Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Mariana del Rosario Ruiz Velasco Leyva, M.Sc.

born in Mexico City, Mexico

Oral-examination: September 18th, 2018

In silico Investigation of Chromatin Organisation in Splicing, Ageing, and Histone Mark Propagation along DNA-Loops

Referees:

Dr. Jan Korbel

Prof. Dr. Benedikt Brors

In silico Investigation of Chromatin Organisation in Splicing, Ageing, and Histone Mark Propagation along DNA-Loops

Mariana Ruiz Velasco Leyva, M.Sc.

Supervised by Dr. Judith B. Zaugg

“Learning is a treasure that will follow its owner everywhere.”

Chinese proverb

Acknowledgements

Terminar el doctorado implica llegar a la cumbre de mi formación académica y poder por fin comenzar a descubrir un mundo de posibilidades en donde aplicar todos los conocimientos que he aprendido de tantas personas en estos años. Mi primer agradecimiento lo dedico a mi familia, quien una vez más valoró y confió en mi decisión de viajar al otro lado del mundo para poder seguir mis sueños. A pesar de la distancia, a pesar de los sacrificios que implica estar lejos de mi tierra, todo ha valido la pena. Muchas gracias por siempre respetar e impulsar mis decisiones y por siempre encontrar la forma de estar presentes en los momentos importantes. Los amo.

Finishing the PhD implies getting to the summit of my studies and opens an exciting world of possibilities for professionally applying the knowledge that I gained in these years. I owe this knowledge to several tutors and colleagues that throughout my studies invested hours of their time to teach me concepts, methods, or skills. My PhD would not be possible without the excellent mentorship that I received from my advisor, Judith Zaugg, who taught me valuable lessons and concepts from *literally* day one of the Zaugg group and during these four years. She prepared me for the next steps by trusting me and motivating me to always do my best and she showed me that working hard, having a positive attitude, and loving what you do is the key to success. I will always be grateful for her advice and her support.

I also want to thank my TAC members: Jan Korbel, Benedikt Brors, and Kiran Patil. I am thankful for the exciting discussions that happened during my TAC meetings, for your feedback and for your time. Thank you Marina Lusic for agreeing to be part of my thesis defence and for the nice conversations that we have shared in the past. Besides them, I want to thank Toby Gibson, Kyung-Min Noh, and Anne-Claude Gavin, who also shared their time, experience, and expertise to shape the projects that I had worked in during these years. Thank you Manjeet Kumar, Alejandro Reyes, Ximing Ding, Brian Lai, Marco Hennrich, Esa Pitkänen, Armando Reyes, Ignacio Ibarra, and Vasavi Sundaram for the exciting discussions and your willingness to collaborate with me.

During my time at the Zaugg group, I was able to work side by side with many amazing colleagues and to receive their feedback. I thank you all -current members and alumni- because besides working together we shared everyday experiences, worries, good, and bad times. First of all, I especially want to thank Armando, Christian, Pooja, Ignacio, Ece, Ana Belén, and Ivan, whom I consider to be the 'pioneers' of the Zaugg group with me, we should consider writing the *Memoirs* of the group some day! Together we experienced many nice moments, trips, coffee breaks, and talks; perhaps also because I shared office with the majority of them. I also want to thank Olga, Irene, Brian, Daria, Giovanni, Raphael, Rocio, and all the other alumni from the group for continuing what I described above. Valuable times came from coffee breaks, informal talks, and group activities. I cannot thank you all enough for all your efforts, your feedback and for igniting so many exciting discussions that make me love biology.

My time at the PhD wouldn't have been half as fun if it wasn't for my friends from EMBL class of 2014: Jessica, Sourabh, Joergen, Sascha, Nade, Marvin, and Lukas. It was always a journey with you, thank you for literally being my crew and for being in my life. I also thank you for extending the group with your amazing families; your friendship is something that I cherish.

Last but never least, I want to thank Alexander. Thanks to you I am able to see this day. Throughout this journey you were my rock, my teammate, and my love. I am also grateful to your family for their continuous support. Vielen Dank für Alles.

Table of Contents

Summary

A long-standing aim in biology is to elucidate how the genome is tightly compacted inside the eukaryotic nucleus while still retaining its capacity to orchestrate the correct functionality of the cell. While years of research have revealed that this three-dimensional structuring of DNA plays a major role in the transcriptional regulation, most of the existing studies have focused on long-range chromatin interactions, which are mainly established by the CCCTC-binding factor (CTCF), rarely centring at the gene level. Furthermore, our current knowledge on the interplay between structure and function remains largely descriptive with little mechanistic insight. In this dissertation I present three distinct computational studies which integrate multiple levels of molecular phenotype data in an attempt to gain further insights into the influence of chromatin organisation in (i) splicing regulation, (ii) in how distal genetic variants convey their signal, (iii) and an overall view of the misregulation of chromatin compaction in ageing stem cells.

Firstly, I describe a novel splicing mechanism whereby CTCF-mediated DNA-loops that are formed within genes facilitate exon inclusion. My results provide substantial evidence that intragenic loops regulate exon usage and that CTCF binding can be affected either by genetic variation across individuals or by the epigenomic landscape in different cell lines. Those exons being CTCF-regulated frequently overlap annotated protein domains and are enriched for being involved in cellular stress-response and signalling pathways. In summary, this study provides strong evidence for alternative exon usage being regulated by chromatin structure, and thus increases our understanding of functional consequences underlying variation in chromatin architecture.

In a second study, I show initial efforts to unravel the mechanisms that allow a genetic variant (distal-QTL) to confer its effect at distant regions through long-range interactions. By measuring allele-specific biases of various molecular phenotypes

occurring along chromatin interactions, I propose two models that intend to explain the propagation of this signal. In the “touch-and-act model” functionality is transmitted through the physical contact of both anchors, independent of the region inside the loop, while in the “spreading model” the function is propagated along the entire loop resulting in a coordinated activation or repression of the whole local neighbourhood. There is evidence for both models occurring at varying proportions, which are partially explained by transcription factor co-enrichments.

Finally, I present a study on how chromatin accessibility impacts the transcriptome and the proteome in mesenchymal stem cells (MSC) from human donors of multiple ages. I also observed a profound misregulation of chromatin organisation occurring with age, possibly due to a decrease in chromatin-related proteins such as histones, CTCF, CENPB, and lamins, which ultimately affect heterochromatin at centromeres and telomeres contributing to genomic instability. By subtle but significant changes in the transcription factor landscape of young and old MSCs, I observe a bias in the differentiation potential. Additionally, I show a loss of bivalent modifications at enhancer and promoter regions that correlate with DNA methylation changes and that could possibly contribute to a decrease in stemness with age.

In summary, I describe a novel splicing mechanism mediated by chromatin intragenic interactions, propose models of how distal-QTLs propagate histone marks, and advance the understanding of chromatin accessibility changes occurring with age in stem cells.

Zusammenfassung

Ein seit langem bestehendes Ziel in der Biologie ist es aufzuklären, wie das Genom innerhalb des eukaryotischen lang-distanzige kompaktiert ist, während es immer noch seine Fähigkeit beibehält, die korrekte Funktionalität der Zelle zu koordinieren. Während Jahre der Forschung gezeigt haben, dass diese dreidimensionale Strukturierung der DNA eine wichtige Rolle bei der Transkriptionsregulation spielt, konzentrierten sich die meisten der vorliegenden Studien auf langreichweitige Chromatininteraktionen, die hauptsächlich durch den CCCTC-Bindungsfaktor (CTCF) erzeugt werden, was nur selten zentriert auf der Genebene ist. Darüber hinaus bleibt unser derzeitiges Wissen über das Zusammenspiel von Struktur und Funktion weitgehend deskriptiv mit wenig mechanistischen Einsichten. In dieser Dissertation präsentiere ich drei verschiedene Computerstudien, die mehrere Ebenen von molekularen Phänotypdaten integrieren, um weitere Einblicke in den Einfluss der Chromatinorganisation zu erhalten bezüglich der (i) Spleißregulation, (ii) Art und Weise, wie distale genetische Varianten ihr Signal übermitteln und (iii) Gesamtansicht der Fehlregulation der Chromatinveraenderung in alternden Stammzellen.

Zunächst beschreibe ich einen neuartigen Spleißmechanismus, bei dem CTCF-regulierte DNA-Schleifen, die in Genen entstehen, die Exon-Inklusion erleichtern. Meine Ergebnisse liefern einen substanziellen Evidenz dafür, dass intragene Schleifen die Exon-Nutzung regulieren und dass die CTCF-Bindung entweder durch genetische Variation über Individuen oder durch die epigenomische Landschaft in verschiedenen Zelllinien beeinflusst werden kann. Diese Exone sind häufig CTCF-reguliert, überlappen häufig mit annotierten Proteindomänen und sind angereichert für Funktionen der zellulären Antwort auf Stress und Signalwegen. Zusammengefasst liefert diese Studie einen starken Evidenz dafür, dass die alternative Exon-Nutzung durch die Chromatinstruktur reguliert wird, und erhöht somit unser Verständnis der funktionellen Konsequenzen, die der Variation in der Chromatinarchitektur zugrunde liegen.

In einer zweiten Studie zeige ich ersten Bemühungen, die Mechanismen zu entschlüsseln, die es einer genetischen Variante (distal-QTL) ermöglichen, ihre Wirkung in distalen Regionen durch weitreichende Interaktionen zu erzielen. Durch die Messung allelsspezifischer Ausprägung verschiedener molekularer Phänotypen bezüglich Chromatin-Wechselwirkungen schlage ich zwei Modelle vor, die die Ausbreitung dieses Signals erklären können. Im "Touch-and-Act-Modell" wird die Funktionalität durch den physischen Kontakt beider Interaktionsanker unabhängig von der Region innerhalb der Schleife übertragen, während sich die Funktion im "Ausbreitungsmodell" entlang der gesamten Schleife ausbreitet, was zu einer koordinierten Aktivierung der gesamten lokalen Nachbarschaft führt. Ich habe Evidenz für beide Modelle beobachtet, die in variierenden Anteilen auftreten und teilweise durch Cofaktoren der Transkriptionsfaktoren erklärt werden können.

Schließlich stelle ich eine Studie vor, in der beschrieben wird, wie die Zugänglichkeit von Chromatin das Transkriptom und das Proteom in mesenchymalen Stammzellen (MSC) von menschlichen Spendern mehrerer Altersgruppen beeinflusst. Ich beobachtete eine tiefgreifende Fehlregulation der Chromatinorganisation mit zunehmendem Alter, möglicherweise aufgrund einer Abnahme von Chromatin-verwandten Proteinen wie Histonen, CTCF, CENPB und Laminen, die letztlich Heterochromatin in Zentromeren und Telomeren beeinflussen, was zur genomischen Instabilität beiträgt. Durch subtile, aber signifikante Veränderungen in der Transkriptionsfaktor-Landschaft von jungen und alten MSCs beobachte ich eine Verzerrung des Differenzierungspotentials. Zusätzlich habe ich einen Verlust an bivalenten Modifikationen in Enhancer- und Promotorregionen, die mit der DNA-Methylierung korrelieren und möglicherweise zu einer Abnahme der Stammzellfähigkeit mit dem Alter beitragen könnten.

Zusammenfassend beschreibe ich einen neuartigen Spleißmechanismus, der durch Chromatin-Interaktionen vermittelt wird, schlage Modelle vor, wie distal-QTL Histone markieren, und verbessere das Verständnis von altersbedingten Zugänglichkeitsveränderungen von Chromatin in Stammzellen.

List of Figures

Abbreviations

Numbers	
3D	Three-dimensional
A	
asCTCF	Allele Specific CTCF
ATAC-seq	Assay for Transposase-Accessible Chromatin
B	
BM	Bone Marrow
bp	base pairs
C	
CARs	Changing Accessibility Regions
CHiC	Capture HiC
CRD	Cis-Regulatory Domains
CTCF	CCCTC-Factor
D	
DE	Differentially Expressed (Genes)
DHS	DNase hypersensitivity
DMR	Differentially Methylated Region
DUE	Differentially Used Exon
E	
ESC	Embryonic Stem Cell
F	
FC	Fold Change
FDR	False Discovery Rate
G	
GO	Gene Ontology
GWAS	Genome Wide Association Studies
H	
HGPS	Hutchinson-Gilford progeria syndrome
hQTLs	Histone marks QTLs
HSC	Hematopoietic Stem Cell
I	
INDEL	Insertion/Deletion
K	
Kb	Kilobases (1000 base pairs)

L	
LADs	Lamina Associated Domains
LCL	Lymphoblastoid Cell Line
lncRNA	long noncoding RNAs
M	
mESC	mouse Embryonic Stem Cell
MSC	Mesenchymal Stem Cell
mNPC	mouse Neural Progenitor Cell
N	
NER	Nucleotide Excision Repair
NPC	Neural Progenitor Cell
nt	nucleotide
O	
OR	Odds Ratio
P	
PCA	Principal Component Analysis
PcG	Polycomb Group
PolII	RNA Polymerase II
Q	
QTL	Quantitative Trait Loci
S	
SD	Standard Deviation
SMC	Structural Maintenance of Chromosomes
SNV	Single Nucleotide Variant
SNP	Single Nucleotide Polymorphism
sQTL	Splicing Quantitative Trait Loci
T	
TAD	Topologically Associated Domain
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
TTS	Transcription Termination Site
W	
WT	Wild-type
Y	
yr	Years
Z	
ZF	Zinc Fingers

“The mind is not a vessel to be filled but a fire to be kindled.”

-Plutarch

1. Introduction: interplay between chromatin organisation and functionality in the mammalian nucleus

This chapter shows an overview of the current knowledge in the fast-paced field of chromatin organisation. Most of the text has been directly taken from a literature review that has been originally written by me:

***Mariana Ruiz-Velasco** and Judith B. Zaugg (2017) Structure meet function: How chromatin structure conveys functionality. Current Opinions in Systems Biology 1, 129-136.*

This article is open access and is available under the terms of the Creative Commons Attribution 4.0 License (CC BY). Parts of the text were also extended to include the literature following its publication. It also includes additional sections that summarise how chromatin organisation is destabilised with age.

The eukaryotic nucleus is the brain of a cell. This organelle is able to store the genetic information by compressing it to fit into a microscopic compartment, while preserving its capacity of dictating the functionality of the cell. Understanding how this compaction occurs, which mechanisms drive this organisation, and how in spite of the intricate condensation there is still order and an accurate molecular regulation, has led to decades of research. In the past years, however, major progress has been accomplished thanks to global consortia which have begun to elucidate and map chromatin interactions in several tissues and across dynamic biological processes. The same interest in gaining additional knowledge about chromatin structure, its relation to function, and how it is affected by dynamic biological processes like ageing or cancer has motivated this doctoral thesis.

1.1 Chromatin organisation at different scales

Chromatin is the physiological structure composed of DNA and histone proteins (Hake and David Allis, 2006). Depending on the cell cycle stage, chromatin is able to linearly compact in the mammalian nucleus about 200-1,000-fold during interphase (Lawrence et al., 1990) and up to about 10,000- to 20,000-fold in the mitotic state to form the 30nm fiber (Swedlow and Hirano, 2003). While this compression is tremendous chromatin 3D-organisation is not random. Years of research in this field have demonstrated that certain genomic locations show a preferential physical interaction with others, as I will discuss below. Therefore, understanding how the genome is compacted in a highly ordered structure that fits in the eukaryotic nucleus, while keeping its capacity of orchestrating the functionality of the cell, is one of the long-standing questions in biology.

Three dimensional chromatin organisation has been extensively studied at multiple scales, ranging from whole chromosome structures to interactions across a few kilobases. In our review (Ruiz-Velasco and Zaugg, 2017), we summarized our current knowledge of

3D-organisation, the mechanisms of how it is established and maintained, and the functional impact it has on gene regulation and complex phenotypes. We also described the various levels of genome architecture by grouping them into (i) large-scale chromatin domains, (ii) intermediate-scale (enhancer-promoter) loops, and (iii) intragenic interactions (Figure 1.1). These three levels are classified by the length of the interaction and encompass various well-described structures in the nucleus, where each level comprises specific characteristics.

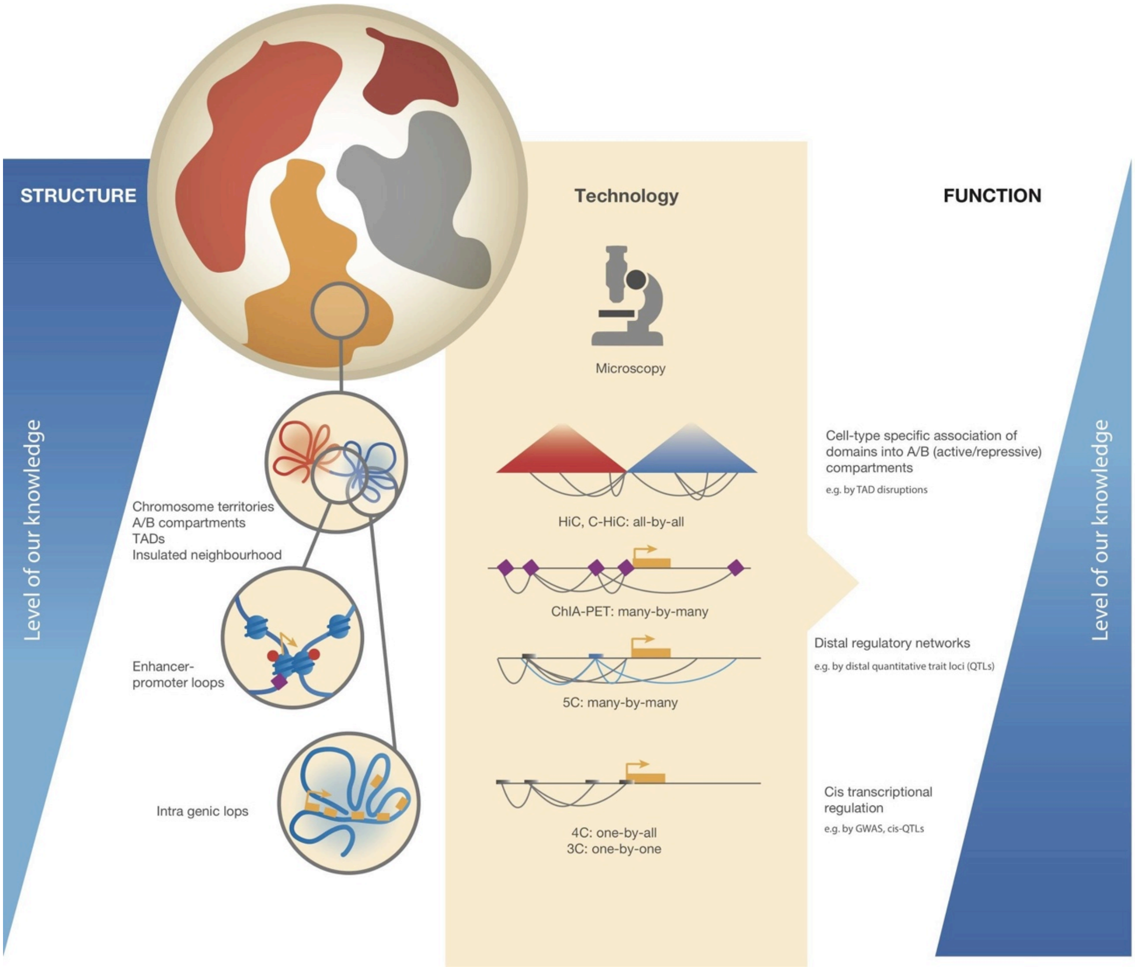


Figure 1.1. Schematic about our current knowledge across the different scales of chromatin organisation for both structure and function. While we know much about the structure of the highest levels and little as we zoom in, the opposite is true for our knowledge about the functional impact of chromatin architecture. The technologies with which each level is typically investigated are mentioned in the middle. This figure was previously published (Ruiz-Velasco and Zaugg, 2017).

1.1.1 Large-scale organisation: topologically associated domains and subcompartments

On the most coarse level genomes are organised into structures known as chromosome territories (Cremer and Cremer, 2001), which separate euchromatic ("active" and "open") from heterochromatic ("repressed" and "closed") regions. While this sub-chromosomal structures were described in the 80s (Cremer and Cremer, 2001), these compartmentalisation was also observed using high-throughput chromosome conformation capture technologies (C-technologies). Euchromatin and heterochromatin domains coincide with the recently termed 'A' and 'B' subcompartments (Lieberman-Aiden et al., 2009). The former is associated to active histone modifications, presence of genes, and higher expression, while the latter pertains to repressive compartments, closed spatial locations, and nuclear lamina (Lieberman-Aiden et al., 2009). These compartments can be further subdivided into A1-A2 and B1-B4 according to their chromatin signature and replication time (Rao et al., 2014).

Chromatin is organised in an almost fractal structure within the chromosome territories (Bancaud et al., 2012) with megabase-sized domains being partitioned into smaller and smaller sub-domains that are currently known as topologically associated domains (TADs) (Dixon et al., 2012; Rao et al., 2014). TADs, were originally proposed based on C-technologies and defined as linear regions that display a high number of physical interactions within and little contact with regions outside (Dixon et al., 2012). TADs seem to behave as "units", both in terms of regulation, as they typically belong to only one compartment (either A or B) in a given cell type (Dixon et al., 2015), and in replication, as they share a near one-to-one correspondence with replication domains (Pope et al., 2014).

Interestingly, whereas TAD boundaries tend to be stable over many cell divisions, invariant across cell types, and conserved across species (Dixon et al., 2016; Melé and

Rinn, 2016; Vietri Rudan et al., 2015), their association with either A or B compartment is highly plastic across cell types (Dixon et al., 2015) and development (Dixon et al., 2015; Pękowska et al., 2014). However, even though it is helpful for data visualisation and interpretation to demarcate chromatin regions into TADs, pinpointing their precise boundaries has proven difficult. This is due to the fractal organisation of chromatin structure, which makes the definition of boundaries highly dependent on arbitrary thresholds and the choice of parameters (Kruse et al., 2016) (Figure 1.2).

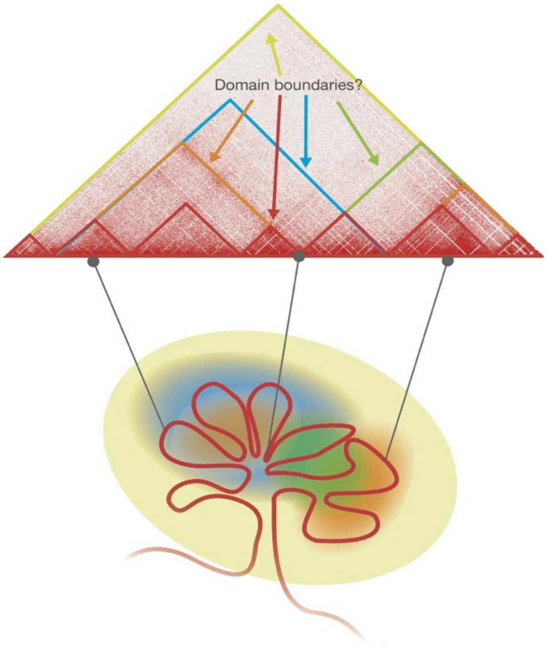


Figure 1.2. The definition of domain boundaries by binning algorithms creates the triangular pattern typical of TADs. However, the precise boundaries are often hard to delimit.

An additional complication is that high-throughput experiments typically pool cells across all cell cycle stages. Since TADs are only being established at early G1, visible during interphase and G2, and lost during mitosis (Dileep et al., 2015) they are thus inferred by averaging across a heterogeneous population of TAD maturation. Despite these difficulties, one general signature of TAD boundaries has emerged: they are strongly bound by cohesin and CTCF (Dixon et al., 2015; Rao et al., 2014), which I will further discuss in this text. Indeed, the nested sets of cohesin-associated CTCF-CTCF loops (recently termed 'insulated neighborhoods' (Dowen et al., 2014; Ji et al., 2016)) can be considered as TADs with well demarcated boundaries.

1.1.2 Intermediate scale organisation: Enhancer-promoter interactions

One of the best studied classes of intra-TAD contacts are enhancer-promoter loops that are typically facilitated by CTCF and cohesin, which enable the interaction of enhancer- and promoter-bound transcription factors (TFs) (Dowen et al., 2014; Hnisz et al., 2016). Enhancer-promoter interactions play an essential role in orchestrating the lineage-specific gene expression programs that underlie cellular identity, recently reviewed in (Gorkin et al., 2014). Interestingly, the physical contacts sometimes precede the activation of the target gene during the developmental process (Ghavi-Helm et al., 2014).

Even though enhancers have the ability to function in either direction and over long distances, their activity is typically restricted to gene targets within the same TAD, as shown by single-loci studies (Lupiáñez et al., 2015; Tsujimura et al., 2015). This has been reproduced in a genome-wide study that identified genetic variants that regulate histone marks (hQTLs) over long distances to reside within the same TAD as their target regulatory elements (Grubert et al., 2015), and in a transposon based screen that measured transcriptional activity at random genomic locations and found enhancers to activate genes pervasively within but very rarely across TADs (Symmons et al., 2014). TADs may thus explain how enhancers are prevented from establishing ectopic interactions.

1.1.3 Intragenic organisation: gene promoter-terminator loops

The chromatin architecture within a single gene represents the smallest level of organisation which, due to limitations in resolution, is difficult to study using an unbiased genome-wide approach. Therefore, insights come mainly from single-locus studies using 3C, 4C, or from factor-specific ChIA-PET experiments. ChIA-PET on CTCF and on Pol II have revealed ubiquitous chromatin looping within genes, in particular between promoters and intronic regions (Mercer et al., 2013). Together with the fact that CTCF is highly prevalent at transcription start sites (Bornelöv et al., 2015; Ruiz-Velasco et al., 2017; Wang et al., 2012), this suggests a role of CTCF in intragenic gene regulation. In particular, this is where I centered a big fraction of my research during the PhD and which I describe in detail in chapters 2 and 3, by describing how intragenic interactions have an effect on alternative splicing, therefore linking structure with function.

1.2 Mechanisms needed to shape and maintain the nuclear architecture

Even though the precise mechanisms of chromatin structure formation are not fully understood, two of the most important players, cohesin and CTCF, are well known.

1.2.1 CTCF and its binding grammar

CTCF is a ubiquitously expressed and embryonic lethal protein that encodes 11 zinc fingers (ZF) and which binds a well-defined core motif of 15 base pairs (Moore et al., 2012; Nakahashi et al., 2013). CTCF-mediated loop formation requires two CTCFs, each bound at one end of the chromatin loop to dimerize, which is only possible if they are facing each other in a convergent orientation (de Wit et al., 2015; Guo et al., 2015; Rao et al., 2014). Interestingly the use of diverse combinations of its ZFs as binding sub-domains has been shown to allow CTCF to interact with a number of diverse DNA sequences, mostly comprising the core motif and either upstream or downstream sub-motifs, in which has been described as the ‘CTCF-binding code’ (Nakahashi et al., 2013).

Importantly the disruption of CTCF can unravel several consequences and ultimately has been linked with disease states and cancer (Norton and Phillips-Cremins, 2017). On the one hand mutations affecting the core motif at the DNA sequence level -

which is recognised by ZF3-ZF7 in human (Hashimoto et al., 2017) - show the most dramatic reductions in CTCF binding, while disruption of sites recognised by ZF9-11 decrease its binding but to a different extent. On the other hand the binding of CTCF can be regulated through epigenetic modifications, in particular through DNA methylation. Of relevance, a recent crystallographic study demonstrated an existing bias for the effect of DNA methylation within the core motif, with CpGs reducing CTCF binding affinity when they occur at position 2, but increasing its association if they fall at position 12 (Hashimoto et al., 2017).

Ultimately the role of CTCF has shown to be at the center of chromatin organisation helping to set up the specific loops which then define chromatin structure from small-scale loops to TAD boundaries, and in particular in the loop extrusion process as evidenced by its preferential location to TAD boundaries (Dixon et al., 2012) and by loss-of-function experiments that demonstrate its involvement in loop anchoring through convergent interactions (de Wit et al., 2015; Guo et al., 2015).

1.2.2 Structural Maintenance of Chromosomes protein complexes

Structural Maintenance of Chromosomes (SMC) complexes comprise multiple subunits that form a ring-like structure and which include cohesin and condensins (Yuen and Gerton, 2018). Cohesin forms and stabilises unspecific chromatin interactions (Zuin et al., 2014) and it typically interacts with CTCF. This interaction is primarily driven through the CTCFs C-terminal tail (Xiao et al., 2011), which may allow it to place cohesin on a particular side of the chromosomal domain and thereby anchor and stabilize the chromatin loop (Vietri Rudan et al., 2015).

Besides cohesin, recent evidence has shown that condensin plays a major role in organising chromatin by serving as a molecular motor for loop extrusion (Ganji et al., 2018). However, major differences exists between cohesin and condensin which makes it necessary to have additional and substantial evidence to proof that cohesin is also capable of functioning as a motor in the loop extrusion process.

Condensin is able to load and unload intrinsically by encircling DNA double helix and securing it with a ‘safety belt’ mechanism to prevent its dissociation in mitotic chromosomes (Kschonsak et al., 2017). Meanwhile, cohesin is loaded onto DNA by the SCC2/SCC4 complex (also known as NIPBL and MAU2), and to released from DNA by WAPL (Haarhuis et al., 2017). Another factor to consider is that interphase, where most cohesin-based assumptions are based, and mitosis, where condensin has been shown to be crucial, are two stages which account for very different fractions of the cell-cycle. Therefore, condensin processivity during loop extrusion must require a faster speed than the one needed for cohesin, as recently discussed by Prof. T. Hirano in a seminar.

1.2.3 Loop extrusion

The most convincing model of chromatin structure formation proposed to date is the “loop extrusion model” (Fudenberg et al., 2016; Sanborn et al., 2015). This model initially stated that an “extrusion complex” - of unknown structure but which contains two tethered DNA-binding subunits - is loaded onto chromatin in close proximity to form a tiny chromatin loop. DNA sliding in opposite directions would make the loop grow bigger until it hits appropriately oriented CTCF motifs where it binds tightly and halts the extrusion complex. The speculated extrusion complex includes a cohesin-CTCF pair that is either loaded simultaneously, or sequentially to slide in opposite directions simultaneously.

Various studies trying to further understand this mechanism and to validate the model have recently emerged. The most conclusive proof to date is a study using real-time imaging in which they visualise the formation of DNA-loops with condensins (Ganji et al., 2018). Here they observe that an "extrusion complex" consisting of condensin, ATP, Mg2+, and DNA, which is stably anchored in one extreme, is asymmetrically translocated by the motor site of condensin, forming a loop that can extend up to several kilobases. The chromatin loop will continue extruding until it hypothetically encounters appropriately oriented CTCF motifs (in a convergent orientation) where it binds tightly and halts the extrusion complex (Fudenberg et al., 2016) or when condensin spontaneously releases the extruded loop (Ganji et al., 2018).

The previous study was able to demonstrate that loop extrusion can occur upon a single condensin complex, which anchors DNA through a charged groove formed by the Ycg1 HEAT-repeat and Brn1 kleinsin subunits as a safety belt to start translocating the DNA at an in vitro rate of up to ~1.5kb/s (Ganji et al., 2018; Kschonsak et al., 2017). While additional studies have proposed that other SMC proteins could potential have extruding capacity, like cohesin or the Smc5/6 complexes (Yuen and Gerton, 2018), the most conclusive evidence that relates cohesins to loop extrusion comes from HiC experiments done in HAP1 cells for wild-type, WAPL-mutants, and SCC4-mutants, which demonstrate that loop size results from the dynamics of loading/unloading cohesin into chromatin (Haarhuis et al., 2017). Therefore further studies need to be done to validate that cohesin is also a loop extrusion motor in interphase chromosomes.

In addition to cohesin and CTCF, other proteins, repetitive elements, and histone marks associated with transcription are enriched at TAD boundaries although cause and effect of these associations often remain speculative. Co-binding of CTCF with YYI (Moore et al., 2015; Schwalie et al., 2013), ZNF143 and Polycomb group proteins (Mourad and Cuvier, 2016) may contribute in the establishment of TAD boundaries. Repetitive elements could act as specific anchor points that spatially organize chromosomes (Cournac et al., 2016; Darrow et al., 2016; Giorgetti et al., 2016). The enrichment for the transcription marks H3K4me3 and H3K36me3 in TAD boundaries suggests an association with highly expressed regions (Moore et al., 2015) and transcription itself has been suggested to play a role in TAD organisation as reviewed in (Melé and Rinn, 2016). One well studied example is the *Xist* RNA, whose transcriptional induction initiates the formation of TAD boundaries and the loss of DNA accessibility in the inactive X chromosome (Giorgetti et al., 2016).

Another mechanistic question is how individual TADs switch between compartments as observed in cellular differentiation processes. One possibility is that chromatin domains get repositioned with respect to chromosome surfaces, with an overall tendency of active genomic regions involved in inter-chromosomal contacts (Osborne et al., 2004; Pękowska et al., 2014) and heterochromatic regions to localise to the nuclear periphery in structures known as 'lamina associated domains' (LADs) (Peric-Hupkes et al., 2010). The direction and degree of repositioning of chromatin domains appears to be determined by associated trans-acting factors and is suggested to depend on the surrounding genomic context, as demonstrated by the repositioning of sub-TADs into different nuclear compartments depending on the recruiting factor (either NANOG, EZH2, or SUV39H1) (Wijchers et al., 2016). Interestingly, establishing cell-type specific chromatin interactions seems to require cell division (Neems et al., 2016). Unsurprisingly, given its prominent role in chromatin structure organisation, CTCF is likely involved in the process of compartment switching, as shown for circadian loci where PARP1 interacts with CTCF to regulate dynamic chromatin interactions with LADs, which promotes transcriptional oscillations (Zhao et al., 2015).

In mammals around 15% of CTCF-binding sites are located at TAD boundaries. The remaining 85% that reside within TADs are equally frequently located in intragenic and intergenic regions. This distribution suggests that CTCF plays an important role in nuclear organisation at all levels of chromatin organisation, and that changes in its binding pattern are likely to contribute to the regulatory programs that control the differences across tissues, species, and developmental stages. Interestingly, we have recently observed that CTCF binding is much more variable between individuals within TADs than at TAD boundaries (Chapter 4, Figure 4.1).

This suggests that the regulation of the highest level of chromatin organisation is less dependent on CTCF binding than on the different factors described above, while intra-TAD interactions are more likely regulated by mechanisms that directly affect CTCF binding affinity, such as DNA methylation. This is supported by the fact that methylation-sensitive CTCF sites are mostly within TADs (Maurano et al., 2015; Wang et al., 2012).

1.3 Elucidating the function of 3D-chromatin architecture

The robust evidence that chromatin topology is pervading cellular function comes from a few recent studies that linked DNA mutations at TAD and loop boundaries to dramatic phenotypic changes. However, overall we know much less about the functional impact of chromatin architecture than what we know about its structure.

1.3.1 Functional impact of TAD rearrangements

As described above, TADs tend to act as regulatory units, either belonging to compartment A (active) or B (inactive); it is thus a tempting hypothesis that genes involved in a specific process might be contained within a single TAD to allow smooth switching between cellular programs. Indeed, some TADs are enriched for lineage-specific genes (Neems et al., 2016) thus potentially facilitating the regulation of cell differentiation. Similarly, poor reconfiguration of specific chromatin contacts was identified as a potentially rate-limiting step in reprogramming of neural progenitor to induced pluripotent stem cells (Beagan et al., 2016).

In addition to these descriptive studies, the recent progress in genetic engineering with the CRISPR/Cas9 system allowed to assess the consequences of removing individual contacts or entire TAD boundaries at specific loci, which revealed that disrupting TAD boundaries can result in aberrant chromatin domain topology. The effects of these disruptions range from transcriptional misregulation (Narendra et al., 2015; Nora et al., 2012; Tsujimura et al., 2015) to whole-organism pathogenic phenotypes, such as polydactyly (Lupiáñez et al., 2015). In fact, already the disruption of specific single intra-TAD interactions has functional consequences on gene expression (de Wit et al., 2015; Guo et al., 2015). Similarly, the disruption of the cohesin complex led to deregulation of gene expression at the *Cd3* super-enhancer region as a result of partial dispersal of enhancer elements (Ing-Simmons et al., 2015).

1.3.2 Distribution of functional genetic variants at TADs and loop boundaries

Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) associated with diseases (Hindorff et al., 2009). GWAS SNPs, most of which lie in the non-coding genome, are highly enriched for quantitative trait loci (QTLs) for molecular phenotypes such as gene expression, histone marks, or DNA methylation ((Grubert et al., 2015; GTEx Consortium, 2015; Karczewski et al., 2013; Lappalainen et al., 2013; Maurano et al., 2012; Waszak et al., 2015), among others). Many of these molecular QTLs, often being far from gene promoters, affect the activity of distal genes and other regulatory elements with which they are physically interacting (Grubert et al., 2015).

A tempting explanation of how GWAS SNPs distal to any gene may affect complex phenotypes is by disrupting either CTCF or cohesin binding sites that are anchoring DNA-loops, which in turn may lead to mis-regulation of various molecular, and, ultimately, complex phenotypes. In support of this hypothesis CTCF and cohesin sites that are variable across individuals (Kasowski et al., 2013), regulated by a QTL (Ding et al., 2014), or physically interacting with promoters (Mifsud et al., 2015) are significantly enriched for GWAS SNPs.

1.3.3 Influence of chromatin architecture in cancer genomes

In addition to common diseases, CTCF binding sites have also been identified as major mutational hotspots in the noncoding cancer genome (Katainen et al., 2015) with recurrent mutations in CTCF sites occurring most frequently at TAD boundaries (Hnisz et al., 2016). Strikingly, in T-cell acute lymphoblastic leukemia, such perturbations have been shown to activate proto-oncogenes due to inappropriate enhancer-promoter interactions (Hnisz et al., 2016). I will further discuss the impact of somatic mutations affecting CTCF motifs and therefore potentially disrupting intragenic loops in chapter 3.

1.4 Understanding how the 3D architecture conveys its functional impact

The general mechanisms of how chromatin architecture impacts on downstream molecular processes, as in the examples cited above, are often still poorly understood. Here we review the current understanding and propose two potential models that could be useful for further studying the mechanism of how chromatin structure exerts its function.

The interplay of CTCF with general transcription factors is one mechanism of how chromatin structure may regulate transcription, as evidenced by the observation that TFII-I is responsible for targeting CTCF to the promoter of metabolic genes, which in turn triggered their expression and altered the cells response to metabolic stress (Peña-Hernández et al., 2015). Another mechanism, as proposed for the *SMAD4A* locus, is a CTCF-dependent promoter-gene body loop that accelerates transcriptional activation e.g. upon stimulation (Larkin et al., 2012).

Apart from isolated examples like these, however, it often remains unclear whether structure precedes function or vice versa. While most observations are consistent with an interdependence of structure and function, a recent study suggested that the picture is even more complex by finding that the transcriptional status of a region and its compartment membership are not always causally related (Wijchers et al., 2016).

On a genome-wide level we and others have recently shown that chromatin variation across individuals, often caused by a genetic perturbation of single TF-DNA interactions, can act as a seed for coordinated regulatory changes mediated largely by long-range chromatin contacts (Grubert et al., 2015; Waszak et al., 2015). In chapter 4 we propose two models that could explain these observations: the “touch-and-act” and the “spreading model”. We also test potential mechanisms that could be affected by distal-QTLs like CTCF dimerisation or protein-protein interactions involved in transcriptional activation.

1.5 Chromatin organisation is affected with age

Ageing is characterized by the progressive functional decline at all organismal levels (Booth and Brunet, 2016). The time-dependent accumulation of cellular damage is widely considered to be the general cause of ageing (López-Otín et al., 2013). Multiple years of research aiming to understand the precise cellular and molecular causes of ageing at multiple biological scales can be summarised in a series of common processes that has ultimately been named as the 'hallmarks of ageing' (Booth and Brunet, 2016; López-Otín et al., 2013). Of important notice, genomic instability, stem cell exhaustion, and telomere attrition are some of these hallmarks, which we will bring back in detail in Chapter 5.

Although many studies have tried to identify genes that confer longevity or robustness in various organisms, some theories on the etiology of ageing challenge this view mainly by arguing that natural causes would eliminate most animals before they grow old and if there are such 'beneficial' genes, they should be pleiotropic and confer additional advantages early in life (Yuen and Gerton, 2018). Consequently, ageing should comprise mostly epigenomic changes and it should be subjected to a great influence from the environment and lifestyle of the organism. In this line, changes occurring at all levels of chromatin organisation have been reported, involving major rearrangements of LADs, changes in the transcriptional regulatory landscape, and even alternative splicing changes within genes, which we discuss in chapter 5.

Firstly, the implication of nuclear lamins in the ageing process came from the Hutchinson-Gilford Progeria Syndrome (HGPS), characterised by the constitutive production of progerin, a mutant form of lamin A, that causes premature ageing (Scaffidi and Misteli, 2006, 2008). HGPS affects several tissues, particularly those of mesenchymal origin and causes changes in the differentiation potential, with enhanced osteogenesis and decreased adipogenesis (Scaffidi and Misteli, 2008). Moreover, the same molecular mechanism responsible for HGPS act at a low level in healthy cells (Scaffidi and Misteli, 2006), implicating lamins in the normal physiological ageing. Loss of lamin B1 has also been shown to decrease with age and proposed as a marker for cellular senescence in vitro and in vivo (Dreesen et al., 2013).

The transcriptional regulation is also affected by age in multiple ways. For example, DNA methylation has also a profound contribution in ageing, as has been described extensively by (Horvath, 2013). Additionally, we know that some TFs show changing levels like FOXO3, which is linked to metabolism and for which a genetic variant is commonly found in centenarian humans (Willcox 2008) or affecting NRF, which causes the activation of genes involved in cellular protection (Booth and Brunet, 2016). Interestingly NRF is affected by DNA methylation (Domcke et al., 2015) and some TFs in this family (NRF2) can bind SWI/SNF nucleosome remodelling complexes (Booth and Brunet, 2016).

Finally specific reports of changing isoforms with age exists for genes such as TP53 (von Muhlinen et al., 2018) or mTOR kinase (Razquin Navas and Thedieck, 2017). However the knowledge of the extent to which alternative splicing is altered with age is still under investigation, some of these findings are described in this dissertation.

1.6 Using 3D-architecture to understand human variation and disease

We predict that unraveling how chromatin structure is formed and maintained, and how it affects downstream molecular processes - either through genetic or epigenetic variation - will provide important insights for understanding common genetic diseases. Towards this end, studying the effect on intermediate molecular phenotypes will be key to understand the consequences of perturbing the 3D architecture on organismal traits and diseases. Fortunately, the increasing availability of chromatin conformation data along with transcriptome, ChIP-seq, and genotype data will provide a vast repertoire of information to answer the questions posed here. The implications of such research will be relevant to study changes in the nuclear organisation not only during development and differentiation, but also in pathogenic cellular processes like cancer, aging, and complex diseases.

2. CTCF-mediated intragenic chromatin loops between promoter and gene body regulate alternative splicing across individuals

This chapter discusses my main project during the PhD and details how we identified, characterised, and explored a new transcriptional regulatory mechanism, where CTCF-mediated interactions within genes are able to affect splicing decisions. The text of the following chapter has been originally written by myself and was taken and adapted from:

Mariana Ruiz-Velasco, *Manjeet Kumar, Mang-Ching Lai, Pooja Bhat, Ana-Belén Solís-Pinson, Alejandro Reyes, Stefan Kleinsorg, Kyung-Min Noh, Toby J. Gibson, and Judith B. Zaugg (2017). CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. Cell Systems 5, 628-637. e6.*

The sections including results from collaborators are explicitly stated and mostly involve Dr. Manjeet Kumar and Dr. Mang Ching Lai with advice from Dr. Toby Gibson and Dr. Kyung-Min Noh (all from EMBL). Additionally, I received support for the analyses from Dr. Alejandro Reyes, Pooja Bhat, Ana-Belén Pinson Solís, and Stefan Kleinsorg. *Cell Systems,* as part of *Elsevier,* states that I retain the right to include text and figures in my thesis, provided it is not published commercially and that the reference is present.

2.1 Introduction

As discussed in the first chapter, it has become evident that the 3D-organisation of chromatin is highly regulated and is thus likely to have a functional role in downstream cellular processes, such as gene regulation (Neems et al., 2016; Nora et al., 2012). However, our knowledge about the functional impact of chromatin architecture has largely remained descriptive, as most studies have focused on describing long-range enhancer-promoter interactions or megabase resolution chromatin domain organisation (Beagan et al., 2016; de Wit et al., 2015; Dixon et al., 2012; Grubert et al., 2015; Ji et al., 2016; Rao et al., 2014), with only a few examples that provide a mechanistic understanding for single loci (Monahan et al., 2012; Ruiz-Velasco and Zaugg, 2017; Symmons et al., 2014).

