding J-H, Yang B-Z, He G-C, Roe C. Characterization of the bidirectional promoter region between the human genes encoding VLCAD and PSD-95. *Genomics*. 2003;82(6):660-668. doi:10.1016/S0888-7543(03)00211-8

57. Yang W, Ng P, Zhao M, Wong TK, Yiu S-M, Lau YL. Promoter-sharing by different genes in human genome – CPNE1 and RBM12 gene pair as an example. *BMC Genomics*. 2008;9(1):1-16. doi:10.1186/1471-2164-9-456

58. Grob S, Cavalli G. Technical Review: A Hitchhiker's Guide to Chromosome Conformation Capture. In: Bemer M, Baroux C, eds. *Plant Chromatin Dynamics: Methods and Protocols*. Methods in Molecular Biology. Springer New York; 2018:233-246. doi:10.1007/978-1-4939-7318-7_14

59. Nagano T, Wingett SW, Fraser P. Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C. In: Kaufmann M, Klinger C, Savelsbergh A, eds. *Functional Genomics: Methods and Protocols*. Methods in Molecular Biology. Springer New York; 2017:79-97. doi:10.1007/978-1-4939-7231-9_6

60. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12). doi:10.1186/s13059-014-0550-8

61. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform*. 2016;17(6):953-966. doi:10.1093/bib/bbv110

62. Lin K-T, Wang L-H. New dimension of glucocorticoids in cancer treatment. *Steroids*. 2016;111:84-88. doi:10.1016/j.steroids.2016.02.019

63. Vockley CM, McDowell IC, D'Ippolito AM, Reddy TE. A long-range flexible billboard model of gene activation. *Transcription*. 2017;8(4):261-267. doi:10.1080/21541264.2017.1317694

64. McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput Biol*. 2018;14(1). doi:10.1371/journal.pcbi.1005896

65. D'Ippolito AM, McDowell IC, Barrera A, et al. Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids. *Cell Systems*. 2018;7(2):146-160.e7. doi:10.1016/j.cels.2018.06.007

66. Mourad R, Hsu P-Y, Juan L, et al. Estrogen Induces Global Reorganization of Chromatin Structure in Human Breast Cancer Cells. *PLOS ONE*. 2014;9(12):e113354. doi:10.1371/journal.pone.0113354

67. Silveira MAD, Bilodeau S. Defining the Transcriptional Ecosystem. *Molecular Cell*. 2018;72(6):920-924. doi:10.1016/j.molcel.2018.11.022

68. Durand NC, Shamim MS, Machol I, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*. 2016;3(1):95-98. doi:10.1016/j.cels.2016.07.002

69. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

70. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288-4297. doi:10.1093/nar/gks042

71. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*. 2016;48(5):488-496. doi:10.1038/ng.3539

72. Hsieh T-HS, Cattoglio C, Slobodyanyuk E, et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell*. 2020;78(3):539-553.e8. doi:10.1016/j.molcel.2020.03.002

73. Krietenstein N, Abraham S, Venev SV, et al. Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*. 2020;78(3):554-565.e7. doi:10.1016/j.molcel.2020.03.003