These studies have revealed an extensive function for CTCF in forming chromatin loops (de Wit et al., 2015; Guo et al., 2015; Rao et al., 2014). Therefore, the relative convergent orientation of a pair of CTCF binding sites restricts its choice of interacting partner and can thus be used to predict whether they can potentially form a chromatin-loop.

Here we proposed a novel functional role for CTCF-mediated intragenic chromatin interactions in regulating alternative exon usage. Overall, by integrating multiple layers of genome-wide molecular phenotype data with genotypes, as well as curated genome and protein annotations, our study proposes a functional role for CTCF-mediated intragenic chromatin looping in alternative splicing, particularly regulating the inclusion of domains in modular proteins involved in signalling and cellular stress response.

2.2 Results

2.2.1 Intragenic CTCF motif orientation suggests pervasive promoter-exon looping

Here we sought to assess the potential of CTCF-dependent chromatin-looping on the intragenic scale. To do so, we first examined the distribution and orientation of CTCF motifs within genes. For this purpose we defined four regions of interest (detailed in section 2.4.1):

- *promoter* window: +/- 1kb from transcription start site (TSS)
- transcription termination site *(TTS)* window: +/- 1kb of TTS
- intronic windows *upstream* of *exons*: 2kb intronic regions upstream of exon 5' end
- intronic windows *downstream* of *exons*: 2kb intronic regions downstream of exon 3' end

First exons, first introns and anything within 5kb from the TSS were filtered out.

We next scanned for CTCF binding sites within these windows (Methods: 2.4.1), and to minimise the number of non-functional CTCF binding sites, we only considered sites that overlap with a CTCF peak in Lymphoblastoid cell lines (LCLs) obtained from a previous study (Kasowski et al., 2013). Directionality was assigned relative to the gene strand defining *sense* CTCF motifs if both the motif and the gene had the same orientation and *antisense* motifs vice versa. Strikingly, we observe a strong bias for *sense* CTCF motifs at promoters (p-value = 7e-7, binomial test) with a very sharp increase within 500bp upstream of the TSS (p-value = 8.3e-11), while the window upstream and, to a much lesser extent, downstream of exons contained preferentially antisense motifs (p-value = 2e-5 and 1e-2, respectively; Figure 2.1A). Consistently, when analysing whole introns (scaled to the same length) we find the fraction of antisense motifs to increase with proximity to the 5' exon boundary (Figure 2.1B). TTS windows showed very few CTCF sites overall and no preference for any motif orientation. Interestingly, we did not find any evidence for such a pattern to occur in long noncoding (lnc)RNAs (Figure S2.1A), indicating that the three-dimensional chromatin structure for lncRNA genes is fundamentally different from protein coding genes. The following analyses are therefore based on protein coding genes only.

The strong enrichment observed for sense CTCF motifs at promoters and antisense motifs proximal to exons suggests pairs of convergent CTCF bound sites between promoter and exon proximal regions, which in turn would trigger the formation of exon-promoter chromatin loops. To evaluate this, we grouped motifs into the following pairs according to their genic region: promoter exon-upstream, promoter exon-downstream, promoter-TTS, exon-upstream TTS, and exon-downstream TTS (schematic in Figure 2.1B). And indeed, when we classified these CTCF-motif pairs as "*convergent*", "*divergent*", "*both sense*", and "*both antisense*", we found both classes of promoter-exon pairs strongly enriched for the convergent configurations (Odds Ratio (OR) =1.5 and 1.3, p-value = 3.4e-4, and 1.4e-2 for exon-upstream and downstream respectively, Fisher's exact test, Figure 2.1C) and significantly depleted for divergent pairs. In line with exons containing preferentially antisense motifs, the exon-TTS motif-pairs were enriched for divergent pairs (OR=1.5 and 1.8, p-value = 8.3e-3 and 8.6e-5 for exon-upstream and -downstream) and depleted for convergent pairs, while very few pairs were observed for the promoter-TTS configuration. Together, these results provide evidence for CTCF-mediated DNA loop formation between promoters and exon proximal regions; overall we identified 181 promoter exon-upstream loops involving 136 promoters. By considering a less stringent threshold (i.e. lowering the motif score and considering a broader promoter region of +/- 1kb), the number of potential exon-promoter loops increases to 1,463 involving 956 promoters (*LCL-specific* set; Figures S2.1B-C, methods: 2.4.1).

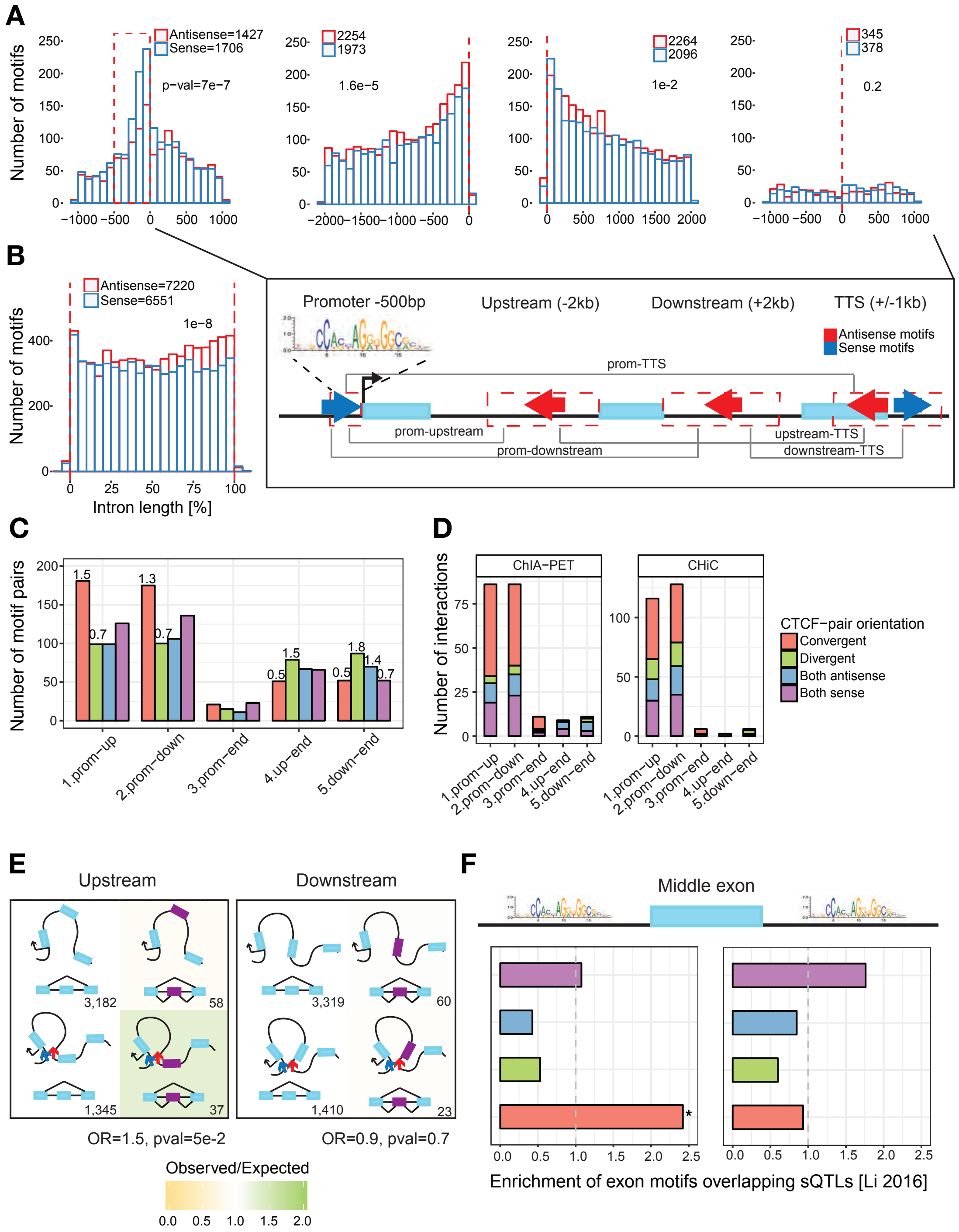


Figure 2.1. CTCF motifs show preferential directionality along transcripts

Distribution of motifs in sense (blue) and antisense (red) orientation for **(A)** four genic regions (see schematic): promoter, upstream of exon, downstream of exon, and TTS, and **(B)** along introns [%]. **(C)** Orientation of CTCF interactions based on the motif pairs: promoter-exon:upstream, promoter-exon:downstream, promoter:TTS, exon-upstream:TTS, and exon-downstream:TTS. **(D)** Number of motif pairs with physical interaction based on ChIA-PET data of Rad21, H3K4me3, and CTCF (left), and CHiC (right). **(E)** Schematic of contingency table for testing the association of DUEs and promoter-exon:upstream (left) and promoter-exon:downstream (right) loops. Background color represents ratio of observed vs expected. **(F)** Enrichment of sQTLs (FDR=10%) overlapping with a particularly oriented CTCF motif (+/-20bp from midpoint) upstream (left) or downstream (right) of a middle exon (star = p-value < 0.05; Fisher's exact test). This figure was produced by myself and is published in (Ruiz-Velasco et al. 2017).

To validate the loops predicted from the motif analysis, we obtained ChIA-PET data for the cohesin subunit Rad21 and histone mark H3K4me3 (Grubert et al., 2015), ChIA-PET for CTCF (Tang et al., 2015), and promoter capture-HiC (CHiC) data (Mifsud et al., 2015). For all datasets, the physical contacts were measured in one of the cell lines that was used in the analysis above (GM12878). Using these contact maps, we captured 57% of our convergent predicted loops, a reasonable number given the high false negative rate in ChIA-PET and CHiC data (Phanstiel et al., 2015). In accordance with our predictions, we observed that most of the interactions occur between promoters and exon-proximal regions and that most of these interactions have a convergent orientation (OR=7.4 for ChIA-PET and 1.6 for CHiC, p-value = 2.5e-9 and 4e-2, Fisher's exact test; Figure 2.1D).

Given this evidence of CTCF-mediated promoter-exon loops along with a previous report associating CTCF with alternative splicing at the *CD45* locus (Shukla et al., 2011), we hypothesised that CTCF might regulate alternative exon usage through the formation of intragenic promoter-exon loops. In support of this hypothesis, we found that exons involved in a predicted promoter exon-upstream loop (convergent CTCF motifs) are more likely to be alternatively used across individuals (OR=1.5, p-value = 5e-2, Fisher's exact test; Figure 2.1E). No such enrichment is observed for promoter exon-downstream loops. Interestingly, a recent study found that genetic variants affecting alternative splicing across individuals (splicing quantitative trait loci, sQTLs) often coincide with QTLs for CTCF (Li et al., 2016), suggesting a genome-wide role of CTCF in alternative splicing. Remarkably, when we overlapped the sQTLs with exon-upstream CTCF binding sites forming pairs with promoters, we observed a strong enrichment for sQTLs to participate in convergent interactions (OR=2.4, p-value = 2e-2, Fisher's exact test), whereas all other configurations - including all of the downstream pairs - were not enriched for sQTLs (Figure 2.1F).

Together, these results suggest a relationship between CTCF-mediated promoter-exon looping and alternative exon usage.

One important prediction of such a mechanism is, that alternative exon usage will occur more often for exons that have the potential to form a loop with their promoter, i.e. convergent pair of CTCF motifs between exon and promoter, than for any other exon. In the following chapter we test this mechanism in the context of human genetic variation and explore the functional consequences on the protein isoforms in the human system.

2.2.2 CTCF-mediated intragenic chromatin loops regulate alternative exon usage

To test the model that CTCF-mediated DNA-loop formation can regulate alternative splicing decisions in a genetic context, we derive two hypotheses: (1) we expect that, across different individuals, CTCF binding correlates with exon usage for exons that can form a loop with their promoters, i.e. where motifs are in a convergent orientation and (2) we expect that such a correlation does not exist for other exons. To perform these tests we employed the CTCF ChIP-seq data for 18 individuals described above with matching RNA-seq data (Kasowski et al., 2013).

To obtain the variation across individuals on the exonic level, we calculated differential exon usage between each individual and the median across all (Figure 2.2A). Briefly, we used Gencode v19 to obtain exons from protein-coding, autosomal transcripts that were expressed in LCLs (as assessed with *kallisto* (Bray et al., 2016)), keeping only exons whose length was within the expected range of 20-5000nt (Derrien et al., 2012) and covered by an average of at least ten reads, leaving 86,285 exons of which 62,663 (73%) were middle exons (Methods: 2.4.3). We used *DEXSeq* (Anders et al., 2012) to calculate differential exon usage (defined as the ratio of exon vs gene expression) across the 18 individuals, testing for the difference between each individual and the median across all individuals. We identified 3,844 significantly **d**ifferentially **u**sed **e**xons (DUEs; FDR 10%), of which 2,081 were middle exons (Figures S2.2A-D for DUE characterisation). The log2 fold-change values of exon usage for each individual against the median were used for the correlation analysis below.

For the variation on the CTCF level we calculated the differential CTCF binding across individuals in a similar way: we averaged the signal of CTCF (extracted with *SNPhood* (Arnold et al., 2016)) within 2kb upstream or downstream of the 5' and 3' boundaries of the exons respectively, and calculated the log2 fold-change between each individual and the median across all individuals (Methods: 2.4.4).

We calculated the Spearman correlation between the variation in CTCF binding at the 5' and 3' boundary and the variation in exon usage for all DUEs (Figure 2.2B). The correlations were generally not driven by outliers as shown for a few representative examples (Figure 2.2C). Importantly, we find that exons with upstream convergent pairs show a strong bias towards positive correlations, whereas the correlations of CTCF with non-looping exons were distributed similar to empirical correlations obtained from 100 permutations (Figures 2.2D, S2.2E), thus supporting the predictions from our model.

In summary, these results based on variation across individuals support the model, that CTCF-dependent alternative exon usage is mediated by intragenic promoter-exon loops. Notably, we find that the evidence holds only for exons where the looping CTCF site is located upstream of its 5' boundary.

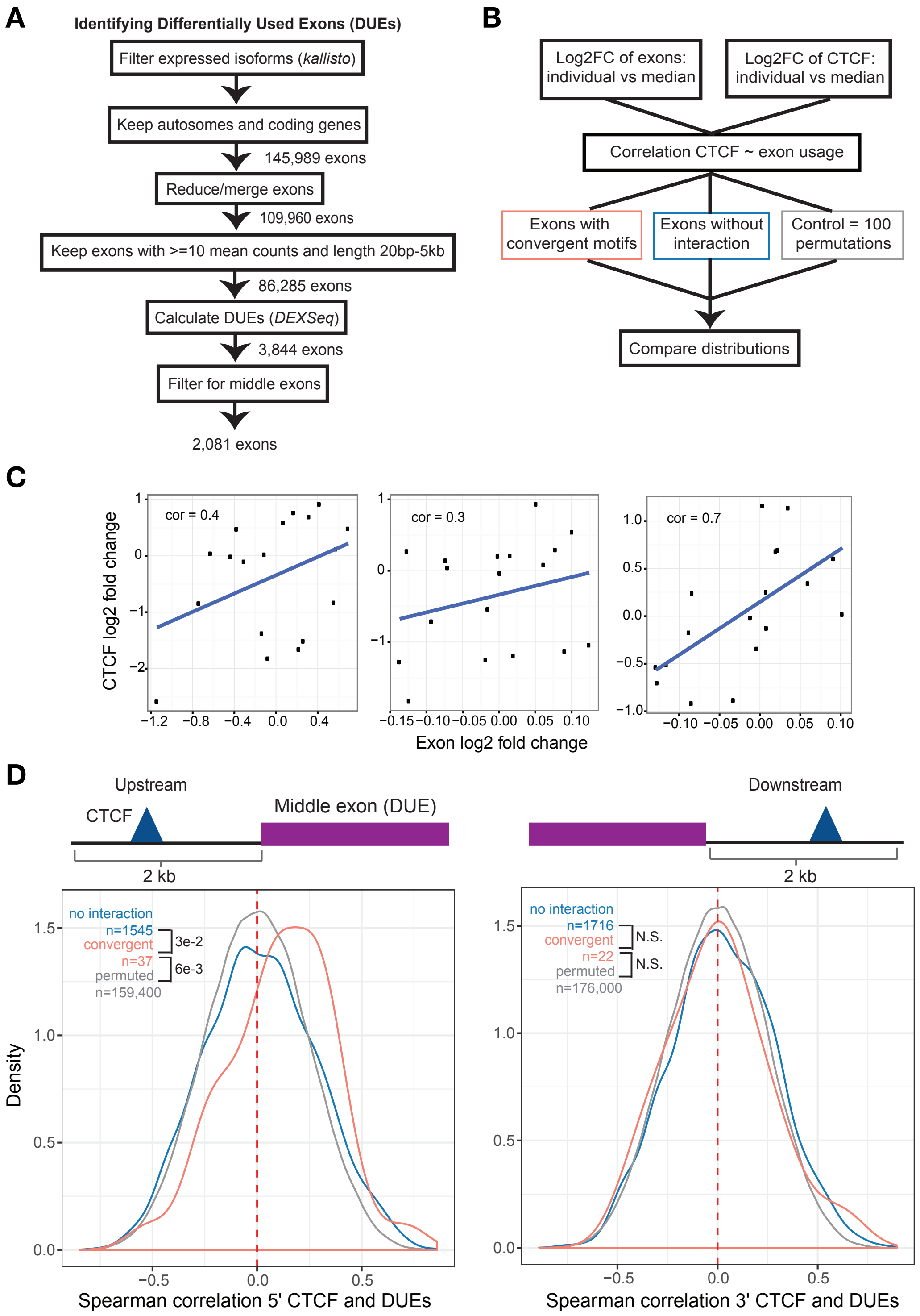


Figure 2.2: Relationship between CTCF and differentially used exons

Workflows used to **(A)** identify differentially used exons (DUEs) across individuals and **(B)** to calculate correlations between CTCF binding and exon usage. Three groups: exons-promoter pairs with convergent motifs (red), without motifs (blue), and 100 permutations (grey). **(C)** Examples of the log2 fold-change of CTCF binding and exon usage are shown. **(D)** Distribution of all the Spearman correlations for CTCF bound upstream (left) and downstream (right) of the exon grouped into convergent (red), exons without motif pairs (blue) and 100 permutations (grey). Convergently oriented upstream-exons show a significant shift towards positive correlations (p-value and N are shown, otherwise stated as N.S.; Wilcoxon test). These panels were generated by myself and are published in (Ruiz-Velasco et al. 2017).

2.2.3 Allele-specific analysis confirms the model of CTCF-dependent alternative exon usage mediated by DNA interactions

In the previous section we show that exon usage is correlated with CTCF binding only for exons predicted to form a loop with their promoter. To directly test the link between chromatin-loop formation and exon usage, we sought to make use of genetic variation to quantify the allelic fractions of chromatin contact frequencies and exonic expression. However, to do so we first had to establish that the correlation of variation in CTCF and exon usage is indeed driven by genetic variants. A strong overlap of CTCF-QTLs and sQTLs has already been reported earlier (Li et al., 2016) and our results showing a significant enrichment for convergent motifs containing sQTLs (Figure 2.1F) provide further evidence.

To obtain direct evidence for or against a genetic basis of CTCF binding and exon expression, we checked whether, at heterozygous positions, they show an allelic bias in the same direction. Briefly, we selected all CTCF peaks within 2kb upstream or downstream of an exon boundary that overlapped with a heterozygous SNP in at least one individual and showed an allelic bias at a nominal p-value of 0.05 (Methods: 2.4.5). In line with genetic variation in the CTCF peak being the driver for the variation in exon usage, we observe a significant positive correlation between the allelic fractions of CTCF and exon expression level for upstream and, to a lesser extent, for downstream regions (Pearson's R = 0.5 and 0.3 respectively; top panel Figure 2.3A). As a more robust quantification of the correlation we binned the exon allelic fraction by the direction of the CTCF allelic bias (allele 1 < allele 2 and vice versa) and assess their difference by a t-test (bottom panel Figure 2.3A). In line with the hypothesis of CTCF-mediated looping affecting alternative exon usage, we found a stronger association between the allelic fraction of CTCF binding and exonic expression for set of exons predicted to form a loop between an upstream

CTCF site and their promoter (Figures 2.3C, S2.3B-C). These findings suggest a mechanism where a genetic variant disrupts CTCF binding upstream of an exon, which in turn prevents the exon from being included in the transcript.

To test whether the difference in CTCF binding indeed translates into differences in chromatin loop formation, and whether these differences translate into alternative exon usage, we performed the same allelic bias associations with exon usage and HiC data for one of the individuals in our study (GM12878) (Rao et al., 2014). As expected from previous studies we found that the allelic bias in HiC and CTCF is positively correlated only when CTCF is in a convergent orientation between the promoter and exon (Figure 2.3C). Notably, and as a strong line of evidence for our model, we found a positive correlation of the allelic bias in exon expression and CTCF binding only for exons in predicted promoter exon-upstream loops (Figures 2.3B, S2.3B). This suggests that indeed exon usage in those sites is dependent on a mechanism that involves a loop between the exon and its promoter.

To validate these global correlations based on allele-specific analyses, Dr. Mang C. Lai and I performed 4C-seq experiments for six individuals in a gene that showed differential usage of exon 5 across these individuals (*THRAP3;* methods: 2.4.6). We found extensive chromatin interactions between the *THRAP3* promoter and several regions within the gene (Figure S2.3C). Importantly, for the differentially used exon 5 we found a stronger 4C-seq signal upstream for the individuals that include the exon than for those that do not include it (Figure 2.3D). No such difference between 4C-seq count was observed for other exons (Figure S2.3D). Together with the global allele-specific results above, this case study adds to the evidence that a difference in CTCF binding leads to changes in exon inclusion through a mechanism involving an intragenic promoter-exon loop.

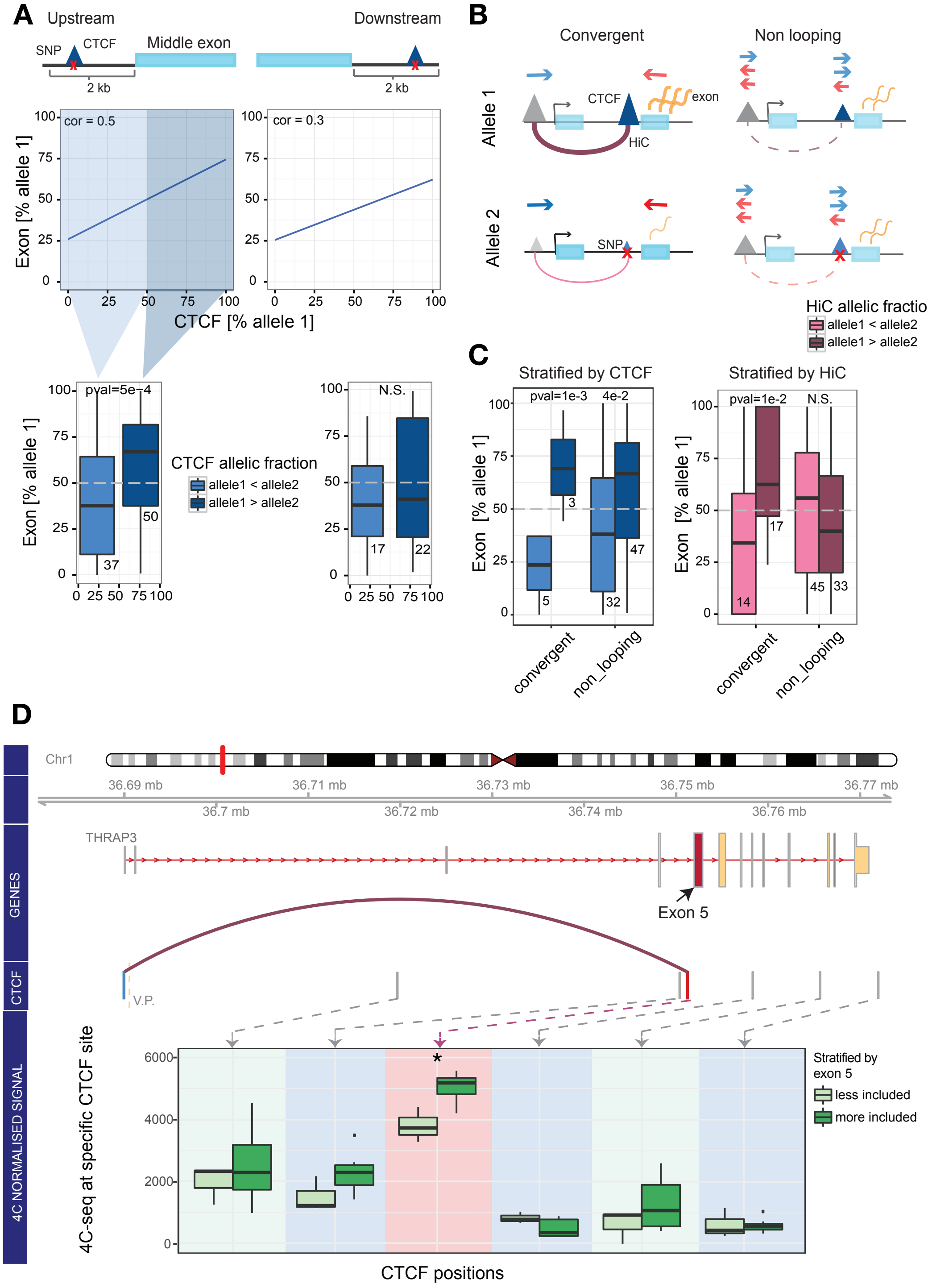


Figure 2.3. Genetic variants jointly affect CTCF-binding and exon inclusion

(A) Correlation of allelic fraction estimates of exon counts and CTCF signal for peaks within 2kb of the exon. Plots show the percentage [%] of reads mapping to allele1. **(B)** Summary of the results of the allele-specific analysis. **(C)** Exon allelic fraction stratified by binned CTCF allelic fraction (left) and binned HiC allelic fraction (right) (allele 1 > allele 2 and vice versa) are shown grouped by the configuration of the exon-promoter motif-pair. **(D)** 4C-seq signals in 6 individuals stratified by amount of inclusion of exon 5 for *THRAP3* are shown for each CTCF site along the transcript. The CTCF peaks predicted to form a loop with the promoter (according to their highest scoring motif) are shown in blue for sense and red for antisense. Exon 5 is highlighted in red. (V.P. = viewpoint, star = p-value < 0.05). Panels A-C were produced by myself and panel D jointly with Dr. Mang Ching Lai; they are published in (Ruiz-Velasco et al. 2017).

2.2.4 Functional consequences of genes and exons with predicted intragenic loops

Finally, we sought to assess the functional impact of exons, that are potentially regulated by a CTCF-loop, on the protein level. These analyses were carried out by Dr. Manjeet Kumar for two different sets of exons: (i) those exons belonging to the *LCL-specific set* described in section 2.2.1 and (ii) the *unbiased set* which comprises all exons with convergent interactions that were solely annotated by the presence of CTCF motif-pairs, regardless of LCL-specific CTCF binding or on the expression status in LCLs.

For a large proportion of the loop-regulated exons, a known protein isoform exists that lacks the respective exon (Figure 2.4A). Moreover, loop exons are more likely than other exons to overlap fully with a missing region of a known alternative protein isoform (OR=1.6, p-value = 4.2e-2, Fisher's exact Test). This provides further evidence that the looping exons are indeed alternatively used and likely translated into distinct protein isoforms.

We also found that both the unbiased and the LCL-specific set were more likely to overlap with a Pfam protein domain than other exons (OR= 1.8, p-value < 2.2e-16 and OR= 3.0, p-value < 2.2e-16 respectively; Fisher's exact test). Additionally, when considering the specific functions of the domains overlapping with the looped exons we found a significant enrichment for kinase domains (OR= 2.3-unbiased and 5.1-*LCL-specific*; adjusted p-value = 2.2e-6 and 3.0e-4). This indicates that alternative usage of the looping exons likely has an impact on the protein function.

We performed a gene ontology (GO) enrichment analysis on the genes containing loop-regulated exons. This revealed a strong enrichment for terms related to signalling and response to stimuli in both sets (Figures 2.4B and S2.4A). Consequently, we found that such genes are also predominantly associated with membrane, cell-periphery, or cell projection structures (Figure S2.4B).

Overall, we identified very interesting examples of well-studied proteins where the looped exon is absent from at least one of the known isoforms (Figures 2.4C-E). As a first example, we show the DNA damage Checkpoint Kinase 2 (Chk2), an enzyme for which over 90 splice variants have been described, 13 of which are annotated in Swiss-Prot/Uniprot, and for five of which the looping exon is missing (Figure 2.4C).

As a second example we show the protein kinase Clk3, for which we identified exon 4 as looping (Figure 2.4D). This protein is known to have a full-length active isoform but also a truncated, catalytically inactive isoform. Active Clk3 regulates alternative splicing by phosphorylating SR proteins of the spliceosomal complex. The inactive isoform is formed by excluding exon 4, which leads to a translational frameshift that induces the formation of a premature stop codon and loss of the kinase domain (Duncan et al., 1995; Hanes et al., 1994). Interestingly, the ratio of inclusion of exon 4, and thereby catalytically active vs inactive isoform, has been shown to control the differentiation process in multiple cell types (murine erythroid cells, neurons and astroglia cells) (García-Sacristán et al., 2005).

As a final example, we have the histone-lysine N-methyltransferase 1 (EHMT1) protein, which is involved in mono- and dimethylation of histone H3 (Lys-9) in euchromatin. Of the four isoforms in Swiss-Prot, two of which involved a loop-regulated exon. Isoform 2 lacks the whole EHMT1 protein due to an alternative splice at loop exon 2, while isoform 4 skips the looping exon 26, thereby removing half of the SET domain methyltransferase (Figure 2.4E), presumably eliminating the catalytic activity while retaining the chromatin location and histone tail-binding properties of the protein.

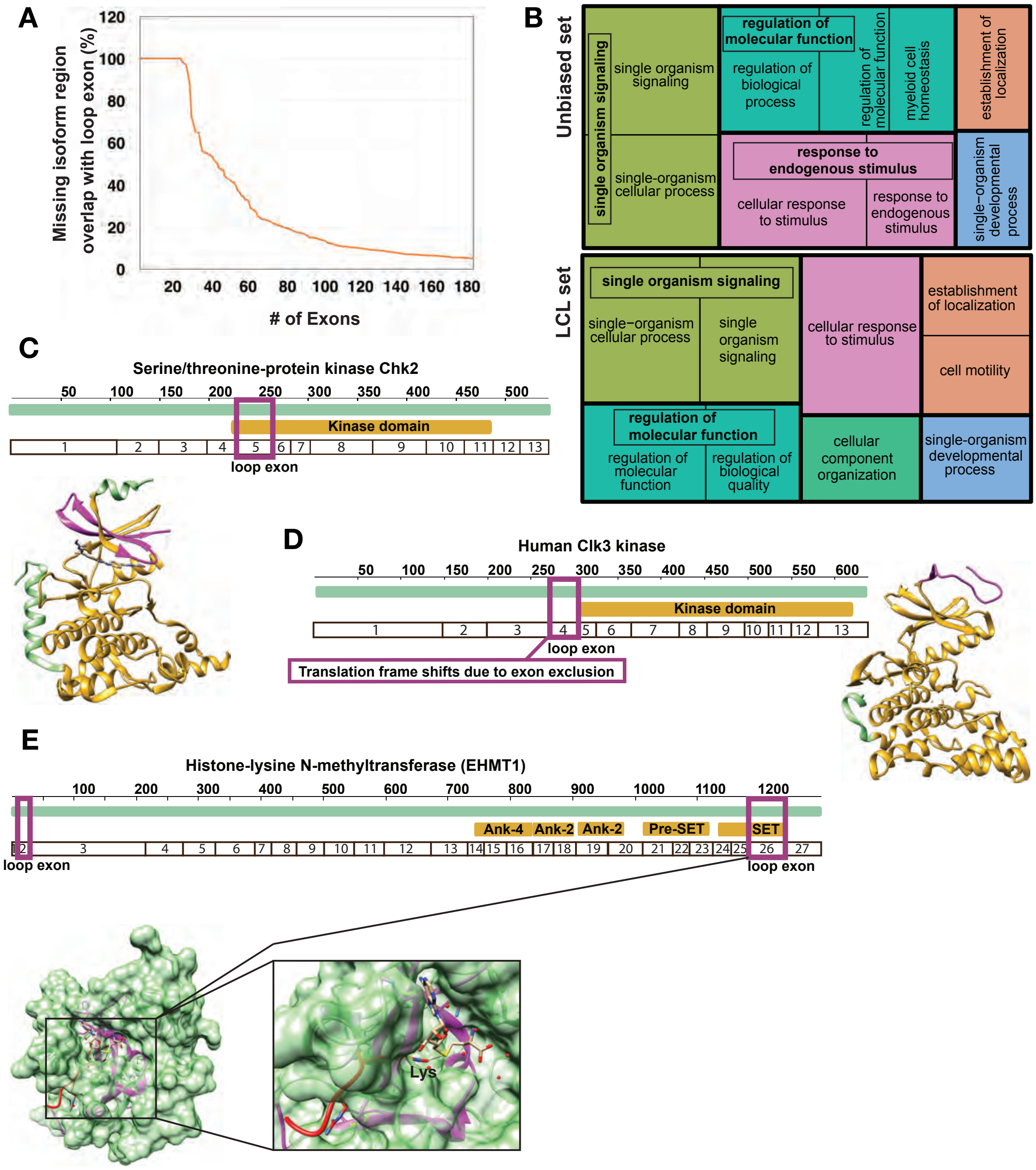


Figure 2.4. Influence of CTCF-mediated intragenic loop exon on protein function

(A) The cumulative distribution of overlaps between looping exons and missing regions of known protein isoforms are shown (as percentage of the missing region). **(B)** Treemap representations from Revigo (Supek et al., 2011) showing significantly enriched biological processes (top 10) for the genes containing loop exons in the unbiased (upper panel) and in the LCL-specific set (bottom panel). **(C, D, E)** Visualisation of exon structure, primary sequence and 3D-structure for selected genes with annotated loopss. Looped-exons (magenta) are shown in the primary and 3D-structure. **(C)** Chk2 (UniProt-ACC:O96017, PDB:2YCF) contains a looping-exon in the kinase domain (exon 5). **(D)** Clk3 kinase (UniProt-ACC:P49761, PDB:2EU9): The loop exon (exon 4) is alternatively used in protein isoforms that then lack the entire kinase domain. **(E)** EHMT1 (UniProt-ACC:Q9H9B1, PDB:3HNA): Loop exons 2 and 26 produce isoforms that affect protein function. This figure was contributed by Dr. Manjeet Kumar and is published in (Ruiz-Velasco et al. 2017).

These examples illustrate that CTCF-mediated intragenic loop formation can influence the ratio of transcript isoforms in many different contexts, and potentially plays a significant role in cellular signalling and decision making, thereby modulating processes such as differentiation, cancer or other forms of pathogenesis.

2.3 Discussion

Chromatin topology and function are tightly linked to ensure proper genomic regulation at all organisation scales; however, there are still very few studies that show a direct link between chromatin organisation and gene regulation (Ruiz-Velasco and Zaugg, 2017). Here we propose a novel functional role of CTCF-mediated intragenic DNA-loops in regulating alternative splicing decisions, that likely have a functional impact on cellular signalling and decision making. The model is derived from integrating multiple layers of genome-scale molecular phenotype data, genetic variation, and functional genome annotation, and was validated by experimental 4C-seq assays.

Individual examples for the involvement of CTCF in alternative exon usage have been described before, e.g. in the *CD45* locus (Shukla et al., 2011), where it was proposed that CTCF promotes inclusion of an alternative exon by binding in its downstream intron and creating a ‘roadblock’ for PolII, which in turn allows splicing factors to assemble the spliceosome (Shukla et al., 2011). However, this model does not take into account the most recent and well studied role of CTCF in chromatin 3D organisation. Furthermore, our data suggests that CTCF bound upstream of an exon also triggers its inclusion thus questioning whether CTCF acting as a roadblock to PolII downstream of the exon is the only mechanism.

The additional mechanism we propose is that CTCF-mediated chromatin loops between the promoter and the upstream region of an exon regulate its inclusion (Figure 2.5). We speculate that the mechanism involves a combination of (i) slowing down PolII and (ii) increasing the local concentration of splicing factors at the exon (see below). However, we cannot exclude the possibility of CTCF-RNA interactions, which are prevalent across the entire transcriptome (Kung et al., 2015; Saldaña-Meyer et al., 2014), also playing a role.

PolII speed has repeatedly been shown to affect splicing decisions (Fong et al., 2014; Jonkers et al., 2014; Oesterreich et al., 2016; Shukla et al., 2011). A recent study, which proposes a link between chromatin loop formation and cohesin pausing at CTCF sites (Haarhuis et al., 2017) opens the intriguing possibility that the process of loop extrusion could also slow down PolII which in turn provides enough time for the spliceosome to include exons that are near to the convergent CTCF sites. Furthermore, splicing factors have been reported to localise to promoters via the Mediator complex and can regulate alternative splicing decisions far from the promoter in a set of alternatively spliced genes (Huang et al., 2012; Kornblihtt et al., 2013; Mikula et al., 2013), it is thus tempting to speculate that this distal splicing regulation might act through CTCF-mediated intragenic chromatin loops. In such a mechanism a transfer of splicing factors from the promoter to the looped exons would make splicing more efficient by increasing their local concentration at the exon.

In support of this we found a correlation between the promoter mark (H3K4me3) at the 5’ end of the exon and its inclusion into the transcript only for exons looping to their promoter (Figure S2.2F). A similar correlation was observed for active mark (H3K27ac) and transcription mark (H3K36me3), but not for the enhancer mark (H3K4me1; Figure S2.2F). These associations of splicing with promoter, active and transcription marks at exons could be explained by the interaction of the exons with their promoter, as in our proposed model.

Several questions remain to be address to better understand the extent of our model. Additionally, the former analyses only demonstrated the mechanism occurring within a single cell type. In the following chapter evidence of intragenic loops affecting splicing are shown during development, across tissues, and briefly for cancer cell lines.

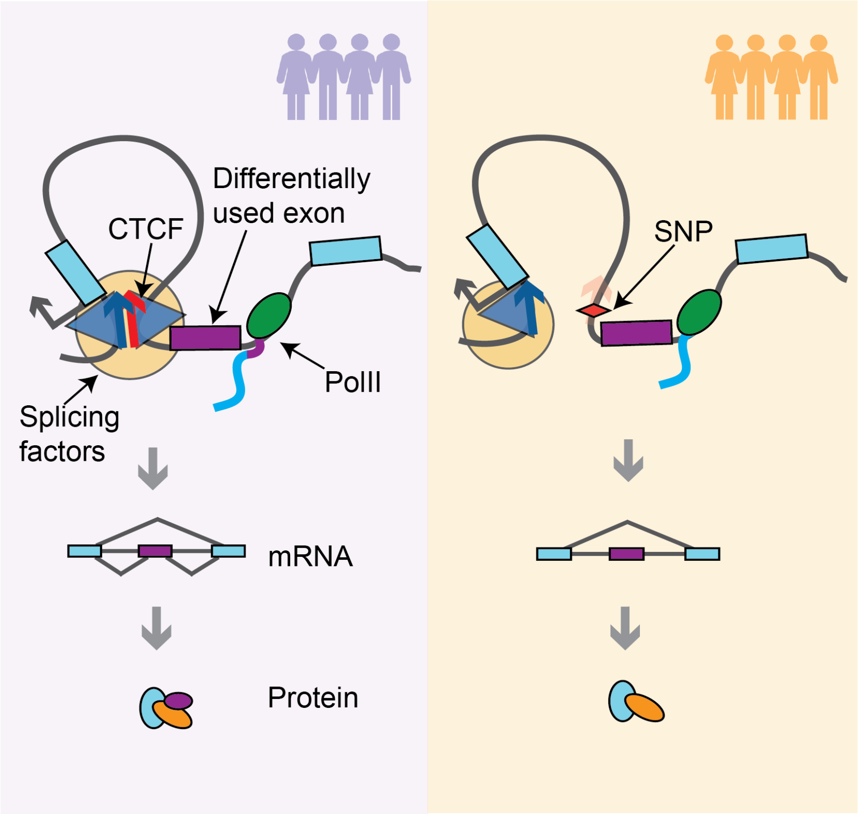


Figure 2.5. CTCF-mediated intragenic loop model

Schematic of the proposed mechanism: CTCF-mediated intragenic chromatin interactions between promoter and gene bodies regulate alternative exon usage by favouring the inclusion of exons, possibly through an increasing concentration of splicing factors. Genetic variants can affect this mechanism by disrupting the binding of CTCF and preventing the loop formation. This figure was produced by myself and published in (Ruiz-Velasco et al., 2017).

2.4 Methods

2.4.1 Annotation of Intragenic Loops

For the identification of CTCF peaks binding along transcripts, we defined four genic regions: promoter (+/- 1kb from transcription start site (TSS)), end of transcript (+/- 1kb of transcription termination site (TTS)), upstream, and downstream regions of exons (either 2 kb from outside the exon boundaries - when the intron was longer than 2kb - or the actual size of the intron in case it was shorter). We expect all promoters to fall within the TSS region and to avoid any overlaps between the promoter and exon regions, we removed any region associated with first exons, first introns and anything within 5kb from the TSS from the set of exonic regions. Additionally, we filtered out exons intersecting with CAGE peaks (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) to prevent mis-classifying non-annotated promoters as exons.

To minimise the number of non-functional CTCF binding sites, we only considered sites that overlap with a CTCF peak in LCLs obtained from our previous study (Kasowski et al., 2013). We then intersected the CTCF peaks with each of these windows and assigned the highest scoring motif (MA0139.1) to each peak by using the PWMScan tool (http://ccg.vital-it.ch/pwmtools/pwmscan.php) (cut-off of 93.13% with non-overlapping matches). Motifs having a score of >= 500 (*LCL-specific set*) or >= 1000 (*unbiased set*) were further used.

Given that CTCF binds to an asymmetric motif, it is possible to assign a directionality. We did so by defining directionality relative to the gene strand thus rendering *sense* CTCF motifs if both the motif and the gene lie in the same strand and *antisense* vice versa. We grouped the CTCF motif-pairs according to their genic region into: promoter:exon-upstream, promoter:exon-downstream, promoter-TTS, exon-upstream:TTS, and exon-downstream:TTS (schematic in Figure 2.1). We next classified the motif pairs based on their relative orientation to each other as "*convergent*", "*divergent*", "*both sense*", and "*both antisense*". For this analysis, we excluded any pair of motifs with a distance smaller than 5kb to allow for potential chromatin looping, and we only kept unique pairs. For the list of pairs visit:

http://www.embl.de/download/zaugg/mariana/intragenic.interactions.unbiased.set.hg19.txt

2.4.2 Overlap with chromatin conformation data and splicing QTLs

For validation we overlapped the predicted loops with existing chromatin conformation data requiring that each motif of a CTCF motif-pair overlapped one side of the same interaction. We used cohesin subunit Rad21 and histone mark H3K4me3 ChIA-PET (Grubert et al., 2015), and CTCF (Tang et al., 2015) ChIA-PET and Capture-HiC (CHiC) data (Mifsud et al., 2015). Briefly, CHiC uses oligonucleotides close to several promoters to amplify the reads obtained close to the TSS.

Based on earlier studies suggesting that CTCF is capable of binding to sequences as long as 40-60bp (Li et al., 2013), we intersected a 40bp CTCF motif (+/-20bp from center of the canonical motif) with the list of splicing QTLs (sQTLs) (Li et al., 2016). We considered all sQTLs above an FDR of 10% as described in their methods.

2.4.3 Annotation of Differentially Used Exons

We obtained RNA-seq data (strand-specific and paired-end reads) for lymphoblastoid cell lines of 18 individuals from (Kasowski et al., 2013). We selected protein coding genes using the genome annotation provided in Gencode v19. We first filtered out lowly expressed isoforms as described in (Soneson et al., 2016) by applying the *kallisto* tool to all RNA-seq files (Bray et al., 2016), only keeping transcripts which were expressed in at least 10 out of the 18 individuals. We used custom scripts to generate a set of non-overlapping exon boundaries. This involved first splitting exons into individual parts (since some exons can be part of multiple transcripts and can have several annotated 5' and 3' boundaries) and then merging them again into units that contain the most 5' and the most 3' annotated boundary for each exon. Chromosomes X, Y and mitochondrial DNA were filtered out.

Based on current knowledge about exon length distribution (Derrien et al., 2012), we further removed exons that were less than 20 or more than 5,000bp long. These annotation and filtering steps resulted in a set of 109,960 exons. We then used *DEXSeq* (Anders et al., 2012) to count the reads falling into each of these exons for each of the replicates of the 18 individuals. As an additional filtering step, we excluded exons with less than 10 counts on average, leaving 86,285 exons for further processing (for a total of 9,198 genes) of which 62,663 (73%) were middle exons.

Differentially **U**sed **E**xons (DUEs) as described in (Anders et al., 2012), are defined as a change in exon usage whereby exon usage is defined as the ratio of reads mapping to a particular exon vs reads mapping anywhere else in a gene. To obtain the DUEs we used *DEXSeq* and compared the exon usage for each individual against the median across all the 18 individuals (Figure 2.2A). The such obtained log2-FC values were used for the correlation between exon and CTCF usage (see CTCF correlation with DUEs below). We considered exons to be differentially used when their adjusted p-value was below 0.1 (Benjamini Hochberg correction), leaving a total of 3,844 DUEs. In order to keep only the middle DUEs, exons were intersected with the TSS and TTS from the Gencode v19 annotation and respectively labelled as first, middle, and last exons. For further analysis, we excluded the first and last exons, considering only the 2,081 middle DUEs.

These DUEs cover 14% (1,148) of expressed genes with multiple exons. From these 1,148 genes, most have less than 10 DUEs (Figure S2.2A), and no strong correlation between the number of middle exons per gene and number of differentially used exons was observed (Figure S2.2B). The size distribution of DUEs is similar to that of all exons and consistent with previous observations (Derrien et al., 2012) (Figure S2.2C). For the list of DUEs visit: https://www.embl.de/download/zaugg/mariana/DUE.annotation.txt

2.4.4 CTCF correlation with DUEs

To calculate the log2 fold-change per exon in each of the individuals against the median expression based on their *DEXSeq* counts, we estimated the exon fold-changes as described above and in (Anders et al., 2012). As a quality control, we checked the distribution of fold-changes with respect to the median for each individual and found that it shows similar patterns for all of them (Figure S2.2D).

For the quantification of the ChIP-seq read counts of CTCF and RNA PolII binding around the exons, we used *SNPhood* (Arnold et al., 2016). We extracted the reads of CTCF falling within 2kb downstream or upstream of the 5' and 3' boundaries of exons respectively. We then calculated the log2 fold-change of CTCF counts between each individual and the median individual, similar as described for the RNA-seq data. These fold-changes were then correlated with the fold-changes of the exon.

The same procedure was applied to other genomic features (PolII, Sa1, H3K4me1, H3K4me3, H3K36me3, and H3K27ac) from paired data (Figure S2.2F). Note that only cases where the median factor read counts were 10 or more were considered. For each exon, we generated 100 sets of permutations for fold-changes in exon usage between individuals, keeping the number of individuals with positive and negative fold-change constant at each exon. This empirical distribution was then used to assess the significance of the correlations by applying Wilcoxon tests.

2.4.5 Allele specific analysis of CTCF binding

We first extracted all heterozygous SNPs from the genotypes of the 14 individuals from the 1000 Genomes project (http://www.internationalgenome.org/) (1000 Genomes Project Consortium et al., 2015), for which genotypes were available, and intersected them with the CTCF peaks from the ChIP-seq data. 80% of all the peaks (158,455) contained 1 or more heterozygous SNPs in at least one individual. We then kept the SNPs closest to the midpoint of the motif that was predicted to form an intragenic interaction.

Using *SNPhood*, we analysed whether there was an allelic bias for the read counts extracted in the region of +/-250 bp surrounding the SNP. Briefly, *SNPhood* performs a binomial test to assess whether the allelic fractions are deviating from the expected 0.5. We selected all CTCF peaks within 2kb upstream or downstream of an exon boundary that overlapped with a heterozygous SNP in at least one individual, calculated the allelic bias and filtered the peaks based on a nominal p-value cutoff of 0.05, which resulted in a set of 9,434 upstream and 7,689 downstream allele-specific bound CTCF peaks. This set of peaks served as a basis for the correlation with exon allelic bias.

To calculate the allelic bias for exons, we split the RNA-seq alignment files by allele to which they map (allele information was obtained from (Kasowski et al., 2013)) using samtools split and then used bedtools (intersectBed) (Quinlan and Hall, 2010) to extract the counts. To assess exon allelic bias we performed a binomial test that resulted in 1,327 and 664 events upstream and downstream respectively (at nominal p-value < 0.05). To avoid low-count artefacts, we only kept SNPs that had at least 5 mean read counts.

To verify whether there was an enrichment of allele specific CTCF binding in the 2kb upstream or downstream of the exons, we identified middle exons that overlap with CTCF peaks using *GenomicRanges* (Lawrence et al., 2013). Only those cases where the individual was heterozygous for the SNP and both CTCF and exon were allele specific were taken into account. As described in the main text, to obtain a more robust quantification we classified the exons according to the directionality of the allelic bias in CTCF binding and assessed differences in exon allelic fractions between the two groups by a t-test (bottom panel Figures 2.3A, S2.3A). This revealed a significant association only for the CTCF sites upstream of the exon.

For the genome-wide correlation of HiC signal with CTCF binding in an allele specific manner, we used the GM12878 diploid HiC maps at 5kb resolution (Rao et al., 2014) and the CTCF and exon counts as mentioned above. We calculated normalised interaction frequencies with Knight-Ruiz normalisation vectors and we followed the same procedure described by the authors in their supplemental methods. We obtained the allelic fraction by calculating the ratio of maternal counts with respect to the total counts. Next, we selected the CTCF peaks binding in +/-2kb from all middle exons that contained SNPs for which GM12878 was heterozygous. We finally stratified either the CTCF or the exon allelic fraction based on the HiC allelic fraction as described above (Figure S2.3B).

2.4.6 Protocol and processing of 4C-seq experiments

4C-seq was performed mostly by Dr. Mang Ching Lai and partially by me in the group of Dr. Kyung-Min Noh using 6 human LCLs and a protocol modified from (van de Werken et al., 2012). All chemicals and reagents were purchased from Sigma Aldrich unless otherwise stated in brackets. To perform crosslinking of chromatin, 10 million cells per cell line were fixed in 5 mL 2% formaldehyde (VWR) at room temperature for 10 minutes. 250 μL of 2.5 M glycine was added to quench the crosslinking reaction for 3 minutes at room temperature. Cells were washed twice in ice cold PBS and were snap frozen using liquid nitrogen prior to 4C experiments.

To perform in nucleus Dpn II restriction digest of crosslinked genomic DNA, 10 million cells per cell line were first lysed in 5 mL lysis buffer [10 mM pH 7.5 Tris-HCl, 10 mM NaCl, 0.2% NP-40, 1X complete EDTA-free protease inhibitor (Roche)] for 30 minutes on ice. Nuclei were centrifuged at 600 rcf for 6 minutes at 4°C. Supernatant was discarded and the nuclei were then resuspended in 1 mL residual supernatant. The samples were then centrifuged again at 600 rcf for 6 minutes at 4°C and the supernatant was discarded. The nuclei pellets were washed in 200 μL of 1.25X Dpn II Buffer (NEB) without resuspension at 600 rcf for 1 minute at 4°C. The nuclei pellets were then resuspended gently in 492.5 μL of 1.25X Dpn II Buffer with 0.3% sodium dodecyl sulfate. The mixture was incubated at 37°C and 950 rpm in a thermomixer (Eppendorf) with heated lid for 1 hour. 50 μL 20% Triton X-100 was then added to the mixture and was further incubated for 1 hour at 37°C and 950 rpm. Dpn II restriction digest was then performed by the addition of 45 μL nuclease free water (Ambion) and 10 μL DpnII enzyme (NEB), and the reaction was incubated at 37°C and 950 rpm for 3 - 5 hours with heated lid in the thermomixer. An addition 10 μL of DpnII was added to the mixture and was further incubated in the thermomixer overnight.

To ligate digested DNA for generating a 3C template, the mixture was heat inactivated for 20 minutes at 65°C, followed by centrifugation at 600 rcf for 6 minutes at 4°C. Supernatant was removed, leaving 50 μL. Ligation reaction mixture of 950 μL [50 mM pH 7.5 Tris-HCl, 10 mM $MgCl^2$, 1 mM ATP (NEB) and 5 mM dithiolthreitol, 0.1 μg/μL bovine serum albumin (NEB) and 150 U/μL T4 DNase ligase HC (ThermoFisher)] was added to the samples and was incubated overnight at 16°C and 600 rpm in the thermomixer.

To reverse crosslink, the samples were incubated with 200 mg proteinase K (ThermoFisher) and 50 μL 20% sodium dodecyl sulfate at 65°C for 1 hour. The samples were then incubated with 100 mg RNase A (Qiagen) for 45 minutes at 37°C.

To purify the 3C template, a phenol chloroform extraction method was performed where 1 volume of phenol:chloroform:isoamyl alcohol 25:24:1 saturated with 10 mM Tris pH 8.0, 1 mM EDTA was added to the samples and were mixed thoroughly by vortexing. The mixture was transferred to a 2 mL heavy phase lock gel tube (5-Prime) and was centrifuged at 16,000 rcf for 5 minutes at room temperature. The top clear aqueous phase containing DNA was transferred to a clean eppendorf. NaCl of a final concentration of 200 mM, together with 1 μL of 20 μg/μL glycogen was added to the aqueous DNA solution, followed by the addition of 2 volume of -20°C 100% ethanol. The mixture was incubated at -80°C for at least 1 hour and was centrifuged at 15,000 rpm for 1 hour at 4°C. The supernatant was removed and 500 μL -20°C 80% ethanol was added to wash the DNA pellet at 15,000 rpm for 60 minutes at 4°C. The supernatant was then removed and the tubes were centrifuged again for 30 minutes at 15,000 rpm at 4°C. The residual supernatant was then removed and the DNA pellet was allowed to be air dried for 5 minutes at room temperature. To resuspend DNA, 50 μL 65°C 10 mM pH 7.5 Tris-HCl was added to the DNA pellet and the solution was incubated at 65°C for 5 minutes to dissolve the pellet. Digestion and ligation efficiencies were determined by agarose gel electrophoresis.

To perform the 2nd round of restriction digestion with NlaIII enzyme (NEB), 5 μL 10 U/μL NlaIII and 6 μL 10X NlaIII buffer (NEB) was added to 50 μL 3C template, and the reaction was incubated at 37°C overnight. NlaIII was heat inactivated at 65°C for 20 minutes.

To generate 4C template, 5X ligation buffer [250 mM Tris-HCl, 50 mM $MgCl_2$, 5 mM ATP, 25 mM dithiolthreitol], 5 µL 10 mg/mL bovine serum albumin (NEB) and 2.5 µL 150 U/µL T4 DNase ligase HC (ThermoFisher) were added to 50 µL NlaIII digested 3C template and the reaction was incubated overnight at 16°C.

To purify 4C DNA template, DNA purification was performed as described above. The final 4C template was further purified using the QIAquick PCR purification kit (Qiagen) according to the manufacturer's instruction. Digestion and ligation efficiencies were assessed by agarose gel electrophoresis.

To design primers for performing 4C-PCR, the viewpoint sequence of the *THRAP3* gene was first chosen using *THRAP3* sequence obtained from the UCSC genome browser (https://genome.ucsc.edu/index.html) on the human GRCh37/hg19 assembly. Restriction site identification and *THRAP3* sequence were processed using ApE-A plasmid editor (http://biologylabs.utah.edu/jorgensen/wayned/ape/). 4C-PCR primers were designed using primer3 v. 0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0/). Viewpoint and primer sequences were chosen and designed, respectively, according to the guidelines as previously published (Splinter et al., 2012.). Viewpoint coordinate correspond to the human GRCh37/hg19 assembly chr1:36690087-36690526. The forward and reverse primer sequences are 5' ctaacttccatcagaggcgctcac 3' and 5' attggcctggttcggtcttctc 3' respectively. Illumina adapters and indexing sequences were incorporated into the PCR primers. High-performance liquid chromatography purification was used in the production of primers (Eurofins Genomics). The primer sequences (5' to 3') are listed here:

Primer: THRAP3 Forward: AATGATACGGCGACCACCGAGATCTctaacttccatcagaggcgctcac
Primer: THRAP3_i2 Reverse: CAAGCAGAAGACGGCATACGAGATACATCGattggcctggttcggtcttctc
Primer: THRAP3_i4 Reverse: CAAGCAGAAGACGGCATACGAGATTGGTCAattggcctggttcggtcttctc

Primer: THRAP3_i5 Reverse: CAAGCAGAAGACGGCATACGAGATCACTGTattggcctggttcggtcttctc
Primer: THRAP3_i6 Reverse: CAAGCAGAAGACGGCATACGAGATATTGGCattggcctggttcggtcttctc
Primer THRAP3_i7 Reverse: CAAGCAGAAGACGGCATACGAGATGATCTGattggcctggttcggtcttctc
Primer THRAP3_i12 Reverse: CAAGCAGAAGACGGCATACGAGATTACAAGattggcctggttcggtcttctc

Oligonucleotide sequences © 2016 Illumina, Inc. All rights reserved. Derivative works created by Illumina customers are authorised for use with Illumina instruments and products only. All other uses are strictly prohibited.

To perform 4C-PCR, 4C template from each of the 6 individuals were amplified using primers above, with each index assigned to 1 individual. 4C-PCR was performed with the following condition:

Temperature (°C)	Time (sec)	Cycle
98	30	1
98	10	30
68	30	
72	180	
4	30	1

For each individual, two 4C-PCRs were pooled together and the PCR products were purified using Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer's instruction. 4C-seq library was prepared by pooling equimolar amounts of the barcoded 4C templates. Sequencing was performed on the MiSeq platform with 75 bp read length for single end reads. Custom sequencing [5' actcacctgggcctaccacagagatc 3'] and indexing [5' gagaagaccgaaccaggccaat 3'] primers were used. Two technical replicates were performed for all individuals.

To process raw 4C-seq data, FastQC was used to verify that all libraries passed the quality controls (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter trimming was not required as sequencing began at the first base after the DpnII site. Sequencing reads were aligned to hg19 using Bowtie2 (Langmead and Salzberg, 2012) with parameters: -p 8, --very-sensitive, --time. Only reads with a MAPQ score >= 10 were used.

The processed data was then analysed using the bioconductor package *Basic4CSeq* according to the author's instructions (Walter, C., 2015). Briefly, the restriction fragments were simulated *in silico* and aligned sequencing reads were mapped to the restriction fragments. Fragment read count was normalised to reads per million (RPM), and 4C interaction was subsequently visualised. We observed a cis (0.2 Mb region adjacent to the viewpoint) to overall reads ratio of about 40% in all of the 4C-seq experiments performed. A cis to overall ratio of 40% or above is considered as a quality control point for 4C experiments (van de Werken et al., 2012), indicating our experiments achieved the recommended quality. 4C interactions were visualised as a plot of normalised reads versus genomic coordinates.

To study the relation between 4C signals and DUEs of the 6 individuals, we used *GenomicRanges* to overlap normalised 4C read counts and CTCF peak signals. 4C reads mapped to fragment ends that overlap a +/- 1kb region of the mid point of the CTCF peaks were considered to be a readout for the CTCF peak's interaction with the viewpoint CTCF. The 4C readout signals was representing CTCF interactions were then stratified by the relative inclusion and exclusion of the exon usage to generate the box plot.

2.4.7 Functional analyses

These analyses were done by Dr. Manjeet Kumar from Dr. Toby J. Gibson's group. Two sets of exons were considered for the analyses: the *unbiased* and the *LCL-specific* set of looped exons. For the *unbiased* set the intragenic loops were annotated for cases where the CTCF motif score was >= 1000; this set reflects the potential of genes to form promoter-exon loops in any given tissue - as the identified loops were based only in the motifs. The *LCL-specific* set is defined above (section 2.4.1) and represents a subset of the *unbiased set*, showing CTCF binding and exon expression in LCLs, along with >= 500 score for the motifs and +/- 1kb for the TSS window.

All exon coordinates were used to retrieve the protein information from Ensembl with the assistance of Perl APIs (Cunningham et al., 2015; Yates et al., 2015). Different features on the retrieved protein regions were mapped using the Uniprot features of the corresponding protein (UniProt Consortium, 2015). Protein domains on the sequence were mapped using a local installation of Pfam (Bateman et al., 2004).

Gene Ontology (GO) term analysis for looping genes was performed using ConsensusPathDB webserver (CPDB) (http://consensuspathdb.org/) (Kamburov et al., 2013). The top 10 GO terms were visualised using Revigo (Supek et al., 2011) and represented as treemaps. All the statistical tests were done in R. The Pfam domain-to-GO term analysis was performed with the dcGOR R package (Fang, 2014). Location of exons on the primary sequence is based on the protein feature view from RCSB-PDB (www.rcsb.org). Domain definitions are based on the Pfam. The structure visualisation and rendering was done with the help of the UCSF Chimera package (Berman et al., 2000).

3. Intragenic CTCF-mediated loops regulate alternative splicing across cell types, are involved in neural differentiation, and misregulated in cancer

The next sections describe follow up analyses performed to better understand the extent to which CTCF-mediated intragenic loops regulate splicing decisions. This mechanism has been thoroughly described in the previous chapter. All of the analyses here are unpublished. The results from section 3.2.1 belong to a collaboration with Dr. Fabian Grubert, Dr. Rohith Srivas, and Dr. Damek Spacek from Dr. Michael Snyder's group at Stanford University and the story is currently under revision. Section 3.2.2 was initially submitted but left out of the final manuscript. Finally, results from section 3.2.3 are part of a collaboration with Dr. Esa Pitkänen and Dr. Jan Korbel and represent an ongoing collaboration.

3.1 Introduction

Our study described in Chapter 2 provides one of the first functional consequences of intragenic chromatin architecture by linking CTCF-mediated DNA loops to the regulation of alternative splicing, ultimately leading to the transcription of various protein isoforms. In addition, the model we propose provides a useful framework for further studying alternative isoform usage, e.g. across tissues, in cancer or across development, by generating hypotheses about alternative exon usage solely based on the analysis of CTCF binding sites, which are easily obtained from the DNA sequence.

To better understand what happens at the genes containing intragenic chromatin loops across tissues, or even in a dynamic process such as during development, we performed various analyses that are described in the next sections. Overall, we followed the same workflow as the one described in the previous chapter, which led us to conclude that intragenic loops are observed not only in Lymphoblastoid cell lines (LCLs), but also in all of the other tissues studied, as well as in mouse.

Finally, we wondered whether frequently mutated regions occurring in cancer, specifically in the PCAWG dataset, showed any of the intragenic loops described by us. The analyses done by our collaborator Dr. Esa Pitkänen showed that convergent interactions, which define intragenic loops, are enriched for frequently occurring somatic SNVs when compared to the other orientations. These findings indicate that CTCF-mediated intragenic loops are a conserved feature in mammals and highlight the importance of such mechanism in the regulation of alternative splicing, along with its misregulation in diseases such as cancer.

3.2 Results

3.2.1 Intragenic loops affect splicing decisions across tissues

An ongoing study from the Snyder lab in Stanford University has generated chromatin interaction maps for 24 cell lines commonly used in the ENCODE project (Grubert, et al., *in revision*). Briefly, the study generated Rad21 (a cohesin subunit) ChIA-PET datasets along with RNA-seq libraries as part of the latest ENCODE release. The main results of this study include the identification of a set of constant and variable interactions across cell lines and the observation that all cell lines clustered into three main groups based on their most variable loops after a principal component analysis (PCA): (i) blood, (ii) stem-cell like, and (iii) solid tissues (Figure S3.1A). These results indicate that chromatin interactions changing across cell types are more likely to be driven by epigenetic variation, rather than genetics. As part of a collaboration, we assessed the extent to which intragenic looping affected splicing regulation among these 22 cell lines (for which both loop and RNA-seq datasets were available).

As a first analysis, we verified which of the chromatin interactions from the pooled ChIA-PET dataset (n=124,830) had one anchor within +/-1kb from the TSS (promoter window) and the second anchor within -5kb of the 5'-boundary (upstream window). In this way, we identified 1,372 loops spanning 1,074 genes. Given that cohesin and CTCF are known to physically interact with each other (Xiao et al., 2011), we consistently observed that the distribution of the Rad21-loop anchors at the promoter and upstream windows recapitulated the observed patterns for CTCF (Figures 3.1A, 2.1A).

Once again, we used *DEXSeq* to identify differentially used exons (DUEs) either among a) the 22 cell lines or b) among the three groups defined by the PCA (Methods: 3.4.2). In this way, we defined a) 95,137 and b) 39,832 DUEs (FDR=10%). We next decided to verify the correlation between the normalised exon counts and the interaction strength of the Rad21 ChIA-PET for the exons next to a loop anchor across the 22 tissues (Figure 3.1B). This analysis showed that there was a shift towards positive correlation values for the exons that were classified as DUEs that was significantly shifted from the distribution observed for the other exons in the same gene (p-value = 9.2x10^{-3}, Wilcoxon test) or whenever the exons were permuted 100 times (p-value = 8.9x10^{-193}).

As a second validation, we compared the exon log2-fold change (log2-FC) against the loop-anchor log2-FC between stem-like and blood cell lines and colored the values based on whether they represented the real DUE-loop pairs (DUE-real) or all of the other exons in the gene with the loops contained for the same gene (DUE-not-real) (Figure 3.1C). The results suggest that the presence of intragenic loops affect the inclusion of the exon next to it, but not the inclusion of other exons in the gene. The same was observed when comparing stem-like and solid tissue cell lines although to a smaller extent (Figure S3.1B).

Finally we characterised gene *ARHGEF7*, where the presence of a loop mostly at blood-related cell types showed a high correlation with the abundance of the exon next to its intragenic anchor (Figures 3.1D, S3.1C). The exon affected by the intragenic loop in *ARHGEF7* was including in some cell lines, with variable abundance, while completely absent in others and this exon was also previously annotated to belong to some but not all isoforms according to ENSEMBL (Figure S3.1D).

Overall, by analysing chromatin loops and exon usage across 22 cell lines from variable tissues, we consistently identified evidence that intragenic interactions affect exon inclusion. With these results, we also demonstrate that this mechanism is not only susceptible to genetic changes, but rather modulated by epigenetic mechanisms that affects the chromatin landscape at a given cell type.

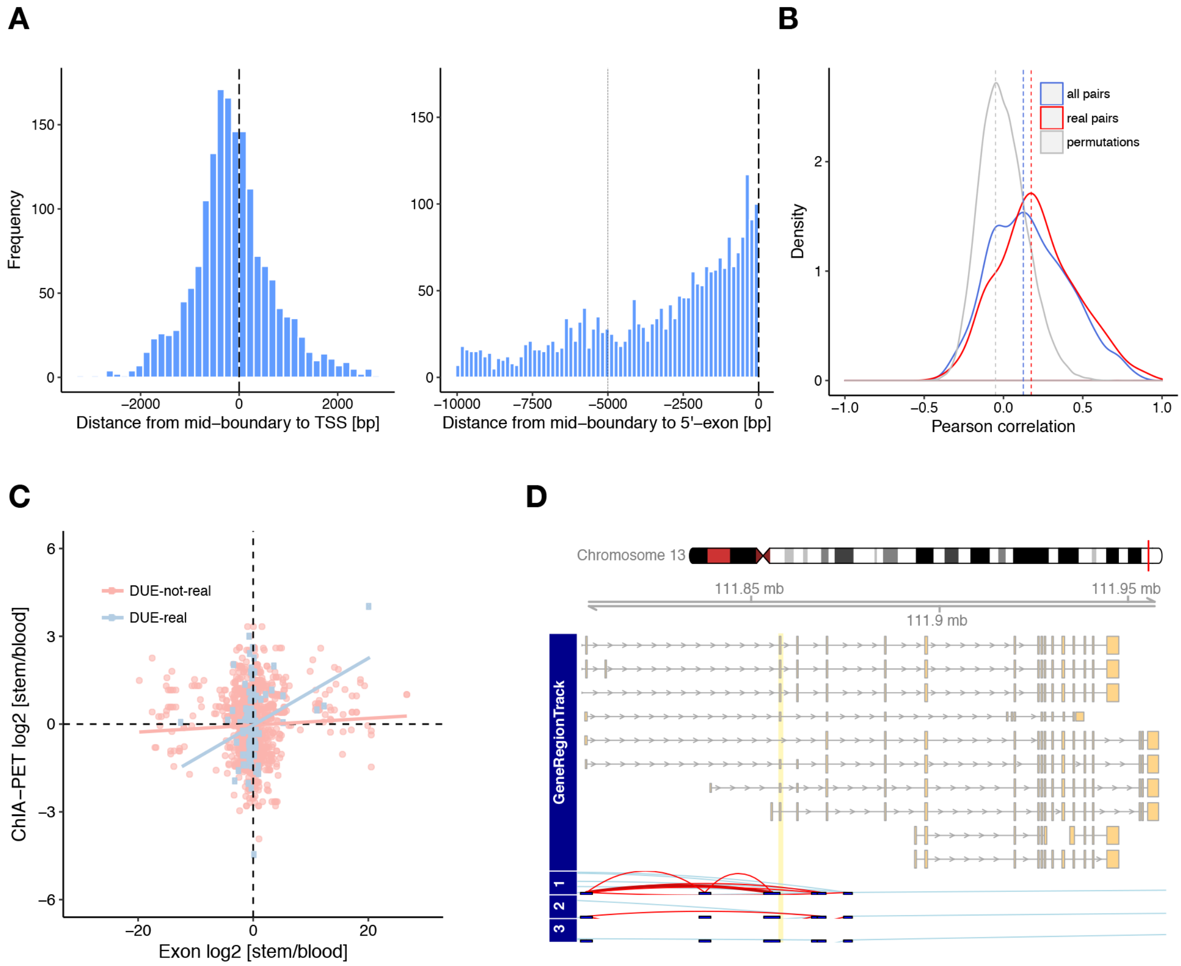


Figure 3.1. Intragenic interactions affect alternative exon usage across tissues

(A) Distribution of the distance (in base pairs) between the center of the anchors and the transcription start site (left) or exons' 5'-boundary (right). **(B)** Pearson correlation of the normalised ChIA-PET anchor counts and the exon counts across 22 cell lines for exon-loop pairs (red, n=277), exon-loops pairs of the same gene (blue, n=1347) and 100 permutations of the exon associated to the anchor (grey, n=27700) (red vs blue: p=9.2x10^{-3}, red vs grey: p=8.90x10^{-193}). **(C)** Scatterplots of the differentially used exon log2-FC and anchor count log2-FC for the real pairs (blue, n=111) and for the other exon-loop pairs within the same gene (pink, n=1347) (Pearson correlation = 0.3 for blue and 0.05 for pink). **(D)** Example of how a loop affects exon usage in the 3 groups for gene *ARHGEF7*. The presence of an intragenic loop from promoter to upstream of exon 6 (highlighted in yellow) is evident for group 1 (blood), but not for 2 (stem) or 3 (solid). All panels were produced by myself.

3.2.2 Evidence for CTCF-loop regulated exon usage in cellular differentiation

While we described intragenic interactions to be a mechanism affecting splicing ubiquitously in human tissues, the question of whether this mechanism is also present in other species still remained. Moreover, we still lacked a biological process where to test our model.

We reasoned that we could use the process of cellular differentiation of mouse embryonic stem cells (ESCs) into neural progenitor cells (NPCs) as a second system to test for the presence of promoter-exon loops. Additionally, a handful of studies have started to characterise the extensive changes in the 3D-chromatin landscape during this process, generating useful datasets and reporting a dynamic change at all chromatin interaction levels that encompass the transition from ESC to NPCs (Beagan et al., 2017; Bonev et al., 2017; Fraser et al., 2015; Pękowska et al., 2014).

To address our hypothesis that CTCF-intragenic loops contributed to neural differentiation, we obtained RNA-seq and CTCF ChIP-seq data for both mESCs and mNPCs from (Pękowska et al., 2014) and (de Wit et al., 2015) respectively, and annotated intragenic loops according to the same rules as described for the LCLs (Methods: 2.4.1 and 3.4.1). Intriguingly, we found that CTCF motifs from peaks occurring either in ESC or in NPC showed an overall similar pattern as the one described for CTCF-motifs (Figure 2.1A) and for anchor boundaries in human cell lines (Figure 3.1A), having highly enriched sense motifs upstream of the TSS and mostly antisense motifs in the introns (Figures 3.2A, S3.2A). Consequently, this pattern could facilitate the formation of convergent interactions between promoters and upstream of exons, suggesting that the intragenic chromatin architecture is conserved between mouse and human.

We next identified the differentially used exons between the two conditions (Methods: 3.4.2) and obtained 1,946 DUEs at FDR=10%, of which 1,080 were included more in mESCs and 866 in mNPCs. Consistently we found that exons in a predicted loop with their promoter are enriched for being differentially used between the two cell types (OR= 2.0, p-value = 1.1e-2 Fisher's exact test). To assessed whether this effect was given through condition-specific chromatin loop formation in neural differentiation, we tested whether exon usage and CTCF binding are correlated for exons that are looping to their promoter. Indeed we found that differential CTCF binding is correlated with differential exon usage only in cases where the exon forms a predicted loop with its promoter, i.e. convergent interactions (OR=3.3, p-value = 3.6e-5), however these was not the case for other configurations (Figure 3.2B, S3.2B).

Importantly, we were able to confirm the influence of CTCF in affecting the splicing decision by calculating DUEs in a dataset which specifically degraded the CTCF protein after four days in mESCs (Nora et al., 2017) (Methods: 3.4.2). With this system, we were able to observe a strong enrichment for the same exons that are alternatively spliced during neural differentiation to also change their inclusion levels upon CTCF knock-down in the same direction (OR=11.04, p-value = 2.24e-05, Fisher's exact test). This study also identified a very sharp enrichment of CTCF just upstream of the TSS for many genes that were misregulated upon CTCF depletion, in agreement with our analyses.

Overall, our findings suggest that CTCF-mediated loops are also present in mouse genes, affecting the inclusion of a group of exons relevant for neural differentiation. Moreover, at least part of the differential exon usage in this process can be explained by changes of CTCF-mediated intragenic loops, highlighting the importance of this mechanism.

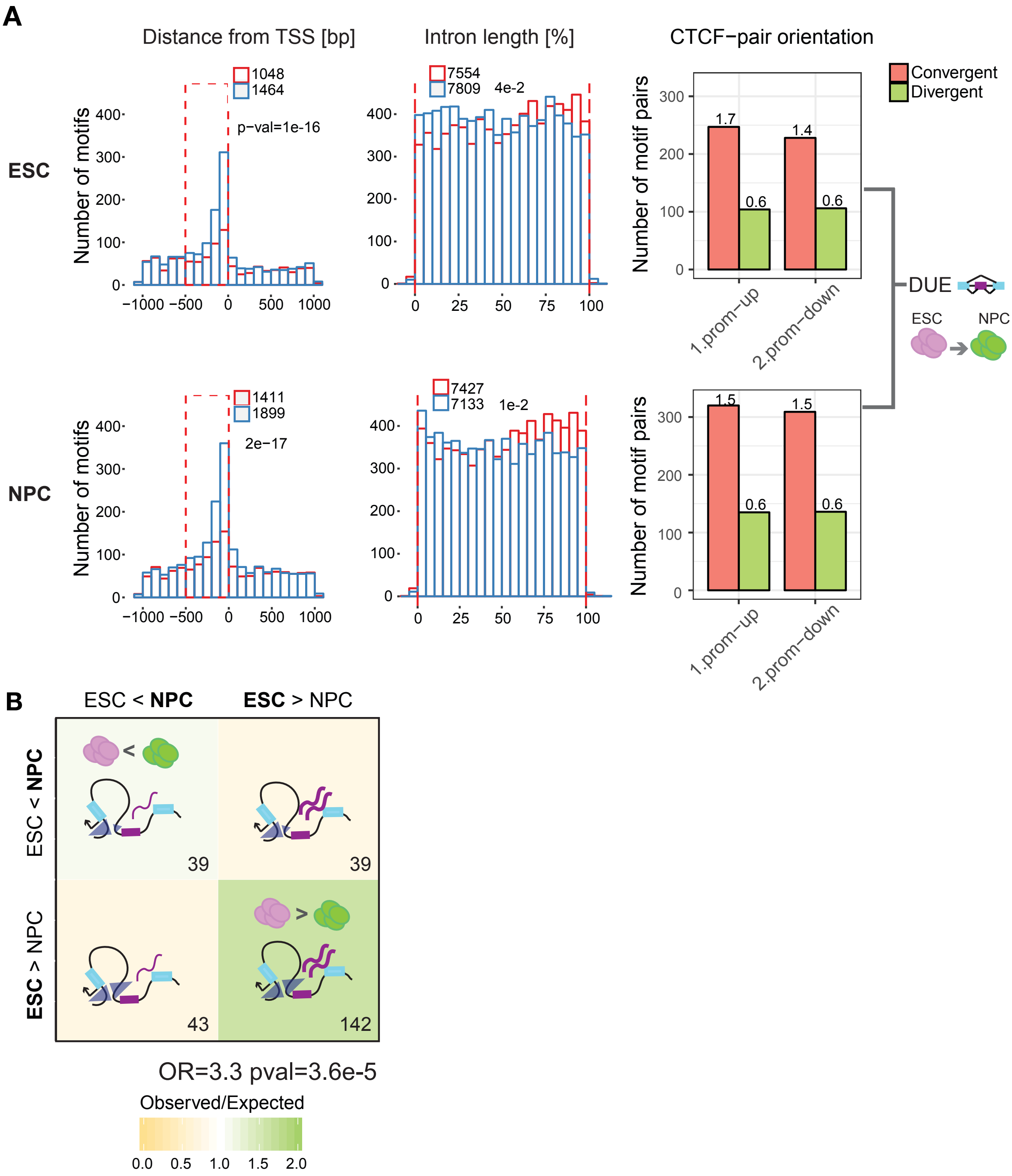


Figure 3.2. CTCF intragenic loops are also present in mouse models

(A) Same as in Figure 2.1A,B distribution of CTCF motifs in sense (blue) and antisense (red) orientation in promoter and intron regions are shown for mouse ESC and NPC (left). Predicted CTCF-motif pairs (convergent and divergent interactions) are summarised for each cell type (right). **(B)** Schematic of the contingency table for testing the association of changes in exon inclusion and CTCF binding in ESCs vs NPCs for exons that have a promoter-exon-upstream loop. The background color represents that observed vs expected value for each square. These figures were produced by myself in collaboration with Stefan Kleinsorg.

3.2.3 Frequently mutated regions in cancer are enriched at CTCF motifs that anchor intragenic interactions

CTCF-mediated intragenic loops were present in various important genes, controlling the splicing of exons that in many cases contained complete protein domains. This made us speculate that such loops might be under a tight regulation which, if modified, could have relevant consequences worth to be explored.

Our observation of an enrichment of sQTLs only at CTCF convergent motifs located upstream of the affected exon supported this idea (Figure 2.1F). In addition, a previous report already identified CTCF/cohesin-binding sites to be frequently mutated in various types of cancer (Katainen et al., 2015). A suggested model trying to explain how this frequently mutated regions arise proposes that the activity of the nucleotide excision repair machinery (NER) is affected by the accessibility of DNA around a TF-binding site, with lower NER directly at the binding motif and which correlates with nucleosome positioning (Sabarinathan et al., 2016), which causes an uneven repair within the CTCF-motif in melanoma cancers for both studies (Poulos et al., 2016). Finally, another study reported chromosome loop anchors bound by CTCF/cohesin to be vulnerable to DNA double strand breaks by topoisomerases at a position adjacent to the CTCF loop anchor, just outside the loop (Canela et al., 2017). For these reasons, we decided to analyse if the intragenic interactions that we identified were misregulated in cancer as part of a collaboration with Dr. Esa Pitkänen and Dr. Jan Korbel. The following analyses were mostly done by Esa, while I characterised the examples.

As a first analysis, we used a set of frequent somatic mutations (SNVs and insertions/deletions (INDELs)) identified in the 2,779 samples from the PCAWG data (Campbell et al., 2017) and verified if any of these mutations overlapped with the CTCF motifs that formed the intragenic loops from the unbiased set (based solely on the motifs, without ChIP-seq evidence, see methods: 2.4.1). Briefly, the 19bp CTCF binding site is referred to as *CTCF motif* while the extension from the motif by +/-100bp is the *CTCF site*, both of which localised either at the promoter or at the intron of the genes.

Overall, we observed that 2,150 out of the 2,779 tumours (77%) had one or more mutated CTCF sites, with 19,501 mutations occurring in total at 14,456 CTCF sites and 2,048 at the CTCF motif (Figure 3.3A). Importantly, sQTLs were significantly enriched at frequently mutated CTCF sites (OR=1.82, p-value = 4.6e-6, Fisher's exact test), which suggests once again that these mutations are functional. Moreover, there were individual CTCF sites with a high mutation load, highlighting that these motifs were repeatedly mutated and pointing to some genes that might be interesting to analyse in depth.

Next, we verified whether the mutation frequency was the same for all CTCF sites or whether there were any changes if the motif participated in a convergent or in any other orientation of the interaction (Figure 3.3B). Remarkably, we observed that convergent sites showed a larger mutation density (9.22e-5 mutations/sites (mut/site)) than non-convergent sites (8.05e-5 mut/site, p-value = 9e-3, U-test). After stratifying by the location of the CTCF site within the gene in either being at the promoter or at the intron, we also observed that the intronic sites were consistently showing a mutation density of 9.07e-5 mut/site for convergent and only of 7.72e-5 mut/site for non-convergent (p-value = 7e-3). Moreover, the convergent intronic CTCF sites mutated more frequently when they co-localised with cohesin (9.95e-5 mut/site for convergent and 8.02e-5 for other orientations, p-value = 3e-3; Figures S3.3A,B). Consistent with these observations, there was a 2.6 fold-enrichment for CTCF motifs to overlap cohesin ChIP-seq peaks at both anchors when the interactions were convergent (using ReMap peaks (Chèneby et al., 2018), p-value = 4.2e-4).

We next decided to analyse the splicing patterns of several genes which contained intragenic loops and where the motif harboured a frequently mutated site (at least 4 mutations were occurring in the same CTCF site for various tumours). Of relevance, we report *PFKFB2* which encodes 6-phosphofructo-2-kinase, a central enzyme of glycolysis (Figure 3.3C). We observed that the CTCF site within the intronic region of *PFKFB2* was mutated in 8 tumours from various tissues (esophagus, stomach, and liver) at multiple positions, having 6 of these SNVs contained within the CTCF motif. We observed that two of the nucleotides in the motif showed two different types of mutations (T > C or T > G), which is compatible with the idea of a motif disruption occurring in this position.

Importantly, by using Rad21 ChIA-PET data from the study of Grubert et al. (see section 3.1 above) we validated that there was a chromatin interaction directly anchored by the CTCF motif under study. In this case, we observed that the mutated CTCF motif in the *PFKFB2* gene localised exactly between two exons that served as alternative end sites for the various isoforms. Of relevance, it is known that the different *PFKFB2* isoenzymes display distinct regulatory and kinetic properties and that cells can coexpress several isoforms suggesting that different isoenzymes cooperate in the regulation of Fru-2,6-P2 levels (Ros and Schulze, 2013).

In addition, we observed that those tumours having the mutated CTCF motif in *PFKFB2* contained an increasing number of interactions that spanned outside of the gene (Figure S3.3C), which made us speculate that perhaps there could be an altered splicing landscape in this locus due to the somatic SNVs occurring in this motif. However, further analyses and a proper quantification need to be done to completely understand what is happening in the samples where the intronic CTCF motif of *PFKFB2* is mutated.

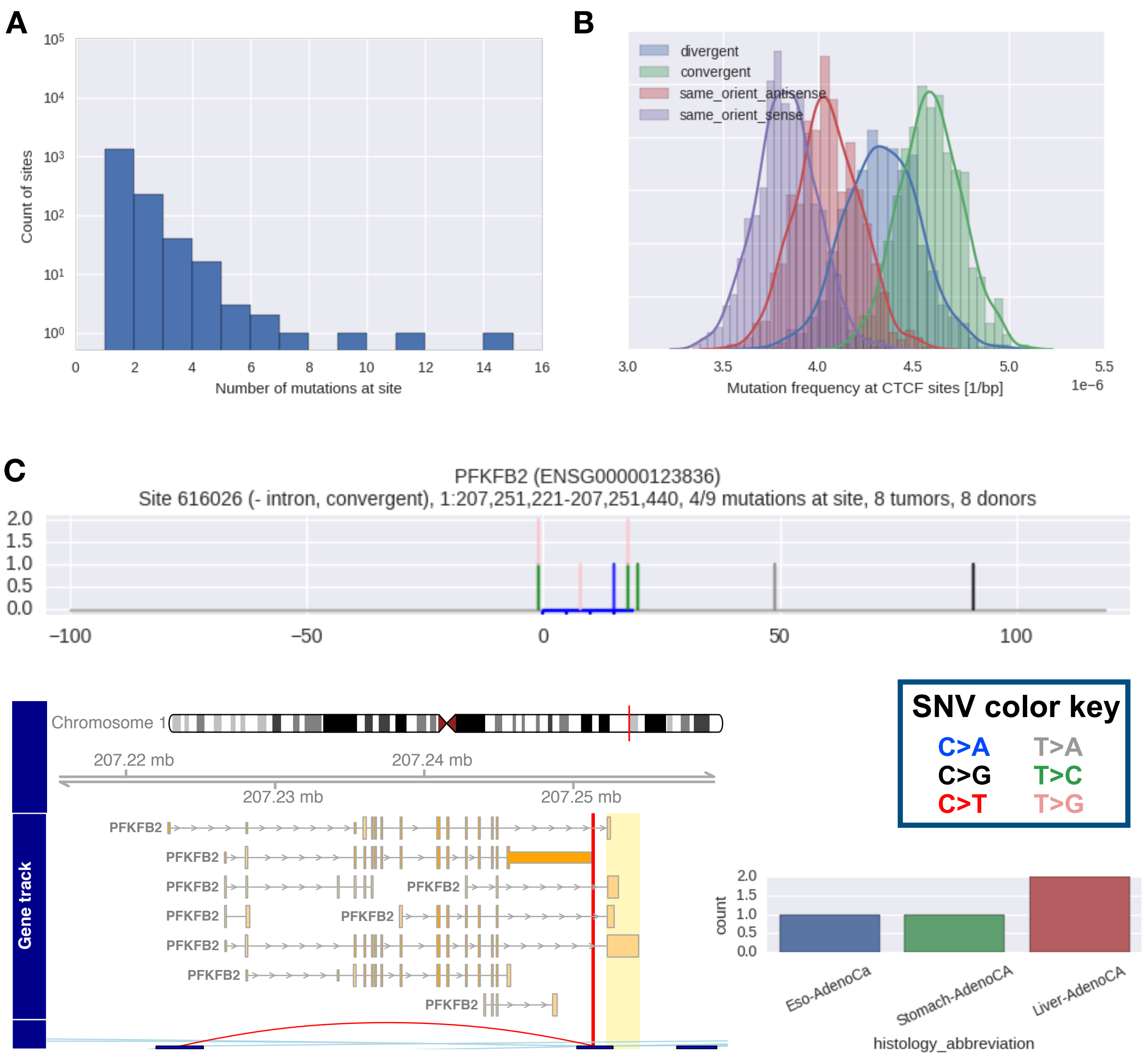


Figure 3.3. Convergent intragenic interactions are the most affected by frequent somatic mutations in cancer
(A) Number of mutations occurring at each CTCF site show that ~10,000 sites have 1 mutation, reaching a plateau around 10 mutations per site. **(B)** Distribution of mutation frequency classified by the directionality of the CTCF motif shows the highest mutation rates for those involved in convergent interactions. The density plots were created by bootstraping the data and plotting the means. **(C)** Gene *PFKFB2* contains a frequently mutated CTCF motif (blue) in its last intron (top). The mutations identified in this site are shown according to the color key and the tissues of the tumours where the motifs were mutated (right). The validated loop along with the precise location of the motif within the gene (red) and the affected exon (yellow) are shown (left). Panels A-C were generated by Dr. Esa Pitkänen, track view of *PFKFB2* was produced by myself.

In summary, by using the CTCF motifs as a predictor of intragenic loops, we decided to analyse if there were observable changes in the mutation frequency at these sites whenever the orientation of the annotated interaction was considered. We have observed a higher mutation rate at convergent sites, which once more suggests that the CTCF sites that anchor intragenic loops are functional, potentially under a tight regulation, and show consequences when altered.

3.3 Discussion

In the previous chapter we have described how we have identified a novel splicing mechanism where CTCF-mediated intragenic chromatin interactions affect the inclusion of the exon next to the intronic anchor of the loop. Nevertheless, we only described our mechanism as occurring in LCLs across individuals, therefore describing inter-individual variation rather than changes occurring between two cell types or two conditions. Given that alternative splicing contributes to the cellular complexity of higher eukaryotic organisms, with nearly 95% of mammalian genes undergoing alternative splicing of pre-mRNA (Kornblihtt et al., 2013; Shukla et al., 2011), we decided to continue with additional studies to find out the extent of our mechanism.

At the beginning of this chapter we describe how we validated that the presence of intragenic loops is ubiquitous across cell types, based on Rad21 ChIA-PET data generated by (Grubert et al., *in revision*). The data produced for this study represents the optimal dataset to test our hypothesis, as RNA-seq and ChIA-PET libraries were available for 22 different cell lines. With both elements we were able to observe an increasing positive correlation for loop strength and exon counts across cell lines and to validate clear examples of intragenic loop-mediated splicing. The major limitation in this study, however, was to delimit loop-exon pairs, as it seems that the mere presence of intragenic loops may still affect splicing in exons surrounding the contacting regions and that the presence of loops could be influenced by gene expression.

Having found this mechanism in multiple cell types, we speculated that the CTCF-mediated loops could also be present in mouse, potentially contributing to biological processes such as neural differentiation. Some previous literature suggested that CTCF was indeed involved in the transition from ESC to NPCs and specific results from these studies hint us to analyse the connection with splicing: a first study described a well-positioned CTCF binding upstream of the TSS in the sense orientation for genes being downregulated upon CTCF degradation (Nora et al., 2017). In agreement with this result, human cells targeted for CTCF degradation also showed CTCF sites close to the promoters of genes that were misregulated (Zuin et al., 2014). Finally, a third study identified multiple interactions going from promoter to gene bodies (although not necessarily mediated by CTCF) by producing high-resolution HiC maps (Bonev et al., 2017). By using CTCF peaks and RNA-seq data we were able to observe similar results as those obtained in human, raising the idea that this mechanism is present in mouse.

Interestingly, while CTCF RNA and protein levels were reported to decrease during mouse neuronal differentiation (Beagan et al., 2017) we observed that the intragenic interactions were more abundant in NPCs, which could suggest that the CTCF remaining throughout differentiation is still retained at these genes as it aids with the splicing transcriptional landscape. Indeed the CTCF motif patterns become sharper and the number of loops increases in NPCs according to our results.

Finally, we decided to investigate if the CTCF-mediated alternative exon usage might be affected during a disease state, in particular in cancer and identified an increased mutation frequency at CTCF motifs involved in convergent interactions. We were also able to identify a central enzyme of glycolysis to have several SNVs in the intronic CTCF-motif, therefore making it an interesting candidate for further studies. It has been shown that mutations in CTCF and cohesin binding sites are often mutated in diverse cancer types (Katainen et al., 2015) and that CTCF is a haploinsufficient tumor suppressor gene which is frequently mutated in several cancer types (Kemp et al., 2014). Furthermore, cancerous cells tend to have a very different DNA methylation landscape than their cell types of origin (Chen et al., 2016). It will therefore be of great interest to investigate whether some of the cancer-driving properties of a cell can be explained by specific transcript isoforms that originate due to differentially methylated regions, or mutations in CTCF binding sites, as initially evidenced by our analysis.

While identifying frequently mutated CTCF motifs in various tumours opens up many relevant cases for further studies, it is still necessary to study the DNA damage pathways that influence such misregulation and the overall consequences of having a disrupted intragenic anchor. Furthermore the identification of an example of CTCF-mediated alternative end usage, rather than the otherwise ‘cassette-exon’ splicing, could imply that our described mechanism has a broader regulatory scope than otherwise thought, indicating once again that we need to investigate this mechanism in further detail.

In summary the mechanism that we have identified, where intragenic CTCF-mediated loops can be found in human and mouse cells, correlates with alternative splicing, and can be regulated either by genetic variants or by the epigenomic landscape therefore influencing the transcriptional landscape of cells.

3.4 Methods

3.4.1 Annotation of Intragenic Loops

The approach described in the previous chapters to identify CTCF-mediated chromatin interactions was used with slight modifications for the analyses done in this chapter:

For the analyses across 22 tissues, we used the Gencode annotation (Release 25; lifted to GRCh37 coordinates), and only kept protein coding genes with at least 1 middle exon. Based on visual inspection, we defined the promoter window as +/- 1kb from the transcription start site and the upstream window as -5kb from the 5'-exon boundary. We then identified intragenic loops as those loops for which one anchor fell in the promoter and the second in the upstream window of the same gene. In this way, we identified 1,372 loops within 1,074 genes. From this set, we identified exon-loop pairs (real pairs) by associating an exon with an anchor of an intragenic loop within 5kb of their 5'-boundaries.

In the case of the mouse data, we followed the exact same approach, leaving +/- 1kb from the TSS and -5kb from upstream regions. We used CTCF ChIP-seq peaks (de Wit et al., 2015) to validated the CTCF motifs calculated with PWM Scan tool (only using those motifs with a score >= 200) (http://ccg.vital-it.ch/pwmtools/pwmscan.php). We once again used CAGE peaks from the FANTOM consortium for mm9 to diminish the alternative start sites in our analysis (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014).

For the cancer section we used the unbiased set of loops, which calculated intragenic loops based solely on CTCF motifs (with a score above a 1000) annotated with PWMscan (http://ccg.vital-it.ch/pwmtools/pwmscan.php). Overlap with cohesin was based on an intersection with SA1 and Rad21 ChIP-seq peaks from ReMap (Chèneby et al., 2018).

3.4.2 Identification of Differentially Used Exons

In the case of the analyses across 22 tissues we flattened the Gencode (Release 25; lifted to GRCh37 coordinates) file with parameters "-r no" and used a modified script to extract counts with subRead (parameters -f -O -s 2 -p -T 40) (Liao et al., 2014) as described in the vignette. We classified the RNA-seq libraries a) according to the 3 clusters identified with the PCA as described above, or b) by the cell line (n=22). Next we normalised for the library size, dispersion, tested for differentially used exons, and estimated the exon log2-FC between a) solid vs blood and stem cell-like vs blood or b) by cell type vs the median exon abundance. In this way, we identified a) 95,137 and b) 39,832 differentially used exons (FDR=10%).

For the mouse data a similar approach as the one described in the previous chapter was used. We used Gencode annotation M1 for mm9 in all cases and the RNA-seq libraries produced by (Pękowska et al., 2014). Briefly, expressed isoforms were retrieved with kallisto for each cell type, *DEXSeq* was run with the cell type as a condition (NPC vs ESC), and a threshold of FDR=10% was used to annotate as DUE. In the case of the CTCF-degron RNA-seq libraries (Nora et al., 2017) we compared all time points of the knock-down (1 day, 2 days, 4 days) against the wild-type (WT) and decided to work with 4 days vs WT, as it maximised the number of DUEs (n=846 at an FDR=10%).

3.4.3 Correlating exon usage and interaction strength

We kept unique exon-loop pairs and correlated the normalised counts of exon and anchor strength across the 22 cell lines. As a control, we permuted all exons 100 times, creating new exon-loop pairs. We also accounted for gene expression by correlating all other exons within the same 'looping' gene and removed any exon within 20kb of the center of the anchor (all pairs). Then we performed a pearson correlation for all complete observations and depicted only the DUEs across the 22 cell lines. For the scatterplot, we used the 3-group classification specified above and we tested for the correlation between real pairs and all pairs of the DUEs.

3.4.4 Enrichment of somatic mutations at intragenic interactions

Dr. Esa Pitkänen from the Korbel group (EMBL) used a set of high-quality somatic mutations (SNVs and INDELs) in cancers (n=2,779) available through the PCAWG project (Campbell et al., 2017). He overlapped these mutations with the CTCF motifs and sites (extending +/-100bp from the CTCF motif) that formed the intragenic loops from the unbiased set (see section 2.4.1). Only those pairs belonging to exactly one orientation were used. The mutation frequencies were calculated as the total number of mutations/site and differences in number of mutations between sites were tested with the U-test.

4. Studying the mechanisms of histone mark propagation along DNA-loops

Knowing that 3D-architecture and function are tightly connected inside the nucleus, we sought to understand the mechanisms by which a distal genetic variant can affect molecular phenotypes which localise megabases away through chromatin contacts in a genome wide manner. In this chapter, we explain our initial attempts to solve this question and propose two models that could explain these observations and that were previously discussed in our review (Ruiz-Velasco and Zaugg, 2017). For this study, we used the data and inspired from the work of Dr. Judith Zaugg and Dr. Fabian Grubert (Grubert et al., 2015). Most of the study was done by me, with analyses of TF co-enrichment and critical input from Ignacio Ibarra. While still at a preliminary stage, this work can lead to exciting results in the near future.

4.1 Introduction

Understanding how genotype encodes phenotype and how genetic variation can impact human traits and disease is still one of the long-standing question in biology. Years of genome-wide association studies (GWAS) have left us with thousands of unique variants associated to traits (~62,000 up to May 2018, https://www.ebi.ac.uk/gwas/home) of which the vast majority are localised at the non-coding regions of the genome. As a consequence, it is now clear that GWAS SNPs frequently affect intermediate molecular phenotypes, bearing quantitative trait loci (QTLs) (Ding et al., 2014; Li et al., 2016; Maurano et al., 2012; Sun et al., 2018). Furthermore, genetic variants can influence distal molecular phenotypes occurring at a considerable distance through chromatin interactions (Delaneau et al., 2017; Grubert et al., 2015).

In particular, a recent study describing the coordinated effect of genetic variation through physical interactions identified local (+/-2kb from TSS) and distal (up to 2Mb from TSS) QTLs (distal-QTLs) which affect the abundance of histone marks and which are abundant and normally contained within topologically associated domains (TADs) (Grubert et al., 2015). Similarly, in an attempt to understand how structure relates to function through genetic variation, another study identified Cis Regulatory Domains (CRDs) which were fully encompassed within TADs, were highly tissue specific, affected gene expression, and in some cases reflected how chromatin is packaged into the nucleus (Delaneau et al., 2017).

These and other studies identified that a common way in which genetic variants affect molecular phenotypes is through the disruption of TF-binding sites (TFBS). Given that CTCF was found to be prevalent at loop boundaries (Rao et al., 2014), we speculated that a possible mechanistic explanation of how genetic variants affect faraway elements could be through the disruption of CTCF binding and the impairment of the loop anchoring. Yet combinatorial binding of TFs could also be affected by distal-QTLs preventing functional genomic elements to properly activate their targets (Zhang et al., 2016).

In this study, we want to understand how distal genetic variants are propagated to distant regions through DNA-loops and how this "transmission of signal" affects other molecular phenotypes contained within these interactions. We propose two models: one in which the close "contact" occurring between the convergent CTCFs anchoring an interaction would explain the transmission of signal, such that only the boundaries would show a coordinated change based on their physical contact. Alternatively, we propose that all elements both within the loop and at its anchors would be affected via "spreading" of the signal along the interaction.

Although the study is still at early stages, preliminary results in a pilot set show that both models are occurring, which could reflect that the process is dynamic and that it is most likely influenced by loop extrusion. Additionally we have gained insights into how the co-abundance of TFs at the anchors of a loop could be disturbed by distal-QTLs and speculate on the consequences that such scenario could generate.

4.2 Results

4.2.1 Characterisation of inter-individual variation at different architectural sites

A first question that we aimed to answer was whether intermediate molecular phenotypes, such as CTCF, cohesin, and histone marks (H3K4me1, H3K4me3, and H3K27ac), would show the same levels of inter-individual variation according to the regions where they localised in the context of chromatin interactions. For this reason, we used ChIA-PET for Rad21 (40,281 loops) and H3K4me3 (10,037 loops) (Grubert et al., 2015) and HiC data

(9,166 loops) (Rao et al., 2014), all of which were done in lymphoblastoid cell lines (LCLs). We split the chromatin contacts into three categories that represent the parts of a loop, while other regions not contained within the interactions were labelled as "inter-loop domains" (schematic on Figure 4.1B):

- *Outer boundaries*: regions of the genome that are physically interacting and form the outermost loop anchors.
- *Inner boundaries*: loop anchors within an outer interaction, that potentially represent intermediate maturation states inside a domain or TAD.
- *Inside regions*: regions within the boundaries that do not overlap any other anchor.

As a way to verify the type of genomic features that were anchoring or captured by the interactions of the different chromatin maps, we visualized the enrichment of chromatin states at the various loop regions (Figure 4.1A). In this way we noted that outer and inner loop boundaries of HiC and Rad21 ChIA-PET were enriched for CTCF, while inside regions were depleted. On the other hand, H3K4me3 ChIA-PET outer and inner boundaries showed enrichment for TSS- and enhancer-related states, but inside regions were only enriched for the latter. These results showed that according to the dataset used we could expect loop populations to capture different interacting functional elements. Specifically, H3K4me3 maps should contain more regulatory interactions formed between enhancers and promoters, while Rad21 and HiC maps contain more topological-like contacts.

We next obtained CTCF, SA1 (cohesin), H3K4me1, H3K4me3, and H3K27ac ChIP-seq peaks from 18 individuals (Kasowski et al., 2013) and we verified the coverage of the peaks per megabase (Mb) according to the loop categories described above (Figure 4.1B, S4.1A). We observed that there are more peaks binding at the loop boundaries when compared to the inside regions. Furthermore CTCF bound more at outer than at inner boundaries, while SA1 showed the opposite pattern. These results are in agreement with a report showing that disruption of CTCF has a greater role in spatial segregation, potentially due to its enriched presence at TAD boundaries, while cohesin was demonstrated to be involved in chromatin interactions within topological domains (Zuin et al., 2014).

Additionally, we questioned whether the inter-individual variability and binding strength would also be affected by the loop region where the factors bound. For this reason we verified the distribution of the median binding strength, as well as the standard deviation (SD), for the peaks in each loop category across the 18 individuals (Figure 4.1C, S4.1B). Importantly, we observed that smaller SD for CTCF and SA1 peaks at loop boundaries which were significantly lower for the outer than for the inner boundaries (CTCF: p-value = 5.56e-11, 1.01e-90, 2.98e-161 for HiC, H3K4me3 and Rad21 ChIA-PET, respectively. SA1: 3.2e-8, 1.01e-90, and 3e-161, Wilcoxon test). Correspondingly, we noticed that the binding strength was significantly higher for CTCF and SA1 peaks at outer boundaries (CTCF: p-value = N.S., 5.4e-49, and 1.8e-140 for HiC, H3K4me3 and Rad21 ChIA-PET, respectively. SA1: 9.5e-3, 3.2e-44, and 9.3e-140, Figure 4.1D).

However, while the stronger binding at boundaries was still noticeable for some histone marks in specific datasets (Figures 4.1D, S4.1B,C), the decrease in SD values at loop boundaries was not so striking except for the H3K4me3 histone mark in the H3K4me3 ChIA-PET dataset. Moreover, the patterns were almost inverted with stronger binding of histone marks at inner boundaries. In agreement with H3K4me3 ChIA-PET capturing mostly enhancer-promoter interactions, we also find the lowest variability and the strongest binding for all three histone marks here.

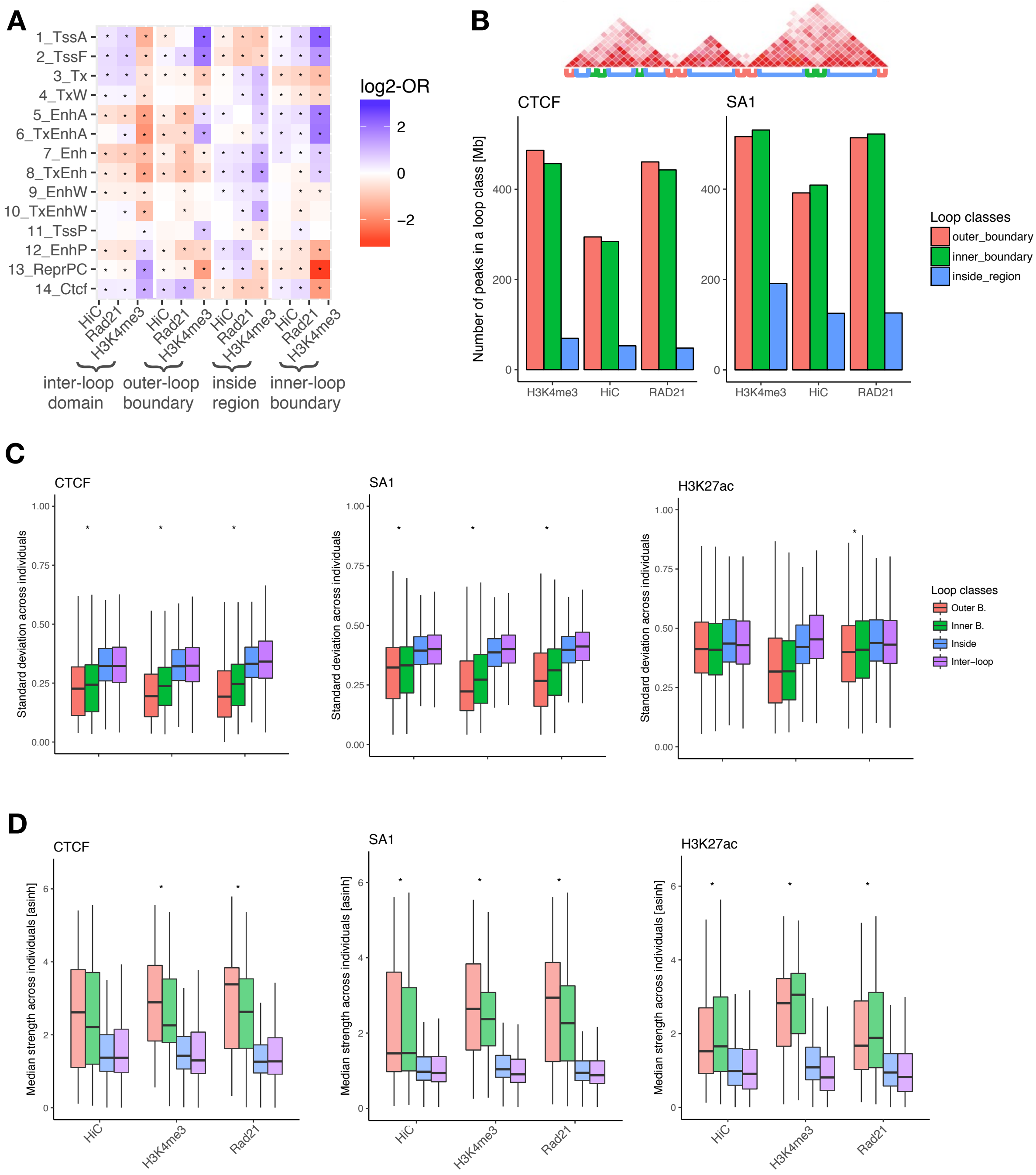


Figure 4.1. Characterisation of loop regions across individuals

(A) Enrichment (blue) and depletion (red) of chromatin states at loop regions (outer-, inner-boundaries, inside-regions, and inter-loop domains) (Significant p-adjusted values = *). **(B)** Peak coverage per Mb for CTCF (left) and SA1 (right) at loop regions of the 3 data sets. A schematic of the parts of the loop is shown on top of the plots according to the color code in the legend. **(C)** Distribution of peak standard deviation and **(D)** median binding strengths for CTCF (left), SA1 (middle), and H3K27ac (right) across 18 LCLs according to where they bind in the context of loop regions. All three maps were visualised and Wilcoxon tests measured differences between outer and inner boundary values (*); peak values were asinh-normalised. All panels were done by myself.

In summary, by classifying three interaction datasets according to their contacts and then verifying the distribution and binding strength of ChIP-seq peaks for various factors across individuals, we were able to characterise and validate that CTCF and SA1 preferentially localise at outer boundaries, possibly serving as TAD boundaries, with a strong binding and small standard deviations. Meanwhile histone marks for active genomic features behave in a different mode and could potentially show functional loops in charge of the transcriptional activation by bridging regulatory elements such as enhancer-promoter contacts.

4.2.2 Architectural and functional loops are likely to be affected by distal genetic variants through the disruption of combinatorial protein binding

Recently, it has been shown that 3D-proximity mediates the influence of genetic variants on distal gene regulation and that diverse pairs of interacting regulatory elements are genetically coordinated (Grubert et al., 2015). In order to better understand how distal-QTLs affect faraway genomic factors (name hereby used in the text to refer to either histone marks, CTCF or cohesin) and the mechanisms that are potentially disrupted, we defined a subset of high-confidence distal-QTL-genomic factor pairs that overlapped with experimentally determined loop anchors and named them loop-QTLs.

We required that each side of the interaction intersected either with the distal-QTL or with the genomic factor and therefore identified 88 loops from HiC, 143 from H3K4me3 ChIA-PET, and 335 for Rad21 ChIA-PET in LCLs (1%, 1.4%, and 0.8% out of the complete set of interactions described in 4.2.1, respectively. Figure 4.2A). Most anchors of loop-QTLs were annotated as inner loop boundaries (for both sides) and only few as outer; this was particularly evident for the Rad21 ChIA-PET loops, where inner boundaries were 1.6-fold enriched to also be loop-QTLs (p-value = 5.76e-5, Fisher's exact test). This result suggests that loop-QTLs preferentially affect functional interactions but not so frequently TAD borders, a result also in line with the reported structural CRD-QTLs which are stated to be under a strong selective pressure and affected only by rare genetic variants (Delaneau et al., 2017).

We speculated that two possible mechanisms affected by distal-QTLs involved either the disruption of a TFBS, which could potentially also alter functional combinatorial binding or alternatively the specific disruption of CTCF which could prevent the formation of a loop (Figure 4.2B). In other words, the presence of a SNP would disrupt the binding of one TF locally and influence the enrichment of a second TF at the other anchor of the loop. We therefore decided to verify if there were any co-abundances of TFs at the boundaries of loop-QTLs, an analysis done by Ignacio Ibarra. By using combinatorial pairs of TFs supported by ChIP-seq peaks from ReMap (Chèneby et al., 2018) present at each of the anchors, we observed that there were various significantly co-enriched pairs for which some of the TFs have been described to affect transcriptional activation (Mourad and Cuvier, 2016) (Figures 4.2C, S4.2B).

Importantly, when checking the top 20 significant co-enrichments in loop-QTLs we were able to observe many interactions that were recently described in a similar study (Zhang et al., 2016). For example, the described DNA-binding protein (DBP) network of PML-FOXM1-STAT5A-RUNX3 showed significant co-enrichments in our hands too for all of these TF-pairs, some of which have validated biochemical interactions. Moreover, the previous study also reported a co-occurrence of PML-STAT5A-CEBPB peaks with cell type specific binding of PAX5-BATF-EBF1-BCL11A-FOXM1-NFIC-ATF2-RUNX3. TFs such as RXRA, BCL11A, ELK1, and P300 were reported to be involved in activation, therefore being depleted for TAD borders (Mourad and Cuvier, 2016).

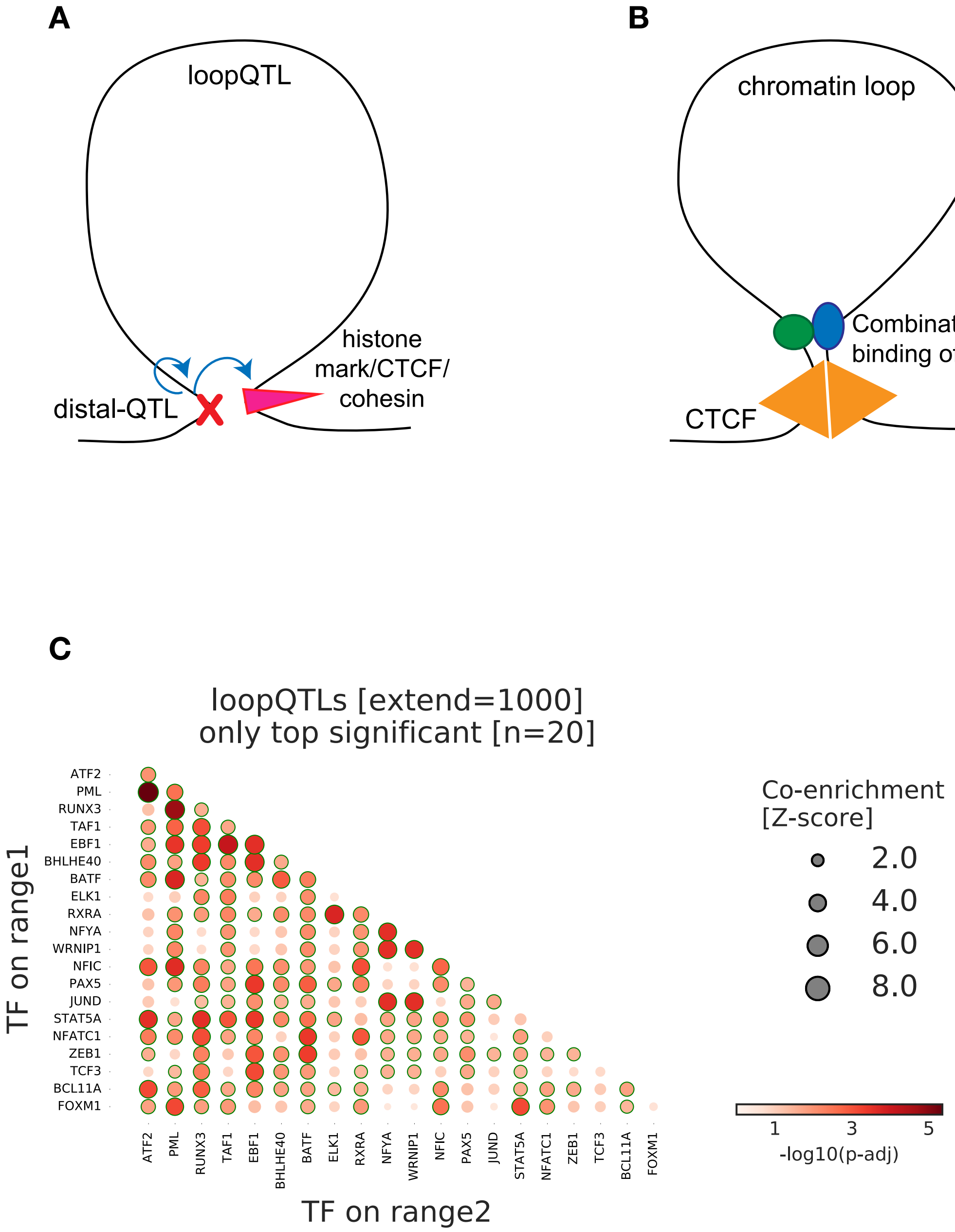


Figure 4.2. Defining loop-QTLs and verifying possible protein co-enrichments

(A) Schematic of a loop-QTL which contains a distal-QTL-genomic factor pair supported by a physical contact. **(B)** Two of the possible mechanisms disrupted by distal-QTLs are a CTCF-dimer anchoring a chromatin loop or a functional interaction of two genomic elements. **(C)** Co-abundance of combinatorial pairs of top 20 most significant TFs occurring at the anchors of a loop-QTL: TF1 at the y-axis and TF2 at the x-axis. The size of the circle shows the adjusted p-values (FDR=5%, green circles are significant) and the color indicates the fold enrichment (Fisher's exact test). Panels A and B were done by myself, panel C was contributed by Ignacio Ibarra.

We also observed NFYA to be highly enriched to interact with itself, with PML, TAF1, BATF, and RXRA as well as NFYA, PML, and RFX factors, which are known to be co-occurring to active the major histocompatibility II genes (Arampatzi et al., 2013). Overall, the combinatorial function of these TFs has been implicated with leukocyte and lymphocyte activation as well as immune response (Zhang et al., 2016), which could explain their prevalence in LCLs. It will be important to have a closer look at the reported co-enrichments to verify if there are any novel associations that could be worth following up.

Our second hypothesis was that distal-QTLs could be explained by the disruption of CTCF on one side, affecting binding on the other side of the interaction and potentially impacting loop formation (Figure 4.2B). This idea is supported by the knowledge that CTCF anchors chromatin loops through dimerisation and that loops can be eliminated in many cases by removing one of the binding sites (de Wit et al., 2015; Guo et al., 2015). With this aim, we retrieved the genotypes of 14 out of the 18 individuals and intersected the heterozygous SNPs with all CTCF peaks. To our surprise, most of the peaks (158,455, corresponding to 80%) contained 1 or more heterozygous SNPs. We then defined which of these were allele-specific CTCF (asCTCF) using *SNPhood* (Arnold et al., 2016) and obtained 93,438 cases (59% of heterozygous peaks).

In agreement with our hypothesis, we observed an enrichment for asCTCF peaks to occur at boundaries of loop-QTLs when compared to the complete set of loops for the three interaction maps (OR= 1.76, 1.5, and 1.3, p-values = 3.21e-4, 6.42e-5, 4.61e-2 for H3K4me3, HiC, and Rad21 respectively). This result suggest the idea that loop-QTLs contain altered levels of CTCF that could explain the disrupted mechanism.

Overall we defined loop-QTLs to be a high-confidence subset of the distal-QTL that have chromatin interaction support and show that they represent mostly inner boundaries and are enriched for containing asCTCF peaks. We also performed a co-enrichment TF analysis that shows several TF-pairs which are co-occurring at loop-QTL anchors, many for which a previous validation exists, opening up exciting hypothesis of the mechanisms disrupted by distal-QTLs.

4.2.3 Proposing possible mechanisms in which 3D-structure conveys functionality

We decided to further study the mechanism by which a distal-QTL could propagate to the other end of a loop and produce a changing phenotype such as a hQTL. We proposed two possibilities that could begin to explain the propagation of signal and that were previously described in our review (Ruiz-Velasco and Zaugg, 2017) (Figure 4.3A):

(i) "touch-and-act model": where the physical contact at the loop anchors would be the only affected, without changing the genomic factors within the DNA-loop to convey its functionality.

(ii) "spreading model": where the complete loop would show an altered function and therefore a coordinated activation or repression of the whole local neighbourhood.

We decided to prove our hypothesis by testing the allelic fraction of intermediate molecular phenotypes along the loops. If the allelic fraction showed a preference to map to one allele only at both boundaries -but not at the inside region- we would predict the touch-and-act model, while if the allelic fraction showed a constant preference along the loop it would be showing the spreading model, which predicts a coordinated activation/repression through the entire domain (Figure 4.3B). Yet those interactions where one anchor showed an opposite allelic fraction to the other end would indicate that the model cannot be tested as the allele-specific events are not linked in the same direction.

Briefly, we followed the same approach as for defining asCTCF by first identifying all heterozygous SNPs per individual, overlapping them with the ChIP-seq peaks (SA1, H3K4me1, H3K4me3, and H3K27ac), extracting the read counts surrounding the SNP (+/- 250bp) using *SNPhood* to calculate the allelic fraction and adjusted p-value (Methods: 4.4.3). We retrieved on average 240,000 unique peaks per individual which contain a heterozygous SNP, all of which were pooled and classified according to the loop regions (Figure S4.3A,B). By keeping those loops having allele-specific factors at both anchors and in total at least ten allele specific events, we were able to obtain 487, 284, and 1909 loops for HiC, H3K4me3, and Rad21 ChIA-PET, respectively. We then normalised the loop length as percentages to visualise all the interactions together and interpolated the data to fill in the blanks (Figure 4.3C).

We decided to check the proportion of loops that fell into each of the models with a heatmap and used k-means to cluster each of the datasets into 6 groups. Strikingly, we observed evidence for the two models in all three datasets at similar proportions. For the HiC data we observed 35% (clusters 5 & 6) of loops to describe the spreading model, 33% (clusters 1 & 4) for the touch-and-act, and 32% (clusters 2 & 3) to be not-linked. Meanwhile H3K4me3 data had 34% (clusters 1 & 2) for the spreading, 34% (clusters 3 & 6) for touch-and-act, and 32% (clusters 4 & 5) for each not-linked while the Rad21 dataset had 33% (clusters 3 & 6), 33% (clusters 1 & 5), and 34% (clusters 2 & 4) respectively. We acknowledge that a major limitation at this point is that the analyses are mostly qualitative and not quantitative and that a larger dataset should be used to gain further insights into the prevalence of each of the models.

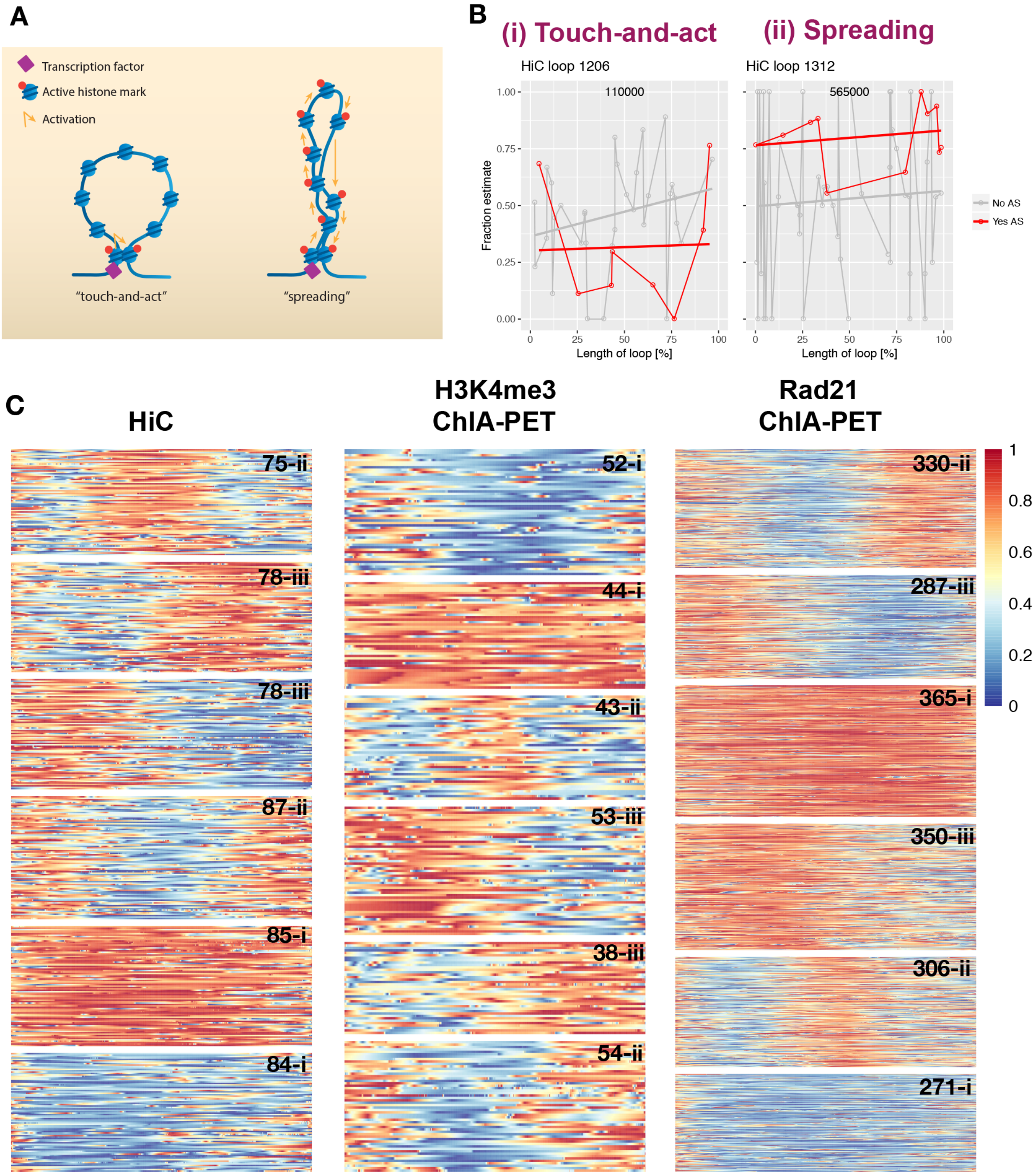


Figure 4.3. Models proposing how functionality is conveyed along DNA-loops

(A) Schematic of the two proposed models for the transmission of signal from a SNP to a distal site. **(B)** Examples of how the two models look like for HiC interactions of one individual when the fraction estimates (y-axis) are visualised along the loop-length [%] for heterozygous (grey) and allele-specific (red) factors. **(C)** Heatmaps grouped with k-means (6 clusters) provide evidence for the presence of both models (i=spreading, ii=touch-and-act; iii=non-linked) at different proportions for the 487, 284, and 1909 loops for HiC, H3K4me3, and Rad21 ChIA-PET respectively. Colour key shows allelic fraction, where red shows reads mapping preferentially to one allele, blue to the other allele and yellow as the 50% of reads contributed by each allele. Panel A was taken from (Ruiz-Velasco and Zaugg 2017) where we briefly mention the models, panels B and C were done by myself and are unpublished.

4.3 Discussion

Multiple studies from the last years have provided us with extensive knowledge and a collection of chromatin conformation data, which has increased our knowledge about chromatin 3D-organisation. However our understanding of how such architecture influence the function of downstream processes, and how structure translates into function, is still scarce and just beginning to emerge (Ruiz-Velasco and Zaugg, 2017). One such question is to understand how genetic variants can affect distant sites through physical contacts and if such transmission of signal would also produce changes in the surrounding intermediate molecular phenotypes.

In this study we took advantage of the prevalent inter-individual variation of intermediate molecular phenotypes in LCLs and tried to exploit it to get mechanistic insights that explained distal-QTLs in the context of 3D-organisation. We defined loop-QTLs as those distal-QTL-genomic factor pairs with a validated physical interaction either in HiC or in ChIA-PET. The finding that the outer loop boundaries most likely represented topological domain boundaries which were less likely to be loop-QTLs could be explained by the stronger binding of CTCF, which implies that it is more stable and less variable based on its lower SD. While we speculated that CTCF-dimer disruption could begin to explain the way in which distal-QTLs exert their function, we did not observe this to occur frequently, which is in line with the idea that these interactions should be the most robust to safeguard the regulatory domains along the genome.

On the other hand, we tested whether a loop-QTL could be explained by the disruption of a combinatorial binding of two TFs in the 3D-context. These analyses provided us with previously described and novel TF pairs which were mostly involved in immune response and lymphocyte-leukocyte activation. We speculate that a possibility is that even when the physical loop is not affected by the distal-QTL, a functional interaction could still be disrupted. However further analyses including extensive characterisation of the inside regions will need to be done to draw any additional conclusion in this regard.

We decided to test whether a distal-QTL would affect only those loci in close physical interaction or whether it would be propagated along the entire loop, which led us to propose two models: the “touch-and-act” and the “spreading” model for the cases just described, respectively. We used allelic biases in genomic factors distributed along loops to verify both models and the proportions to which they happened. Interestingly there is evidence for both of these mechanisms, which could either suggest that there are multiple mechanism of how 3D structure conveys its function that depend on the combination of loci and regulatory factors (Ruiz-Velasco and Zaugg, 2017) or represent the various maturation states during loop extrusion.

While this study is still at its early stages and more improvements and analyses can be done, we are excited about the prevalence of both models and the possibility of gaining functional insights to the mechanisms by which genetic variants affect the genome. Although a couple of studies with varying approaches have been published since we started our study, an advantage in our analyses is that we are using also ChIA-PET data to test our models, which should capture more functional interactions given that it filters for specific proteins know to be present at loop anchors, and that we plan to extend our study to the existing cohorts of LCLs to maximise our results. Finally, we would need to also validate that such models apply only when the loop is present and not for peaks outside chromatin interactions.

4.4 Methods

4.4.1 Overview of the samples and datasets used

We obtained ChIP-seq data (CTCF, SA1, H3K4me1, H3K4me3 and H3K27ac) for LCLs of 18 individuals from (Kasowski et al., 2013). We used Rad21 and H3K4me3 ChIA-PET (Grubert et al., 2015) and HiC data (Rao et al., 2014). The genotypes of 14 individuals were retrieved from the 1000 Genomes project (http://www.internationalgenome.org/) (1000 Genomes Project Consortium et al., 2015). Only autosomes were used for all the analyses.

4.4.2 Defining loop-QTLs

We intersected the HiC (boundary width of 5 kb) and ChIA-PET (boundary width of 3 kb) loops with the local and distal QTLs pairs defined by (Grubert et al., 2015). We extended a window of +/- 5kb from the local and the distal-QTL pairs for the HiC and for the ChIA-PET loops.

4.4.3 Co-enrichment of TFs at loop-QTLs

We checked if there were co-enrichments of TF pairs in loop-QTL regions by interrogating whether ChIP-seq peaks for both transcription factors are more frequent in both sides of the loop, with respect to loops with one or no TF peaks. ChIP-seq peaks were collected from ReMap (Chèneby et al., 2018) and a distance of 1kb was defined as a cutoff to define if a ChIP-seq peak was overlapping a loop anchor.

For a pair of TFs named TF1 and TF2, its co-enrichment in loops with anchor sites 1 and 2 can be interrogated by counting four possible overlaps between ChIP-seq peaks and anchors: (a) TF1 in anchor 1 and TF2 in anchor 2, (b) TF1 in anchor 1 and TF2 not in anchor 2, (c) TF1 not in anchor 1 and TF2 in anchor 2, (d) TF1 not in anchor 1 and TF2 not in anchor 2. From these numbers, an odds ratio is defined via Fisher's exact test. Due to the dependency of this odds ratio to the assignment of TF pairs in anchor sites 1 or 2, reported co-enrichments are the highest observed value between both possible assignments. P-values were corrected for multiple testing with Benjamini Hochberg and significance was assigned at an FDR=5%. Z-scores were calculated through permuting the observed occurrences for both TFs in the interrogated anchor sites 10000 times, and re-estimation of (a).

4.4.4 Identification of allele-specific genomic factors

We first identified all heterozygous SNPs per individual and overlapped their coordinates with the ChIP-seq peaks, keeping only the SNP closest to the center of the peak. We then ran *SNPhood* (Arnold et al., 2016) to extract the read counts surrounding the SNP (+/-250bp) for each factor and each of the 14 individuals and calculated the allelic bias (binomial test, see methods: 2.4.5), corrected for multiple testing with Benjamini Hochberg, and classified as allele specific (FDR=10%).

For each individual, a table containing the collection of all peaks-SNPs that are heterozygous along with the information from *SNPhood* was created. The peak-SNP pairs were also then overlapped with Rad21, H3K4me3 ChIA-PETs and HiC datasets, where each peak-SNP event was classified according to whether it overlapped any of the loop categories: 'outer/inner_boundary', 'inside_loop', and 'inter-loop-domain'.

4.4.5 Classifying loop-QTLs into models

Only those loops that have >= 1 heterozygous peak_SNP event in both of the boundaries were saved. As in several cases the same SNP was overlapping various genomic features, we reduced the points and only retained the SNP that was allele specific or the one with the lowest p-adjusted value. We applied this function to all loops that had at least one allele specific event in each of the boundaries of the interaction.

We normalised the loop length by using percentages and to allow the visualisation in a heatmap, we filled the missing values with a linear interpolation. We next clustered the signal along loops to view the extent to which each model is related to using k-means (k=6) and used pheatmap for visualisation.

5. Global chromatin changes alter regulatory landscape in ageing stem cells

The other chapters have described the intricate link between structure and function in the mammalian nucleus of healthy cells. Here we investigate how chromatin structure is affected by ageing. This study is a collaborative project with Dr. Anne-Claude Gavin, Dr. Mang Ching Lai, and Dr. Ximing Ding (EMBL) and aims to identify chromatin accessibility changes occurring with age, which can ultimately be reflected in gene and protein abundance changes in mesenchymal stem cells (MSCs). All the computational analyses shown here, along with the text and figures, have been done by me unless explicitly stated. Experimental work has been done by Dr. Ximing Ding, Dr. Marco Hennrich, and Dr. Mang Ching Lai.

5.1 Introduction

Aging is a gradual process that ultimately ends in the loss of physiological integrity and death of an organism (Booth and Brunet, 2016; López-Otín et al., 2013). While several years of research have elucidated key features of aging, mostly at the transcriptome and proteome level, studies are only starting to assess the contribution of epigenetics in such process (Booth and Brunet, 2016; Liu et al., 2013; Sun et al., 2014). Moreover, most of the existing reports have studied senescence or replicative-ageing and only a handful have focused on healthy human ageing.

Two of the hallmarks of aging are the decline in the regenerative potential of tissues due to stem cell exhaustion, which can lead to malignancies and to an increasing genomic instability caused by lesions in DNA and defects in nuclear architecture (López-Otín et al., 2013). Recently, the SyStemAge consortium generated a valuable cohort across 59 individuals to study age-related changes occuring at the protein level for the various cell types of the bone marrow (BM) niche, including hematopoietic and mesenchymal stem cells (HSCs and MSCs) (Hennrich et al., *in revision*). MSCs have been recognised as a critical component for the maintenance of a healthy BM niche (Reagan and Rosen, 2016) as they are able to differentiate into various mesenchymal tissues including bone and fat throughout life (Scaffidi and Misteli, 2008). While Hennrich's study have revealed several insights about the ageing process, many proteins, and in particular transcription factors (TFs), were not abundant enough to be captured by mass spectrometry, resulting in an incomplete view and leaving a "blindspot" of the changing regulatory landscape with age.

In this study we have obtained data for RNA and proteins from the SyStemAge project in an attempt to integrate both molecular levels. Surprisingly, we observed that chromatin related proteins were significantly downregulated with age, which made us wonder whether we could identify major changes in chromatin associated with age in healthy human MSCs. For this aim and due to the limited amount of available cells, we used an assay to profile open chromatin.

By performing comparisons between older and younger individuals, we were able to elucidate global and specific changes occurring in chromatin accessibility, as well as changes in gene expression and isoform usage which correlated with age. As a way to fill in this “blindspot” of proteomics, we successfully applied a method that verifies the TF activity changes between age groups (Berest et al. *in revision*). By doing such analysis we were able to observe that there are changes in the regulatory landscape of MSCs that result in a differentiation bias towards bone in younger and fat in older individuals. In addition, by looking at the TF activity we were able to identify (i) changes in Polycomb group (PcG) proteins that correlated with the down-regulation of accessibility signal at bivalent promoters and (ii) changes in centromere-associated proteins, which suggests an increased genomic instability with age.

Through a multi-omic integration approach, our study has validated multiple changes observed in human stem cells with age, at the time it has identified new ageing-associated genes which still need to be studied. Overall these analyses shed some light on how time affects chromatin organisation, ultimately leading to changes at other biological levels which sum up to the impaired functionality observed in the ageing process.

5.2 Results

5.2.1 Establishment of a bone-marrow mesenchymal stem cell dataset to assess age-related changes in chromatin

The SyStemAge consortium (Hennrich et al. *in revision*) collected BM from 59 healthy donors ranging 20-60 years old (yr), to later study changes at the transcriptomic and proteomic level with respect to age in the various cell populations of the BM-niche. While valuable insights were obtained from this study, non-targeted proteomic techniques still face some limitations, especially when the abundance of the peptides is scarce like for TFs.

We reasoned that both transcription and chromatin accessibility could help us to fill in the "blindspot" of proteomics by providing additional genes that were not captured in the proteomics assay and also by showing changes in accessibility that can result from preferential TF occupancy at a given condition (Figure 5.1A). Furthermore, a multi-omics integration could provide a more solid view of how specific genes are altered during human ageing. We then selected a subset of the SyStemAge samples for which proteomics (n=45 with MSC data), and in some cases RNA-seq datasets (n=19), were available and generated chromatin accessibility with ATAC-seq (n=16) (Figure 5.1B). We centered on MSCs as their commitment into adipocytes and osteoblasts has been previously implicated in pathological conditions and their differentiation potential can be affected with age (Moerman et al., 2004).

We verified that all the datasets were of good quality by doing Principal Component Analyses (PCAs) and additional controls per technique, which left us with 45 samples for proteomics, 18 for transcriptomics, and 15 for chromatin accessibility (Figures 5C-F, S5.1A-C).

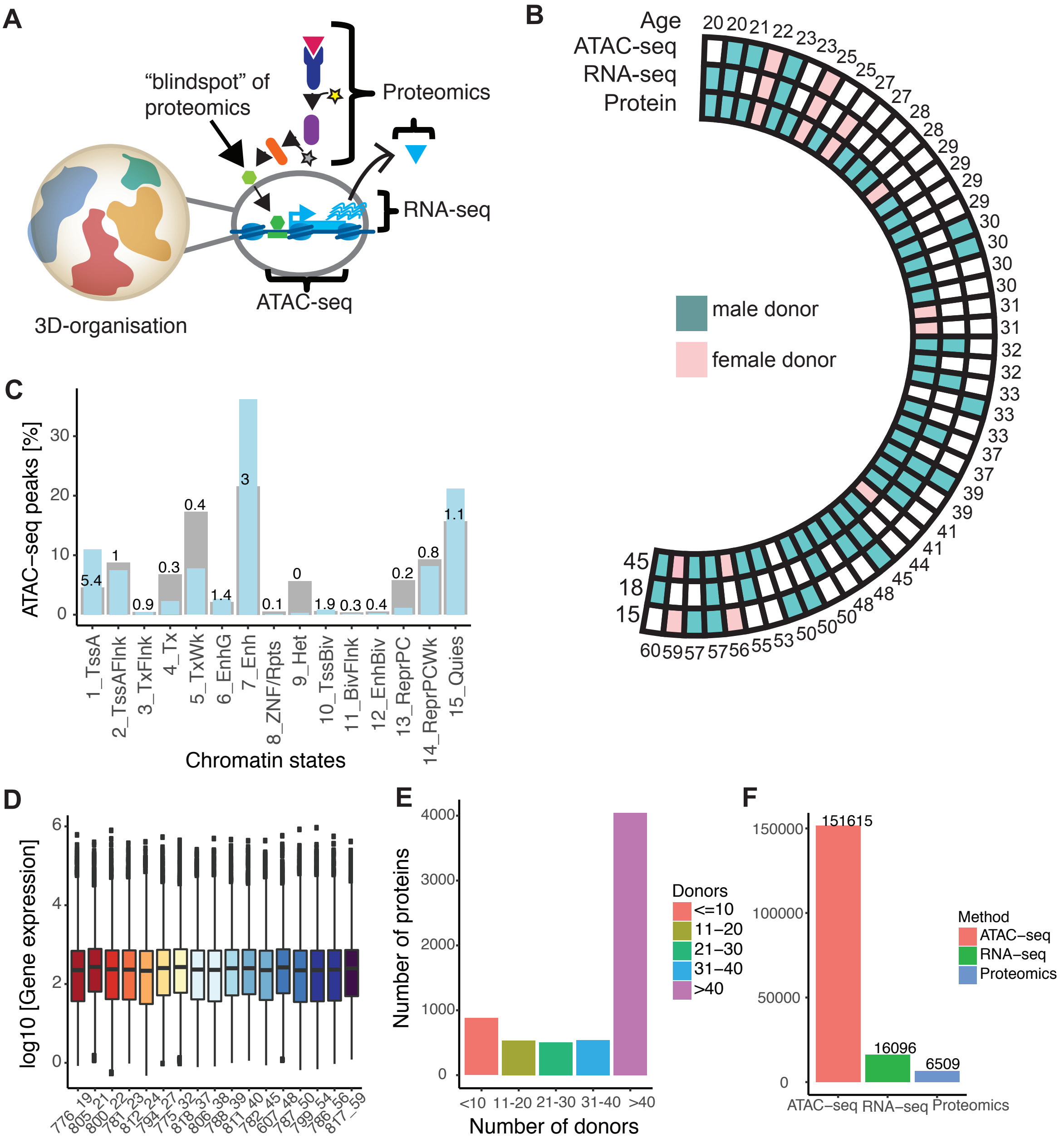


Figure 5.1. Overview and quality control of datasets

(A) Depiction of the integration analyses for the three datasets: By using RNA-seq and ATAC-seq data we intend to fill in the ‘blindspot’ of proteomics and also to gain some knowledge of global changes occurring in chromatin structure with age. **(B)** Summary of the type of data available from each of the MSC donors along with their age and sex. **(C)** Distribution of ATAC-seq peaks [%] per chromatin state (blue) vs the total percentage of each state as the background (grey). **(D)** Distribution of gene expression values (log10) from RNA-seq libraries for all the good quality donors. **(E)** The number of donors where a protein was detected. For most of the analyses only proteins detected in more than 10 donors were considered. **(F)** Number of peaks (ATAC-seq), genes (RNA-seq), and proteins (proteomics) defined in our study. Panel A was produced in collaboration with Dr. Judith Zaugg, panel B was done in collaboration with Dr. Ximing Ding, panels C-F were produced by myself.

We required ATAC-seq libraries to have a good TSS-enrichment score (enrichment of read counts at +/-2kb from TSS vs 4kb outside of this window, methods 5.4.3, Figure S5.1A), which reflected in the majority of peaks occurring at open regions (enhancers and promoters), a pattern that is expected and resembles the genomic distribution of DNase hypersensitivity (DHS) data (Figures 5.1C, S5.1D). For RNA-seq, we verified that the distribution of the gene expression per sample was not highly variable for any given library (Figures 5.1D, S5.1B,C) and for proteomics we checked that most of the proteins used were detected in a large number of samples (Figure 5.1E). This filtering left us with 151,615 ATAC-seq consensus peaks, 16,096 expressed genes, and 6509 quantified proteins (Figure 5.1F).

5.2.2 Integration of transcriptomics and proteomics data reveal global changes in chromatin related proteins upon ageing

We first identified differentially expressed (DE) genes in the RNA-seq data using *DESeq2* (Love et al., 2014), which rendered 275 genes more abundant in young and 269 in old individuals at an FDR=10% (Figure 5.2A, Methods: 5.4.6). We then performed a Gene Ontology (GO) enrichment analysis for the two groups with *clusterProfiler* (Yu et al., 2012). Interestingly, we observed that the cellular compartments for the genes that were downregulated with age related to the nucleus (such as “chromosomal region” or “centromeric region”), while the genes that were upregulated related more to the “extracellular matrix” (ECM) and the “membrane” (Figure 5.2B). Biological processes such as “DNA replication” and “chromosome segregation” were therefore present in the younger and “extracellular matrix organisation” in the older samples (Figure S5.2A), which is in agreement with previous reports of DE genes in mouse stem cells (Sun et al., 2014).

Next we verified how the proteomics dataset behaved by visualizing the distribution of the spearman correlation for the protein abundance with respect to age (Figure 5.2C). This showed that 287 proteins were significantly more abundant in younger and 436 in older individuals at a nominal p-value < 0.05.

One of our first aims was to analyse whether changes in gene expression that are observed with age are consistent at the protein level in MSCs. For this we compared the log2-fold change (FC) abundance of the genes [old/young] and the spearman correlation for their encoded proteins (Figure 5.2D). By doing this comparison, we observed that there is an overall positive correlation (R=0.3, pearson correlation) between the level of the gene and the protein abundance, and there is an enrichment for a gene to be more expressed if there is a larger protein abundance with age (OR=1.95, p-value = 5.0e-39, Fisher's exact test). This trend was even more striking when we only focused on the DE genes (OR=13.4, p-value = 4.2e-21, Fisher's exact test), showing also a strong positive correlation (R=0.61, pearson correlation). The correlations further improved when filtering for the number of donors where the protein was detected (Figure S5.2B).

In agreement with previous studies, we observed that histone H1 was more abundant in younger individuals (Feser et al., 2010; Funayama et al., 2006), while fibronectin in older individuals (Kular et al., 2014). Notably, besides histone H1, we observed that most of the histones, lamins, condensins, and nucleoporins, along with many other chromatin related proteins annotated with the EpiFactors database (Medvedeva et al., 2015) (e.g. EZH2, DNMT1, and CTCF) were negatively correlated with age (Figure 5.2D). The observed histone loss made us wonder whether this would lead to global chromatin changes in ageing. Indeed, when comparing the abundance of chromatin related genes/proteins with all other transcripts/proteins quantified they were significantly downregulated with age at both levels (p-value = 3.3e-5 and 2.2e-16 respectively, t.test) (Figure 5.2E).

To verify that what we were observing was not only due to senescence but rather due to ageing, we made a similar comparison for the RNA and the protein levels of HSCs from the same study (Hennrich et al. *in revision*). While the shift of chromatin related proteins was still significantly biased towards the younger individuals, it was clear that HSCs don't show such a striking pattern as MSCs especially at the RNA level (p-value = 0.41 and 2.2e-16, RNA and protein levels, t.test), indicating that a fraction of the chromatin factors downregulated with age resulted from ageing and another from senescence (Figures S5.2C,D, Table S5.1). However further analyses need to be done to classify the latter factors as it has been described that adult stem cells are not subject to replicative senescence while they can still accumulate damage with increasing age (Schultz and Sinclair, 2016).

In summary, by comparing RNA log2-FC [old/young] with correlations of protein abundance with age, we observed the downregulation of several chromatin-related proteins. This finding led us wondering whether we could also observe these regulatory changes at the chromatin accessibility level.

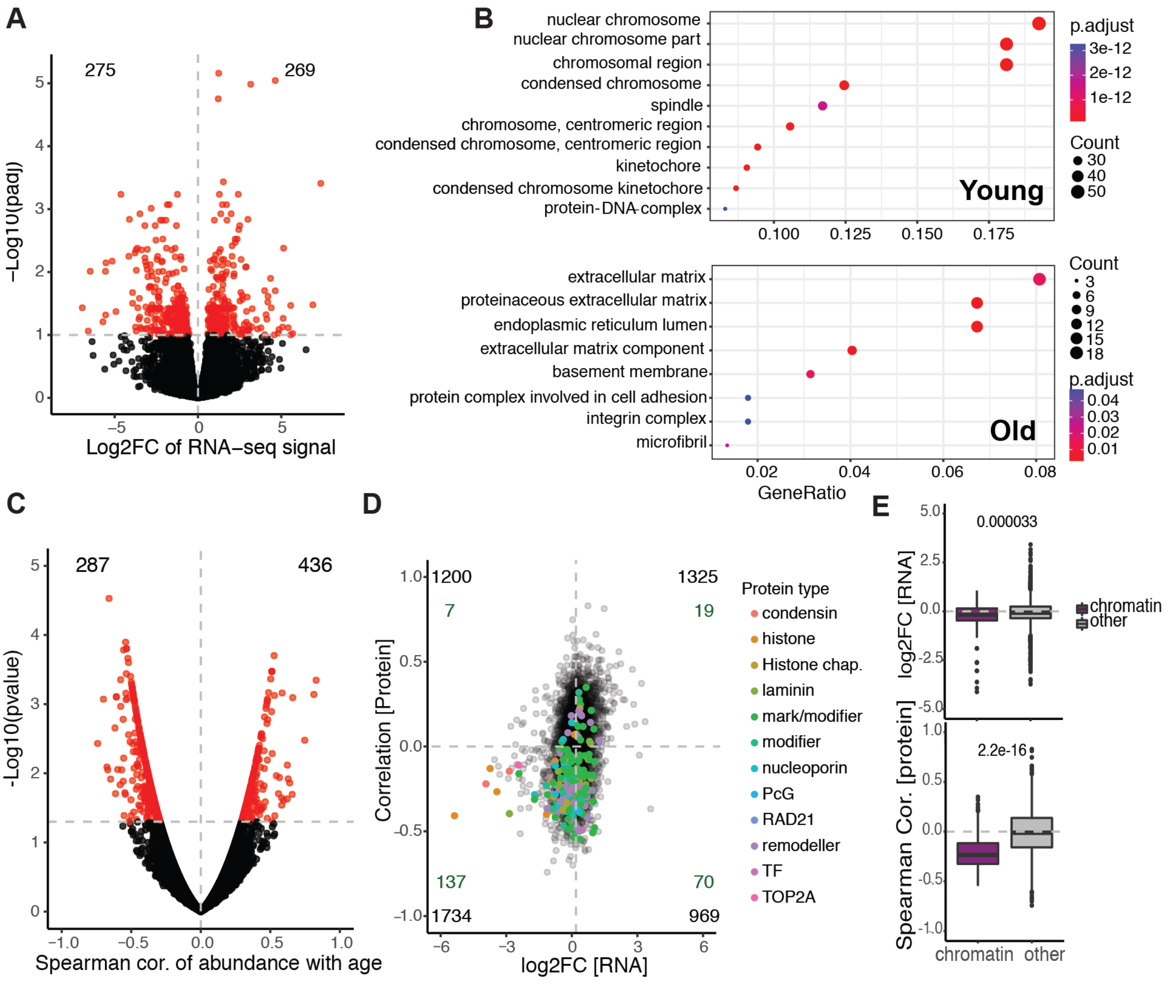


Figure 5.2. Integration of transcriptome and proteome shows downregulation of chromatin-related proteins

Volcano plot showing **(A)** genes that are down- (left, n=275) or up-regulated (right, n=269) according to their log2-FC or **(C)** proteins that are more abundant in younger (left, n=287) or in older (right, n=436) donors based on their spearman correlation with age. **(B)** Gene ontology enrichment analysis of cellular compartments shows terms as 'nucleus' or 'chromosome' for down- and 'ECM' or 'membrane' for up-regulated genes. **(D)** Scatterplot of log2-FC (gene) vs spearman correlation (protein) values coloured by whether they are related to chromatin (Medvedeva et al. 2015) or not along with their protein type. **(E)** Distribution of log2-FC and spearman correlation of chromatin-related (purple) and other proteins (grey) shows that only the first ones are downregulated with age. All panels were done by myself.

5.2.3 A global chromatin rearrangement is observed with age

We wanted to analyse whether a global rearrangement of chromatin architecture would be expected from the downregulation of proteins such as histones, histone remodelers, and by the PcG, among other chromatin regulators. Additionally, we asked whether a global change would be rather heterogeneous or whether it would be consistently originating at specific locations during ageing.

We generated chromatin accessibility data with ATAC-seq, a method capable of capturing open chromatin and nucleosome positioning using as little as 500-50,000 cells. The method is reported to allow TF-footprinting in a genome-wide manner at high resolution (Tsompana and Buck, 2014). Briefly, MSCs were isolated from healthy donors, cultured and ATAC-seq libraries were constructed and sequenced for each individual (Methods: 5.4.3). The dataset was processed with an in-house pipeline leaving us with 15 out of 16 individuals with two replicates each that passed the quality controls (TSS enrichment > 7.5 and a good correspondence of age with principal component 1, Figure S5.2A). By counting the number of reads per peak per individual, we defined a consensus set of autosomal peaks (n=151,615) using *DiffBind* (Ross-Innes et al., 2012). We kept only those peaks that were common in at least 2 different individuals and to ensure that the highest signal was extracted, we recentered the summits and extended +/-250bp for the consensus peaks (Methods: 5.4.3).

As a first approach we performed a PCA of the top 500 variable peaks in order to understand the impact of ageing on chromatin accessibility. Importantly, we observed that PC1 correlated well with the age of donor, showing older individuals to one side of the plot and opposite from the younger donors indicating that ageing is a major driver of variation (PC1=24.9%) for the ATAC-seq dataset (Figure 5.3A).

We wondered whether peak width contributed to the variability observed with age. For this we used the original peak-width of each of the samples and grouped them by age. Overall, we observed an increase in the width of the peaks that was proportional with age (p-value < 2.2e-16 for youngest vs oldest samples, t.test) (Figure 5.3B). Given that peak calling is done based on the read counts accumulating in certain regions and broader peaks would indicate an increase in reads, our results suggests that the accessibility of the chromatin increased for the older samples.

As an additional quantification we extracted the normalised signal in a window of +/-500bp from the re-centered summit, where a peak is +/-250bp, and we calculated the ratio of the sum of signal within the peaks vs the sum of signal in the complete window ($\Sigma\, normalised\ reads\ in\, peak\ \div\ \Sigma\, normalised\ reads\ in\ window$) per individual. Once again, we observed that there was a positive correlation that resulted in higher ratios with increasing age (R=0.5, Pearson correlation, Figure 5.3C). With these results we noticed that the additional reads were not only occurring inside the peaks, but also in their close surroundings suggesting once again a global increase of accessibility with age.

Given these changes in chromatin accessibility with age, we wondered whether any region in the chromosomes would be preferentially enriched for accessibility peaks which could serve as "hotspots" for ageing changes. We first visualised the distribution of the consensus peakset per chromosome and observed no obvious pattern besides a larger fraction of peaks having more accessibility in the younger donors (Figure 5.3D). Afterwards we defined a set of changing accessibility regions (CARs, see below and in Methods) and visualised the peak distribution in the linear chromosomes. Strikingly, we observed that there were loci that contained a cluster of downregulated peaks in very close proximity to each other with large log2-FCs at various chromosomes (Figures 5.3E, S5.3A). While we don't know exactly what these regions are, they don't seem to affect gene levels, but rather to delimit heterochromatin blocks or sometimes to localise next to the centromeres, indicating that measurable, sharp signal is present in younger individuals, which is not observed for older donors (Figure S5.3B).

Heterochromatin disorganisation with age has been previously reported for human aged MSCs (Ren et al., 2017; Scaffidi and Misteli, 2008). We speculated that one possible contributor is CTCF in its role of insulator, preventing heterochromatin variegation (Guerrero and Maggert, 2011) as its abundance was diminished with age in our dataset. We decided to visualise the peaks containing the CTCF motif in the same way as we did before. Notably, we observed that many of the peaks within one of the 'hotspots' in chromosome 1 contained multiple CTCF motifs (Figure S5.3B). However, other hotspots, like the one in chromosome 21, didn't show this enrichment which suggests that alternative explanations exist for the accumulation of significant CARs in a given condition, that are only sometimes affected by CTCF binding.

In summary, by performing ATAC-seq in MSCs from donors ranging in age (21-59 yr), we have observed that chromatin accessibility changes throughout time and represents a major source of variability. Our initial analyses show that there are global changes associated with age, such as increasing width and accessibility signal in older donors and specific accessibility hotspots where the changing peaks concentrate. We next decided to center in those regions that were significantly changing their accessibility with age and to study them in the context of functional genomic annotations, along with their consequences in other biological layers.

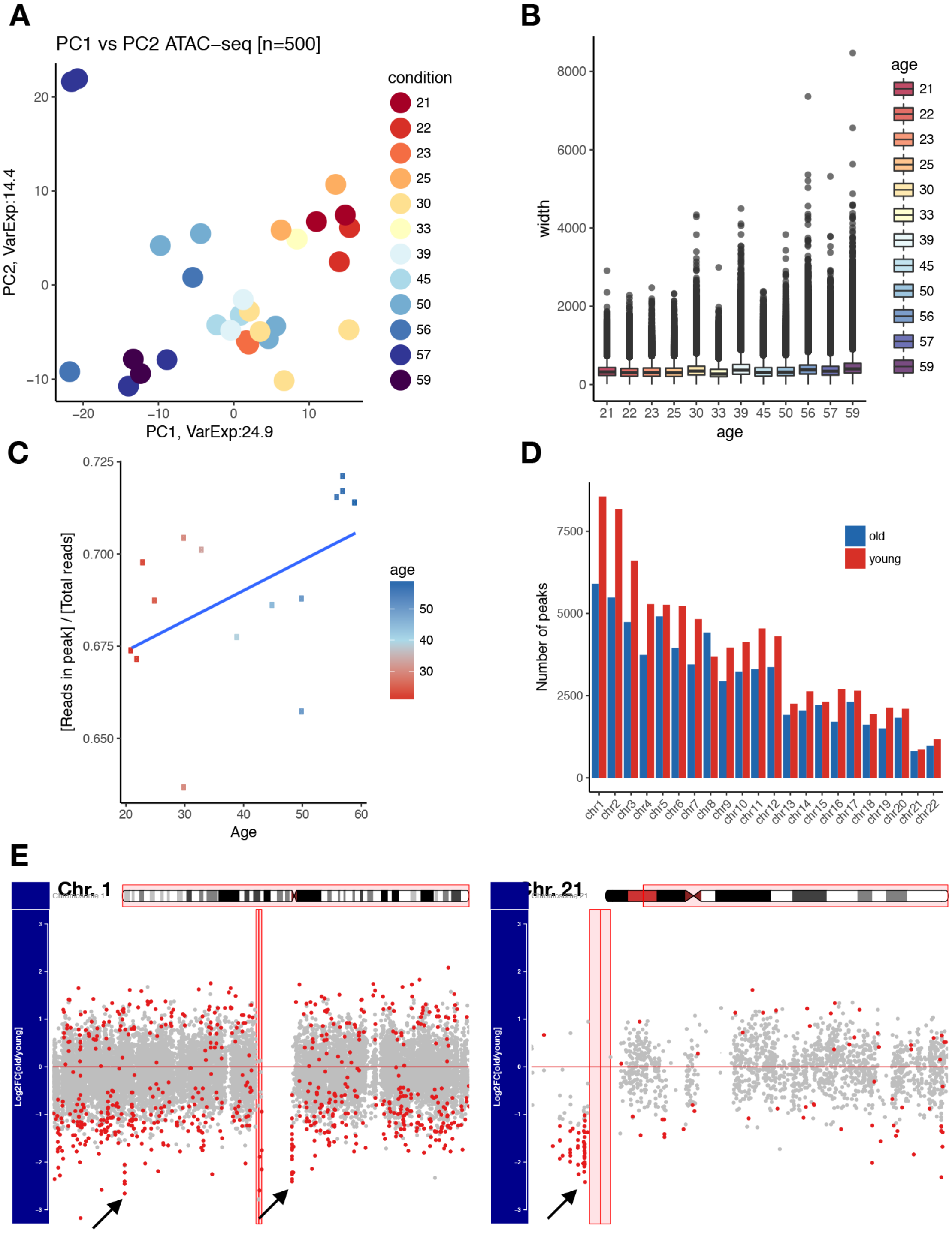


Figure 5.3. ATAC-seq profiling identifies global chromatin accessibility changes with age

(A) Principal component analysis of top 500 variable peaks in the ATAC-seq datasets shows that PC1 [24.9%] separates samples by age. **(B)** Distribution of original peak width by age shows increasing width with age. **(C)** Scores showing an increase in the overall accessibility with age calculated as (sum(normalised reads in peak) / sum(normalised reads in window)). Peak=+/-250bp and window=+/-500bp. **(D)** Distribution of consensus peaks per chromosome stratified by log2-FC as more accessible in younger (red) or in older (blue) samples. **(E)** Depiction of accessibility peaks along the linear chromosomes stratified by log2-FC and coloured by significantly changing with age (red) or not significant (grey). These panels were generated by myself.

5.2.4 Specific changes in accessibility are observed with age mostly at heterochromatin, enhancer, promoter, and bivalent chromatin states and correlate with DNA methylation

In order to identify relevant changing accessibility regions (CARs) with age, we used two approaches: (i) a *linear regression* of the signal with age and (ii) a *log2-fold change* of the 4 youngest and the 4 oldest individuals in our dataset. For both cases we used *DESeq2* (Love et al., 2014); each approach rendered 4,786 and 5,248 CARs respectively, which we later combined to have 6,746 peaks at FDR=10% (Methods: 5.4.3).

We obtained 4,244 peaks down- and 2,477 peaks upregulated with age (Figures 5.4B, S5.4A). Interestingly, we observed that most of the CARs were decreasing (69% and 66% in linear and log2FC approach, respectively) rather than increasing their accessibility with age (31% and 34%). When plotting the distribution of CARs per chromosomes it was also very visible that there were more CARs for younger than for older donors, with only chromosome 8 showing the opposite pattern (Figure S5.4B).

We next clustered the CARs normalised counts based on their accessibility level and observed that while the youngest (21-33 yr) and the oldest (56-59 yr) group of individuals showed an overall stable signal, a third group of intermediate ages (39-50 yr) showed rather a gradient or a “transitioning”, undefined accessibility between being open or being closed. This observation may imply that there is a time in life where the changes start accumulating before they switch completely, similar to the idea of the “point of no return” in ageing described by (Booth and Brunet, 2016).

To test whether there was a preferential distribution of the CARs with respect to genomic features, we used the chromHMM annotation defined in the Roadmap Epigenome consortium (Roadmap Epigenomics Consortium et al., 2015) for bone marrow-MSCs. To address whether there was an enrichment or depletion of any state for the CARs, we used the distribution of the consensus peaks per chromatin state as the background and tested the distribution of the CARs. We observed that the peaks gaining signal with age were enriched at enhancer (Enh, EnhG) regions (Figure 5.4A). Meanwhile, the regions that lose signal with age are enriched at active promoters (TssA), heterochromatin (Het, ZNF/Rpts), Polycomb-related (ReprPC, ReprPCWK), quiescent (Quies), and bivalent (TssBiv, BivFlnk, EnhBiv) states but depleted at enhancers. These findings suggest that changes in chromatin accessibility are not homogeneous but rather occurring at specific genomic regions that may eventually lead to a misregulated transcriptional landscape.

We speculate that DNA methylation could be associated to the observed changes in accessibility either by increasing (hyper-methylation) and therefore further compacting the chromatin or by decreasing (hypo-methylation) and decompacting specific regions. After all age has been reported to have a profound effect on DNA methylation and previous studies quantified and annotated differentially methylated regions (DMRs) in various tissues (Heyn et al., 2012; Rakyan et al., 2010) including MSCs (Fernández et al., 2015). We then tested whether the DMRs that were hyper- or hypomethylated with age in MSCs preferentially associated with any of the chromatin states from the ATAC-seq data. Consistent with previous reports (Wang et al., 2016), we observed that DNA hypermethylation occurs predominantly at bivalent chromatin regions, regions repressed by PcG, flanking promoters, and at enhancers (Figure 5.4C).

Importantly we identified some genes (n=30) with bivalent promoters and hypermethylated DMRs, of which a couple of them were already validated to be modified with age in MSCs (Fernández et al., 2015). For example *TBX18* is a known cardiac-specific developmental TF that is necessary for the formation of the head region that generates the electrical impulses that induce heart contraction (Li et al., 2018). In addition it has been described to have Mb-hypermethylation immediately upstream of its promoter, adjacent to repressive PcG-chromatin (Baribault et al., 2018). A second example is the G-protein coupled receptor 37 (*GPR37*), a negative regulator of oligodendrocyte differentiation and myelination (Yang et al., 2016) that has been associated to Parkinson's disease (Morató et al., 2017), to stem cell maintenance (Choi et al., 2015), and to have age-related hypermethylation and expression changes (Maegawa et al., 2010). Finally *PPM1E* is a protein phosphatase that inactivates CaM kinases and which has been shown to have altered DNA methylation levels in colorectal cancer (Yi et al., 2011), loss of histone acetylation levels at its promoter with age, and a lower expression in patients with schizophrenia and bipolar disorder (Tang et al., 2011).

In contrast, hypomethylation was not significantly changing in bivalent regions of the genome and it was depleted from heterochromatin and active promoters. All CARs gaining and losing signal, hypo-, and hyper-methylated peaks were significantly changing with age at enhancer and at repressed by PcG regions, which suggests that these loci are subjected to major changes in accessibility throughout life.

Finally, we reasoned that the impact of CARs would be reflected in the transcriptional program of MSCs. To test this idea, we checked for the distribution of the log2-FC of CARs vs the log2-FC of the expression of the closest gene within 100kb of distance (Figure 5.4D). Overall we observed no global correlation between the accessibility of the consensus peaks and the expression of their closest genes (R=0.006). However when selecting for specific features where the CARs were significantly associated with age (FDR=10%) we observed a better correlation that was most obvious for TSS flanking regions (R=0.3) and bivalent promoters (R=0.2) (Figure 5.4E). Two potential explanations for the lack of correlation could be (i) the presence of both activator and repressor factors that are present in the accessible regions and (ii) the CAR being at any chromatin state and not necessarily affecting transcription of its closest gene.

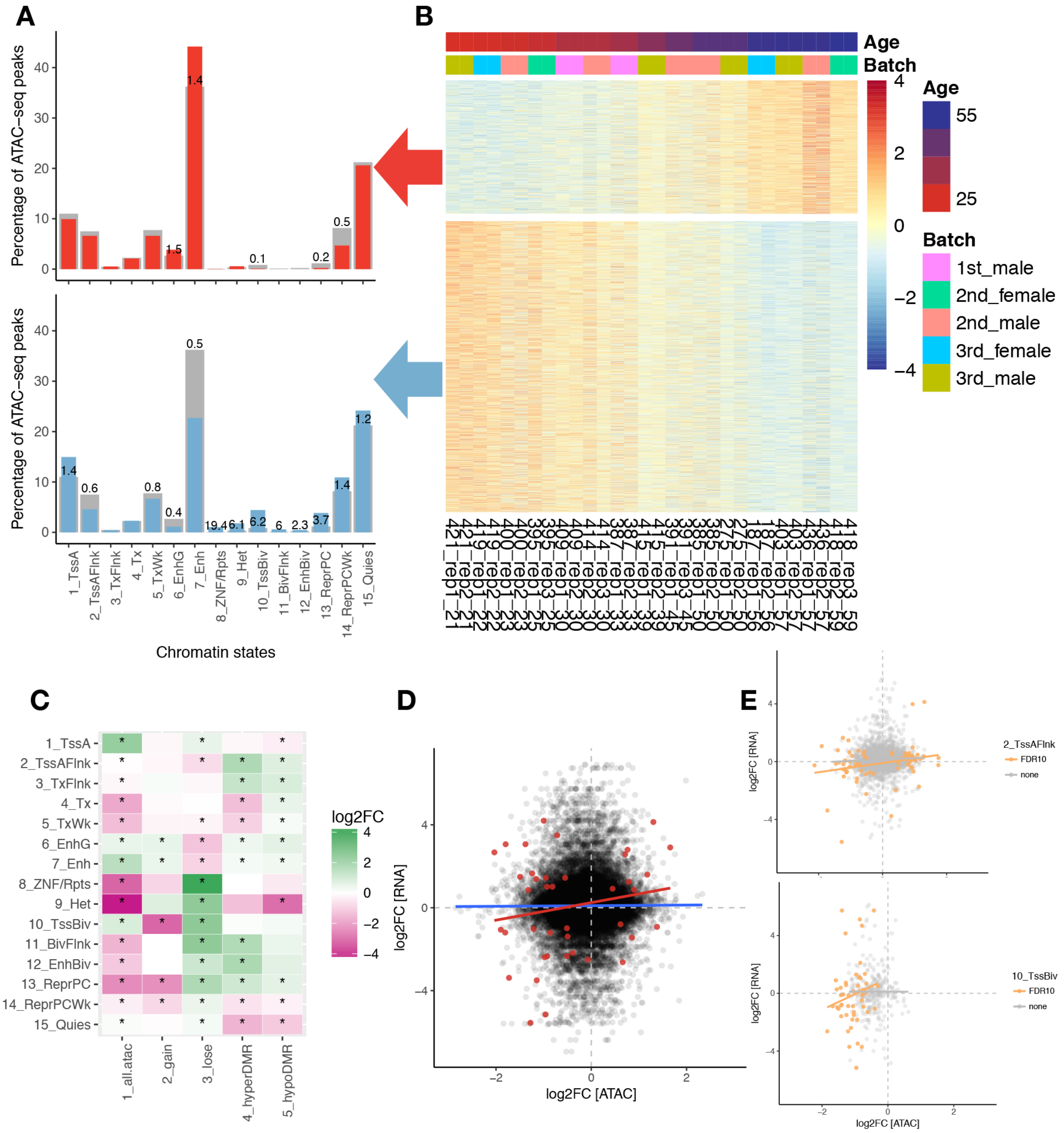


Figure 5.4. Changing accessibility regions (CARs) localise differentially across functional states and are reflected in transcriptional changes

(A) Distribution and enrichment [%] of CARs gaining (red) or losing signal with age (blue) per chromatin state with respect to that of consensus peaks contained at each region (grey). Only significant odds ratios are shown on top of the bars. **(B)** Heatmap showing the normalised ATAC-seq signal at CARs. **(C)** Heatmap coloured by log2-FC of consensus peaks, CARs, hyper- and hypomethylated regions per chromatin state. Significant values (*). Scatterplot of the log2-FC of ATAC-seq vs log2-FC of RNA for the closest gene to each peak (up to 100kb in distance) for **(D)** all ATAC-RNA pairs (black) and pairs with significant accessibility and expression changes (red). **(E)** Signal showed only for pairs where the ATAC-seq peak occurs at a specific chromatin state (Top = active TSS-flanking, bottom, bivalent TSS). This figure was produced by myself.

By identifying age-related changing accessibility regions and annotating them based on where they lie with respect to functional genomic features, we have been able to observe that chromatin accessibility is mostly gained at enhancer regions but lost at promoters, heterochromatin (both constitutive and facultative), and bivalent states throughout life. In particular the observation that bivalent states lose signal with age was in line with having enriched hypermethylation and downregulated PcG proteins. This could indicate that these regions, which usually are at promoters of genes involved in differentiation and are characterised by the presence of both H3K4me3 and H3K27me3, are repressed by DNA methylation in later stages of life, potentially decreasing the ‘stemness’ and thereby differentiation potential of MSCs.

5.2.5 Altered transcription factor landscape recapitulate changes in the differentiation fate with age

One of the advantages of ATAC-seq is that it allows to make inferences about TF-binding in a genome-wide manner. For investigating differences in the TF binding activity across ageing we used *diffTF* (Berest, Arnold, et al., *in revision*), which identifies changes in TF usage between two conditions (i.e. young vs old age group) by aggregating the differential ATAC-Seq signal across all putative binding sites of each TF.

We ran this analysis using the 4 youngest and the 4 oldest individuals in our ATAC-seq dataset to identify TFs that showed altered activity with age in MSCs. We observed that ZEB1 and DLXs TFs, a known adipogenic repressor (Gubelmann et al., 2014) and distal homeobox genes involved in osteogenesis (Frith and Genever, 2008) respectively, to be more active in younger donors. Meanwhile factors from the CEBP family, which orchestrate adipogenesis (Gubelmann et al., 2014), were more active in older individuals, reinforcing the described observation that the differentiation potential of

MSCs is altered with age from bone (for younger) to fat (for older individuals) (Moerman et al., 2004) (Figure 5.5A). Additionally, CENPB and CTCF were more active in the young.

As this analysis only accounted for TF binding sites (TFBS), we decided to verify the changes in activity for other factors for which ChIP-seq data exists, such as DNA binding proteins or chromatin remodelers. With this aim we used the ReMap database, which is a collection of all ChIP-seq experiments done in multiple cell types for various factors (Chèneby et al., 2018). We first filtered the 485 factors by gene expression in MSCs, leaving us with 407 factors to run in *diffTF,* many of which were common proteins to the previous analysis and some which were additional ones. The results were consistent and showed CEBPs and PPARG factors to be more active in the old condition while RUNXs (a known osteogenic TF) and HES1, a know repressor of adipogenesis in mammals (Lei et al., 2013) were more active in the young (Figure 5.5B). Notably, components of the PRC2 complex (EZH2, SUZ12, and EED, part of the PcG) were consistently more active in the young condition, in agreement with their higher abundance at the transcript and protein level (see above).

To gain more confidence on our results we checked whether the TF activities were supported by significant changes at the RNA and protein abundance. This analysis validated that CENPB, ZEB1, CTCF, HES1, EZH2 and DLX2 were downregulated and less active, while CEBPs and PPARG were upregulated with age (Figures 5.5C-F). Overall, by using a novel approach that identifies genome wide changes in TF binding, we have examined how the regulatory landscape of MSCs is altered with age.

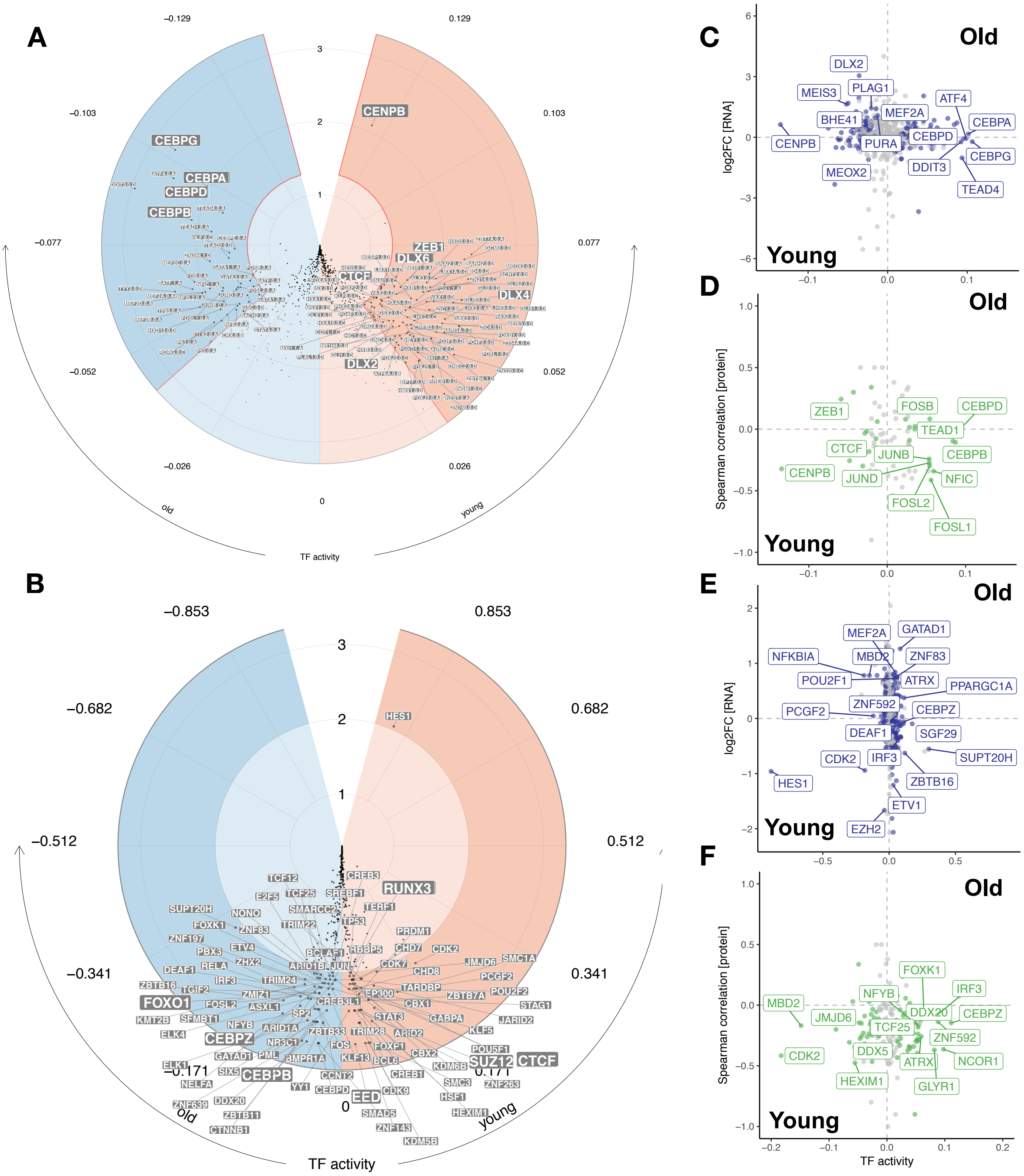


Figure 5.5. Changes in the TF-regulatory landscape could begin to explain differentiation potential bias and loss of stemness with age

Circular plots showing changing TF activity in younger or older donors **(A)** using TFBS or **(B)** ChIP-seq experiments (see methods 5.4.7). **(C-F)** Scatterplots of TF activity vs **(C, E)** log2-FC of RNA or **(D, F)** spearman correlation of protein and coloured by significant (blue for RNA and green for protein) or none-significant (grey) TF activity changes. All figures were done by myself.

Our results have been able to validate the known change in differentiation potential associated with age described for MSCs (Moerman et al., 2004) as well as to point to various other factors that are changing their activity with age, such as CENPB, CTCF, or PcG proteins. In addition we found some TFs that have not yet been implicated with any ageing phenotype and which might have a relevant function in ageing.

5.2.6 Multi-omic integration shows corresponding patterns across different layers of organisation

As a way to exploit our datasets, along with the novel TF-activity approach, we decided to test whether we could fill the "blindspot" of proteomics by integrating proteomics with chromatin accessibility as TF-activity. We therefore retrieved the available information on protein-protein interactions as collected by STRING (Szklarczyk et al., 2015) and gathered each of the quantifications per gene per technique, keeping only those interactions relevant to MSCs (where at least one of the quantifications was significantly changing with age). As expected, we found that if the activity of a TF was significantly changing with age both the transcript abundance (OR=3.82, p-value = 9.5e-109, Fisher's exact test) and the protein levels of the same gene/protein (OR=1.3, p-value = 6.93e-5) were also significantly affected.

Given a TF_n-$protein_n$ interaction, where TF_n was described by *diffTF* to significantly change its activity with age and $protein_n$ corresponds to a described direct interactor, we decided to test whether all proteins interacting with a TF would show a significant change at the protein level and in the same direction. For this we subset the MSC-specific network of interactions by whether the TF-activity was significantly (FDR=1%) changing in young or in old donors.

By doing so we observed a preference for protein levels to be more abundant in younger donors if its interacting TF was also more active in the younger group (OR=1.7, p-value = 6e-3, Fisher's exact test) and the same hold true for the proteins that were more abundant in the older group (OR=1.6, p-value = 6.7e-5).

We also decided to visualise the protein-protein interactions' network of the top 50 down- and up-regulated genes for which we ranked the genes by the largest or smallest log2-FC of the RNA. We observed a highly interacting network for the most downregulated genes that was consistent with what we have reported throughout this text and with previous reports (Figure 5.6A). Some interactions related to the centromere associated proteins (CENPA, NDC80, BUB1), some to the histones and lamin (HIST1Hs, LMNB1), and some others with cell cycle progression (AURKA, E2F8). On the contrary, genes that were upregulated were not so interacting with each other and participated in different processes like immune response and inflammation (IL6, IL32), oxidative phosphorylation (SOD2), or were related to the ECM and to collagen (COL8A2, MMP1).

Through the integration of various layers of biological quantifications, we have been able to take a glimpse at genes that are changing their abundance and activity with age. It is from here that we can start to propose which observed changes are relevant to continue studying or even to further validate, as to completely understand their involvement in the biology of ageing.

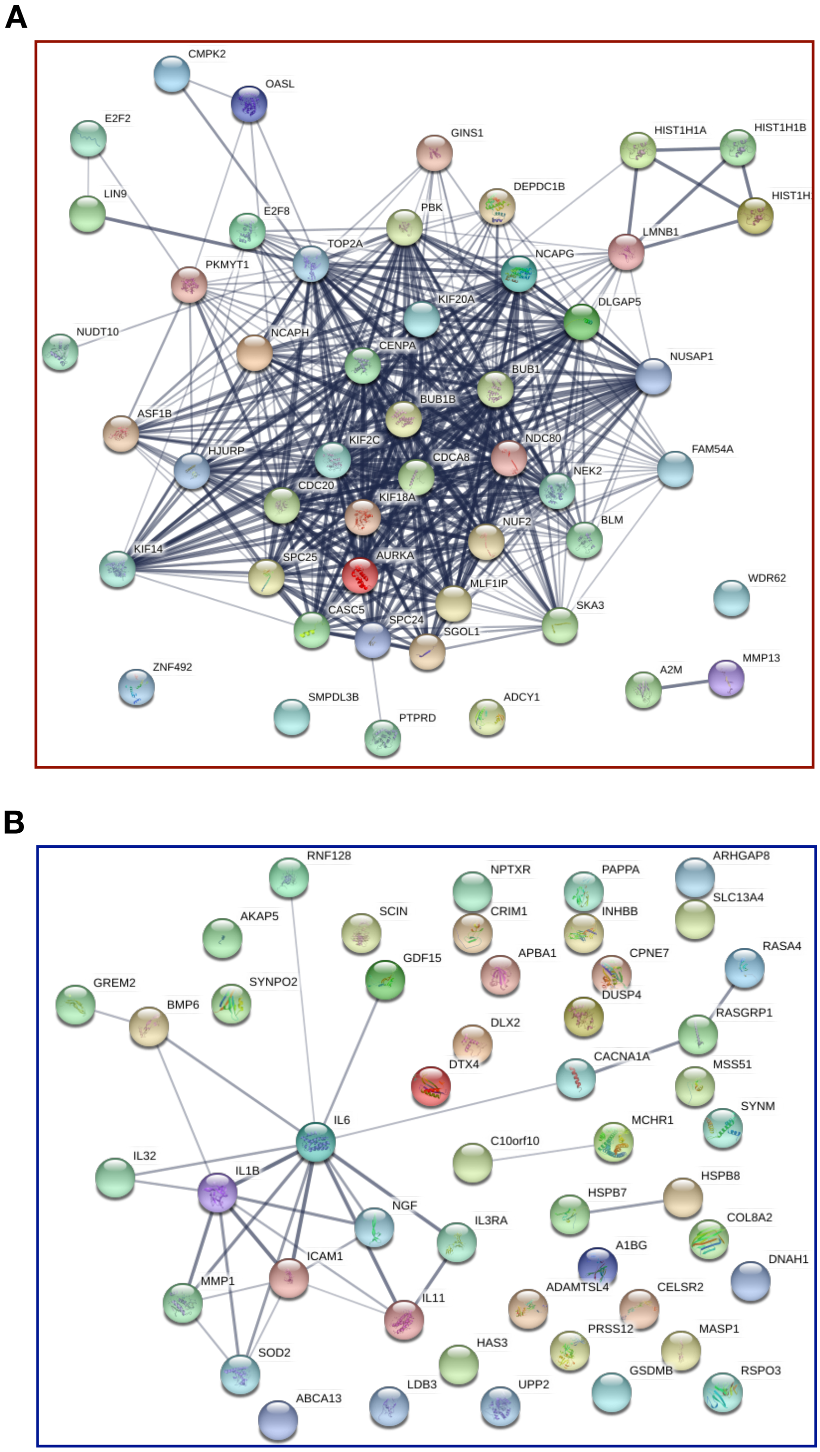


Figure 5.6. Filling the blindspot of proteomics unravels an increased number of interactions

STRING interaction networks for the top 50 genes that are **(A)** downregulated and **(B)** upregulated with age. This figure was done by myself.

5.2.7 Extensive changes in splicing regulation are observed with age

Finally, we decided to investigate whether there were splicing changes related to age following the existing reports showing changes in abundance of splicing factors, isoform changes in proteins as relevant as p53 (Deschênes and Chabot, 2017), and also based on our results showing a downregulation of *SF3B3* and *SRSF1*. For that, we used the same 8 individuals that served to identify DE genes and followed the *DEXSeq* workflow that we have described in Chapters 2 and 3 (Methods: 2.4.3 and 3.4.2).

To our surprise, we identified 6,761 differentially used exons (DUEs) in 3,461 genes (FDR= 10%) of which 3,190 were more abundant in younger and 3,570 in older donors. This number was even larger than the DUEs identified across individuals in chapter 2, showing that age affects the splicing mechanism too, potentially also increasing the inter-individual variation.

We next checked whether the genes containing DUEs were preferentially enriched for any biological processes by doing a gene ontology enrichment analysis. Surprisingly, we observed that genes containing exons that were mostly expressed in younger samples (n=1897) were enriched for biological processes such as mRNA processing and RNA splicing, suggesting that DUEs affecting splicing regulators could also potentially explain the big amount of changes observed with age in this process (Figure 5.7A). On the other hand, genes containing exons mostly expressed in older individuals (n=2512) were enriched for response to oxygen levels and metabolic processes. However, it is important to note that a large fraction of the genes contained DUEs from the two classes, which may explain the retrieved common terms (n=949).

Once again, we wondered whether any of the DUEs could be regulated by CTCF-mediated intragenic loops based on the consensus peaks containing CTCF-motifs. With this aim we annotated the loops in MSCs (see 2.4.1) and identified 488 promoter-upstream convergent interactions in 305 genes. As an internal validation, we predicted *THRAP3*, the gene that we used as an experimental validation for chapter 2.2.3, to form a loop.

Some of the intragenic loops were next to a DUE, showing once again that chromatin organisation is likely changing and impacting splicing throughout human lifetime. As an example we show gene *LRIG1*, a negative regulator of tyrosine kinase receptors involved in accelerated intracellular degradation and which was also reported to affect the plasticity of the ageing hippocampus (Trinchero et al., 2017). *LRIG1* has two DUEs, one of which is next to a convergent interaction (Figure 5.7B). Moreover, interactions within this gene are present for the Rad21 ChIA-PET data that we describe in 3.2.1. While these results seem promising, further analyses and potentially experimental validations will need to be done to further understand the extent of the CTCF-loop mechanism affecting this gene and the overall ageing splicing landscape.

In summary, we were able to observe splicing to be greatly affected with age and once again that intragenic loops also relate to alternative splicing in this process. Additional analyses need to be done to understand the extent of change of intragenic loops and whether DNA methylation could also be responsible for the splicing changes.

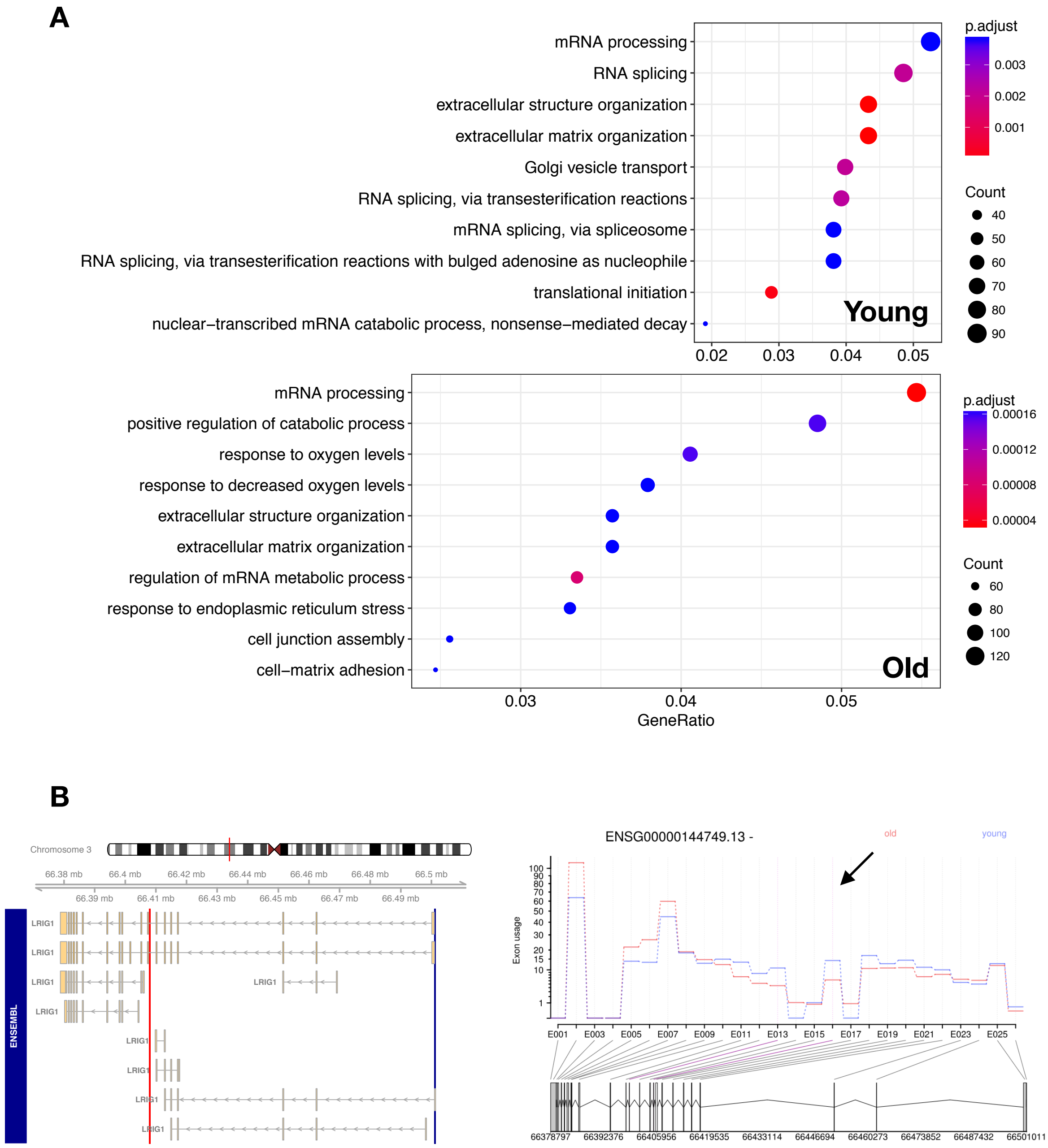


Figure 5.7. Intragenic loops affecting alternative splicing are also found with age

(A) Gene ontology enrichment analysis (for biological processes) of the genes containing exons that are included in young or in old donors. **(B)** *LRIG1* shows CTCF motif in sense orientation at promoter (blue) and antisense motif in the intron (red), both of which are validated by ATAC-seq peaks (left). Exon usage stratified by condition for *LRIG1* shows two DUEs, the one indicated by the arrow coincides in location with the anchor for a predicted intragenic CTCF-mediated loop. This figure was produced by myself.

5.3 Discussion

The BM is a multifunctional tissue of crucial importance which sustains the stem cell pool of an organism. It contains cells such as HSCs and MSCs, which will differentiate throughout the lifetime into mature cells such as haematopoietic elements, osteoblasts, or adipocytes (Reagan and Rosen, 2016). With the pass of time multiple types of damages, such as genomic instability and epigenetic drift, unfold into stem cell exhaustion which likely constitutes one of the ultimate culprits of tissue and organismal aging (López-Otín et al., 2013). Given that a tight link exists between age and the onset of several diseases, it is necessary to gain more understanding about the underlying changes occurring in BM stem cells throughout life with a systemic approach. Such studies will provide a more comprehensive view of how the ageing process affects the BM niche from the molecular to the cellular level.

Here we present a study that combines chromatin accessibility, transcriptomic, and proteomic data in MSCs from human donors of multiple ages. By integrating the previous data we were able to clearly observe that there is profound misregulation of chromatin organisation at the long, medium, and short scales with age, which accounts for the largest percentage of inter-individual variation, and which ultimately has consequences in a variety of processes at the transcriptomic and proteomic level.

Why is chromatin organisation changing with age?

The notion that chromatin structure is affected with age has been described by the most obvious phenotype of premature ageing observed in the Hutchinson-Gilford progeria syndrome (HGPS), which is characterised by mutations in the lamin A/C gene (Dreesen et al., 2013; Scaffidi and Misteli, 2006, 2008). HGPS affects several tissues, particularly those of mesenchymal origin and causes changes in the differentiation potential of MSCs, with enhanced osteogenesis and decreased adipogenesis. Additionally, it has been described that the same molecular mechanism responsible for HGPS acts at a low level in healthy cells (Scaffidi and Misteli, 2008).

One of the most striking results from our study was observing that chromatin related proteins were significantly downregulated with age. Among the changing ageing proteins we observed a collection of proteins including all the lamins, several histones, nucleoporines, chromatin remodelers, and PcG subunits, to name a few. We therefore speculate that the changing chromatin structure landscape is a sum of the decrease in abundance and activity from all of the previously mentioned factors, along with decreased CTCF, methyl-CpG binding proteins, DNA methylation, telomere shortening, and centromere associated factors.

The chromatin changes seem to occur for both senescence and ageing. For example, MeCP2 was consistently upregulated with age in HSCs and MSCs. MeCP2 is a methyl-CpG binding protein known to repress transcription from methylated promoters through its interaction with HDACs (such as SIN3A which was previously implicated with ageing) (Fuks et al., 2003). On the contrary, proteins like EED (subunit of the PcG complex), splicing factors such as SF3B3 and SRSF1, and TOP2 (DNA topoisomerase) were downregulated on HSCs and MSCs.

In addition, the consistent loss of CTCF at the transcript, protein, and TF-activity level in MSCs suggested changes in global chromatin organisation. Most strikingly, multiple cohesin subunits were downregulated with age in MSCs as well as condensin in both cell types. Given the known roles that the previous proteins play in loop extrusion to establish the nuclear architecture, as amply described in chapter 1, it is tempting to speculate that there will be differences in chromatin structure occurring with age at high-, medium-, and short-range interactions. The only evidence supporting this idea is the study done in oncogene-induced senescent cells where HiC maps revealed a loss of within TAD interactions and a corresponding gain of cross-boundary interactions (Chandra et al., 2015).

Where can we observe these age-related changes?

Reports exists showing heterochromatin disorganisation with age, specifically at centromeres, telomeres, and at the nucleolus in human aged MSCs (Ren et al., 2017). The centromere is a chromosomal locus that ensures delivery of one copy of each chromosome during cell mitosis through a protein complex known as the kinetochore (Cleveland et al., 2003). In line with these observations, we observed a significant decrease in the TF-activity and protein abundance of CENPB and a three-fold downregulation of the NDC80 complex, which is the key site for kinetochore-microtubule attachment in anaphase. A significant decrease in TELO2 abundance was also observed in our dataset.

By looking at the linear distribution of the CARs, we were able to observe some 'hotspots' where multiple peaks were located in close proximity to each other and had high log2-FC. These regions were preferentially localised next to heterochromatic blocks or close to the centromeres, and they seemed to be present in the younger, but not the older individuals. It is important to note that more questions became open after these initial analyses and more work needs to be done to completely address what these 'hotspots' are and the effect that they have on chromatin mis-regulation. Experimental validations may also be necessary to have a better understanding of the extent of such changes.

We also observed that the CARs localised in diverse chromatin states depending on whether they gained or lost signal with age. The upregulated peaks were enriched at enhancer regions, while the downregulated regions occurred at active promoters, heterochromatin (both facultative and constitutive), and bivalent states. We observed that some of the changes in these distribution could be attributed to DNA methylation. For example, the bivalent enhancer and promoter-flanking regions were enriched for downregulated CARs and for hypermethylated CpGs from (Fernández et al., 2015), consistent with their report. However, although they also identified hypomethylation primarily occurring at repetitive DNA sequences, we couldn't validate this finding perhaps due to the very small number of peaks falling in this state. Finding CARs at quiescent regions, which have been shown to be enriched for laminB (Roadmap Epigenomics Consortium et al., 2015), could also suggest an altered attachment of the chromatin to the nuclear lamina with age.

Finally, as we were able to identify intragenic loops and abundant alternative splicing in the ageing MSCs, it could be possible that CTCF binding changes, reflected as changes in chromatin accessibility, also have an effect at the gene level and contribute to the ageing phenotype.

What are the consequences of changing chromatin accessibility in the ageing cell?

We hypothesised that global chromatin changes would impact (i) the transcriptional and translational program of MSCs and (ii) the structural organisation of the genome, which we already described above. One of the advantages of having such a multi-omics approach is that we can have a broader quantification of the thousands of genes and proteins that are present at any given condition by filling the "blindspot" of proteomics.

A first outcome of a changing regulatory landscape with age is the bias in the differentiation potential for MSCs, where younger cells will preferentially lead to osteogenesis but older to adipogenesis. Although the change has been described (Moerman et al., 2004), our study shows the bias from the chromatin landscape to the protein level.

A second possible consequence is the potential loss or slowing of stemness that results from repressed bivalent regulatory regions. Bivalent promoters have been described to occur only at stem and cancer cells and are characterised by the simultaneous presence of activating and repressive histone marks (Liu et al., 2013). In our study we found that PRC2, the complex responsible for adding the repressive modifications, was downregulated with age. We also observed an enrichment of bivalent chromatin states only in younger stages of life. Lastly, hypermethylated DMRs were also enriched at these same regions, raising the possibility that later in life the repression of bivalent states becomes stronger through DNA methylation.

Finally, the changing accessibility and abundance of centromeric and telomeric regions could lead to a plethora of changes that include increased DNA damage, misregulation of cell cycle, aneuploidies, and ultimately diseases like cancer.

In summary, we have analysed, characterised, and integrated various layers of information that made us unravel extensive changes in chromatin structure with age. Some of these changes were accompanied by functional consequences. The variety of identified changing genes/proteins show that the ageing process is a sum of several mis-regulated pathways which contribute to the attrition of cells with time, i.e. through changes in chromatin accessibility, differences in splicing, or in proteins that ensure a tight regulation of replication.

5.4 Methods

5.4.1 Overview of the samples and datasets used

We selected a subset of the SyStemAge samples (Henrrich et al. *in revision*) with ages ranging from 20-60 y.o. for which proteomics (45 samples), and in some cases RNA-seq datasets, were available. RNA-seq data consisted of 19 MSCs paired-end libraries with donors ranging from 20-59 y.o. (one replicate per donor). ATAC-seq was done for 16 MSCs donors, most of which had proteomics and some RNA-seq data available (Figure 5.1B).

We used the EpiFactors database (http://epifactors.autosome.ru/) (Medvedeva et al., 2015) to identify chromatin interacting proteins in human. To identify curated ageing genes, we used GenAge in the HAGR database (genomics.senescence.info/) (Tacutu et al., 2018). For the DMRs we used the sites identified by (Fernández et al., 2015) and for the genomic annotation we used the chromatin states from BM-MSCs of the Roadmap Epigenome Project (Roadmap Epigenomics Consortium et al., 2015). Finally, we used the STRING database (Szklarczyk et al., 2015) to verify protein interactions.

5.4.2 Culture, isolation, and construction of ATAC-seq libraries for MSCs

Bone marrow aspirates from the iliac crest of healthy human subjects were washed with phosphate-buffered saline (PBS, Sigma-Aldrich #D8537). The eluates were laid on Ficoll-Paque (15 mL, Biochrom), centrifuged (800g, 30 minutes) to separate mononuclear cells (MNCs). The MNC fraction was seeded in T75 cell culture flasks with Verfaillie medium (at a density of ~106 cells/mL), and cultured in fibronectin coated flasks until adherent colonies formed (at least 4 days). The adherent colonies were MSCs at passage 0 and for expansion of the cells, freshly isolated or thawed MSCs were culture-expanded in VM for one passage and then seeded in parallel into VM at 37 °C with 5% CO2.

For library preparation, an aliquot of 50,000 cells in exponential growth phase was harvested 48h after the passage. The cells were washed once with 50 µl of cold PBS and subsequently lysed (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630). Immediately after lysis, nuclei were spun at 500g for 10 min. The pellet was resuspended in the transposase reaction mix (25µl Transposase Reaction Buffer (TD), 2.5µl Nextera Tn5 Transposase (TDE1), 22.5µl H2O, and 0.5 µl of 0.1% digitonin) and incubated at 37 °C for 30 min. After the tagmentation step, the first 5 cycles of PCR amplification were performed using Nextera Barcoded Primer set and NEBNext High Fidelity PCR mix as described in the original protocol (Buenrostro et al., 2013).

In order to avoid the saturation in the amplification of prepared libraries, the appropriate number of cycles to amplify the ATAC-seq library was determined based on a qPCR test using 5 µl of previously PCR-amplified DNA as template and SYBR Green as fluorescent dye. Once the library was additionally PCR-amplified, it was ready to be sequenced.

5.4.3 ATAC-seq data analysis and identification of CARs

Libraries (n=38) for 16 individuals were processed with an in-house pipeline. Briefly, adapters were trimmed with Trimmomatic (Bolger et al., 2014) and reads were aligned to hg38 using Bowtie2 (Langmead and Salzberg, 2012) with parameters --very-sensitive -X 2000 and filtered based on good quality (MAPQ>=10). Mitochondrial, duplicated, and INDEL-containing reads were removed using samtools (Li et al., 2009) and Picard tools (https://broadinstitute.github.io/picard/). Quality controls were carried out at multiple stages of the pipeline, ending up with a TSS enrichment score which verifies that the majority of ATAC-seq reads fall within +/-2kb from TSS vs 4kb outside of the TSS (upstream of the -2kb from TSS). Good TSS enrichments are considered to be above ~7-

fold, we decided to use 7.5-fold on both libraries. For peak calling MACS2 was used (--nolambda --nomodel --qval 0.1 --slocal 10000) and those peaks falling in blacklist regions removed.

In order to identify relevant changing accessibility regions (CARs) with age, we used the *DiffBind (Ross-Innes et al., 2012)* and *DESeq2* (Love et al., 2014) bioconductor packages. We first defined a consensus set of autosomal peaks (n=151,615) for those peaks that were present in at least 2 individuals and counted the number of reads per peak per individual. Based on the TSS enrichment we were able to pinpoint some libraries which could contain quality issues (both libraries from male-21 yr batch_2, male-22 yr batch_2, male-50 yr batch_2). We also identified a donor which behaved as an outlier in the PCA (male-22 yr batch_1, Figure S5.1B). After a thorough series of analyses where we gradually removed these libraries from the complete dataset, we observed the best results by keeping only 15 donors, each with 2 replicates.

To ensure that the highest signal was extracted, peak-summits were recentered and extended +/-250bp. We then followed the *DESeq2* workflow. After initial exploratory analysis, we decided to identify the differential accessibility regions through two approaches: 1) fitting the signal to a linear regression as a function of age and 2) defining three age categories and calculating log2 fold-change values.

For the linear regression approach we used a design = ~ Batch_Sex + Age for the 15 ATAC-seq libraries of our dataset. After normalising for library size and estimating the dispersion, we applied a Likelihood Ratio Test to obtain a set of peaks with increasing or decreasing signal as a function of age (4,221 peaks at FDR 10%). Based on an unsupervised hierarchical clustering, we observed that samples fell mostly in three age-groups: 21-33 y.o. (first), 39-50 y.o. (second), and 56-59 y.o. (third). We therefore selected four samples from each of these groups and ran *DESeq2* with the design = ~ Batch_Sex +

age-group, using the same consensus set of peaks. We identified 5,270 differential peaks (FDR 10%) between the first and third groups and labelled all peaks as having differential signal with the linear and/or the log2 fold-change approach. In total, we obtained 4,244 peaks losing and 2,477 peaks gaining signal with age (6,746 peaks FDR 10%) and labelled all peaks as being differential with the linear, the log2 fold-change, or both approaches.

5.4.4 Assessing changes in the shape of the peaks

We measured if there were width changes for those peaks overlapping the consensus peak set. For that we used the original peak width of each of the samples and grouped them by age. Additionally, we used *SNPhood* (Arnold et al., 2016) to extract the normalised read counts in 50 bins of 20bp, leaving us with +/-500bp surrounding the re-centered peak summit of the differential peaks.

To check the overall changes in chromatin accessibility we calculated a ratio of the sum the normalised reads within the peaks vs the total normalised reads in the +/-500bp window. Then we visualised the ratios against the age of the donor.

5.4.5 Analysis of chromatin states in the accessibility regions

To investigate the distribution of accessibility regions in the genome, we used the chromHMM annotation defined in the Roadmap Epigenome consortium (Roadmap Epigenomics Consortium et al., 2015) for BM-MSCs. To address whether there was an enrichment or depletion of such states for the peaks that were significantly changing their signal with age, we used the Fisher's exact test and verified the distribution of the consensus peak per chromatin state as the background and that of the differential peaks that gained or lost signal with age, correcting for multiple testing with Benjamini-Hochberg.

For testing the distribution of DMRs we applied a similar approach to the MSC DNA methylation data from (Fernández et al., 2015).

5.4.6 Differential expression and exon usage analyses

In order to identify genes that were differentially expressed with age in MSCs, we selected the four youngest (20, 20, 23, 32) and the four oldest (50, 55, 57, 60) libraries of the dataset, all of which were males. We then ran a *DESeq2* (Love et al., 2014) analysis using only expressed genes which were previously filtered with *kallisto* (Bray et al., 2016) using Gencode v27 as annotation (https://www.gencodegenes.org/). We also filtered all the transcripts for which there were less than 10 total reads. Given that the eight libraries used were produced in the same batch, our design only considered the condition (design = ~ age-group). We verified that selecting these eight libraries was robust with other *DESeq2* analyses that used different numbers of individuals per group, while maximizing the output. With this approach, we identified 544 differentially expressed genes at an FDR=10% (275 more abundant in young vs 269 more abundant in old).

Gene Ontology over-representation analysis was done with *ClusterProfiler* (Yu et al., 2012)*,* in particular with enrichGO. We tested for 'Cellular Compartment', 'Biological Process', and 'Molecular Function' (not shown).

As a way to identify changes in splicing with age in MSCs, we used *DEXSeq* (Anders et al., 2012) and we followed the standard workflow described in their vignette and in detail in chapters 2 and 3. Briefly, after filtering for the used isoforms in MSCs, we flattened the Gencode v27 file and extracted the exon counts from the RNA-seq libraries using the *HT-seq* scripts provided with the package. We next compared exon fold change between the youngest group and the oldest group of individuals that we used for the DE analysis above. This rendered 6,761 DUEs at an FDR= 10% of which 3,190 were more abundant in younger and 3,570 in older donors.

5.4.7 Identification of a changing transcription factor landscape

To identify TFs that are significantly changing with age, we used *diffTF* (Berest, Arnold, et al. *in revision*) and compared the distribution of TFBS from HOCOMOCO (Kulakovskiy et al., 2018) in ATAC-seq peaks for the first and the third group. We use an extension window of 100bp and default parameters. We took an FDR= 10% to call significance of TF activity.

As a way to gain insights not only of the TF activity, but also of other factors for which ChIP-seq exists, we filtered the 485 ReMap 2018 (Chèneby et al., 2018) (http://tagc.univ-mrs.fr/remap/index.php) based on median gene expression, leaving us with 407 factors. We then followed the *diffTF* pipeline with default arguments using the three groups defined above to measure old vs young.

5.4.8 Integration of the ATAC-seq, RNA-seq, and proteomics datasets

To gain further insights into the protein-protein interactions that have been described and validated, we used the STRING database v.10 (Szklarczyk et al., 2015). We filtered for interactions with a combined score of 400 and above and added the quantifications from the diverse layers to the database (TF-activity, accessibility at promoter, transcript, and protein abundance). We then filtered to keep only those interactions where at least one of the genes was significantly changing with age at any of the layers. We later stratified by young and old based on the significant TF-activity as elucidated by *diffTF* (Berest et al. *in revision*).

6. Conclusions and perspectives

The interplay between chromatin organisation and the correct transcriptional regulation of the cell has become evident in the last years. However, studies have majorly focused on long-range chromatin interactions and only rarely analysed the 3D-organisation at the gene level. Furthermore, mechanistic insights relating chromatin structure and function remain scarce, while paradoxically increasing evidence supports a strong link between both levels.

In this dissertation I have undertaken a computational investigation to understand how nuclear architecture conveys functionality, mainly at the splicing process, but also by asking how 3D-organisation allows the propagation of molecular intermediate phenotypes - in particular histone marks, CTCF and cohesin - from distal genetic variants. Finally, I have also looked at the effects of increasing age in chromatin accessibility and how these changes reflect in RNA and protein abundance. While each of the chapters discusses the findings and limitations of the separate studies, in the next lines I sum up the major conclusions of my thesis regarding chromatin organisation in splicing and ageing, both of which were my most advanced projects. I also speculate how these research projects may continue in the future.

6.1. Chromatin organisation at the gene level conveys functionality by affecting alternative exon usage

Alternative splicing contributes to the cellular complexity of higher eukaryotic organisms, with nearly 95% of mammalian genes undergoing alternative splicing of pre-mRNA (Kornblihtt et al., 2013; Shukla et al., 2011). Aberrant splicing has been implicated in a number of human diseases, thus showing the importance of its correct regulation (Blencowe, 2006; Kornblihtt et al., 2013).

During my doctoral studies I was able to unravel a new mechanism regulating alternative splicing that consists on the formation of a CTCF-mediated chromatin loop, from promoter to upstream of the exon to be spliced in, that occurs both in human and in mouse cells. By starting with DNA motifs for CTCF and genomic annotations, I was able to predict the presence of intragenic loops. In a second phase, I exploited several datasets for lymphoblastoid cell lines from diverse biological levels to test the contribution of loops in splicing and gained mechanistic insights about this process. Finally, by performing some chromatin conformation experiments in collaboration with colleagues from EMBL, we were able to validate the presence of these intragenic loops and the correlation in exon and CTCF levels across individuals in lymphoblastoid cell lines. Through various collaborations, we verified the extent of the identified mechanism and found similar results in different cell lines and in mouse. Moreover, by using a list of frequently mutated regions in cancerous tumours we were able to observe that the CTCF motifs anchoring intragenic interactions bear a higher mutation frequency, particularly at intronic sites overlapping cohesin and that significantly overlap with splicing-QTLs. Overall, this mechanism is affected by natural genetic variation and changes in the epigenomic landscape.

I acknowledge that, in spite of the clear evidence for a mechanism where CTCF-looping affects alternative exon usage, it is still one of the many mechanisms that regulate alternative splicing, as also reflected by the overall moderate effect sizes reported in our studies. Furthermore, we have not discarded that the gene expression levels affect the presence and strength of CTCF-intragenic loops or that these interactions could affect more than one exon at a time. In line with this claim, a study describing intragenic loops in mouse that were claimed to be independent of CTCF was able to correlate the presence of loops with increasing number of exons and with the expression levels of the gene itself (Bonev et al., 2017). I still speculate that many of the loops identified in the previous study could be mediated by CTCF as the quality of the ChIP-seq data was higher in other studies, as in the ones used in this dissertation. Furthermore, while we controlled for analysing "cassette-exon" splicing, some of my results also hinted to cases where the intragenic loop was formed next to transcription start or end site, which suggests that we will need further analyses to know to which extent this mechanism is affecting splicing.

Nevertheless, a mechanism that regulates exon inclusion through intragenic CTCF loops offers some intriguing possibilities: since most CTCF binding events are invariant across many cell types (Li et al., 2013). A proportion of CTCF binding however, is sensitive to DNA methylation (Marina et al., 2016; Maurano et al., 2015), which in turn is known to be highly cell-type specific. Thus, one could conceive a mechanism where DNA methylation regulated CTCF binding could explain a portion of alternative isoform usage observed across different tissues or during development, as we have observed for the differentiation of mouse ESCs to NPCs. As a first attempt to start understanding the interplay between alternative splicing, CTCF intragenic loops, and DNA methylation, the group of Dr. Vessela Kristensen and in particular Dr. Sunniva Stordal from Oslo University Hospital in Norway, started to tackle this question. We recently started collaborating with them to analyse the ATCG dataset. Finally, we also know that this mechanism is affected by genetic variation as shown in section 2.2.4 and by the epigenetic landscape across tissues as in section 3.2.1.

Overall, our study provides one of the first functional interpretations of intragenic chromatin architecture by linking CTCF-mediated DNA-loops to the regulation of alternative isoform usage and protein isoforms across human tissues, in mouse neural differentiation, and in cancer. In addition, the model we propose provides a useful framework for further studying alternative isoform usage by generating hypotheses about exon usage solely based on the analysis of CTCF motif, which are easily obtained from the DNA sequence.

6.2 Chromatin architecture is misregulated at all levels and through multiple mechanisms during human ageing in stem cells

Ageing is the progressive physiological deterioration that ultimately ends in the death of an organism and that arises through the accumulation of multiple biological misregulated processes (Booth and Brunet, 2016; López-Otín et al., 2013). While several studies have attempted to understand the molecular causes of ageing, especially at the transcriptome and proteome level, the motivation of my study came from the persistent lack of understanding about the influence of chromatin accessibility in ageing and ultimately of the consequences that such differences would lead to. Moreover, it is known that stem cells are subjected to a different ageing influence and that they comprise a very specific niche that when affected contributes to the overall organismal decline (Schultz and Sinclair, 2016).

Although my analyses are still ongoing, preliminary results recapitulate some of the hallmarks of ageing (as described by López-Otín et al., 2013) and have resulted in the identification of specific chromatin regions affected with age and also in a description of global chromatin accessibility differences that are mainly observed at centromeres and telomeres for multiple chromosomes. It will be important to establish better quantitative metrics to classify those "hotspots" of changing accessibility regions (CARs) and probably to also link them with chromatin states or genomic features. For example, given the changes in DNA methylation with age and the observed enrichment of CARs that lose signal with age in heterochromatin, it would be relevant to check if transposable elements are significantly affected. This question is currently being addressed by Dr. Vasavi Sundaram, a shared postdoc between our group, Dr. Paul Flicek (EMBL-EBI), and Dr. Duncan Odom (MRC in the University of Cambridge).

In addition one of our most relevant results so far is that multiple chromatin related proteins are downregulated with age which once more hints to the relevance of the epigenome in ageing. The epigenetic regulation has been evidenced to be affected with age, for example through DNA methylation changes. Importantly, the environment has a strong influence in ageing, which has been demonstrated by altered lifespans as a consequence of caloric restriction or exercise (Booth and Brunet, 2016). Now that we have identified chromatin-related ageing proteins that occur in both hematopoietic and mesenchymal stem cells, it will be crucial to have a closer look at them and to test if they can bring some novelty into the ageing field.

Finally, I am currently assessing the best way to integrate the various layers of molecular data that are available in an attempt to have a more systemic view of the ageing changes going from chromatin to transcript and to protein. Additionally, as increasing cell-to-cell variability has been proposed as a consequences of ageing in the transcriptome (Martinez-Jimenez et al., 2017), it will be important to assess if such variability already comes from chromatin or whether it exists at all at this layer. If it is indeed the case that certain accessibility regions in chromatin show increased variability with age, it will be relevant to understand the extent to which it can affect the regulation of specific chromatin states and possibly if it contributes to disease states. These questions are now being studied by Dr. Mang Ching Lai, a postdoc in our group. Ultimately my aim is to have a manuscript explaining our findings in the near future, therefore contributing to a deeper understanding of the molecular mechanisms misregulated with age.

6.2 Concluding remarks

With this dissertation I have intended to gain further understanding on the intricate relation between chromatin organisation, its mechanisms that convey function, and its influence in cellular processes such as in splicing. Furthermore, our results confirm that the misregulation of chromatin accessibility also leads to changes at the RNA and protein level as a result of ageing. The next years should see an increase in genome-wide high-throughput datasets and also of an improving biostatistical framework to integrate the multiple layers of regulation, such as the application of machine learning approaches and reductionality algorithms, among others. The possibility to also integrate phenotypic data as measured by microscopy or screening assays with CRISPR-Cas9 should also open an exciting path to study the interplay of chromatin structure and function. To conclude, the findings that I generated throughout four years of intensive research hopefully add up to the increasing research of inter-individual variability, the link between DNA and phenotype, and overall represent a grain of sand in the knowledge of human epigenomics.

Appendix

A. Supplementary information for chapter 2

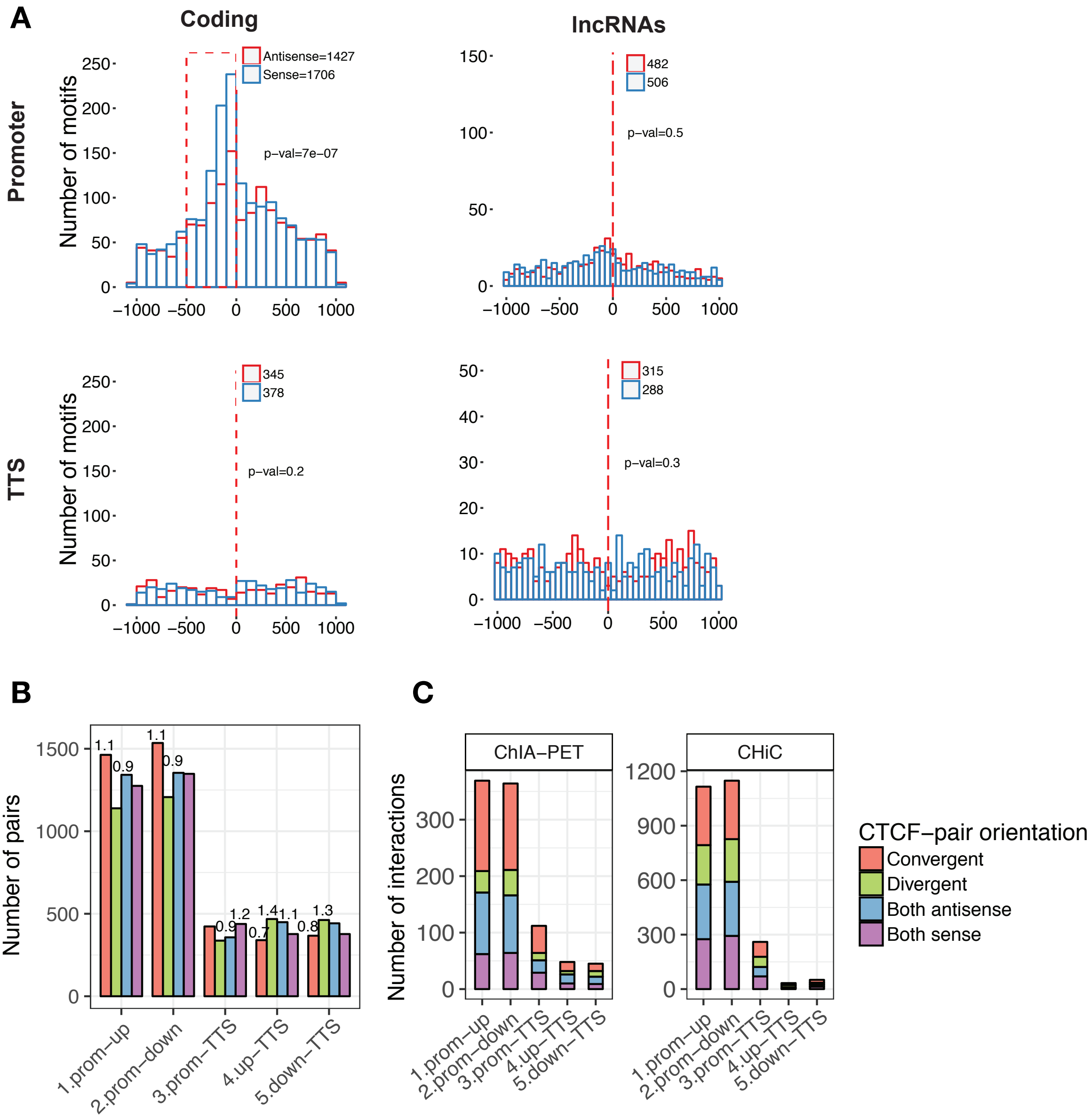
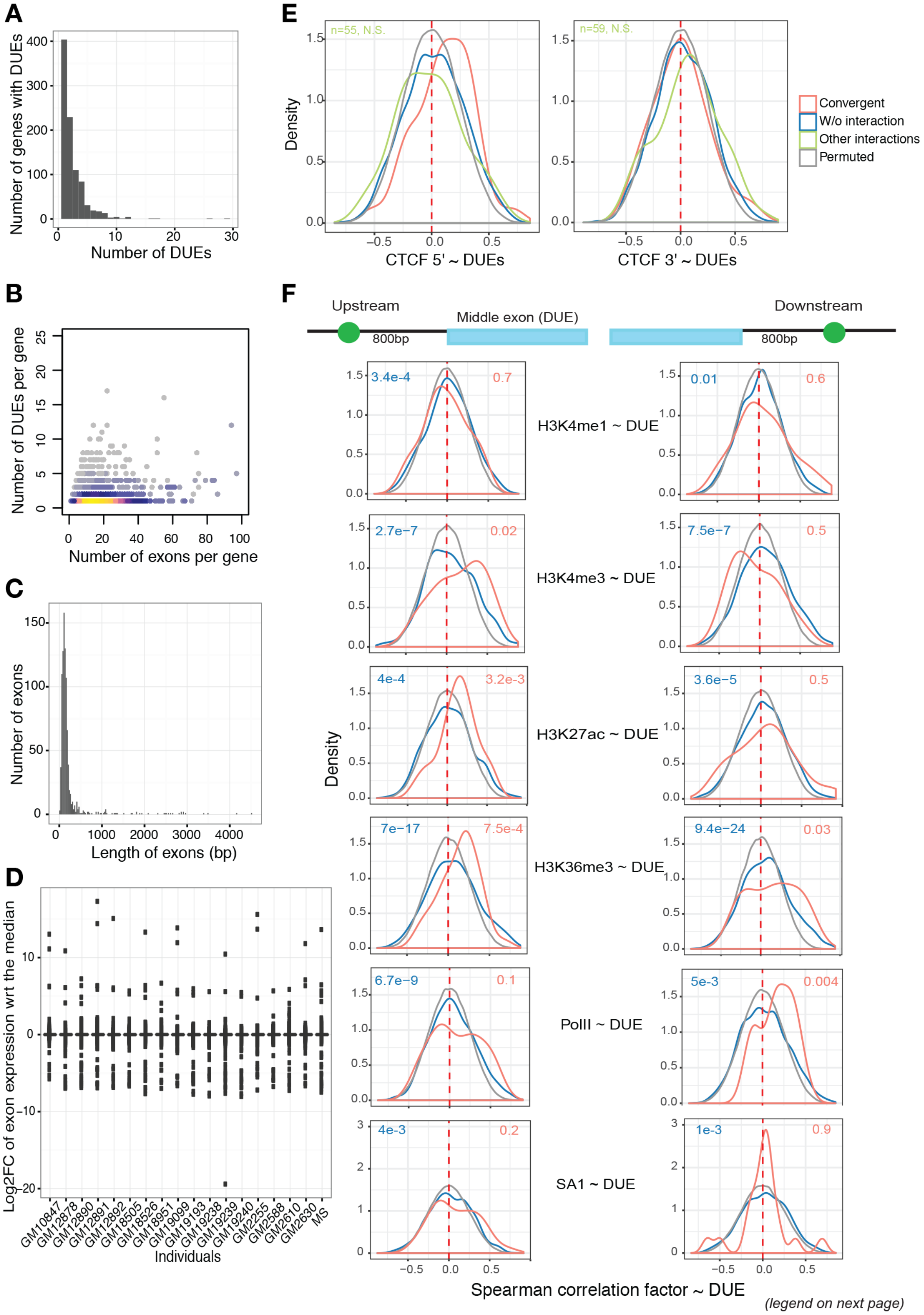


Figure S2.1, related to figure 2.1. CTCF-motif pair pattern is preserved with changing parameters, but absent in long noncoding RNAs

(A) Distribution of motifs, as shown in Figure 2.1A,B for the promoter (top) and TTS (bottom) for coding transcripts (left) and lnc-RNAs (right). Note that the sharp peak of sense motifs observed for protein-coding promoters is lost for lncRNAs (only transcripts with 2 or more exons were considered). **(B)** Distribution of motif-pairs for the LCL-specific set, as in Figure 2.1C. Note that changing the score of the motifs from 1000 to 500 and broadening the promoter window to +/-1kb allowed more CTCF motif-pairs to be annotated. **(C)** Same as in Figure 2.1D for the LCL-specific set. This figure was generated by myself and is published in (Ruiz-Velasco et al. 2017).



(legend on next page)

Figure S2.2, related to figure 2.2. Characterisation of differentially used exons and their correlation with diverse genomic factors

(A) Number of genes containing one or more differentially used exons. **(B)** Scatterplot of the correlation between the number of middle exons per gene and number of DUEs in that gene. **(C)** Distribution of the length of the middle exons. Note that most of them fall in a range of 20-200bp. **(D)** Log2-FCs for middle-exon expression with respect to the median for the 18 individuals. **(E)** Same as in Figure 2.2D, the spearman correlation between exon log2-FC and CTCF log2-FC is significantly higher for exons involved in promoter exon-upstream interactions than for those exons in other orientations (green) or than for promoter exon-downstream CTCF motif-pairs (numbers and p-values for the other categories are reported in Figure 2.2D). **(F)** Spearman correlation between DUEs and different histone marks, PolII, and cohesin (SA1) shows a higher correlation for looping upstream-exon in some (H3K4me3, H3K27ac, H3K36me3) but not all (H3K4me1, PolII, SA1) factors. PolII and H3K36me3 also have higher correlation for looping downstream-exons. Panels A-D were generated by myself in collaboration with Ana Belén Pinson-Solís and Pooja Bhat, panels E and F were done by myself and they are published in (Ruiz-Velasco et al. 2017).

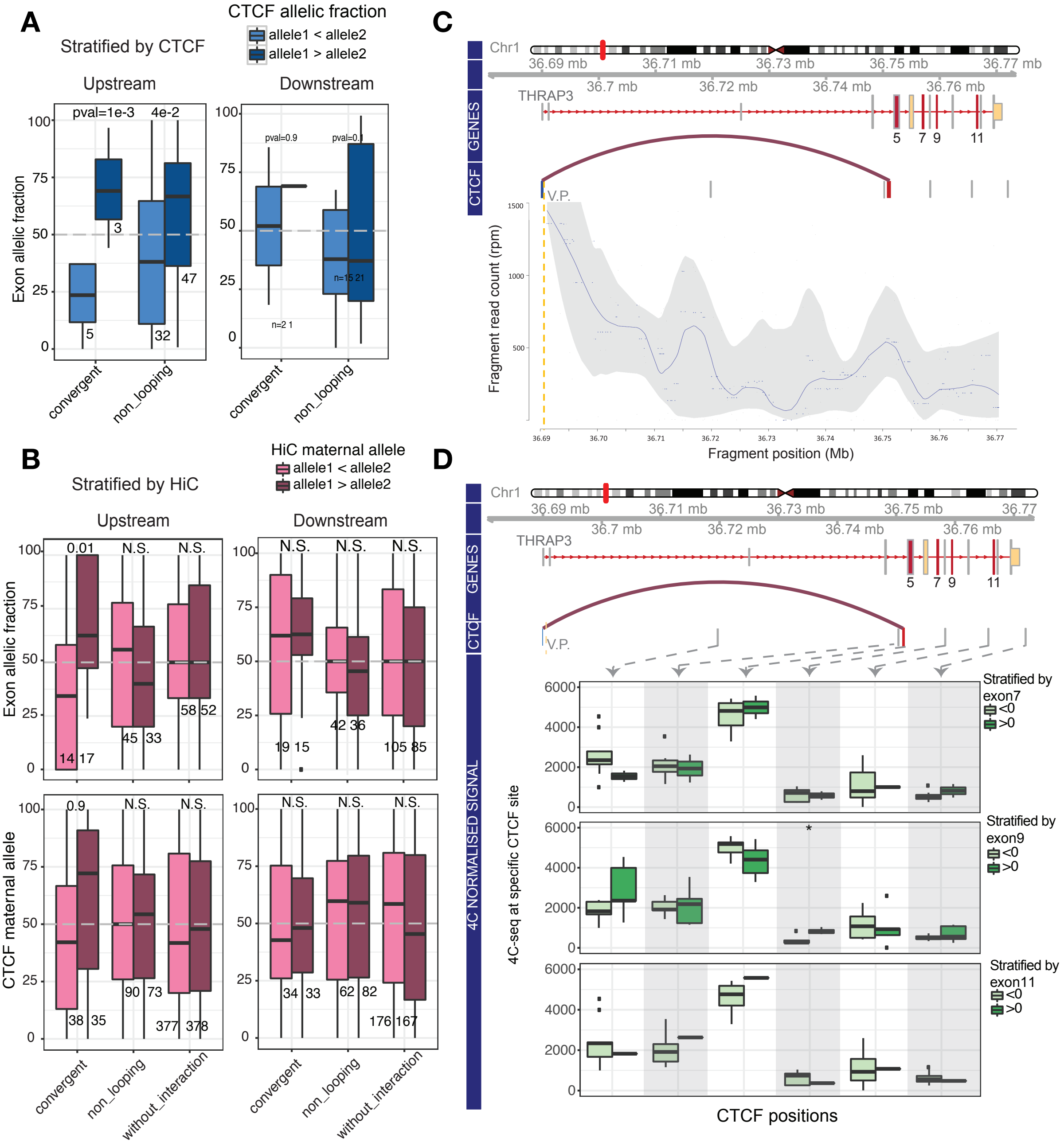


Figure S2.3, related to figure 2.3. Genetic basis of CTCF binding and exon inclusion

Exon allelic fraction stratified by **(A)** CTCF or **(B)** HiC for upstream (left) and downstream (right). Allele 1 > Allele 2 and vice versa are shown grouped by the configuration of the exon-promoter motif-pair. **(C)** 4C-seq normalised signal (RPM) at the *THRAP3* gene showing various interactions within the gene and in particular the predicted promoter exon-upstream interaction that we predicted. **(D)** Same as in Figure 2.3D, analysis of 4C-seq signal for exons 7,9, and 11. Note that the significant shift for 4C signal is only observed when stratified by exon 5. Panels A and B were generated by myself, panels C and D were done in collaboration with Dr. Mang Ching Lai and are published in (Ruiz-Velasco et al. 2017).

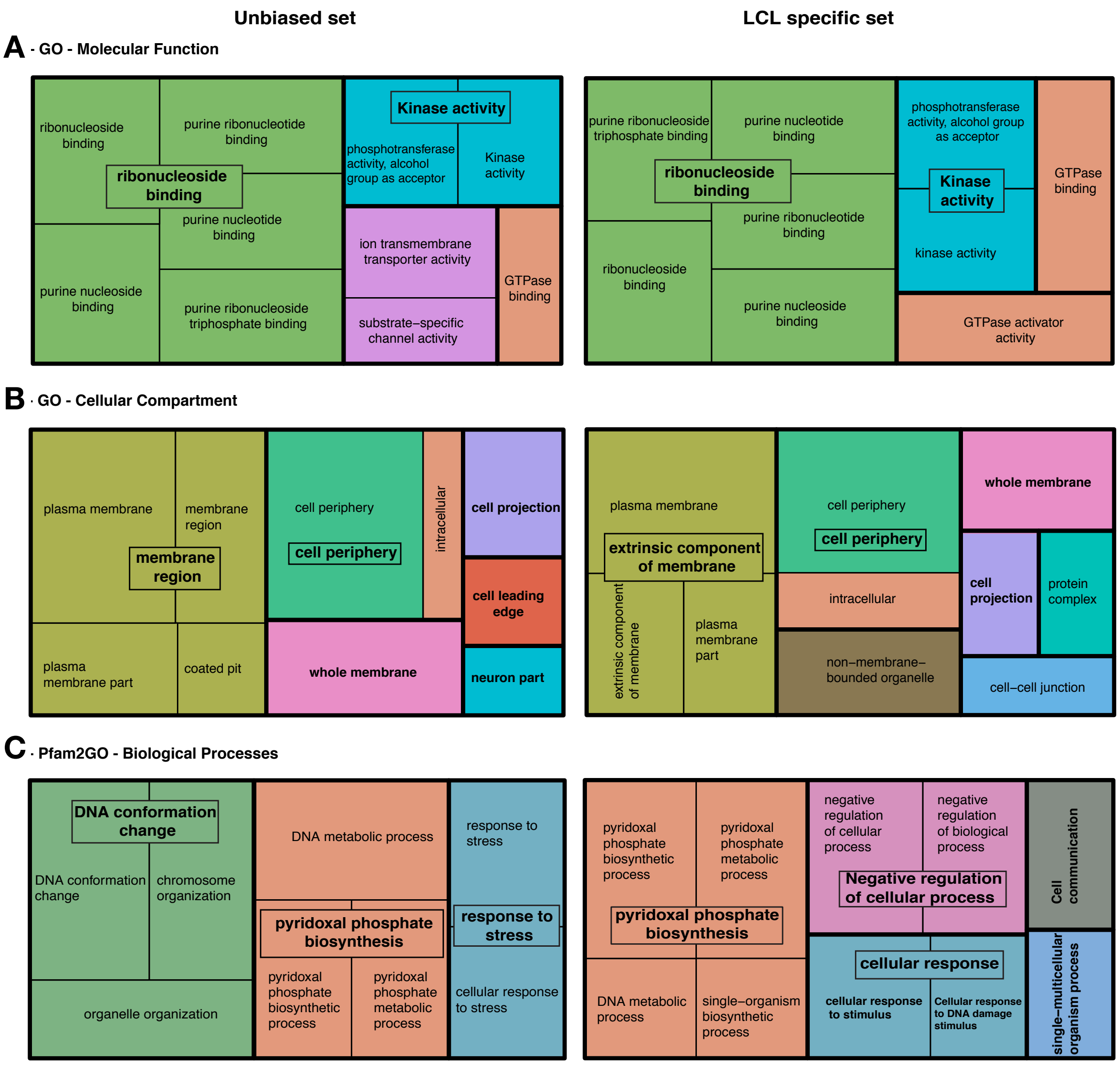


Figure S2.4, related to figure 2.4. Gene ontology (GO) based functional analysis

Treemap showing the top 10 enriched terms for **(A)** Molecular Function, **(B)** Cellular compartment, or **(C)** Biological processes (Pfam protein domains) for the genes with looping exons. Both the unbiased (left) and the LCL-specific (right) sets are shown. This figure was contributed by Dr. Manjeet Kumar and is published in (Ruiz-Velasco et al. 2017).

B. Supplementary information for chapter 3

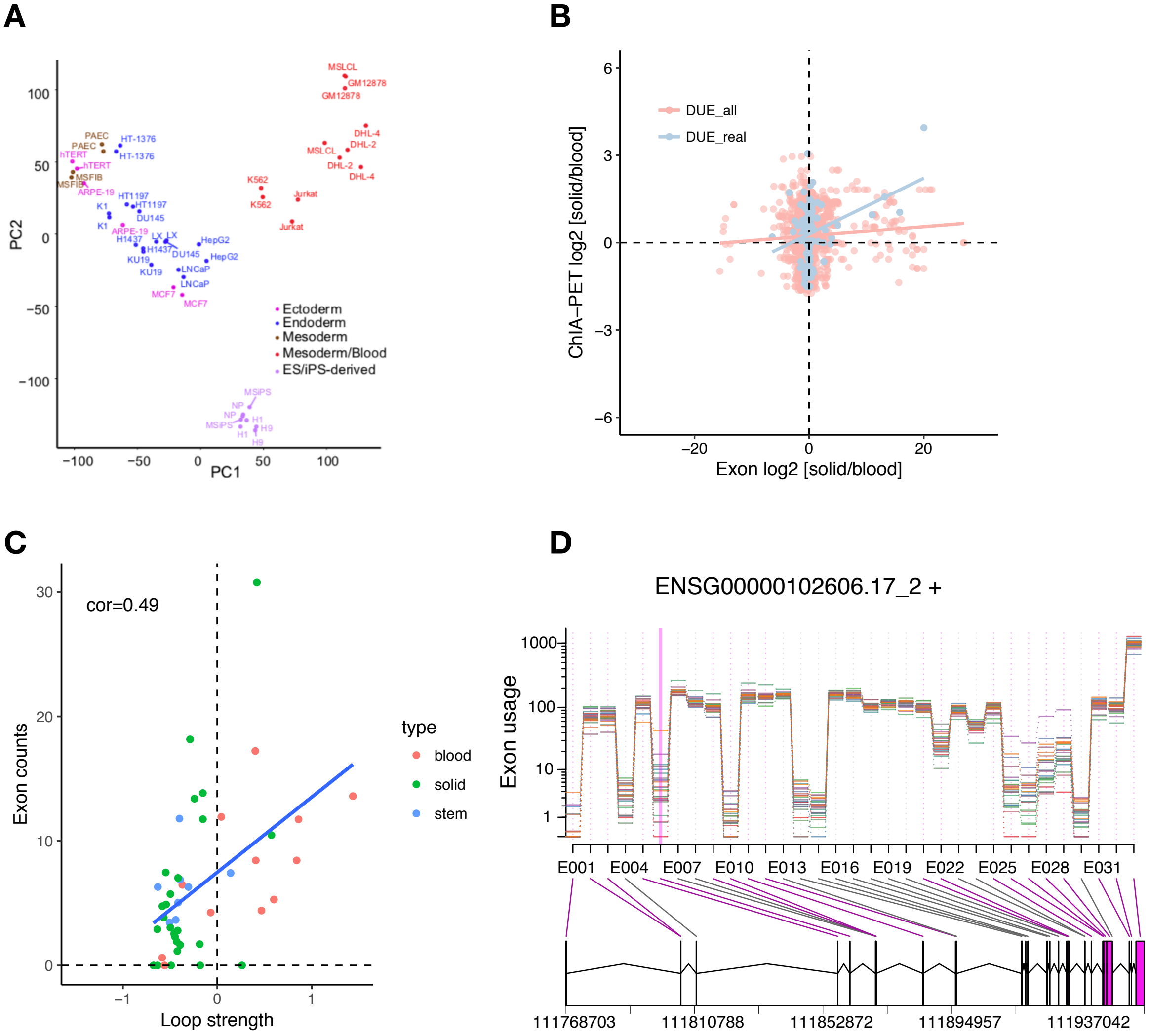


Figure S3.1, related to figure 3.1. Intragenic interactions affect alternative exon usage across tissues
(A) PCA of chromatin loop interaction frequencies across all samples. Colours denote the germ layer origin of each sample; x-axis indicates the percentage of variance explained by PC1, and y-axis by PC2. **(B)** Same as figure 3.1C, scatterplots of the differentially used exon and anchor counts for the real pairs (blue) and other exon-loop pairs within the same gene (pink) in solid tissues. **(C)** Normalised counts of exon 6 in *ARHGEF7* according to the normalised loop strength show a positive correlation (Pearson cor. = 0.5) for all 22 cell lines. The highest exon abundance is for blood tissues and a couple of outliers. **(D)** *DEXSeq* plot showing the differential exon usage of all exons for gene *ARHGEF7*, highlighting exon 6, which is affected by an intragenic loop. Panel A was taken with permission from Grubert et al. (*in revision*), panels B-D were produced by myself.

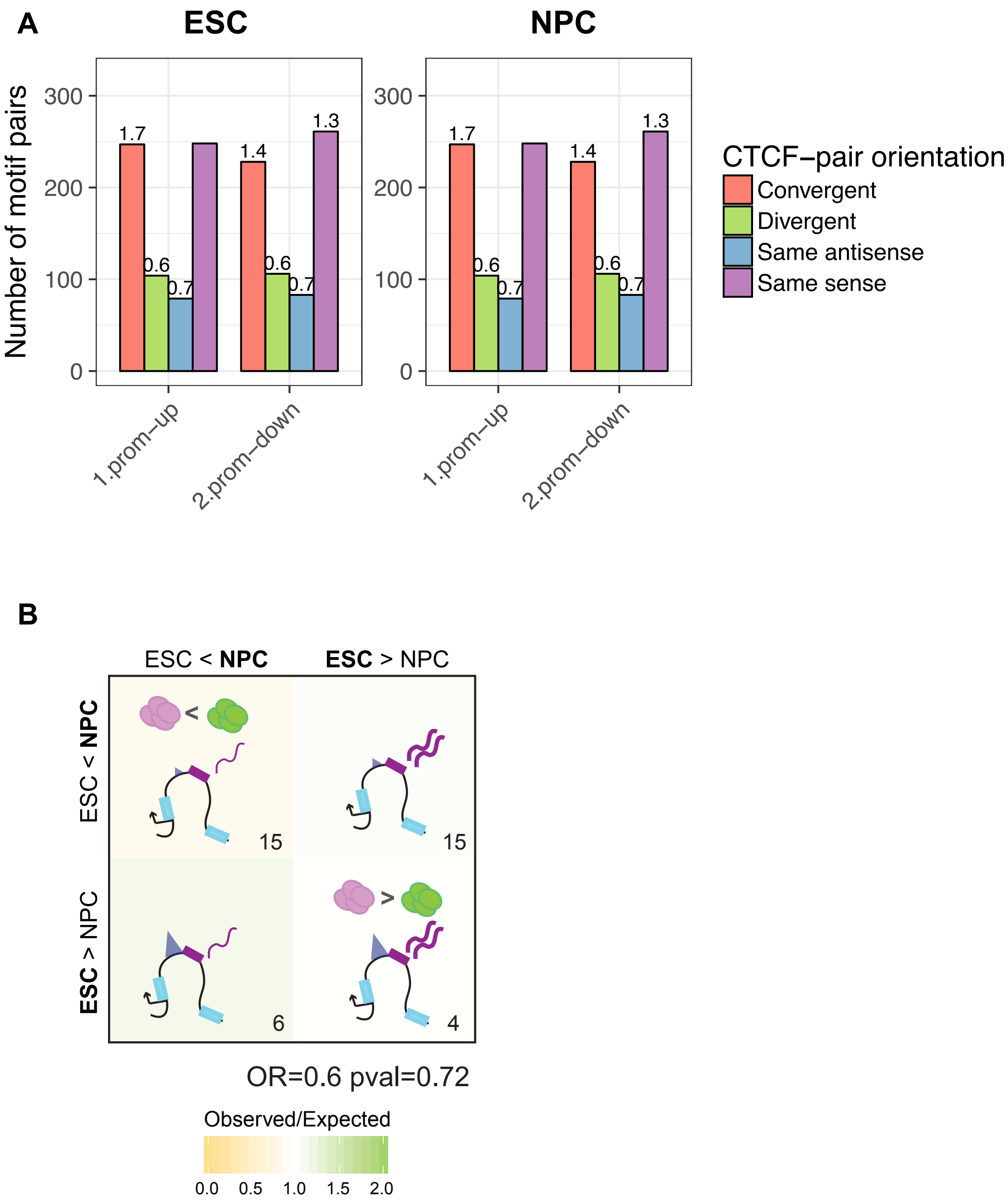


Figure S3.2, related to figure 3.2. Intragenic loops are also prevalent in mouse

(A) Intragenic interactions predicted for mouse ESC (left) and NPCs (right). **(B)** Similar to figure 3.2B, the schematic shows the test for exons to be included more in ESC or in NPC and to have a stronger CTCF binding in ESC or in NPC when the tested exons do not have a promoter-exon-upstream loop. The complete figure was produced by myself.

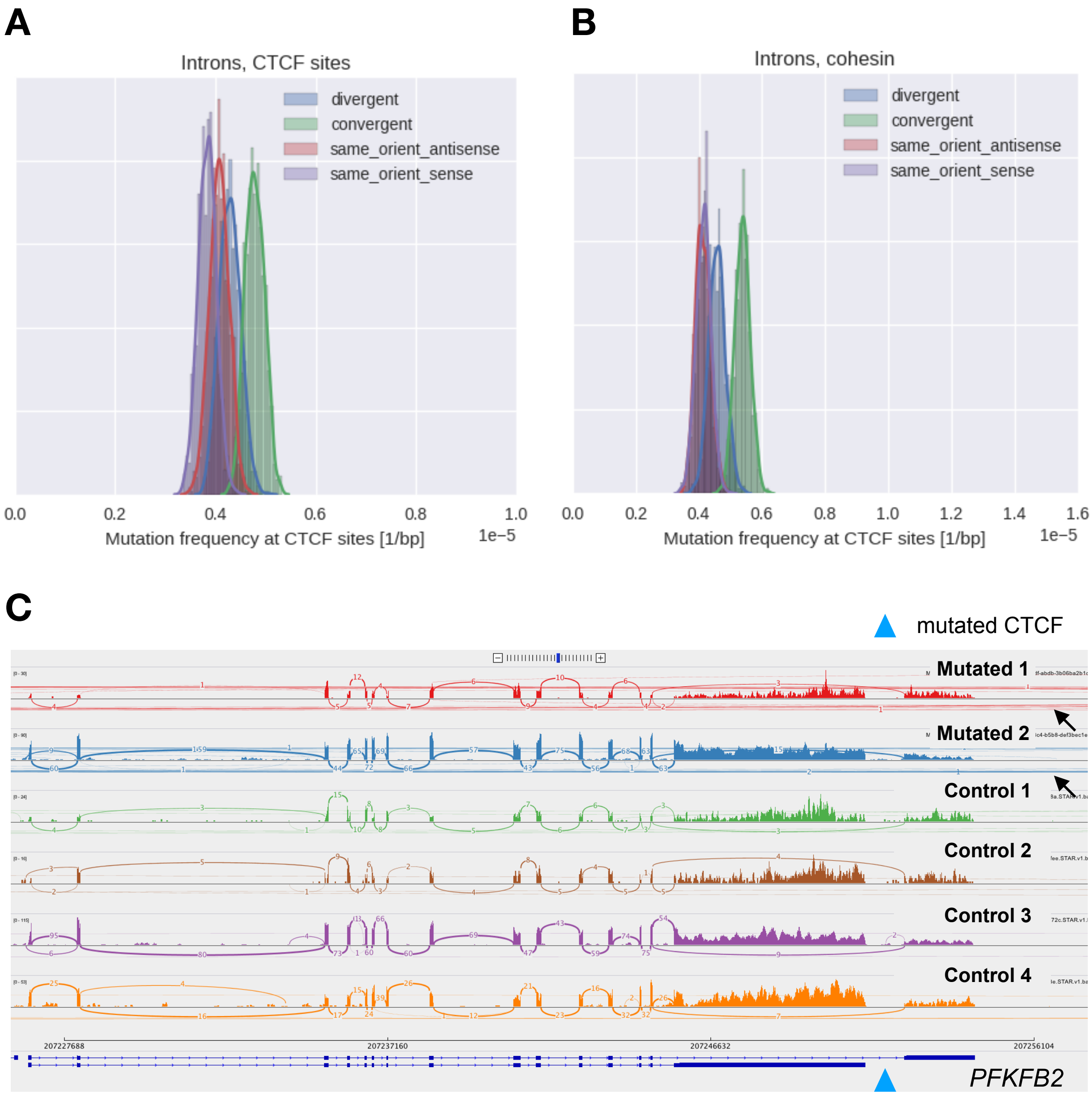


Figure S3.3, related to figure 3.3. Intronic CTCF sites involved in convergent interactions are frequently mutated when overlapping cohesin

(A, B) Similar to figure 3.3B, intronic CTCF motifs **(A)** and in particular those overlapping cohesin **(B)** show an increased mutation frequency in convergent interactions. **(C)** Sashimi plot produced using IGV for *PFKFB2* showing that mutated tumours have an increased number of reads that span outside of the gene (black arrows), which is rarely observed for the non-mutated samples (blue triangle = mutated CTCF). Panels A and B were produced by Dr. Esa Pitkänen and panel C by myself.

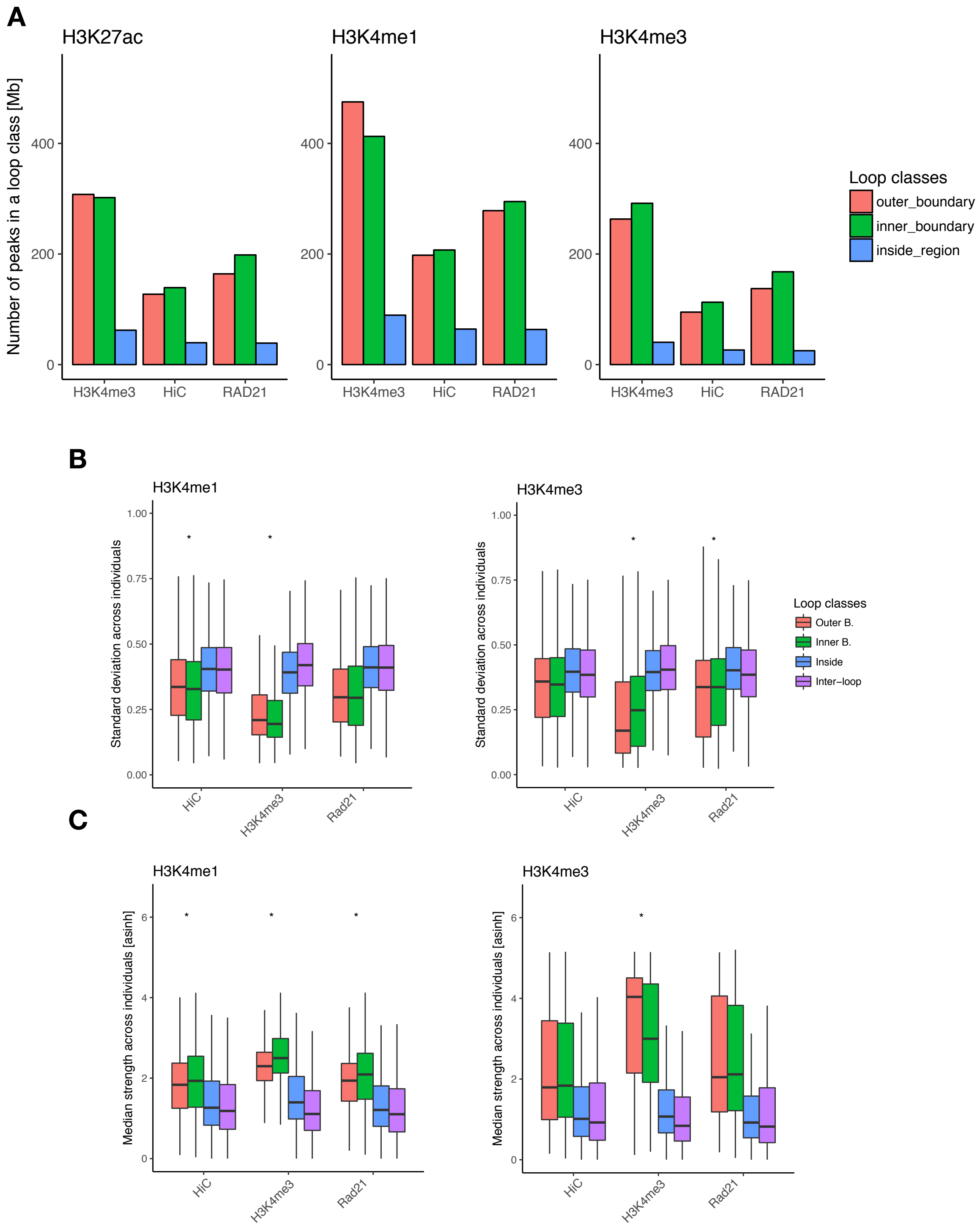


Figure S4.1, related to figure 4.1. Characterisation of loop regions across individuals for histone marks

(A) Same as in figure 4.1B for H3K27ac (left), H3K4me1 (middle), and H3K4me3 (right) at loop regions of the 3 data sets. **(B, C)** Same as in Fig. 4.1C for H3K4me1 (left) and H3K4me3 (right) across 18 LCLs according to where they bind in the context of loop regions. All three maps were visualised and Wilcoxon tests measured differences between outer and inner boundary values (*). The figure was produced by myself.

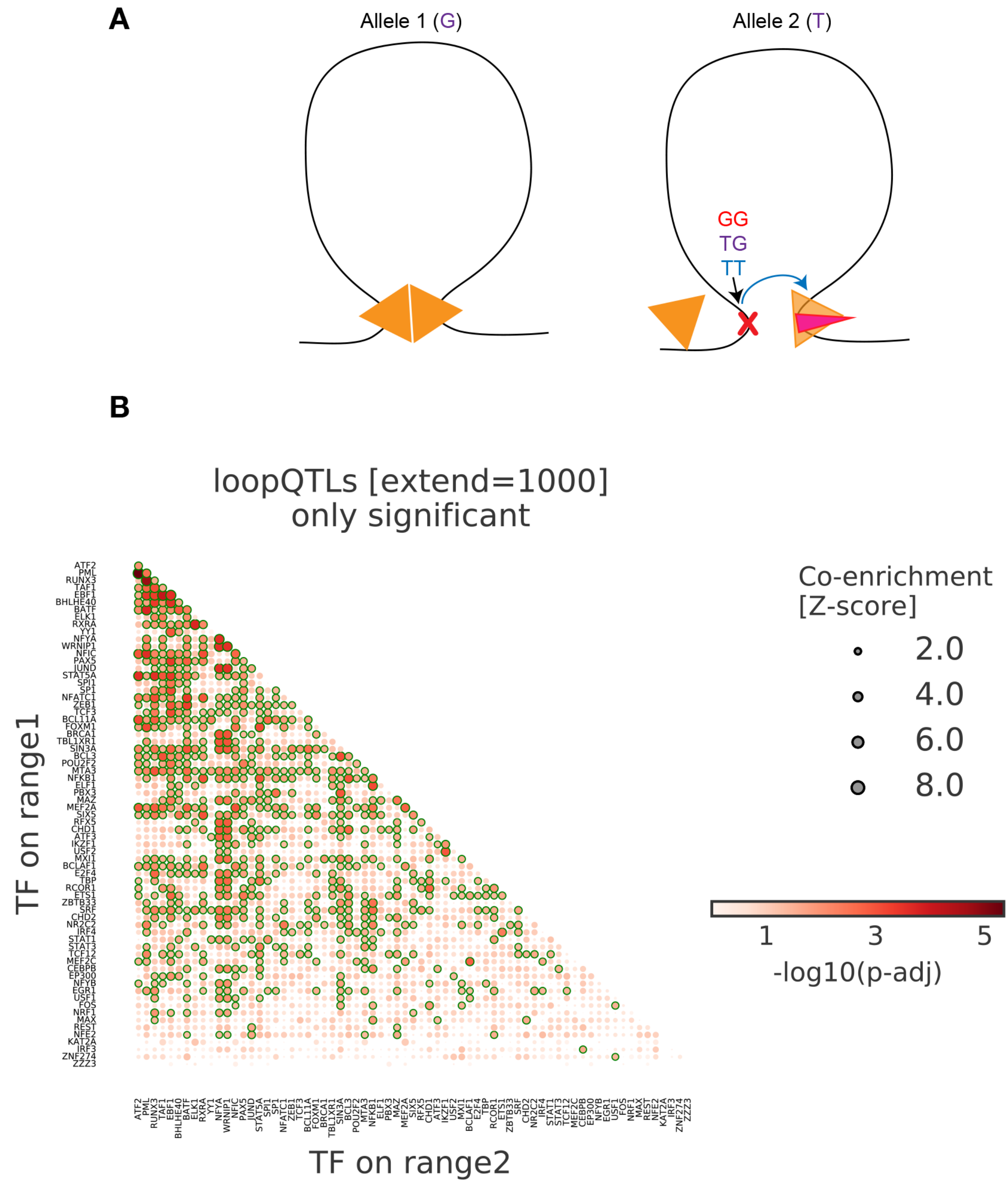


Figure S4.2, related to figure 4.2. TF-co-enrichments in loopQTLs

(A) Schematic of how a SNP could affect the binding of the protein on one side and affect the binding of the other side of the contact. The concept of a heterozygous SNP is also depicted. **(B)** Same as in Figure 4.2C but for all significant co-enriched pairs. Panel A was done by myself, panel B was contributed by Ignacio Ibarra.

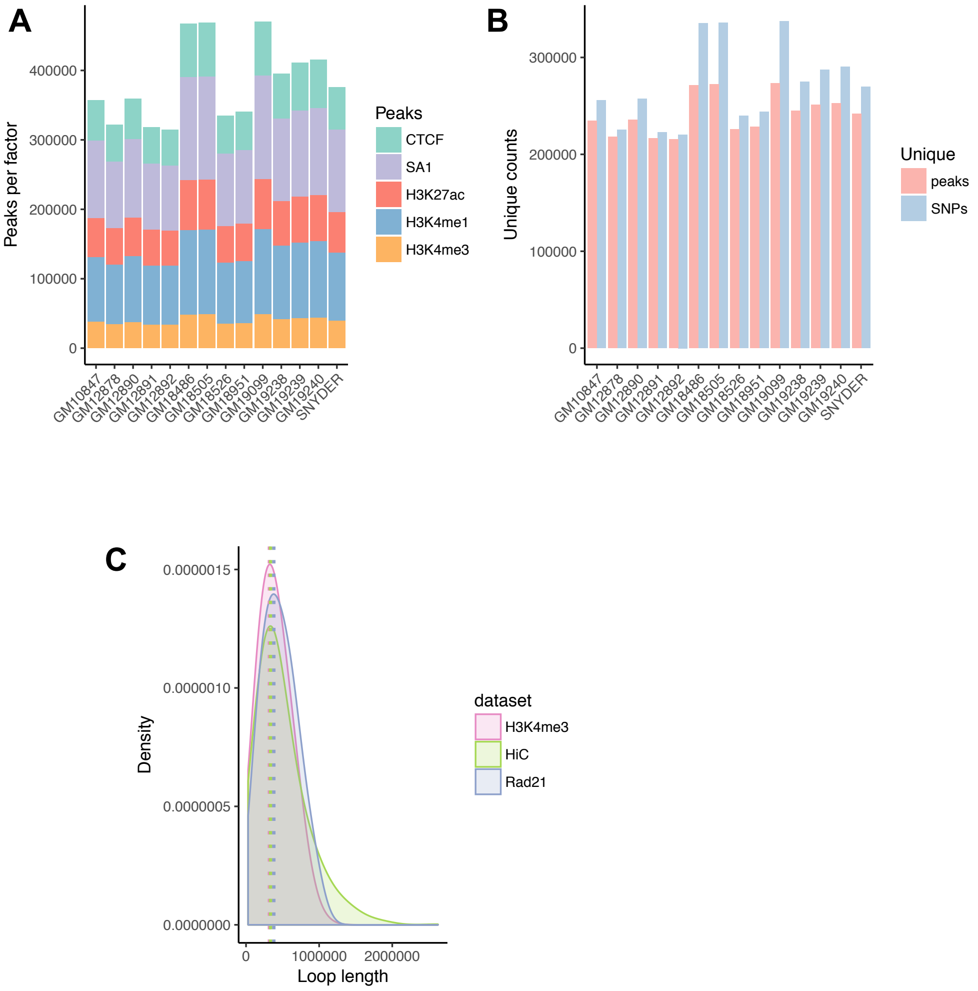


Figure S4.3, related to figure 4.3. Characterisation of allele-specific factors across individuals
Number of **(A)** peaks per genomic factor, **(B)** unique peaks and SNPs per individual. **(C)** Average loop length for the three datasets. All panels were done by myself.

D. Supplementary information for chapter 5

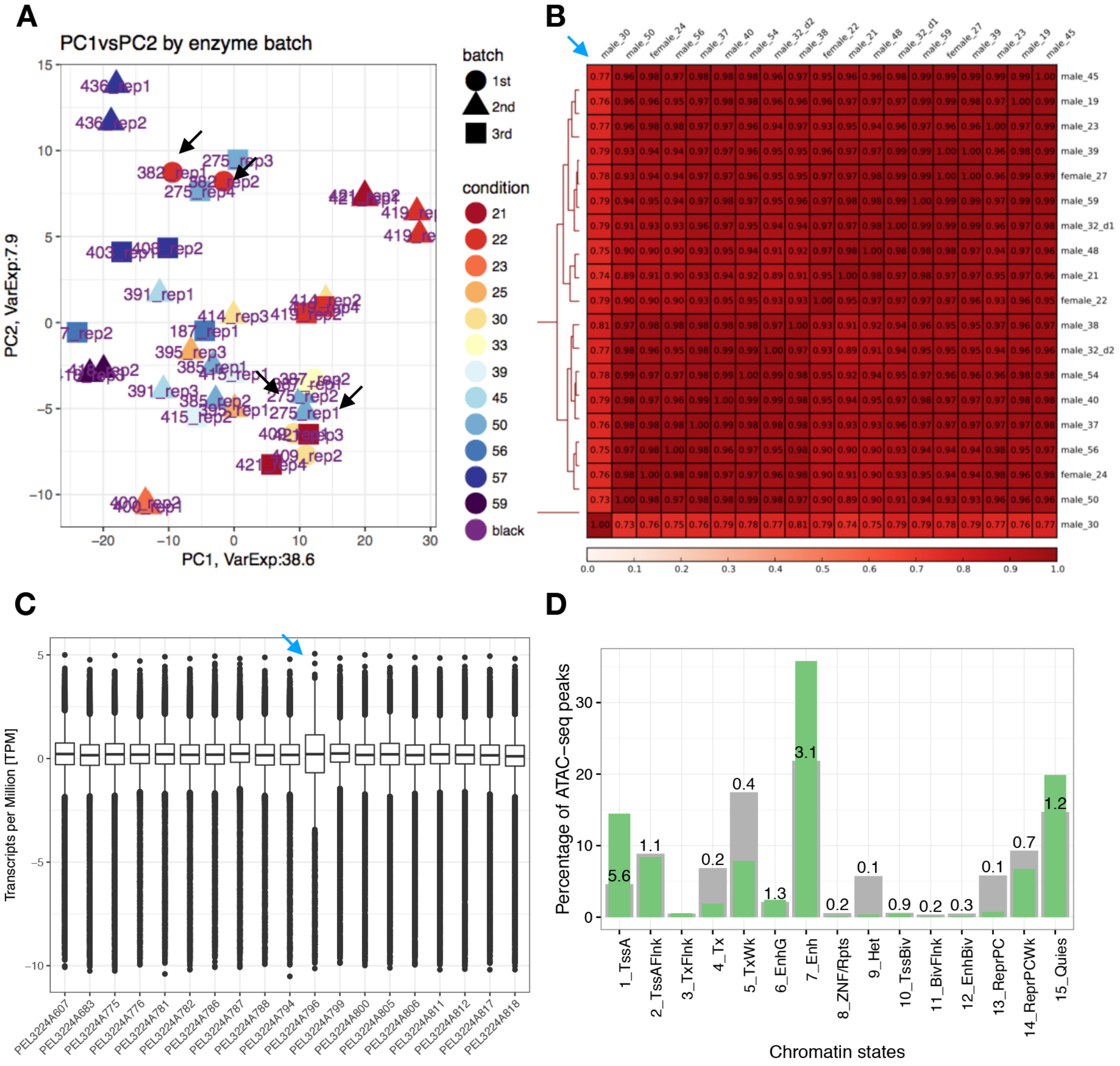


Figure S5.1, related to figure 5.1. Overview of quality data and outliers per technique

(A) PCA of the 38 ATAC-seq libraries coloured by age shows some outliers (black arrows). In addition, based on the TSS enrichments we also removed 4 more libraries, leaving us with 30 libraries corresponding to 15 donors. **(B)** Heatmap for the pearson correlation of RNA-seq bam files produced with bamCorrelate from deeptools (Ramírez et al. 2014) and **(C)** distribution of transcripts per million shows one outlier (blue arrows). **(D)** Distribution of DHS-seq peaks (%) per chromatin state (green) vs the total percentage of each state as the background (grey). All panels were produced by myself.

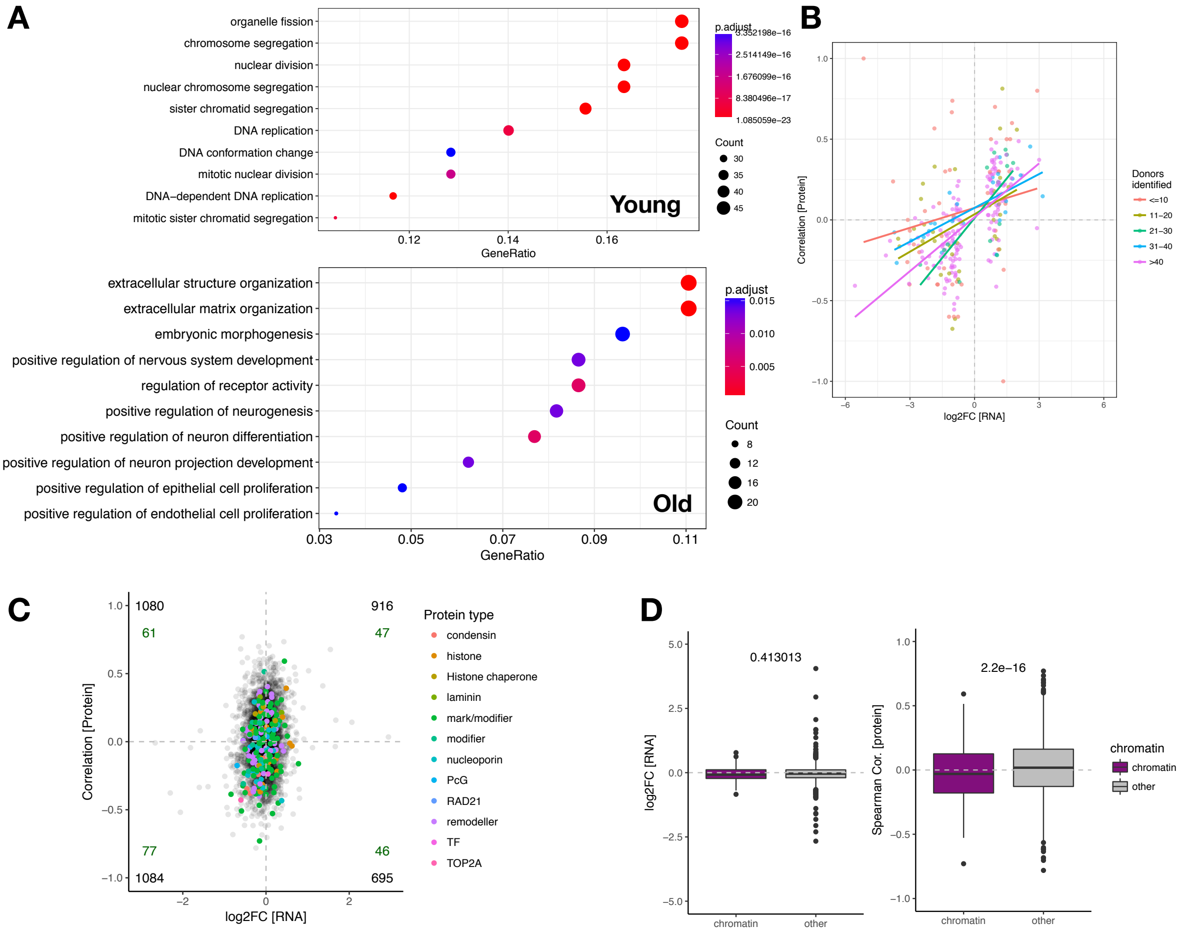
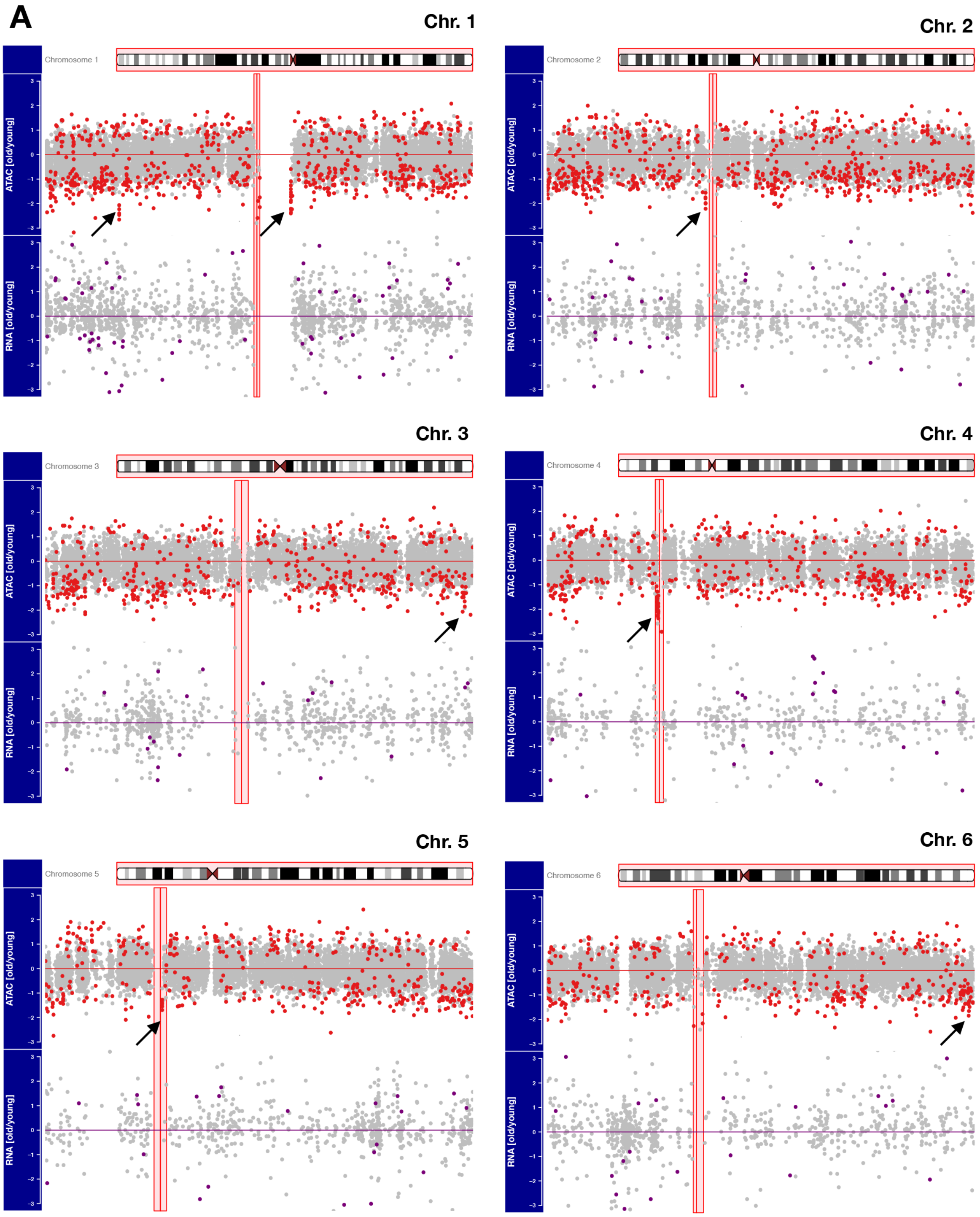


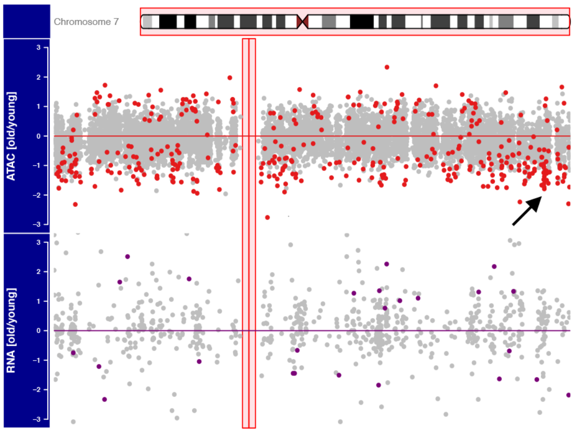
Figure S5.2, related to figure 5.2. Integration of RNA and protein in HSCs

(A) Gene ontology enrichment analysis of biological processes shows terms as ‘nuclear division’ or ‘chromosome segregation’ for down- and ‘ECM organization’ for up-regulated genes. **(B, C)** Scatterplot of log2-FC (gene) vs spearman correlation (protein) values **(B)** coloured by number of donors where the protein was detected for DEGs in MSCs or **(C)** for all genes with both values in HSCs coloured whether they interact with chromatin. **(D)** Distribution of log2-FC and spearman correlation of chromatin-related (purple) and other proteins (grey) in HSCs shows that the chromatin-proteins downregulation with age is not as strong as in MSCs. This figure was done by myself.

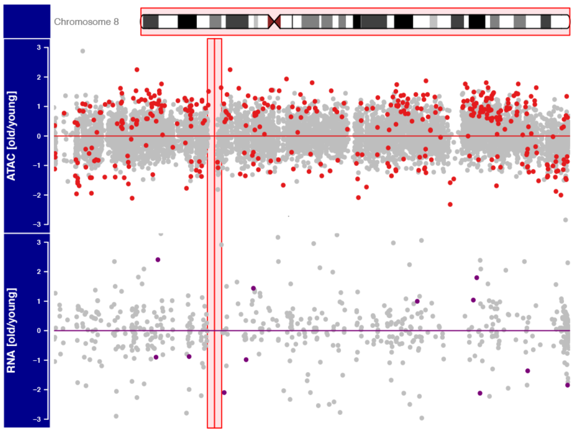


(Continue in the following page)

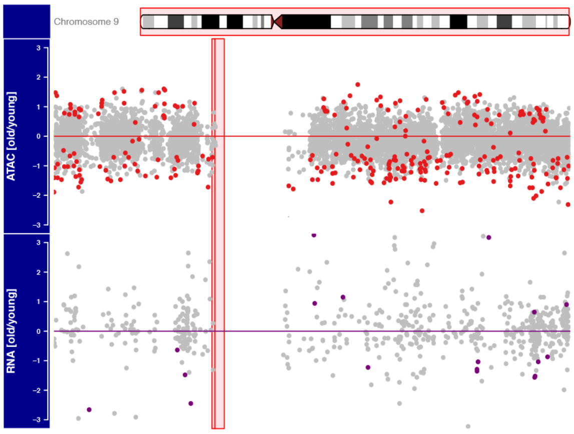
Chr. 7



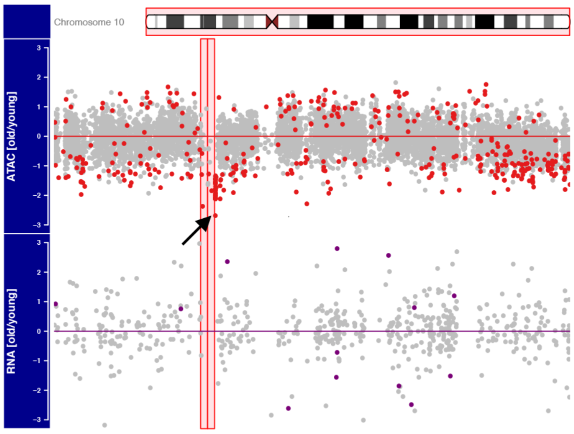
Chr. 8



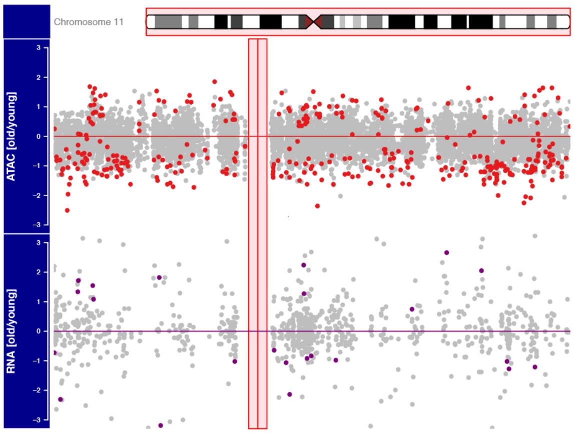
Chr. 9



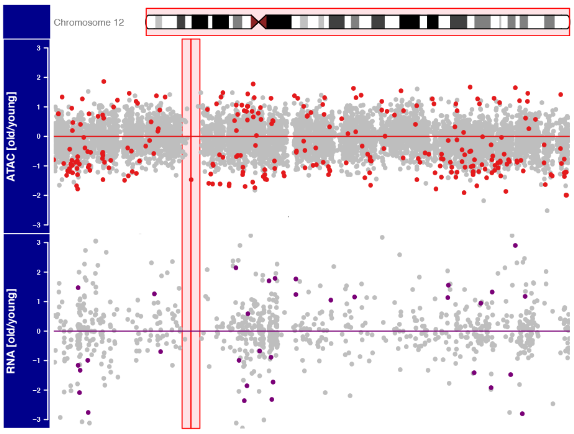
Chr. 10



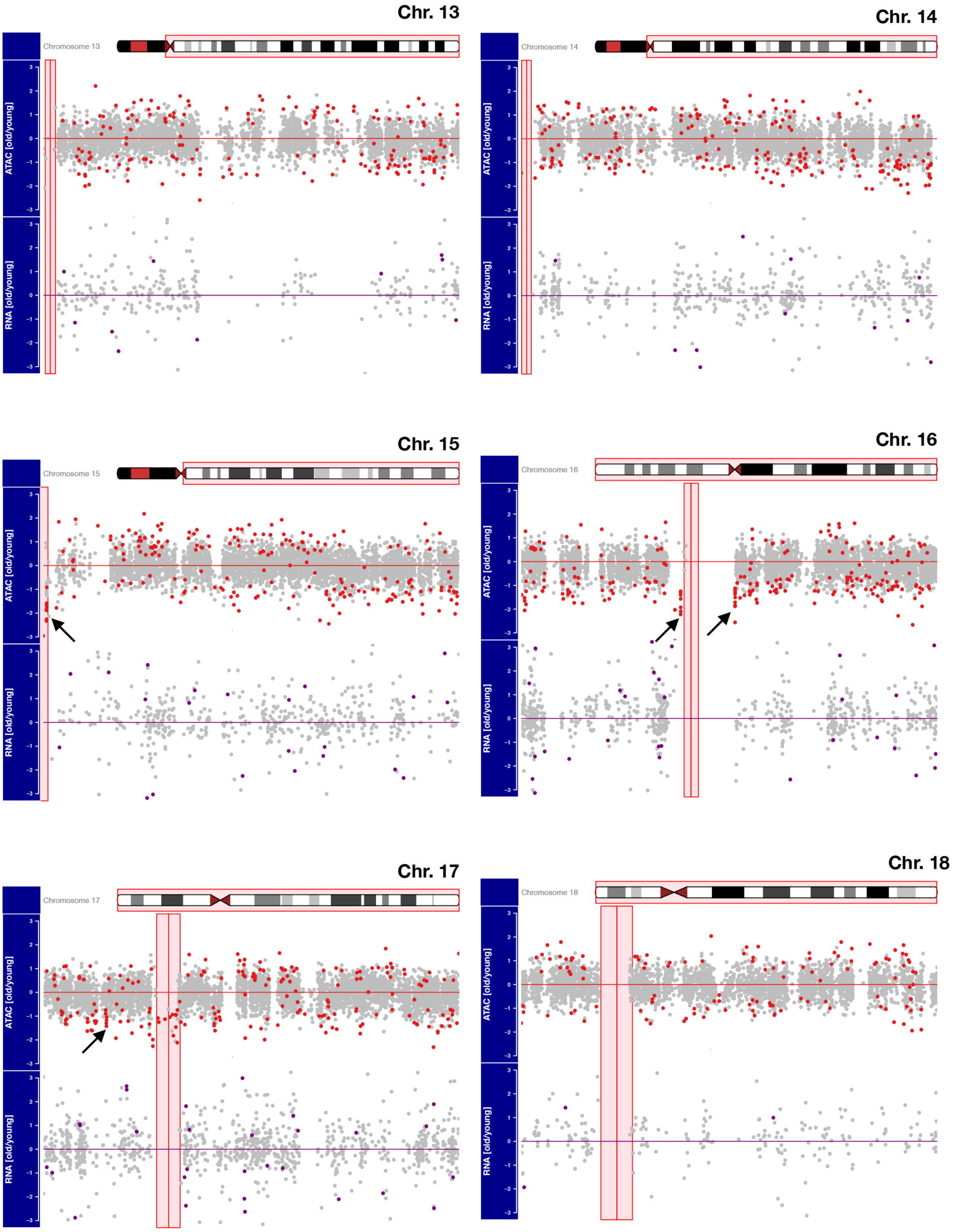
Chr. 11



Chr. 12



(Continue in the following page)



(Continue in the following page)

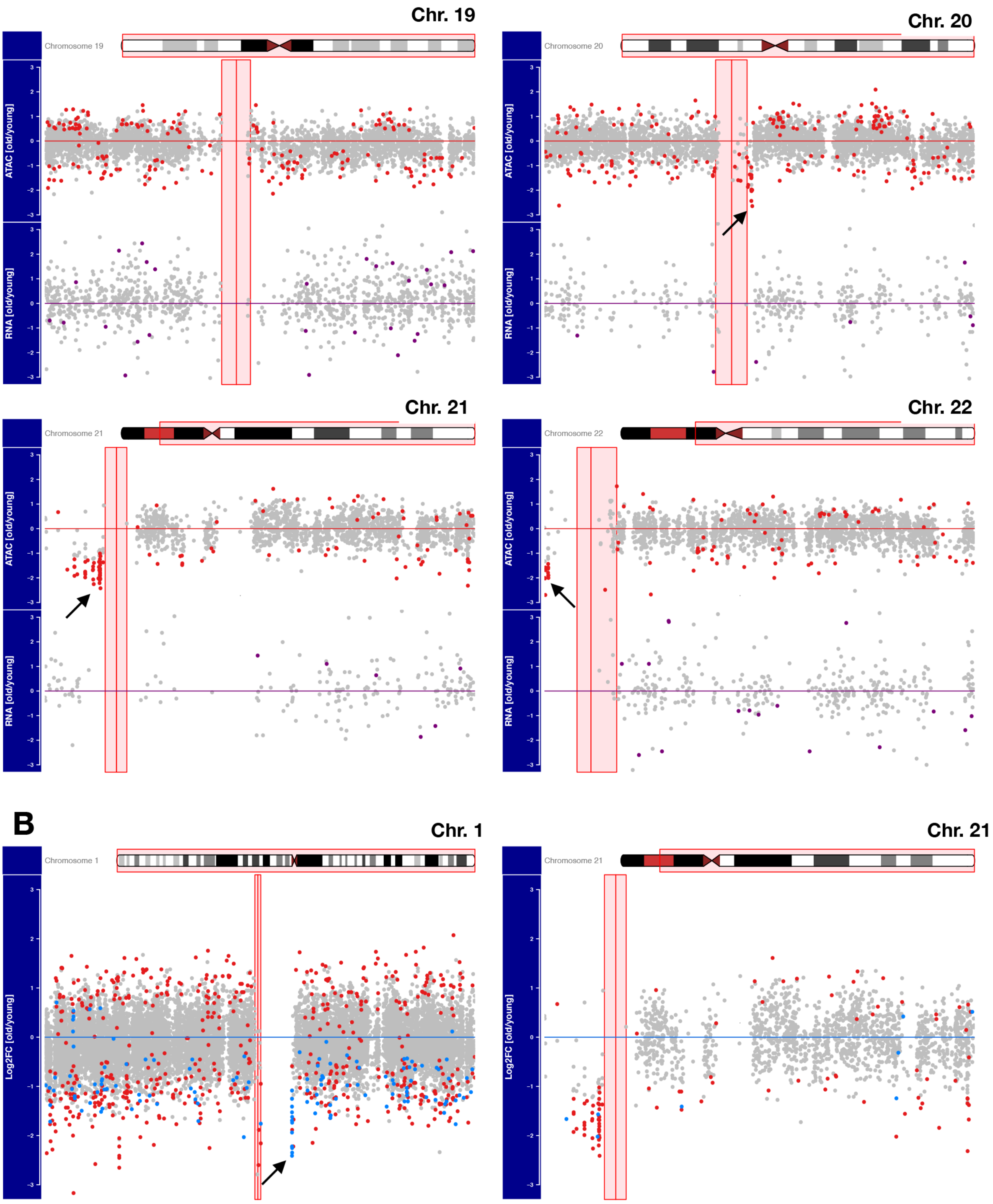


Figure S5.3, related to figure 5.3. Linear visualization of chromatin accessibility changes with age

(A) Depiction of accessibility peaks along the linear chromosomes stratified by log2-FC and coloured by significantly changing with age (red) or not significant (grey). The RNA log2-FC is plotted below and show no specific enrichment of misregulated genes in the chromatin hotspots (indicated by the black arrows), **(B)** CARs containing CTCF motifs are coloured (blue) and show accumulation only at certain hotspots. All figures were produced by myself.

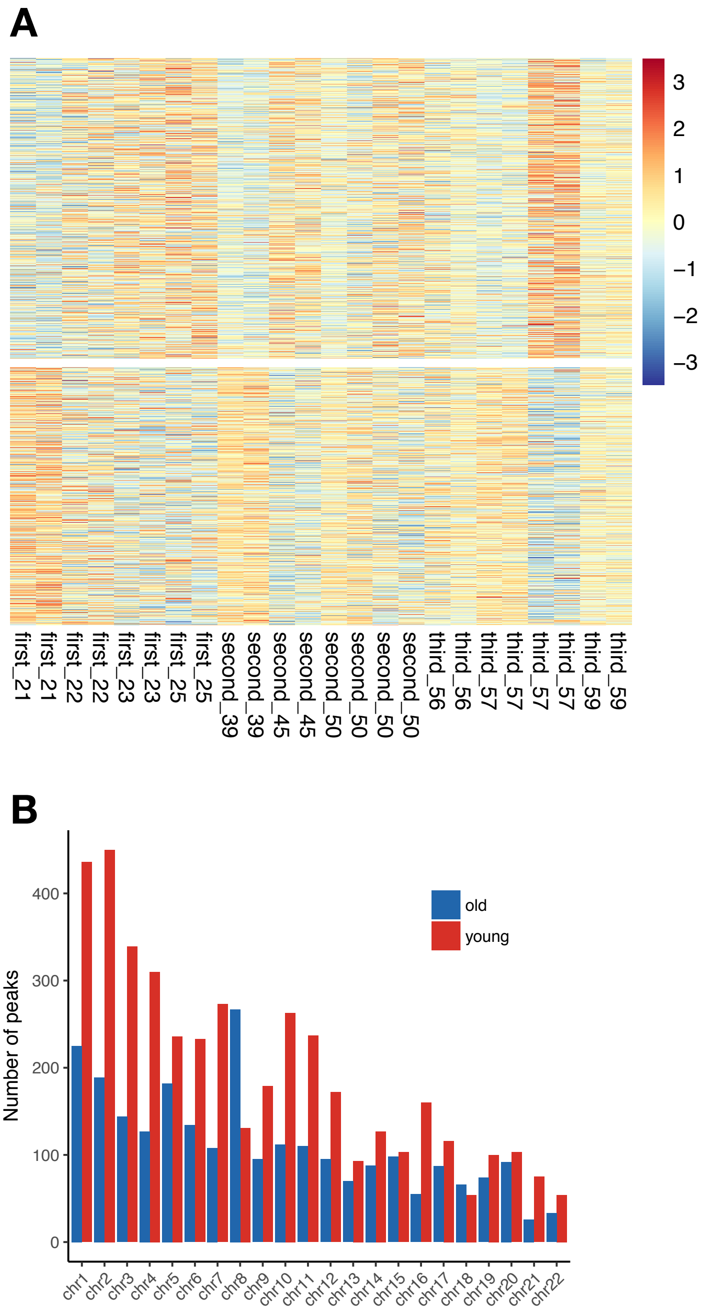


Figure S5.4. Heatmap of CARs according to the 3 groups and distributed by chromosome, related to figure 5.4
(A) Heatmap showing the normalised ATAC-seq signal at CARs calculated from the comparison of young vs old (log2-FC approach). **(B)** Distribution of CARs per chromosome stratified by log2-FC as more accessible in younger (red) or in older (blue) samples. This figure was generated by myself.

Table S5.1. List of ageing proteins associated to chromatin common for MSCs and HSCs

Gene name	Ensembl_id	log2FC RNA MSC	Spearman cor. MSC	log2FC RNA HSC	Spearman cor HSC
APEX1	ENSG00000100823	-0.56	-0.47	-0.04	-0.19
ATAD2	ENSG00000156802	-1.31	-0.3	-0.22	-0.3
ATF2	ENSG00000115966	-0.43	-0.34	-0.16	-0.73
BRD7	ENSG00000166164	-0.32	-0.49	-0.04	-0.19
BRD8	ENSG00000112983	-0.57	-0.38	-0.06	-0.17
CBX5	ENSG00000094916	-0.43	-0.36	-0.16	-0.03
CDK1	ENSG00000170312	-2.61	-0.16	-0.56	-0.32
CDK2	ENSG00000123374	-0.94	-0.42	-0.41	-0.14
CHD2	ENSG00000173575	1.06	0.03	0.10	0.20
DAPK3	ENSG00000167657	0.09	0.12	0.18	0.9
EED	ENSG00000074266	-0.1	-0.10	-0.69	-0.18
ELP4	ENSG00000109911	-0.23	-0.24	-0.16	-0.04
H2AFX	ENSG00000188486	-1.28	-0.38	-0.30	-0.15
HDAC1	ENSG00000116478	-0.43	-0.26	-0.11	-0.12
HDAC3	ENSG00000171720	-0.39	-0.33	-0.09	-0.31
HELLS	ENSG00000119969	-1.57	-0.3	-0.20	-0.17
HMGN2	ENSG00000198830	-1.05	-0.28	-0.28	-0.34
HMGN3	ENSG00000118418	-0.25	-0.22	-0.41	-0.10
HMGN4	ENSG00000182952	-0.09	-0.08	-0.27	-0.09
LEO1	ENSG00000166477	-0.42	-0.26	-0.07	-0.21
MAPK3	ENSG00000102882	0.26	0.18	0.24	0.09
MDC1	ENSG00000137337	-0.33	-0.27	-0.13	-0.16
MECP2	ENSG00000169057	0.7	0.21	0.44	0.59
MTA2	ENSG00000149480	-0.47	-0.29	-0.08	-0.002
NASP	ENSG00000132780	-1.03	-0.29	-0.14	-0.37
NBN	ENSG00000104320	-0.46	-0.29	-0.32	-0.11

NCAPD2	ENSG00000010292	-1.163	-0.21	-0.24	-0.30
NCAPG	ENSG00000109805	-3.05	-0.15	-0.46	-0.35
NCAPG2	ENSG00000146918	-2.18	-0.3	-0.03	-0.12
NCAPH	ENSG00000121152	-4.12	-0.22	-0.39	-0.39
NCAPH2	ENSG00000025770	-1.03	-0.4	-0.18	-0.24
NDC1	ENSG00000058804	-0.94	-0.11	-0.21	-0.03
NUP107	ENSG00000111581	-0.32	-0.31	-0.10	-0.16
NUP35	ENSG00000163002	-1.9	-0.28	-0.23	-0.22
NUP37	ENSG00000075188	-0.18	-0.4	-0.44	-0.25
NUP43	ENSG00000120253	-0.39	-0.12	-0.15	-0.12
NUP50	ENSG00000093000	-0.39	-0.38	-0.12	-0.28
NUP62	ENSG00000213024	-0.44	-0.12	-0.08	-0.34
NUP85	ENSG00000125450	-0.57	-0.30	-0.19	-0.03
NUP88	ENSG00000108559	-0.82	-0.35	-0.01	-0.15
NUP93	ENSG00000102900	-0.95	-0.28	-0.23	-0.29
PCNA	ENSG00000132646	-1.32	-0.37	-0.31	-0.37
PPM1G	ENSG00000115241	-0.48	-0.28	-0.06	-0.23
PSIP1	ENSG00000164985	-0.44	-0.23	-0.45	-0.06
RBBP7	ENSG00000102054	-0.69	-0.19	-0.17	-0.18
SAP18	ENSG00000150459	-0.16	-0.32	-0.3	-0.25
SF3B3	ENSG00000189091	-0.27	-0.39	-0.12	-0.13
SKP1	ENSG00000113558	-0.18	-0.08	-0.12	-0.01
SRSF1	ENSG00000136450	-0.59	-0.27	-0.42	-0.22
TFDP1	ENSG00000198176	-1.04	-0.21	-0.13	-0.3
TOP2A	ENSG00000131747	-2.63	-0.11	-0.60	-0.43
UBE2N	ENSG00000177889	-0.49	-0.13	-0.50	-0.06
UBR7	ENSG00000012963	-1.35	-0.23	-0.32	-0.02
UHRF1	ENSG00000276043	-1.89	-0.31	-0.59	-0.51

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Arampatzi, P., Gialitakis, M., Makatounakis, T., and Papamatheakis, J. (2013). Gene-specific factors determine mitotic expression and bookmarking via alternate regulatory elements. Nucleic Acids Res. *41*, 2202–2215.

Arnold, C., Bhat, P., and Zaugg, J.B. (2016). SNPhood: investigate, quantify and visualise the epigenomic neighbourhood of SNPs using NGS data. Bioinformatics *32*, 2359–2360.

Bancaud, A., Lavelle, C., Huet, S., and Ellenberg, J. (2012). A fractal model for nuclear organization: current evidence and biological implications. Nucleic Acids Res. *40*, 8783–8792.

Baribault, C., Ehrlich, K.C., Ponnaluri, V.K.C., Pradhan, S., Lacey, M., and Ehrlich, M. (2018). Developmentally linked human DNA hypermethylation is associated with down-modulation, repression, and upregulation of transcription. Epigenetics *13*, 275–289.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

Beagan, J.A., Gilgenast, T.G., Kim, J., Plona, Z., Norton, H.K., Hu, G., Hsu, S.C., Shields, E.J., Lyu, X., Apostolou, E., et al. (2016). Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State during Somatic Cell Reprogramming. Cell Stem Cell *18*, 611–624.

Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell *126*, 37–47.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonev, B., Cohen, N.M., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557–572.e24.

Booth, L.N., and Brunet, A. (2016). The Aging Epigenome. Mol. Cell *62*, 728–744.

Bornelöv, S., Komorowski, J., and Wadelius, C. (2015). Different distribution of histone modifications in genes with unidirectional and bidirectional transcription and a role of CTCF and cohesin in directing transcription. BMC Genomics *16*, 300.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., Stein, L.D., and - ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net (2017). Pan-cancer analysis of whole genomes.

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., et al. (2017). Genome Organization Drives Chromosome Fragility. Cell.

Chandra, T., Ewels, P.A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S.W., Kirschner, K., Thuret, J.-Y., Andrews, S., Fraser, P., and Reik, W. (2015). Global reorganization of the nuclear landscape in senescent cells. Cell Rep. *10*, 471–483.

Chen, Y., Breeze, C.E., Zhen, S., Beck, S., and Teschendorff, A.E. (2016). Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. Epigenetics Chromatin *9*, 10.

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. *46*, D267–D275.

Choi, H.Y., Saha, S.K., Kim, K., Kim, S., Yang, G.-M., Kim, B., Kim, J.-H., and Cho, S.-G. (2015). G protein-coupled receptors in stem cell maintenance and somatic reprogramming to pluripotent or cancer stem cells. BMB Rep. *48*, 68–80.

Cleveland, D.W., Mao, Y., and Sullivan, K.F. (2003). Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell *112*, 407–421.

Cournac, A., Koszul, R., and Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic Acids Res. *44*, 245–255.

Cremer, T., and Cremer, C. (2001). Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells. Nature Reviews Genetics *2*, 292–301.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., Shamim, M., et al. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proc. Natl. Acad. Sci. U. S. A.

Delaneau, O., Zazhytska, M., Borel, C., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D., Ambrosini, G., Bielse, D., et al. (2017). Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Deschênes, M., and Chabot, B. (2017). The emerging role of alternative splicing in senescence and aging. Aging Cell *16*, 918–933.

de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell *60*, 676–684.

Dileep, V., Ay, F., Sima, J., Vera, D.L., Noble, W.S., and Gilbert, D.M. (2015). Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. Genome Res. *25*, 1104–1113.

Ding, Z., Ni, Y., Timmer, S.W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G.E., Lieb, J.D., et al. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. PLoS Genet. *10*, e1004798.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

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

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Dreesen, O., Ong, P.F., Chojnowski, A., and Colman, A. (2013). The contrasting roles of lamin B1 in cellular aging and human disease. Nucleus *4*, 283–290.

Duncan, P.I., Howell, B.W., Marius, R.M., Drmanic, S., Douville, E.M., and Bell, J.C. (1995). Alternative splicing of STY, a nuclear dual specificity kinase. J. Biol. Chem. *270*, 21524–21531.

Fang, H. (2014). dcGOR: an R package for analysing ontologies and protein domain annotations. PLoS Comput. Biol. *10*, e1003929.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Fernández, A.F., Bayón, G.F., Urdinguio, R.G., Toraño, E.G., García, M.G., Carella, A., Petrus-Reurer, S., Ferrero, C., Martinez-Camblor, P., Cubillo, I., et al. (2015). H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. Genome Res. *25*, 27–40.

Feser, J., Truong, D., Das, C., Carson, J.J., Kieft, J., Harkness, T., and Tyler, J.K. (2010). Elevated histone expression promotes life span extension. Mol. Cell *39*, 724–735.

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.-D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. *28*, 2663–2676.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C.A., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852.

Frith, J., and Genever, P. (2008). Transcriptional control of mesenchymal stem cell differentiation. Transfus. Med. Hemother. *35*, 216–227.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. *15*, 2038–2049.

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., and Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J. Biol. Chem. *278*, 4035–4040.

Funayama, R., Saito, M., Tanobe, H., and Ishikawa, F. (2006). Loss of linker histone H1 in cellular senescence. J. Cell Biol. *175*, 869–880.

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. Science.

García-Sacristán, A., Fernández-Nestosa, M.J., Hernández, P., Schvartzman, J.B., and Krimer, D.B. (2005). Protein kinase clk/STY is differentially regulated during erythroleukemia cell differentiation: a bias toward the skipped splice variant characterizes postcommitment stages. Cell Res. *15*, 495–503.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575–579.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell *14*, 762–775.

Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell *162*, 1051–1065.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Gubelmann, C., Schwalie, P.C., Raghav, S.K., Röder, E., Delessa, T., Kiehlmann, E., Waszak, S.M., Corsinotti, A., Udin, G., Holcombe, W., et al. (2014). Identification of the transcription factor ZEB1 as a central component of the adipogenic gene regulatory network. Elife *3*, e03346.

Guerrero, P.A., and Maggert, K.A. (2011). The CCCTC-binding factor (CTCF) of Drosophila contributes to the regulation of the ribosomal DNA and nucleolar stability. PLoS One *6*, e16401.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

Hake, S.B., and David Allis, C. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: The "H3 barcode hypothesis." Proc. Natl. Acad. Sci. U. S. A. *103*, 6428–6435.

Hanes, J., von der Kammer, H., Klaudiny, J., and Scheit, K.H. (1994). Characterization by cDNA cloning of two new human protein kinases. Evidence by sequence comparison of a new family of mammalian protein kinases. J. Mol. Biol. *244*, 665–672.

Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G., and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol. Cell *66*, 711–720.e3.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., et al. (2012). Distinct DNA methylomes of newborns and centenarians. Proc. Natl. Acad. Sci. U. S. A. *109*, 10522–10527.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Hnisz, D., S. Weintraub, A., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Peng Fan, Z., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Huang, Y., Li, W., Yao, X., Lin, Q.-J., Yin, J.-W., Liang, Y., Heiner, M., Tian, B., Hui, J., and Wang, G. (2012). Mediator complex regulates alternative mRNA processing via the MED23 subunit. Mol. Cell *45*, 459–469.

Ing-Simmons, E., Seitan, V.C., Faure, A.J., Flicek, P., Carroll, T., Dekker, J., Fisher, A.G., Lenhard, B., and Merkenschlager, M. (2015). Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res. *25*, 504–513.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262–275.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. *41*, D793–D800.

Karczewski, K.J., Dudley, J.T., Kukurba, K.R., Chen, R., Butte, A.J., Montgomery, S.B., and Snyder, M. (2013). Systematic functional regulatory assessment of disease-associated variants. Proc. Natl. Acad. Sci. U. S. A. *110*, 9607–9612.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

Kemp, C.J., Moore, J.M., Moser, R., Bernard, B., Teater, M., Smith, L.E., Rabaia, N.A., Gurley, K.E., Guinney, J., Busch, S.E., et al. (2014). CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Rep. *7*, 1020–1029.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013).

Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat. Rev. Mol. Cell Biol. *14*, 153–165.

Kruse, K., Hug, C.B., Hernández-Rodríguez, B., and Vaquerizas, J.M. (2016). TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics.

Kschonsak, M., Merkel, F., Bisht, S., Metz, J., Rybin, V., Hassler, M., and Haering, C.H. (2017). Structural Basis for a Safety-Belt Mechanism That Anchors Condensin to Chromosomes. Cell *171*, 588–600.e24.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259.

Kular, J.K., Basu, S., and Sharma, R.I. (2014). The extracellular matrix: Structure, composition, age-related differences, tools for analysis and applications for tissue engineering. J. Tissue Eng. *5*, 2041731414557112.

Kung, J.T., Kesner, B., An, J.Y., Ahn, J.Y., Cifuentes-Rojas, C., Colognori, D., Jeon, Y., Szanto, A., del Rosario, B.C., Pinter, S.F., et al. (2015). Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. Mol. Cell *57*, 361–375.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Larkin, J.D., Cook, P.R., and Papantonis, A. (2012). Dynamic reconfiguration of long human genes during one transcription cycle. Mol. Cell. Biol. *32*, 2738–2747.

Lawrence, J.B., Singer, R.H., and McNeil, J.A. (1990). Interphase and metaphase resolution of different distances within the human dystrophin gene. Science *249*, 928–932.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

Lei, T., Bi, Y., Gao, M.J., Gao, S.M., Zhou, L.L., Zheng, H.L., and Chen, X.D. (2013). HES1 inhibits adipogenesis of porcine mesenchymal stem cells via transcriptional repression of FAD24. Domest. Anim. Endocrinol. *45*, 28–32.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Y., Huang, W., Niu, L., Umbach, D.M., Covo, S., and Li, L. (2013). Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC Genomics *14*, 553.

Li, Y., Yang, M., Zhang, G., Li, L., Ye, B., Huang, C., and Tang, Y. (2018). Transcription factor TBX18 promotes adult rat bone mesenchymal stem cell differentiation to biological pacemaker cells. Int. J. Mol. Med. *41*, 845–851.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Liu, L., Cheung, T.H., Charville, G.W., Hurgo, B.M.C., Leavitt, T., Shih, J., Brunet, A., and Rando, T.A. (2013). Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. Cell Rep. *4*, 189–204.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A., and Issa, J.-P.J. (2010). Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. *20*, 332–340.

Marina, R.J., Sturgill, D., Bailly, M.A., Thenoz, M., Varma, G., Prigge, M.F., Nanan, K.K., Shukla, S., Haque, N., and Oberdoerffer, S. (2016). TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. EMBO J. *35*, 335–355.

Martinez-Jimenez, C.P., Eling, N., Chen, H.-C., Vallejos, C.A., Kolodziejczyk, A.A., Connor, F., Stojic, L., Rayner, T.F., Stubbington, M.J.T., Teichmann, S.A., et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science *355*, 1433–1436.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Maurano, M.T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J.A. (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell Rep. *12*, 1184–1195.

Medvedeva, Y.A., Lennartsson, A., Ehsani, R., Kulakovskiy, I.V., Vorontsov, I.E., Panahandeh, P., Khimulya, G., Kasukawa, T., FANTOM Consortium, and Drabløs, F. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database *2015*, bav067.

Melé, M., and Rinn, J.L. (2016). “Cat’s Cradling” the 3D Genome by the Act of LncRNA Transcription. Mol. Cell *62*, 657–664.

Mercer, T.R., Edwards, S.L., Clark, M.B., Neph, S.J., Wang, H., Stergachis, A.B., John, S., Sandstrom, R., Li, G., Sandhu, K.S., et al. (2013). DNase I–hypersensitive exons colocalize with promoters and distal regulatory elements. Nat. Genet. *45*, 852–859.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Mikula, M., Bomsztyk, K., Goryca, K., Chojnowski, K., and Ostrowski, J. (2013). Heterogeneous nuclear ribonucleoprotein (HnRNP) K genome-wide binding survey reveals its role in regulating 3'-end RNA processing and transcription termination at the early growth response 1 (EGR1) gene through XRN2 exonuclease. J. Biol. Chem. *288*, 24788–24798.

Moerman, E.J., Teng, K., Lipschitz, D.A., and Lecka-Czernik, B. (2004). Aging activates adipogenic and suppresses osteogenic programs in mesenchymal marrow stroma/stem cells: the role of PPAR-gamma2 transcription factor and TGF-beta/BMP signaling pathways. Aging Cell *3*, 379–389.

Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M., and Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. Proc. Natl. Acad. Sci. U. S. A. *109*, 9125–9130.

Moore, B.L., Aitken, S., and Semple, C.A. (2015). Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. *16*, 1–14.

Moore, J.M., Rabaia, N.A., Smith, L.E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C.M., Collins, S.J., Kemp, C.J., Lobanenkov, V.V., et al. (2012). Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. PLoS One *7*, e34915.

Morató, X., Luján, R., López-Cano, M., Gandía, J., Stagljar, I., Watanabe, M., Cunha, R.A., Fernández-Dueñas, V., and Ciruela, F. (2017). The Parkinson's disease-associated GPR37 receptor interacts with striatal adenosine A2A receptor controlling its cell surface expression and function in vivo. Sci. Rep. *7*, 9452.

Mourad, R., and Cuvier, O. (2016). Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation. PLoS Comput. Biol. *12*, e1004908.

von Muhlinen, N., Horikawa, I., Alam, F., Isogaya, K., Lissa, D., Vojtesek, B., Lane, D.P., and Harris, C.C. (2018). p53 isoforms regulate premature aging in human cells. Oncogene *37*, 2379–2393.

Nakahashi, H., Kieffer Kwon, K.-R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Rep. *3*, 1678–1689.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Danny, A.R. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017–1021.

Neems, D.S., Garza-Gongora, A.G., Smith, E.D., and Kosak, S.T. (2016). Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. Proc. Natl. Acad. Sci. U. S. A. *113*, E1691–E1700.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22.

Norton, H.K., and Phillips-Cremins, J.E. (2017). Crossed wires: 3D genome misfolding in human disease. J. Cell Biol.

Oesterreich, F.C., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat. Genet. *36*, 1065–1071.

Pękowska, A., Klaus, B., Klein, F.A., Anders, S., Oleś, M., Steinmetz, L.M., Bertone, P., and Huber, W. (2014). Neural lineage induction reveals multi-scale dynamics of 3D chromatin organization. bioRxiv.

Peña-Hernández, R., Marques, M., Hilmi, K., Zhao, T., Saad, A., Alaoui-Jamali, M.A., del Rincon, S.V., Ashworth, T., Roy, A.L., Emerson, B.M., et al. (2015). Genome-wide targeting of the epigenetic regulatory protein CTCF to gene promoters by the transcription factor TFII-I. Proc. Natl. Acad. Sci. U. S. A. *112*, E677–E686.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell *38*, 603–613.

Phanstiel, D.H., Boyle, A.P., Heidari, N., and Snyder, M.P. (2015). Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics *31*, 3092–3098.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong, J.W.H. (2016). Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. Cell Rep. *17*, 2865–2872.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., et al. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. *20*, 434–439.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Razquin Navas, P., and Thedieck, K. (2017). Differential control of ageing and lifespan by isoforms and splice variants across the mTOR network. Essays Biochem. *61*, 349–368.

Reagan, M.R., and Rosen, C.J. (2016). Navigating the bone marrow niche: translational insights and cancer-driven dysfunction. Nat. Rev. Rheumatol. *12*, 154–168.

Ren, R., Deng, L., Xue, Y., Suzuki, K., Zhang, W., Yu, Y., Wu, J., Sun, L., Gong, X., Luan, H., et al. (2017). Visualization of aging-associated chromatin alterations with an engineered TALE system. Cell Res. *27*, 483–504.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Ros, S., and Schulze, A. (2013). Balancing glycolytic flux: the role of 6-phosphofructo-2-kinase/fructose 2,6-bisphosphatases in cancer metabolism. Cancer Metab *1*, 8.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389–393.

Ruiz-Velasco, M., and Zaugg, J.B. (2017). Structure meets function: How chromatin organisation conveys functionality. Current Opinion in Systems Biology *1*, 129–136.

Ruiz-Velasco, M., Kumar, M., Lai, M.C., Bhat, P., Solis-Pinson, A.B., Reyes, A., Kleinsorg, S., Noh, K.-M., Gibson, T.J., and Zaugg, J.B. (2017). CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. Cell Syst *5*, 628–637.e6.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature *532*, 264–267.

Saldaña-Meyer, R., González-Buendía, E., Guerrero, G., Narendra, V., Bonasio, R., Recillas-Targa, F., and Reinberg, D. (2014). CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. Genes Dev. *28*, 723–734.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A. *112*, E6456–E6465.

Scaffidi, P., and Misteli, T. (2006). Lamin A-dependent nuclear defects in human aging. Science *312*, 1059–1063.

Scaffidi, P., and Misteli, T. (2008). Lamin A-dependent misregulation of adult stem cells associated with accelerated ageing. Nat. Cell Biol. *10*, 452–459.

Schultz, M.B., and Sinclair, D.A. (2016). When stem cells grow old: phenotypes and mechanisms of stem cell aging. Development *143*, 3–14.

Schwalie, P.C., Ward, M.C., Cain, C.E., Faure, A.J., Gilad, Y., Odom, D.T., and Flicek, P. (2013). Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. Genome Biol. *14*, R148.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.

Soneson, C., Matthes, K.L., Nowicka, M., Law, C.W., and Robinson, M.D. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biol. *17*, 12.

Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature *558*, 73–79.

Sun, D., Luo, M., Jeong, M., Rodriguez, B., Xia, Z., Hannah, R., Wang, H., Le, T., Faull, K.F., Chen, R., et al. (2014). Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. Cell Stem Cell *14*, 673–688.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

Swedlow, J.R., and Hirano, T. (2003). The making of the mitotic chromosome: modern insights into classical questions. Mol. Cell *11*, 557–569.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res. *24*, 390–400.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., Diana, E., Lehmann, G., Toren, D., Wang, J., et al. (2018). Human Ageing Genomic Resources: new and updated databases. Nucleic Acids Res. *46*, D1083–D1090.

Tang, B., Dean, B., and Thomas, E.A. (2011). Disease- and age-related changes in histone acetylation at gene promoters in psychiatric disorders. Transl. Psychiatry *1*, e64.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611–1627.

Trinchero, M.F., Buttner, K.A., Sulkes Cuevas, J.N., Temprana, S.G., Fontanet, P.A., Monzón-Salinas, M.C., Ledda, F., Paratcha, G., and Schinder, A.F. (2017). High Plasticity of New Granule Cells in the Aging Hippocampus. Cell Rep. *21*, 1129–1139.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics Chromatin *7*, 33.

Tsujimura, T., Klein, F.A., Langenfeld, K., Glaser, J., Huber, W., and Spitz, F. (2015). A discrete transition zone organizes the topological and regulatory autonomy of the adjacent tfap2c and bmp7 genes. PLoS Genet. *11*, e1004897.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Rep. *10*, 1297–1309.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

Wang, Y., Zhang, J., Xiao, X., Liu, H., Wang, F., Li, S., Wen, Y., Wei, Y., Su, J., Zhang, Y., et al. (2016). The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. Sci. Rep. *6*, 22722.

Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050.

van de Werken, H.J.G., Landan, G., Holwerda, S.J.B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A.M., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods *9*, 969–972.

Wijchers, P.J., Krijger, P.H.L., Geeven, G., Zhu, Y., Denker, A., Verstegen, M.J.A.M., Valdes-Quezada, C., Vermeulen, C., Janssen, M., Teunissen, H., et al. (2016). Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. Mol. Cell *61*, 461–473.

Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. Mol. Cell. Biol. *31*, 2174–2183.

Yang, H.-J., Vainshtein, A., Maik-Rachline, G., and Peles, E. (2016). G protein-coupled receptor 37 is a negative regulator of oligodendrocyte differentiation and myelination. Nat. Commun. *7*, 10884.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G.R.S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2015). The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics *31*, 143–145.

Yi, J.M., Dhir, M., Van Neste, L., Downing, S.R., Jeschke, J., Glöckner, S.C., de Freitas Calmon, M., Hooker, C.M., Funes, J.M., Boshoff, C., et al. (2011). Genomic and epigenomic integration identifies a prognostic signature in colon cancer. Clin. Cancer Res. *17*, 1535–1545.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Yuen, K.C., and Gerton, J.L. (2018). Taking cohesin and condensin in context. PLoS Genet. *14*, e1007118.

Zhang, K., Li, N., Ainsworth, R.I., and Wang, W. (2016). Systematic identification of protein combinations mediating chromatin looping. Nat. Commun. *7*, 12249.

Zhao, H., Sifakis, E.G., Sumida, N., Millán-Ariño, L., Scholz, B.A., Svensson, J.P., Chen, X., Ronnegren, A.L., Mallet de Lima, C.D., Varnoosfaderani, F.S., et al. (2015). PARP1- and CTCF-Mediated Interactions between Active and Repressed Chromatin at the Lamina Promote Oscillating Transcription. Mol. Cell *59*, 984–997.

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 996–1001.